Physicochemical foundations of life that direct evolution

Why chance and natural selection cannot explain evolution

Didier Auboeuf
Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 Allée d'Italie, Site Jacques Monod, F-69007, Lyon, France ; Mail: didier.auboeuf@inserm.fr

Abstract

The current framework of evolutionary theory postulates that evolution relies on random mutations generating a diversity of phenotypes on which natural selection acts. This framework was established using a top-down approach as it originated from Darwinism, which is based on observations made on complex multicellular organisms, and then modified to fit a DNA-centric view. In this article, I argue that, based on a bottom-up approach starting from the physicochemical properties of nucleic and amino acid polymers, we should reject the facts that: i) natural selection plays a dominant role in evolution, and ii) the probability of mutations is independent of the generated phenotype. I will show that the adaptation of a phenotype to an environment does not correspond to organism fitness but rather corresponds to maintaining the genome stability and integrity. In a stable environment, the phenotype maintains the stability of its originating genome, and both (genome and phenotype) are reproduced identically. In an unstable environment (i.e., corresponding to variations in physicochemical parameters above a physiological range), the phenotype no longer maintains the stability of its originating genome but instead influences its variations. Indeed, environment- and cellular-dependent physicochemical parameters define the probability of mutations in terms of frequency, nature and location in a genome. Evolution is non-deterministic because it relies on probabilistic physicochemical rules, and evolution is driven by a bidirectional interplay between genome and phenotype, the phenotype ensuring the stability of the genotype in a cellular and environment physicochemical parameter-depending manner.

Introduction

The current framework of evolutionary theory coming from the Modern Synthetic Theory of Evolution postulates that evolution relies on random mutations, generating a diversity of phenotypes on which natural selection acts. This concept is widely accepted in the scientific community despite the fact that some important issues have been raised about it[1-3]. The notion of random mutations can lead to multiple interpretations. It can mean that the nature or location of mutations is random, and indeed, mutations are often described as “errors” during replication[4,5]. However, many factors influence the rate, nature and location of mutations[5-10]. Thus, the appearance of a mutation is probabilistic and depends on multiple cellular- and environment-dependent physicochemical parameters. The notion of random mutations can also mean that the probability of a mutation is independent of the phenotype it generates. One of the objectives of this article is to show that if cellular- and environment-dependent physicochemical parameters influence the frequency, nature and location of mutations, the probability of a mutation should depend on the phenotype it generates. Indeed, I will show a continuum between physiological and genetic adaptation, as already has been proposed[11-13], which indicates that physiological adaptation facilitates and guides genetic variations in an environment-depending manner (see below).

The notion of natural selection is also subject to multiple interpretations as it can be negative, positive or neutral[14-17]. How can evolution be explained, if chance generates a large number of possibilities, and natural selection can be positive, negative or neutral? In addition, the notion of natural selection has a limited interest because a living organism observable over more or less long time periods is necessarily adapted to its environment, otherwise it disappears without leaving any descendants. The second objective of this article is to replace the notion of natural selection with the notion that the role of the phenotype in evolution is to maintain the integrity and stability of its originating genome.

Another issue raised by the current model of evolution relies on the fact that the combination of chance and natural selection does not provide a single conceptual framework that simultaneously explains both evolution and organism activities, since evolution but not organism activities would rely on chance and natural selection. However, life and evolution are inseparable and must depend on the same fundamental principle. Indeed, living organisms are hierarchically structured since multicellular organisms are composed of cells that are composed of molecules. The lower levels of organization have necessarily appeared before more complex living forms and fundamental principles of evolution must be applicable from molecules to complex organisms. In this setting, the current model of evolution was established using a "top-down" approach originating from Darwinism that is based on observations made on complex multicellular organisms[18] and that was modified after the discovery of DNA. In this article, I will use a "bottom-up" approach accordingly to the facts that i) life started with the emergence of nucleic and amino acid polymers before the emergence of more complex forms of life; ii) physicochemical laws are the foundations of cellular processes and complex organism activities. A "bottom-up" approach will

allow us to redefine i) the notion of chance in a precise context of physicochemical laws, and ii) the notion of adapted phenotype not in terms of organism fitness but rather in terms of impact on genome stability.

Life is based on two types of polymers: nucleic acid polymers (DNA and RNAs) and amino acid polymers (proteins). Emphasis has been given over the last decades on a functional dichotomy in which nucleic acids are described as the support of the genetic information, while proteins perform the cellular activities. As a consequence, nucleic acids are often "simply" considered as the carrier (DNA) or vectors (RNAs) of genetic information and are represented in the form of a suite of letters A, C, G, and T, corresponding to the four nucleotides that composed them. Thus, the physicochemical properties of nucleic acids are obscured in the context of the current Synthetic Theory of Evolution, in which evolution corresponds basically to the random substitutions of one letter by another one. In the first part, I will underline that the physicochemical properties of nucleic and amino acid polymers depend on their composition, which is constrained by cellular- and environment-dependent physicochemical parameters. I will next show that this principle has consequences on the way evolution proceeds.

Indeed, I will highlight in the second part that the emergence of life probably corresponds to the establishment of the interdependency between RNAs (or similar molecules) and proteins (or similar molecules). This means that, as observed in modern organisms, protein synthesis depends on RNAs as templates and RNA synthesis depends on proteins. This interdependency generates a positive feedback loop, as an RNA (the proto-genome) generates a protein (the proto-phenotype) that contributes to the synthesis of the RNA on which it depends (Fig. 1A). Moreover, RNAs are unstable and are rapidly degraded by hydrolysis. Therefore, the RNA/protein system can only self-reproduce and self-amplify if proteins (the proto-phenotypes) maintain the physical integrity of the template (the proto-genome) from which they are derived. This fundamental principle can be expressed as follows: the genome (in this case, an RNA) generates a phenotype (here, a protein) that contributes to the reproduction and stability of its originating genome by relaxing environment-dependent physicochemical constraints (Fig. 1A).

I will show in the third part that this principle captures the mechanisms by which cellular- and environment-dependent physicochemical parameters direct evolution. Indeed, I will show that specific cellular- and environment-dependent physicochemical parameters exert constraints on some nucleic and amino acid polymers, which triggers a specific cellular response through the regulation of the expression of a selected set of genes. If the resulting cellular activities relax the initiating constraints (i.e., return to equilibrium), this corresponds to physiological adaptation (Fig. 1B, path 1). Otherwise, I will show that physicochemical constraints persist and challenge the physical integrity of nucleic and amino acid polymers inducing more-or-less directly mutations in the corresponding genome locations. This process, corresponding to genetic adaptation, only stops when new sequences directly or indirectly relax the initiating constraints (Fig. 1B, path 2). Therefore, cell activity (or physiological adaptation) and evolution (genetic adaptation) are based on the same principle: the phenotype is derived from a genome in

response to physicochemical constraints and relaxes the initial constraints. If the phenotype is adapted to a given environment (i.e., a set of physicochemical constraints), the genome is stable and will be reproduced identically (physiological adaptation, Fig. 1C). If the phenotype is unsuitable, the genome is unstable and will be modified until generating a phenotype that will ensure its stability (genetic adaptation, Fig. 1C).

If the main role of a phenotype is to maintain the stability (see Glossary) of its originating genome, the phenotype (as the sum of all cellular activities) in turn creates physicochemical constraints on the genome. For example, the metabolic activities of cells produce a diversity of molecules (e.g., reactive oxygen species) that can interact with DNA and induce mutations. This means that a genome can generate a phenotype that can in turn generate physicochemical constraints on its originating genome (Fig. 1D). I will illustrate in the fourth part this feedback loop by showing that UV-radiation triggered the emergence of photosynthesis, which then triggered the emergence of oxidative metabolism followed by eukaryogenesis. In addition, I will show that this principle explains the emergence of epigenetic modifications, multicellular organisms, and germline cells. I will finally highlight the fact that the interplay between the genome and the phenotype described above at the RNA/protein level is still operating in multicellular organisms when considering that the phenotype generated from the germline cell DNA corresponds to the production of somatic cells whose function is to protect the originating genome (i.e., the germline cell DNA) from environment-dependent physicochemical constraints.

To summarize, in a stable environment, a genome generates a phenotype that maintains the stability of its originating genome and both (genome and phenotype) are reproduced identically (Fig. 1E, left panel). In an unstable environment (corresponding to variations in physicochemical parameters above a physiological range), the genome generates a phenotype that no longer maintains the stability of its originating genome and instead triggers mutations those rate, nature and location dependent on the initial constrains and the phenotype. This process occurs until new genetic variants generate a phenotype that maintains the stability of its originating genome (Fig. 1E, right panel). I will conclude that in disagreement with the current framework of evolutionary theory, the activities and evolution of living organisms are governed by the same physicochemical rule. Indeed, accordingly to the current framework there is no direct relationship between physiological and genetic adaptation since physiological adaptation is based on physicochemical principles of homeostasis, while genetic adaptation would be fueled by random mutations (Fig. 1F, left panel). In contrast, in the model proposed in this article, genetic adaptation is the consequence of physiological adaptation that takes place as long as fluctuations in environment-dependent physicochemical parameters do not exceed a certain threshold. Above this physiological threshold, the integrity of nucleic and amino acid polymers, in particular DNA, is challenged leading to targeted mutations until the emergence of a phenotype that maintains the integrity of its originating DNA with regard to environmental constraints (Fig. 1F, right panel).

Part 1. Environment-dependent physicochemical constraints on nucleic and amino acid polymer composition

Nucleic and amino acid polymers have emergent physicochemical properties (e.g., solubility, folding, and stability) that depend on intrinsic parameters relying on their composition and extrinsic parameters (e.g., temperature). The aim of this part is to show that cellular- and environment-dependent physicochemical parameters constrain the composition of cellular nucleic and amino acid polymers. I will show in the next par that this principle has consequences on the way evolution proceeds.

1.1. Physicochemical constraints on protein composition

The amino acid composition of proteins is constrained by intrinsic parameters impacting on protein solubility, folding, and stability; for instance, proteins with too many hydrophilic amino acids tend to unfold, while proteins with too many hydrophobic amino acids tend to aggregate[20,21]. Extrinsic physicochemical parameters, such as temperature or cellular and environment chemical composition, also constrain the amino acid composition of proteins, notably owing to the chemical modifications of amino acid side chains. Indeed, protein amino acids can react with and be modified by various chemical compounds (e.g., lipids, sugars, and reactive oxygen species or ROS)[22]. Protein chemical modifications (see Glossary): i) are spontaneously or enzymatically generated; ii) change the physicochemical properties of the modified amino acids; and iii) contribute to cellular regulatory processes or induce protein mis-folding and aggregation[22]. For example, amino acid oxidation in proteins plays a role in many cellular processes yet high rates of it induces protein damage and aggregation[22]. In agreement with the fact that (bio)chemical compounds constrain the amino acid composition of proteins, organisms growing under different levels of salt or oxygen, or having different metabolic activities, produce proteins with different amino acid composition biases[21,23] (see Glossary). Finally, the amino acid composition of proteins depends on the environmental supply of key elements that are required for the biogenesis of amino acids. For example, organisms growing in nitrogen- or sulphur-poor environments produce proteins that contain low amounts of nitrogen-rich amino acids (e.g., arginine) or sulphur-containing amino acids (e.g., cysteine), respectively[24,25]. In summary, the amino acid composition of proteins obeys physicochemical laws and on cellular- and environment-dependent physicochemical parameters.

1.2. Physicochemical constraints on nucleic acid polymer composition

RNA and DNA molecules are polymers composed of four monomers or nucleotides that interact with each other when they follow each other in a sequence (i.e., stacking interactions) or when they face each other in two different strands (i.e., base-pairing interactions). These interactions have consequences on the structural and physicochemical properties of nucleic acid polymers. For example, G:C pairs form stronger base-pair interactions than A:T pairs, purine-purine dinucleotides (e.g., GpA) form strong

stacking interactions, and GC dinucleotides form polymorphic structures[26-28]. As a consequence, GC- and purine-rich polymers are thermodynamically stable, and the genome of thermophilic organisms is enriched in GC and/or purine nucleotides[26,29]. Nucleotide composition also determines DNA mechanical properties (e.g., flexibility, bendability), contributing to its cellular functions and its resistance to torsional stresses (Fig. 2A and 2B, see Box 1). For example, increasing the GC content increases the B- to Z-DNA conformational transition, DNA-bendability, and DNA resistance to torsional stresses; accordingly, highly transcribed genes are GC-rich[27,30-32].

Like proteins, DNA and RNA molecules undergo dozens of spontaneous or enzyme-dependent chemical modifications (including oxidation, methylation, deamination, alkylation, and glycation) that i) affect the physicochemical properties of nucleic acid polymers; and ii) contribute to regulatory processes but can also induce damages[33-35](Fig. 2C). For example, DNA oxidation not only plays a role in gene expression regulation but also increases DNA damages and mutations[33,34]. The reason why DNA chemical modifications can generate mutations during replication is because a chemically modified nucleotide can "mimic" another one; for example, an oxidized-guanine can base-pair with an adenine instead of a cytosine, which can result in mutations during replication[33,34]. As for proteins, environmental physicochemical parameters (temperature, chemical composition) constrain the nucleotide composition of genomes; this can be observed in thermophilic, halophilic, acidophilic, aerobic, and radiation-exposed organisms[29,36]. Finally, the amount of DNA per cell (genome size and ploidy) is constrained by the environmental availability of phosphorus, and the genomes of organisms growing in nitrogen-poor environment are enriched in A:T pairs, which require 7 nitrogen instead of 8 in G:C pairs[37-39].

In summary, the properties of nucleic and amino acid polymers depend on intrinsic parameters and these polymers can undergo reversible structural and chemical modifications triggered by extrinsic physicochemical constraints; above a certain threshold, these modifications can challenge their integrity (Fig. 2A-D). This raises a major issue for coding sequences, which are under both nucleic acid- and protein-related constraints (Fig. 2E).

1.3. Interdependency between the physicochemical properties of nucleic acid polymers and their cognate amino acid polymers

Coding sequences accommodate different constraints, not only by the encoded protein sequence but also by cellular processes such as chromatin organisation (e.g., nucleosome positioning), transcription (e.g., DNA flexibility), RNA folding, splicing, and RNA-RNA or RNA-protein interactions[40-42]. It has been assumed until now that variations of the third nucleotide of codons (the "wobble" position, which is usually the only variable between all codons encoding an amino acid) solves nucleic acid–related constraints without affecting the encoded amino acids[40]. However, this assumption is challenged by the diversity of the constraints described above, and by the fact that the third nucleotide of codons changes

the thermodynamic property of codon–anticodon interactions, with consequences on translation fidelity, speed, and co-translational protein folding[40,43]. Accommodation of different constraints in coding sequences also relies on the fact that the genetic code is not randomly organized, as amino acids that share physicochemical properties (e.g., hydropathy) correspond to codons with a similar nucleotide composition bias. For example, hydrophilic or hydrophobic amino acids correspond to A- or T-rich codons, respectively, and small or large amino acids are encoded by GC-rich or GC-poor codons, respectively[44,45]. Consequently, two codons with only one nucleotide difference (either at the first or third position) can encode either the same amino acid or different amino acids with similar physicochemical properties.

However, the organization of the genetic code implies that the nucleotide composition bias of a nucleic acid polymer affects the physicochemical properties of the encoded protein and, conversely, that the amino acid composition bias in a protein affects the physicochemical properties of the cognate nucleic acid polymer. For instance, compact proteins comprising small amino acids correspond to GC-rich coding sequences (as small amino acids correspond to GC-rich codons), while proteins comprising hydrophobic regions (i.e., containing stretches of hydrophobic amino acids) correspond to T-rich coding sequences (as hydrophobic amino acids correspond to T-rich codons)[31,46,47]. Along the same lines, the global nucleotide composition bias of a genome (e.g., GC content) is associated with a global amino acid composition bias of the encoded proteome[48,49].

Supporting the notion that constraints on the physicochemical properties of one kind of polymer affects composition biases of the cognate polymers are the following observations: i) nucleosome positioning leaves a footprint in protein sequences; ii) splicing sites and splicing factor–binding motifs constrain the amino acid composition of peptides encoded by splicing-regulated exons; iii) mRNA secondary structures depending on base complementarity have consequences on the secondary structures of the encoded protein[41,50-52]. Conversely, protein secondary structures leave a footprint in the nucleotide composition bias of coding sequences, as for example amino acids that favor alpha-helices and beta-sheets correspond to codons ending with purines and pyrimidines, respectively[53]. Along the same line, alternation of hydrophobic and hydrophilic amino acids in amphipathic alpha-helices that rely on a periodicity of ~3.5 amino acids corresponds to a specific detectable ~10 bp periodicity in DNA, with consequences for the helical pitch of nucleosome-wrapped DNA[54]. Finally, purine enrichment in coding sequences is determined by protein-related physicochemical constraints, such as solubility and folding[55].

In summary, the organization of the genetic code “buffers” nucleotide- and amino acid–related constraints; however, above a threshold, nucleotide and amino acid composition biases affect each other. As nucleotide or amino acid composition biases determine the physicochemical properties of the nucleic or amino acid polymers, the composition-dependent physicochemical properties of these interdependent polymers must be adapted to the same fundamental physicochemical constraints (Fig. 2E), as detailed below.

Part 2: Evolution of the genetic code

Life relies on the interdependency between nucleic and amino acid polymers, since the biogenesis of proteins requires a nucleic acid polymer as a template (RNA) while biogenesis of nucleic acid polymers requires proteins. In addition, nucleic and amino acid polymers are in "competition" with each other, as their biogenesis requires the same elements (e.g., nitrogen) and required the same template—single-stranded RNA (ssRNA) before the emergence of DNA. The aim of this section is to propose that these fundamental principles (interdependency and competition) constrain the genetic code evolution to match the fundamental physicochemical properties of both polymers. In other words, specific codons correspond to specific amino acids because their presence in nucleic acids/cognate proteins allows both polymers to deal with the same fundamental physicochemical parameters.

2.1. Interdependency and competition between RNAs and proteins

While there is an ongoing debate about whether the origin of life started with only RNAs ("RNA world") or with RNAs and peptides ("RNP world"), there is a consensus that the activity of ribozymes (e.g., RNAs that catalyze nucleic acid polymerization) has been enhanced at some points of evolution by their interactions with amino acids or with randomly generated small peptides that may have also stabilized RNAs, which would otherwise be rapidly degraded by hydrolysis (Fig. 3A)[56-58]. For example, randomly generated peptides composed of abiogenetically-produced amino acids, such as Gly and Asp, may increase the efficiency of replicating ribozymes, as these amino acids play a very important role in the AsnAlaAspPheAspGlyAsp (NADPDGD) peptides found in all polymerases. In particular, binding of these amino acids to the catalytic metal ion Mg^{2+} could have enhanced polymerization and protected RNA from Mg^{2+}-dependent hydrolysis[59]. Coincidentally, a primeval genetic code corresponding to GC-rich and RNY codons (e.g., GGC, GCC, GAC, GTC) has been proposed, because these codons are the most frequent in coding sequences and correspond to the most metabolically simple amino acids (Gly, Asp, Ala, and Val) that are sufficient to produce stable, folded, and functional proteins[57,60]. A positive feedback loop between nucleic and amino acid polymers could have been initiated by primeval GC-rich RNAs, by first interacting with randomly-generated peptides (e.g., made of Gly and Asp), which would then favor the production of more complex peptides through a primeval GC-rich genetic code.

However, such a cooperation would have been inefficient if RNA and protein polymerization were physically uncoupled, for two main reasons. First, ssRNA templates give rise to stable double-stranded RNAs (dsRNAs) after replication, which decreases the rate of other rounds of replication as well as of protein synthesis that requires ssRNA templates[61,62]. Second, freely diffusible RNAs and proteins would limit their cooperation, as freely diffusible peptides generated from an RNA template can stabilize and enhance the replication of potential "parasitic" or mutated RNA replicators[63,64]. Simulation and *in vitro* selection experiments demonstrate that ribozyme-dependent replication cycles rapidly end when

molecules freely diffuse, due to the appearance of mutated RNAs, which become smaller and therefore are more rapidly amplified while simultaneously losing their enzymatic activity[63,64]. Proto-cell compartmentalization and physical coupling between replication (i.e., RNA production in an RNP world) and amino acid polymerization (i.e., protein production) could solve these two main issues[65-69]. First, interactions between the nascent RNA and the nascent peptide could decrease the formation of stable dsRNAs while protecting ssRNAs from degradation. This is observed in modern prokaryotes or eukaryotes: protein binding to nascent RNAs during co-transcriptional–translation or co-RNA processing prevents RNAs from interacting with the DNA template and increases transcription[70-73](Fig. 3B). Second, the physical proximity between replication and translation would increase the probability that the neo-synthetized proteins "protect" and enhance replication of its originating RNAs, which would have facilitated the increase in replication and translation fidelity[65-69].

While the physical coupling between RNA and protein polymerization may seem to be a "sophisticated" molecular process, several authors have proposed straightforward models[66-69,74]. For example, proto-tRNAs (tRNA ancestors) comprising three nucleotides (the proto-anticodons) bound by one amino acid could have been used simultaneously for replication and translation: the three anti-codon nucleotides could have been used as "building" blocks during replication, while the attached amino acids could have chemically enhanced the triplet polymerization and be used as building blocks for protein production (Fig. 3C). Of note, i) the use of tri-nucleotides rather than mono-nucleotides increases the efficiency and fidelity of replication by ribozymes[75]; ii) phylogeny analyses suggest that tRNAs originated in replication[76]; and iii) the coupling between transcription (i.e., RNA biogenesis) and translation is still operating in prokaryotes, and many features are shared by transcription and translation in modern cells[70-73](Box 2).

To summarize, life likely emerged from the cooperation between nucleic and amino acid polymers. This has consequences on the nature of amino acids that can be assigned to specific nucleotide sequences, as detailed below.

2.2. Co-adaptation of nucleic acid polymers and their encoded proteins to the same fundamental physicochemical parameters

In a proto-cell without complex compartments and protein-dependent compensatory mechanisms, nucleic acids and proteins are exposed to the same physicochemical parameters (e.g., chemical compounds). As both polymers depend on each other, their composition and the processes that lead to their biogenesis need to satisfy the same physicochemical constraints. Supporting this model, GC- and/or purine-composition biases increases the thermostability of nucleic acid polymers and in turn correspond to codons of amino acids that increase protein thermostability[29,47,77].

Along the same line, the organization of the genetic code allows nucleic and amino acid polymers to co-adapt to the bioavailability of nitrogen. Indeed, A:T pairs that require fewer nitrogen atoms than G:C pairs (7 vs. 8, respectively) correspond to amino acids that also requires fewer nitrogen atoms[24,37,38]. Accordingly, plant genomes and proteomes are AT-rich and contain amino acids that requires fewer nitrogen atoms as compared to animal genomes and proteomes; this fits well with the fact that nitrogen sources are limited for plants, while animals can access organic sources of nitrogen[24,37,38].

Oxygen is highly toxic, as oxygen derivatives (e.g., ROS) can damage nucleic and amino acid polymers and induce their cleavage or aggregation[22,33,34]. Stepwise increases of oxygen in the biosphere over time (see Part 4 for more details) may have given the impulse for the late incorporation into the genetic code of amino acids that can act as “ROS scavengers” (including Trp, Tyr, Met, Cys and His) and thereby protect biopolymers from oxidative damages[78]. Interestingly, the most frequent mutations in ROS-producing cancer cells affect Arg codons (and in particular, the CGN codons), with mutations producing codons for Cys (TGY), Trp (TGG), stop codon (TGA), and His (CAR)[79-83]. It has been proposed that these mutations i) are induced by ROS-mediated deamination of (methyl)cytosine, leading to C>T mutations, or incorporation of oxidized guanine during replication, leading to G>A mutations; and ii) protect cancer cells from the high levels of ROS by increasing the global anti-oxidant capacity of the cancer proteome[79-83]. Interestingly, CGN codons (that encode Arg) seem to be particularly sensitive to oxidation because of their physicochemical properties[79,84]. Further, (methyl)cytosine deamination produces CG>TA mutations and may thereby increase the nitrogen availability required by proliferative cancer cells, as it reduces the nitrogen-richer C:G pairs in favor of the nitrogen-poorer T:A pairs as well as the GC-rich sequences that contain codons (e.g., CGN) corresponding to nitrogen-rich amino acids (e.g., Arg)[79]. Therefore, ROS-induced cytosine deamination would simultaneously save nitrogen atoms (at both genome and proteome levels) and protect cells from ROS. Although speculative, one possibility is that the assignment of “ROS scavenger” amino acids to codons that originate from the oxidation of Arg codons generates anti-oxidant proteins from oxidized nucleic acids. Supporting such a possibility, assignment of Met to the ATA codon (in addition to the ATG codon) in most animal mitochondria lineages explains the high frequency of Met in mitochondria-encoded respiratory chain complexes, and it represents an adaption to high ROS level in mitochondria[85].

If physicochemical constraints shaped the genetic code, it is very likely that the universal genetic code is the result of horizontal transfers of "code fragments", as proposed by several authors[58,86]. For example, proto-cells present in extreme environments (e.g., hot or cold, nitrogen-rich or -poor, oxygen-rich or –poor) could have developed "code fragments” adapted to their specific environment. At the frontiers of these habitats where physicochemical parameters fluctuate, genetic horizontal transfers between cells using different “code fragments” may have led to the emergence of the modern genetic code (Fig. 3D).

To summarize, the organization of the genetic code was constrained to match general properties of the interdependent nucleic and amino acid polymers, as well as to favor their interaction and coordinate their biogenesis (Box 2). As a consequence, nucleic acids and their cognate proteins may share more physicochemical and structural properties than previously anticipated (Fig. 3E), which has consequences on evolutionary mechanisms, as discussed in Part 3. However, it must be first emphasized that the cell metabolism is another important factor in the genetic code evolution.

2.3 Feedforward and feedback loops between gene products and their products (i.e., metabolites)

Nucleic and amino acid polymers depend on the availability of molecules containing elements, such as nitrogen (see Part 1). This dependency would have favored the emergence of polymers with metabolic activities that modify environmental resources and allow elements to be incorporated into amino acids and nucleotides. In an RNP world, the proto-cell phenotype would be a positive feedback loop between bio-synthetic pathways (i.e., metabolism in modern cells) and polymerization of RNAs and proteins (i.e., the gene expression process in modern cells): polymer production depends on bio-synthetic pathways that in turn depend on polymerization products (Fig. 3F). Supporting this interplay between gene expression and biosynthetic pathways, the genetic code likely co-evolved with amino acid biosynthetic pathways, as codons starting with A, C, U, or G correspond to amino acids synthetized from oxaloacetate, α-ketoglutarate, pyruvate, or from the reductive amination α-keto acid, respectively[57,87-89]. A possibility is that simple amino acids or metabolites covalently attached to polynucleotides (e.g., proto-tRNAs made of three nucleotides) could have been chemically transformed to give rise to more complex amino acids[89]. Another possibility is that chemical modifications of simple amino acids occurred after their incorporation into proteins, and that chemically modified amino acids (i.e., complex amino acids) were thus available after protein hydrolysis. Complex amino acids would have next been incorporated into the genetic code.

In this context, it must be underscored that the chemical modifications of proteins and nucleic acids establish a direct bridge between cell metabolic activities and gene expression/gene product functions (i.e., the cellular physiological adaptation). Indeed, chemical modifications of nucleic and amino acid polymers (e.g., post-translational modifications, DNA or RNA methylation) that change their physicochemical properties also affect their cellular activities, and these chemical modifications depend on the cell metabolic activities. For example, DNA, RNA, and protein oxidation depend on cellular oxidative metabolism, while DNA, RNA, and protein methylation depend on the production of *S*-adenosyl methionine (SAM), the universal methyl-group donor produced in the one-carbon cycle[90-92]. Finally, as chemical modifications can also have deleterious effects on DNA, RNA, and proteins by challenging their physical integrity (see Part 1), I will show in the next section that metabolism-depend chemical modifications of nucleic and amino acid polymers establish a continuum between physiological and genetic adaptation (Fig. 2B, 2C and 3G).

Part 3. Continuum between physiological and genetic adaptation

The aim of this section is to show that environmental fluctuations induce physical and chemical constraints on nucleic and amino acid polymers, which triggers the cellular physiological adaptation that maintains cellular homeostasis (Fig. 1B, 1). However, if the cellular response to environmental fluctuations does not allow a return to equilibrium, constraints persist and can challenge the physical integrity of DNA, leading to DNA damages, mutations, and genetic variations. This process only ends when mutated sequences allow the direct or indirect relaxation of the initial constraints (Fig. 1B, 2).

3.1. Genetic adaptation directed by transcription: transcription-replication conflicts

Several authors have already proposed that cellular stresses (i.e., environmental fluctuations above a physiological range) induce "adaptive mutations" (i.e., mutations that occur at a high rate) and/or "directed mutations" (i.e., mutations occurring at specific genomic locations)[93-95]. A possible underlying mechanism is that stress-directed transcriptional activation of specific loci (as part of the cellular physiological adaptation) increases the probability of mutations to occur within these loci because transcription induces mechanical stresses (e.g., formation of supercoiling) that challenge the physical integrity of transcribed DNA[12,35,94-97]. I will describe below how this straightforward principle establishes a continuum between physiological and genetic adaption.

Highly transcribed genes are enriched in GC nucleotides that can "absorb" mechanical stresses by favoring the B- to Z-DNA transition, owing to the physical properties of both base-pairing and base-stacking interactions between G and C nucleotides (see Part 1). Critically, this G/C enrichment can be explained by several transcriptional-dependent mutational biases (see Glossary), which result in part from conflicts between transcription and replication[97,98]. Indeed, while the act of transcription increases DNA accessibility to DNA polymerases, the simultaneous transcription and replication of a locus creates conflictual physical stresses leading to DNA breaks[97-99]. DNA breaks can be repaired by heteroduplex DNA recombination, a process known to favor G:C over A:T base pairs[100-103]. This phenomenon may in part rely on the fact that a T in a mismatched base-pair (T:G or T:C) within heteroduplex DNA can spontaneously flip out of the dsDNA, increasing the probability of its removal[104]. In addition, transcription-replication conflicts have been shown to induce adenine deamination, giving rise to hypoxanthine, a nucleotide that mimics guanine and leads to A:T>G:C mutations during replication[105]. Different mutational biases also occur in early and late replicating regions[7]. For example, the accumulation of free oxidized-dGTP before replication may result in its incorporation in place of Ts (as oxidized-dGTP mimics adenine), leading to A:T>G:C mutations during the early phase of replication[106]. Additionally, as DNA cytosine methylation is associated with transcription repression, heavily methylated regions correspond to late replicating regions; these regions could more frequently undergo C:G>T:A mutations because of the high rate of spontaneous deamination of methylcytosine[107]. Replication-dependent mutational bias can also be due

to decreases during the cell cycle of concentrations of free dGTPs and dCTPs (as producing these nucleotides requires more energy than producing dATP and dTTP), which leads to a higher incorporation rate of A or T nucleotides in late-replicating regions[108]. Therefore, mutational biases associated with replication timing and replication-transcription conflicts could increase the GC- and AT-content in early and late replicating regions, respectively.

While the GC-content of loci may increase as a consequence of DNA breaks resulting from transcription-replication conflicts, this increase next exerts positive feedback loops by: i) synchronizing transcription and replication; ii) increasing local DNA stability during transcription and/or replication; and iii) increasing the transcription activity of modified loci as well as many downstream steps of the gene expression process[28,32,42,109]. Indeed, the mechanical properties of GC-rich DNA regions favors transcription efficiency (but not elongation speed) and avoids nascent RNAs from interacting with the DNA template due to formation of stable secondary structures in the nascent GC-rich RNAs[70,71]. High GC-rich content may also increase local rate of recombination/deletion, leading eukaryotic GC-rich genes to bear smaller introns (as compared to AT-rich genes)[30,31,102,110]. Of note, small GC-rich introns are more efficiently spliced, likely because of intronic RNA secondary structures[111,112]. In addition, high GC content in RNAs increases the efficiency and fidelity of translation by smoothing translation elongation and favoring co-translational protein folding[52,113-115]. And finally, as the genetic code is not random, increasing gene GC content leads to the biogenesis of proteins with small amino acids, which in turn leads to decreases in: i) protein volume, ii) concentration-dependent aggregate formation, and iii) the energetic cost of protein production[31,116-118]. Therefore, the “over-stimulation” of the transcriptional activity of some genes under sustained stresses could result in i) replication-dependent mutational bias, ii) increases in the GC-content of stress-induced genes; and iii) decreases in the local transcription-dependent DNA instability, while improving gene product biogenesis at multiple levels (Fig. 4A). Another genetic process that increases the biogenesis of specific gene products under stress situations is gene duplication, which is also linked to transcription-replication conflicts[12,119]. For example, stress-induced promoter activity can stimulate gene duplication by destabilizing stalled replication forks[119].

These observations therefore support a model in which environmental fluctuations induce physical constraints on DNA during transcription, leading to the biogenesis of RNAs and proteins, and then leading to re-establishing equilibrium through the cellular physiological adaptation (Fig. 4A, 1). However, if the constraints persist, sustained transcriptional activation can result in transcription-replication conflicts resulting in GC-mutational biased or gene duplication, which could relax the initiating constraints by increasing gene product levels (Fig. 4A, 2).

3.2. Genetic adaptation directed by transcription: role of ssDNA formation and DNA folding

Transcription-dependent chromatin relaxation and ssDNA formation increase the accessibility of transcribed DNA regions to mutational agents—so-called "transcription-associated mutations"[5,12,35,95-97,120]. For example, ROS-mediated oxidation of cytosines and methylcytosines is enhanced in ssDNA, which can lead to C>T mutations during replication[35]. As a consequence, CGN codons (corresponding to Arg) of transcriptionally activated genes under ROS exposure have a higher probability to be mutated into TGG, TGA, and TGY codons (corresponding to Trp, stop, and Cys codons, respectively), allowing the synthesis of proteins containing amino acids that protect cells from oxidation (see Part 2 and Fig. 4B).

Remarkably, genome-wide analysis of mutations occurring in organisms growing under different environmental constraints (e.g., different metabolic resources or temperatures) show that each challenging condition is associated with a specific mutational bias[121-125] (see Glossary). This is in agreement with the fact that each mutagenic agent i) affects sequences with specific physicochemical properties, and ii) induces nucleotide modifications toward a particular pattern, as has now been well established in cancer genetics[126]. Of note, in addition to ROS-associated mutational signatures in cancers (see above), mutational bias induced by high intracellular pH has recently been shown to favor Arg (CGY) > His (CAY) mutations that confer pH-regulated protein functions[127]. It will be very interesting to systematically characterize the relationship between i) the physicochemical properties of mutagenic agents as well as of their most affected sequences; and ii) the nature of the induced mutations and the resulting physicochemical consequences at the nucleotide and amino acid levels[8]. Each mutagenic agent could trigger a specific mutation bias that more or less directly relaxes the initial physicochemical constraints (Fig. 1B, 2).

Of particular interest, transcription-dependent ssDNA formation increases also the probability of insertion of repeated elements, such as transposons and retrotransposons, a process known to play a major role in the genomic plasticity under sustained stresses[128-132]. It has been proposed that cellular stresses leading to a global chromatin relaxation could on the one hand de-repress (retro)transposon activity and, on the other hand, increase the likelihood of their insertion in specific stress-transcriptionally activated genes[128-132]. Insertion of (retro)transposons in stress-activated genes can influence gene expression at multiple levels, in particular by playing a role in the spatial genome organization. Indeed, it is now well recognized that regulation of gene transcription is based on the three-dimensional (3D) genome organization, which roughly corresponds to DNA folding (Box 1). DNA folding plays a critical role in co-regulating genes by bringing them closer together in space[133-137]. Factors binding to repeated sequences dispersed in various genomic locations could facilitate 3D genome organization and promote co-regulation of repeated sequence–hosting genes[138,139](Fig. 4C). Spatial clustering of co-regulated genes could also increase the probability of their recombination, a process facilitated by the presence of repeated elements[140-143]. Recombination between transcriptionally co-regulated genomic regions can

lead to the formation of new gene products but also can facilitate the expression coordination of co-stimulated genes (Fig. 4C). The importance of gene position in genomes with respect to their regulation and cellular functions is clearly established in operons from bacteria and in the so-called topologically-associated domains (TADs) in eukaryotes[144-147]. An emerging theme is that the 1D and 3D locations of genes play a major role in coordinating the expression of genes whose products are involved in the same cellular pathways (Box 1).

In sum, environment-dependent physicochemical constraints on DNA trigger cellular physiological adaptation and a continuum between physiological and genetic adaptation is established when environment-dependent physicochemical parameters above a physiological range challenge DNA physicochemical integrity.

3.3. Physiological adaptation facilitates genetic adaptation: role of RNAs

DNA chemical modifications can contribute to physiological adaptation (e.g., role of DNA methylation and oxidation in transcription regulation) and induce mutations during replication when a chemically modified nucleotide "mimics" another nucleotide (see Part 1). This mimicry process can also result in transcription "infidelity" by affecting the base-pair rules, thereby leading to biogenesis of RNAs with different sequences than encoded by the DNA template[148-150]. This so-called "transcriptional mutagenesis" is more frequent than previously anticipated and could play a major role in both physiological and genetic adaptations. For example, chemical modifications in transcribed DNA may lead to the biogenesis of new RNAs and proteins, which could contribute to cell survival. As a consequence, only cells bearing the DNA chemical modifications can divide, so that having the DNA chemical modifications would increase the probability to give rise to genetically adapted daughter cells [148-153] (Fig. 4D). The same principle can be used to establish a direct link between genetic adaptation and epigenetic modifications (i.e., DNA or histone chemical modifications that impact gene expression), as environmental fluctuations induce epigenetic modifications of specific loci as part of the cell´s physiological adaptation. If transcription-dependent epigenetic modifications at specific loci increase cell survival, the surviving cells may have a higher probability to undergo mutations within these genes, as chromatin organization and epigenetic marks more or less directly impact both DNA damage and DNA repair[35,154,155].

Along the same line, RNAs produced from loci as part of the cellular physiological adaptation could under certain circumstances induce mutations in the loci they originated from. For example, the spatial proximity of co-regulated genes (see above) could promote the biogenesis of chimeric RNAs via a mechanism called trans-splicing, thereby "fusing" RNAs produced from two genes[156]. These chimeric RNAs can give rise to new proteins that could contribute to the survival of cells in a stressful situation. As RNAs can be used as a matrix during DNA break repair, surviving cells expressing chimeric RNAs could use

these RNAs during the repair of loci broken by the stress-induced transcription, increasing the probability of recombination between specific loci[157,158].

RNAs can increase the likelihood of genetic variations in numerous ways[159,160]. I have recently described a spectra of molecular mechanisms by which physicochemical constraints on RNAs and proteins at the time of their synthesis could trigger mutations in their originating genes through the biogenesis of RNA fragments[73,161,162]. Briefly, environmental fluctuations can on the one hand induce the transcriptional activity of target genes, thereby generating a greater amount of mRNAs and proteins, and on the other hand generate constraints on nascent RNAs / nascent proteins during transcription / translation (Fig. 4E, 1 & 2). Perturbation of mRNA or protein synthesis leads to the biogenesis of RNA fragments; for example, mRNA cleavage occurs when the dynamics of ribosomes (e.g., as a consequence of nascent protein misfolding) along the mRNA template is disturbed[163,164]. RNA fragments generated during transcription and/or translation could then interact with their originating genomic regions and induce genomic instability and mutations in the targeted regions (Fig. 4E, 3). Therefore, RNA-directed mutations could increase the likelihood of mutations to occur in specific loci when cells experienced constraints during the biogenesis of specific RNAs or proteins (see also Part 4).

In sum, environment-dependent physicochemical parameters trigger cellular physiological adaptation through changes of the cellular activities that leave traces / footprints on nucleic and amino acid polymers (e.g., DNA breaks, chemical modifications, RNA biogenesis). If the physiological adaptation allows a return to equilibrium, polymer modifications are temporary and reversible (Fig. 1B, 1). If not, these modifications may have a more-or-less direct effect on replication, potentially leading to mutations at specific loci. This mutational process triggered by physicochemical constraints only stops when new sequences relax the initiating constraints (Fig. 1B, 2). This principle is well suited to unicellular organisms in which the same DNA molecule is used as a template during i) the physiological response to environmental fluctuations and ii) replication and transmission across generations.

Part 4: Physicochemical rules: from genes to ecosystems

The objective of this section is to illustrate, by examining the emergence over evolutionary time of photosynthesis, oxidative metabolism, eukaryogenesis, multicellularity, and gametes, that i) a phenotype is adapted to an environment if it maintains the integrity of its originating genome with respect to environment physicochemical constraints (Fig. 1D, 1); and ii) a phenotype (i.e., the sum of all cellular activities) generates physicochemical constraints on its originating genome (Fig. 1D, 2). Indeed, numerous compounds are produced by the cell in response to environment-dependent physicochemical variations, which can interact and react with nucleic and amino acid polymers with potentially deleterious effects on their stability, folding, or solubility[22,165] (see Part 1). Therefore, cellular activities represent a source of physicochemical constraints on the originating genome (Fig. 5A).

4.1. Feedforward and feedback loops between toxicity ("constraints") and detoxification ("adaptation")

Before the emergence of the ozone layer, cells were exposed to strong solar UV radiations[166-168]. By inducing genomic instability, UV favored the emergence of genomes that produce pigments absorbing UV-damaging radiations, which would give these genomes a greater probability to be "accurately" reproduced[166-168](Fig. 5B, 1). Interestingly, it has been proposed that the molecular ancestors of photosynthetic light acceptors (e.g., chlorophyll) were pigments that protected nucleic and amino acid polymers from UV irradiation[166-168]. One possibility is that UV-absorbing pigments generated genomic instability because of the free dispersion of energy (heat) released from UV-absorbing pigments, favoring the emergence of pigment-interacting proteins containing photosynthetic reaction centers that can concentrate energy into complex molecules (e.g., sugars or "energy tanks") by fixing CO2[166-168] (Fig. 5B, 2). This gave rise to photosynthesis that in turn generated new constraints, as photosynthesis produces oxygen, a highly toxic compound that damages nucleic and amino acid polymers. In this setting, the ancestors of the gene products and metabolites involved in cell respiration have been proposed to have originated from oxygen scavenger compounds or oxidases that do not conserve energy and that protected their originating genome from the rise of oxygen[169-171]. Therefore, the emergence of cellular respiration could have resulted from a "detoxification" process concentrating oxygen-derived energy in biosynthetic pathways, in the same way that photosynthesis emerged from energy concentration from UV radiations[169-171] (Fig. 5B, 3).

The rise of oxygen produced by photosynthetic cyanobacteria in the earth atmosphere may also have impelled anaerobic cells (e.g., archaeas) and aerobic cells (e.g., alphaproteobacteria) to cooperate, as aerobic cells could protect anaerobic cells by scavenging environmental oxygen and/or because both cell types exchanged intermediate metabolites[172-174]. This cooperation might have resulted in the internalization of aerobic bacteria (the ancestors of mitochondria) by anaerobic archaea, resulting in the emergence of eukaryotes[172-174]. However, as mitochondria subsequently generated intracellular toxicity by producing intracellular ROS, this detoxification would have favored the biogenesis of new cellular compounds[175,176]. In this setting, biogenesis of sterols that requires oxygen-dependent enzymes could have first played a role in oxygen detoxification[176-178]. In addition, these molecules have specific properties when incorporated into membranes that contributes to the development of eukaryotic intra-cellular membranes, like the nuclear membrane, which could have initially protected intracellular polymers (e.g., DNA) from ROS[176-179]. Indeed, intracellular bio-membranes can fold up into three-dimensional periodic arrangements ('cubic membranes'), representing antioxidant defense[180].

The increase in intracellular ROS levels produced by mitochondria may also have been relavent for the origin of eukaryotic spliceosomal introns from group II introns found in archaea and bacteria, which are in fact mobile retro-elements that use the combined activities of an autocatalytic RNA and an

intron-encoded reverse transcriptase to propagate within genomes[181,182]. It has been proposed that retromobility of group II introns can be stimulated by oxidative stress and that the presence of introns may have contributed to "trapping" ROS in introns, thereby decreasing the probability of nucleotide oxidation in coding exons[181,183-185]. Supporting this model, nucleotide oxidation increases the probability of GC>AT mutations, and the frequency of GC nucleotides is higher in exons than in introns[185,186]. This implies that exons have been protected from ROS, which could have been achieved by histones. Indeed, histones are preferentially found in GC-rich sequences because of the flexibility of the G-C stacking interactions (see Part 1), and they protect DNA from a variety of mutagenic stresses by i) binding to and stabilizing dsDNA, ii) compacting DNA, and iii) providing a "shield" through their C-terminal tails that are rich in Arg and Lys, amino acids that are basic and positively-charged and can act as ROS scavenger[187-191]. Therefore, binding of histones to GC-rich exons may protect them from oxidation, while intronic GC>AT mutations induced by ROS would ultimately lead to exclude histones from introns. A consequence of intron invasion was an increased DNA size, leading to the formation of linear DNA from an ancestor circular DNA, as increasing the size of a circular DNA favors the formation of linear DNA because of intramolecular recombination[182]. Of note, group II introns that converted to linear DNA through expansion are also likely the ancestors of telomeres that protect linear DNA extremities[182].

In summary, photosynthesis, oxidative metabolism, and eukaryogenesis may have been triggered by extracellular (e.g., UV radiation) and intracellular (e.g., ROS) physicochemical constraints that destabilize genomes (i.e., inducing genetic variations). New genomes produce molecules that relax the initiating constraints, therefore stabilize their originating genomes but simultaneously generate new constraints, constantly pushing organism evolution (Fig. 5B and Fig. 1D).

4.2. Adaptation to a diversity of environment-dependent constraints: epigenetics and multicellularity

The cell physiological adaptation to environmental fluctuations relies on the biogenesis of gene products and metabolites. Nevertheless, all possible biochemical reactions cannot take place simultaneously in a cell as i) each reaction depends on specific physicochemical conditions and ii) the diversity of the generated biochemical products would be highly toxic, due to their ability to interact and react spontaneously with each other[22,165]. In addition, each nucleic acid and amino acid polymer can only be adapted to a limited number of constraints (e.g., salt or oxygen concentrations) (see Part 1). Therefore, a cell can physiologically adapt to a limited number of fluctuating physicochemical parameters that push towards the cooperation between unicellular organisms performing complementary biochemical reactions, thereby protecting each other from environmental fluctuations[192] (Fig. 5C).

In this context, histones may have allowed an increase in the diversity of metabolic activities encoded by a single genome[193]. Eukaryotic histones evolved by compacting DNA and by acting as a "chemical" shield and therefore maintaining the stability of the genome that produces them (see above).

Histone chemical modifications (i.e., epigenetic marks) could first have been triggered as a chemical "shields" and played a role in maintaining DNA stability against chemical DNA "attacks"[194]. However, by affecting DNA accessibility, histone chemical modifications would not only have protected specific genes but also coordinated their activity depending on the intracellular chemical composition. Indeed, different epigenetic marks may protect different parts of the genome from different chemical compounds, while simultaneously adapting gene transcriptional activities with respect to these chemical compounds (Fig. 5D). Two pieces of evidence support such a possibility. First, epigenetic marks are directly dependent on the cell metabolism: for example, methylation depends on SAM produced by the one carbon cycle, and demethylation relies on oxidation of methylated residues and therefore on the cellular oxidative metabolism[90-92,193]. Second, histone chemical modifications either reduce or facilitate DNA access to RNA polymerases and to potentially genotoxic molecules[91,92,154,194,195]. By selectively protecting and regulating gene expression, histone chemical modifications allowed the emergence of different cell types (i.e., multicellularity) containing the same genome but performing different metabolic activities (Fig. 5D).

Multicellularity that emerged several times in the course of evolution corresponds to a "metabolic division of labor", with cells containing the same genome performing specific sets of chemical reactions and protecting each other's by exchanging metabolites[196-199]. However, increasing the diversity of metabolite exchanges within a group of cells would have increased the probability of DNA damage events and could have favored the emergence of sexual reproduction. Indeed, meiosis, the cellular process involved in the production of haploid gametes, likely emerged over evolutionary time as a mechanism to erase DNA oxidation resulting from cellular activities[200-202]. Supporting this model, meiosis is associated in many species with homologous recombination, a process primarily involved in DNA repair. As already mentioned, homologous recombination favors GC over AT nucleotides (i.e., the so-called GC-bias gene conversion) because of physical properties of DNA; for example T (corresponding to an oxidized (met)C) flips outside the DNA molecule in T:G mismatches (see above)[100-104]. In so doing, homologous recombination limits the load of GC>AT mutations triggered by C oxidation. Further supporting a role of meiosis in erasing ROS-dependent DNA damages, oxidative stresses trigger meiotic homologous recombination (even in organisms such as male Drosophila, which normally do not perform meiotic homologous recombination), and meiotic recombination and crossovers do not occur randomly but in DNA regions localized between methylated nucleosomes that are less protected from ROS[200-202]. In addition, meiosis that results in the formation of haploid cells is essential to eliminate damaged alleles that can be otherwise masked in diploid cells[200-203]. A side effect of meiotic homologous recombination is that different parts of different DNA molecules are mixed, so that this process increases the genetic diversity without the need of *de novo* mutations.

To sum up, the increase in cellular metabolic activities favored by i) the raise of oxygen and ii) the emergence of molecules protecting cells from DNA oxidation (e.g., histones) would have triggered the

emergence of processes like meiosis, which "counteract" and "erase" DNA damages triggered by an increasing diversity of biochemical processes. In the model depicted above, known as the "dirty work hypothesis"[204,205], somatic cells performed metabolic activities that maintain the integrity of the germline DNA molecule, which is transmitted to the next generation and which (after fertilization) can generate the same phenotype—in other words, production of somatic cells protecting germline DNA molecules (Fig. 5E). Therefore, the principle described in Sections 2 and 3 at the molecular and cellular levels, respectively, is still operating in multicellular organisms, considering that the germline genome generates a phenotype (somatic cells) that maintain the stability of the originating genome (i.e., those of germ cells). However, what happens if somatic cellular activities do not maintain the organism homeostasis in response to environmental fluctuations?

4.3. Evolution of multicellular organisms

In Section 3, I proposed that environment-dependent physicochemical constraints on DNA trigger a cellular physiological adaptation that can leave "marks" (e.g., DNA chemical modifications or RNAs), which next potentially results in mutations during replication. As the replicated-DNA that is transmitted across generations in multicellular organisms is no longer directly involved in physiological adaptation, do physicochemical constrains exerted on somatic cell DNA induce modifications on germline cell DNA?

To answer this question, it must first be stressed that germline cells depend on and are exposed to the activities of somatic cells. For example, metabolic disorders are associated with the overproduction by somatic cells of small sugars or lipids that can react with and induce damages in the germline cell DNA. Further, metabolic activity of somatic cells, for example under nutrient constraints of the parents, can induce epigenetic changes on specific loci in germline cell DNA with consequences on the development and activity of the offspring[206-209]. There is also a consensus regarding the fact that molecular exchanges between somatic and germline cells are more complex than previously anticipated. For example, somatic cells produce extracellular vesicles that i) contain proteins, metabolites, and a diversity of small RNAs (e.g., microRNAs, tRNA-derived small RNAs, [tsRNAs]); and ii) are internalized by germline cells with consequences on the development and activity of offspring after fertilization[208,210-214] (Fig. 5F, 1). These processes, collectively called "transgenerational epigenetic inheritance", establish that a somatic cell's "experiences" in the parent organism can be transmitted at the molecular level to germline cells, with consequences on the offspring's phenotype[208,210-215]. Another process, called "genomic imprinting" that also depends on germline epigenetic marks and small RNAs leads to the selective expression of alleles transmitted from one of the two parents sometimes in an environment depending manner[216]. Epigenetic marks of one allele can also be transferred to the other allele, a process called "paramutation", which is based on the biogenesis of biochemically-modified small RNAs and which underlines the plasticity associated with epigenetic mechanisms[217,218]. All the processes described above maintain a diversity of

alleles by transcriptionally activating or repressing some of them in the offspring depending on the parental somatic cells' experiences (Fig. 5F, 1 & 2); combined with meiosis, these processes could "purge" some deleterious alleles[200-203,219-221].

Indeed, the frequency of transmission of some alleles in offspring can vary depending on their parental origin. This mechanism, known as the "transmission ratio distortion", is an exception to Mendel's laws of equal segregation and seems to rely on a diversity of mechanisms, such as "selection" among chromosomes during meiosis (e.g., non-random crossovers during meiotic recombination), allele-dependent elimination of gametes, or selective elimination during early zygote development[219-221]. Although the underlying molecular mechanisms of transmission ratio distortion are not yet fully understood, such a process establishes that sexual reproduction allows the selective elimination of some alleles. One possibility is that genes targeted by somatic RNAs in germline cells have a lower probability to be transmitted in next generations depending on the parents' experiences (Fig. 5F, 3 & 4). Therefore, sexual reproduction allows genes to be turned on/off, and/or for some alleles, to be eliminated, across generations without the need of *de novo* mutations in germline cells.

In this context, several studies have shown that even though the number of germline *de novo* mutations per generation is very low in some species, their distribution across genomes is biased[9,10,222,223]. For example, an association between *de novo* mutation occurrence, replication timing, transcription, and chromatin organization has been observed in germline DNA[9,10,222,223]. Furthermore, *de novo* "mutational clusters" corresponding to multiple *de novo* mutations in very close vicinity in a single individual, as well as "mutational hotspots" corresponding to *de novo* mutations occurring at the same location in several individuals, have been reported[222-224]. As the distribution of *de novo* mutations across genomes is biased, an important issue is to decipher whether their occurrence could depend on the somatic cell experiences? One possibility could rely on the fact that RNAs produced by somatic cells induce local and targeted epigenetic modifications in the germline DNA, which next induces more or less directly targeted *de novo* mutations because of the interplay between the chromatin environment and the local mutational rate (Fig. 5G, blue arrows)[35,154,155,225,226]. Although very speculative, there is also the possibility that the DNA of somatic cells challenged by environment-dependent physicochemical parameters produces "parasite-mimicking RNAs" that could form DNA-RNA hybrids in their complementary loci in the germline DNA and locally trigger mutations (Fig. 5G, red arrow). In support of this possibility, i) a convergence between RNA-containing extracellular vesicles and viral particles has been described; ii) RNAs are widely used in all living organisms to cleave or mutate parasitic nucleic acids; iii) RNA:DNA hybrids can be genotoxic, as for example RNA:DNA hybrids can induce DNA adenine deamination[73,159-162,217,218,227-232]. It would be very interesting in the future to investigate whether some of the 150 chemical modifications of RNAs identified so far could trigger selective mutations in RNA:DNA hybrids[73,159-162,217,218,229,230,232]. To summarize, the role of RNAs produced by somatic cells in the origin of *de novo*

mutations in germline cells still remaining to be determined, especially considering i) the production of extracellular vesicles containing RNAs that are produced by somatic cells; ii) the observations that germline *de novo* mutations are not randomly distributed across genomes; iii) the involvement of RNAs in epigenetic and genetic modifications; and iv) the diversity of RNA modifications associated in particular with the elimination and mutation of parasitic sequences.

If *de novo* mutations in germline cells are triggered by RNAs produced by somatic cells, how this could then lead to the emergence of complex phenotypes? Adaptation of multicellular organisms to cold stress provides some clues. Cold may lead to genetic instability by decreasing the kinetics of (bio)chemical reactions, and this cold-induced genomic instability might be relaxed by increasing the activity of genes encoding enzymes involved in cellular respiration, therefore increasing heat production[233]. However, increasing cellular respiration increases the production of ROS that induces genetic instability and therefore may lead to increase the activity of genes involved in the detoxification of ROS, such as those coding for uncoupling proteins (UCPs)[234,235] (Fig. 5H, 1 & 2). Indeed, uncoupling proteins, such as UCP1, are anion carriers that mediate proton leaks across the inner mitochondrial membrane, which i) mitigate ROS production and ii) simultaneously lead to cellular heat production[234,235]. The UCP-dependent heat production has been proposed to contribute to the emergence of heat-producing muscle cells, and next the emergence of mammalian brown adipose cells that i) express UCP1; ii) derive from skeletal muscle progenitor cells; and iii) play an important role in heat production in mammals [236,237](Fig. 5H, 2 & 3). Interestingly, it has been proposed that the loss of genes like UCP1 in bird ancestors (due to yet unknown mechanisms) did not allow the emergence of brown adipose cells but instead led to hyperplasia of heat-producing muscle cells. Thermoregulation depending on muscle hyperplasia (vs. brown adipose cells) has been proposed to generate constraints during development, with consequences on the body plan organization of bird (vs. mammal)[238,239] (Fig. 5H, 4). Thus, environment-dependent physicochemical parameters (e.g., cold) could create constraints on somatic and germ cells, triggering genetic instability and leading to the emergence of new polymers (in this case, UCPs) and new cell types (muscle cells) that relax the initial constraints (by heat production) yet inducing new constraints (muscle hyperplasia). In other words, the relationship between RNA (proto-genome) and proteins (proto-phenotype) in proto-cells (Section 2) is similar to the relationship between DNA and the unicellular cell phenotype (Section 3), as well as to the relationship between germ cell DNA and somatic cells that correspond to the phenotype generated by the germ cell DNA (Fig. 5I).

Conclusion

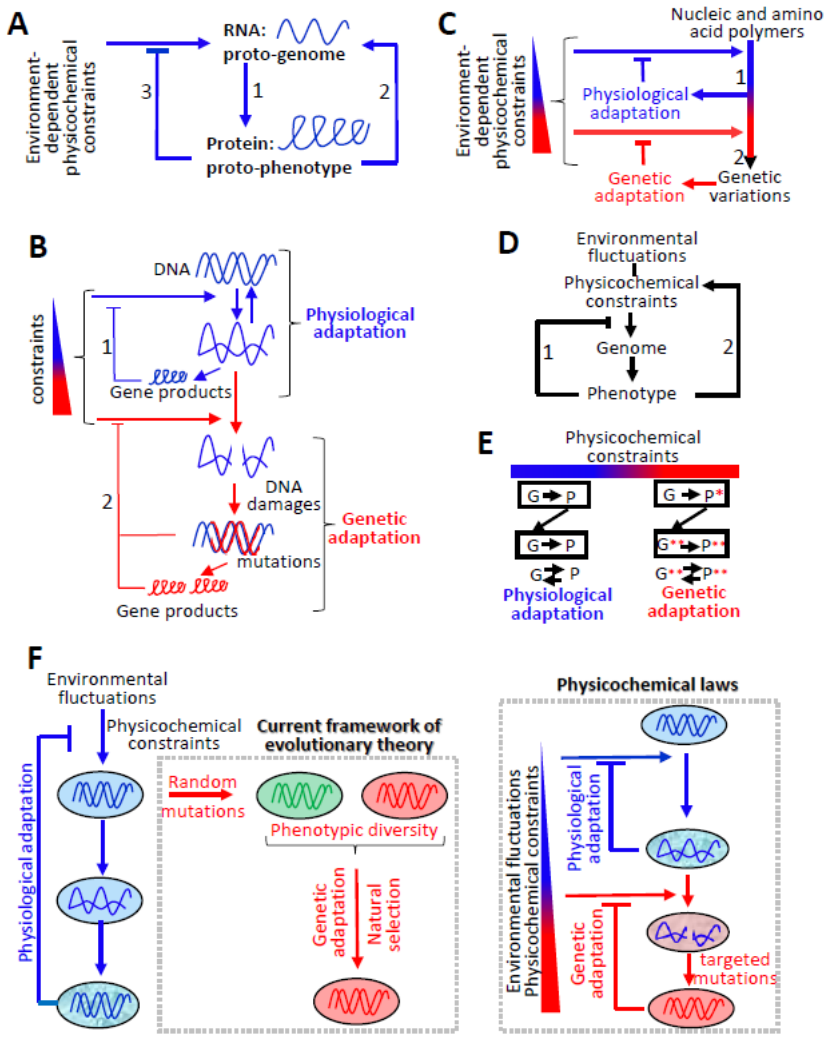
The fundamental physicochemical principles on which life and evolution rely have been established owing to the interdependence of two types of polymers (RNAs and proteins) that are complementary and share properties thanks to the evolution of the genetic code. The complexity of life organization has increased, starting from RNAs and proteins, and next evolving to unicellular (DNA, RNA, proteins, cell) and multicellular organisms (germline and somatic cell DNAs, RNAs, proteins, cell populations), accordingly to the facts that i) the physicochemical properties of each level of life organization depend on those of the lower levels and determine those of the upper levels; ii) each new level of life organization allows an increasing number and wider environment-dependent physicochemical constraints to be relaxed. For example, if muscle cells have emerged by limiting cold damages by producing heat, the energy they generate has also allowed animals to move. Mobility allows animals to be less under the yoke of physicochemical constraints in their immediate environment.

The notions of chance and natural selection are useful to highlight the fact that life has not been "created" by or for something, and that organisms are necessarily adapted to their environment. However, these notions are useless to explain how evolution proceeds. In the framework proposed in this article, evolution is not fueled by random mutations, as mutations i) are directed by cellular- and environment-dependent physicochemical constraints on biopolymers involved in organismal physiological adaptations and ii) are passed across generations when they drive a phenotype that relaxes the initiating constraints (Fig. 1D, 1). Natural selection based on organism fitness acts only as a filter, as an adapted phenotype is a phenotype that maintains the stability of its originating genome in a given environment. Finally, each new phenotype generates physicochemical constraints on its own genome. This feedback loop feeds evolution and allows complex phenotypes to emerge (Fig. 1D, 2 and Fig. 5B).

To summarize the proposed model, a genome generates a phenotype that maintains the stability of its originating genome in a stable environment where both (genome and phenotype) are reproduced identically (Fig. 1E, left panel). In an unstable environment (corresponding to variations in physicochemical parameters above a physiological range), the genome generates a phenotype that no longer maintains the stability of its originating genome and instead triggers mutations those rate, nature and location dependent on the initial physicochemical constrains and the phenotype. This process occurs until new genetic variants generate a phenotype that maintains the stability of its originating genome (Fig. 1E, right panel). This process can be slow or fast depending on the initial constraints, as the main factor relies on the generated phenotype maintaining the stability of its originating genome and therefore determining its transmission rate across generations. The more a phenotype relaxes the physicochemical constraints of the environment on biopolymers, the more stable the genome/phenotype pair is and the more likely it will be transmitted across generations in a stable environment.

Figure legends

Figure 1



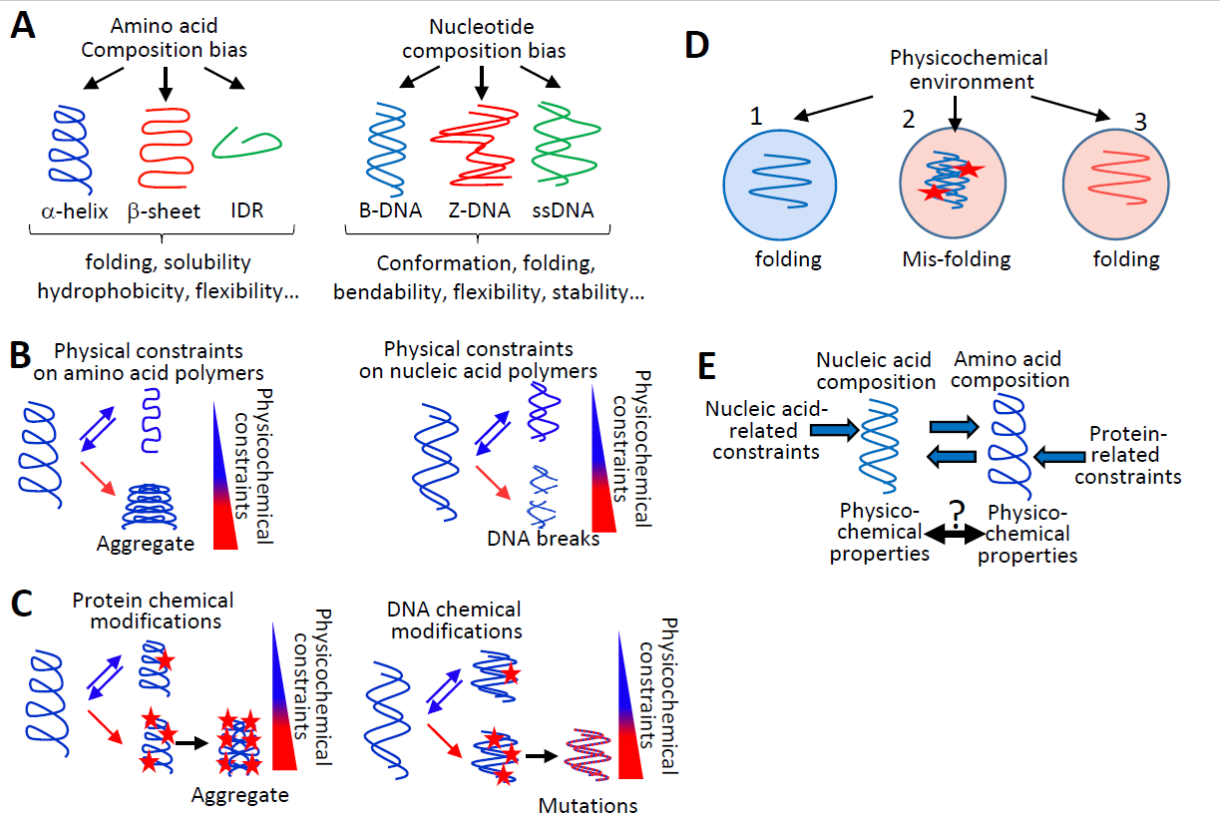
A. The origin of life likely relies on simpler forms of organization than those observed in modern living organisms. The commonly accepted hypothesis postulates that the origin of life corresponds to the emergence of polymers, such as RNAs and proteins. These two polymers are interdependent as RNAs serve as templates for protein synthesis (1) and proteins are necessary for RNA synthesis (2). This interdependence, which can be represented in the form of feedforward and feedback loops between the proto-genome (RNAs) and the proto-phenotype (proteins) can only be maintained if proteins relax environment-dependent physicochemical constraints triggering for example RNA degradation (3). This interdependence is the foundation of life and evolution.

B. Variations in the extracellular environment induce constraints on cellular polymers sensitive to these variations, triggering the cellular response that results in the biogenesis of gene products, whose activity relaxes the initial constraints (1). However, if these variations exceed a certain amplitude or persist over time, they challenge the integrity of the targeted polymers, which ultimately lead to mutations. This process stops only when new sequences (directly or indirectly) relax the initial constraints (2).

C. Variations in physicochemical parameters lead to constraints on nucleic or amino acid polymers, resulting for example in the production of new gene products whose activities relax the initial constraints (return to equilibrium). This feedback loop corresponds to the cellular physiological adaptation (1). If the initial constraints persist, they may challenge the integrity of nucleic or amino acid polymers, leading *in fine* to mutations. This process stops when new sequences (new polymers) relax (directly or indirectly) the initial constraints, corresponding to genetic adaptation (2).

D. A DNA molecule is subjected to environmental fluctuations of physicochemical parameters, which triggers the biogenesis of polymers (gene products) whose activities correspond to the phenotype. Cellular activities (the phenotype) allow a return to equilibrium by relaxing the initial constraints (1). Nevertheless, these activities also generate constraints directly or indirectly on their originating genome, meaning that a genome is adapted to the constraints generated by its own activities (2).
E. A genome (G) generates a phenotype (P) that maintains the stability of its originating genome in a stable environment and both (genome and phenotype) are reproduced identically (left panel). In an unstable environment (corresponding to variations in physicochemical parameters above a physiological range), the genome generates a phenotype (P*) that no longer maintains the stability of its originating genome and instead triggers mutations those rate, nature and location dependent on the initial constrains and the phenotype (right panel). This process occurs until new genetic variants (G**) generate a phenotype (P**) that maintains the stability of its originating genome.
F. According to the current framework of evolutionary theory (left panel), there is no direct relationship between physiological and genetic adaptation because physiological adaptation is based on physicochemical principles of homeostasis as a function of environmental fluctuations, while genetic adaptation would be fueled by random mutations generating a diversity of phenotypes on which natural selection acts. In contrast, in the model proposed in this article (right panel), genetic adaptation is the consequence of physiological adaptation. Indeed, physiological adaptation can take place as long as fluctuations in environment-dependent physicochemical parameters do not exceed a certain threshold. Above this physiological threshold, the integrity of nucleic and amino acid polymers, in particular DNA, is challenged which leads to targeted mutations. This mutational process stops when the mutations generate a phenotype that maintains the integrity of the DNA with regard to environmental constraints (genetic adaptation).

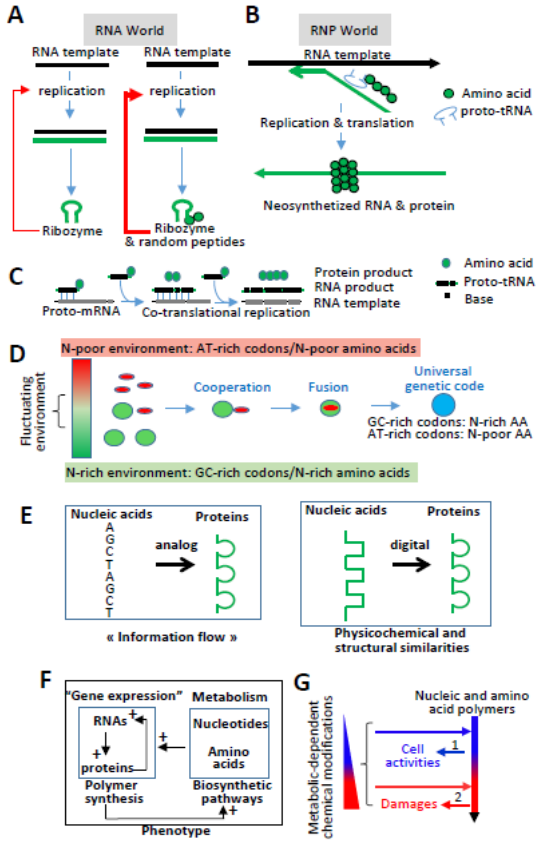
Figure 2



A. The amino acid or nucleotide composition of proteins (left) or DNA (right), respectively, determines the physicochemical properties of these polymers with consequences on their physical and chemical properties. IDR: Intrinsically Disordered Region; ssDNA: single-stranded DNA.
B. The composition of a polymer determines its physicochemical properties and therefore its folding and physical resistance to specific constraints. Depending on the composition of a given polymer, some

physicochemical constraints induce reversible structural changes (blue arrows), and others, irreversible damages (e.g., aggregation, breaks) (red arrows).
C. Chemical modifications of amino acids (left) or nucleotides (right) change the physicochemical properties of polymers. These chemical modifications are reversible (blue arrows) or induce irreversible damages (red arrows).
D. Any given polymer (blue) is stable in a physicochemical environment and unstable in another one (1 vs 2). A different polymer (red) differently reacts under the same constraints (3 vs 2).
E. Composition determines the physicochemical properties of nucleic or amino acid polymers. Nucleotide or amino acid composition is constrained by physicochemical parameters, which suggests that their composition must correspond to the same fundamental physicochemical parameters as the sequence of nucleic acid polymers determines the composition of proteins.

Figure 3



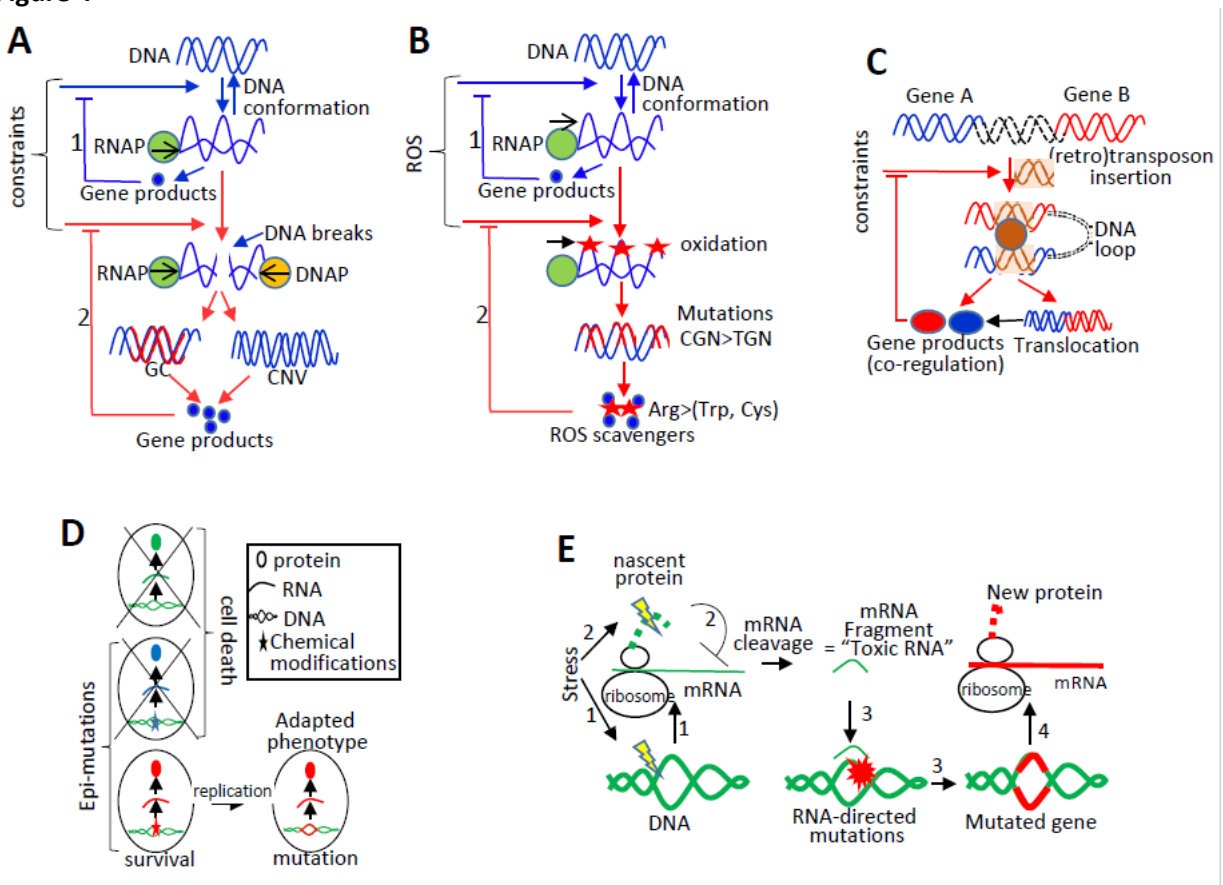
A. In an RNA world, an RNA molecule is replicated thanks to the product of replication (e.g., ribozyme, left panel). This process is enhanced by amino acids or small peptides (right panel).
B. In an RNP world, replication (RNA production) and translation (protein production) may have been performed simultaneously as in modern prokaryotes, thereby avoiding the nascent RNA to interact back to the template and increasing the probability that the neo-synthetized protein interacts with the RNA replication product.
C. Co-translational replication in an RNP world could have be performed owing to amino acids attached to proto-tRNAs that enhanced the polymerization of proto-tRNAs and that were simultaneously incorporated into the nascent protein.
D. Two different proto-cells (red and green circles) growing in different chemical environments (e.g., N-poor vs N-rich environment) may have developed different proto-genetic codes. Their cooperation in fluctuating environments could have led to horizontal transfer, leading to the emergence of the universal genetic code.
E. Nucleic acid polymers are often represented as linear strings of letters that are translated into proteins with physicochemical properties unrelated to those of nucleic acid polymers. If the genetic code has been

constrained over evolutionary time to match nucleic acid polymers and their cognate amino acid polymers to the same fundamental physicochemical constraints (e.g., temperature, element availability), nucleic and amino acid polymers share more physicochemical properties than previously anticipated.

F. In an RNP world, cooperation between interdependent polymers (i.e., RNAs and proteins) relies on their activities towards the biogenesis of nucleotides and amino acids necessary for their synthesis. The phenotype of a proto-cell in an RNP world corresponds to the polymerization of nucleotides and amino acids. The polymerization products produce nucleotides and amino acids.

G. Metabolic-dependent chemical modifications of nucleic and amino acid polymers contribute to the cell activities and to the cellular physiological adaptation in response to environment-dependent constraints (1). However, metabolic-dependent chemical modifications can also induce irreversible damages of nucleic and amino acid polymers (2).

Figure 4



A. Variations in the extracellular environment increase the expression of target genes, which are associated with physical constraints on DNA generated by RNA polymerases (RNAP). These physical constraints are transient if the gene activity relaxes the initial constraints (e.g., through the biogenesis of gene products) (1). If not, physical constraints persist, and conflicts between RNAP and DNA polymerases (DNAP) induce DNA damages. DNA damage is repaired by homologous recombination, which favors GC over AT nucleotides but which can also induce gene copy number variation (CNV). Both of these processes in turn generate more gene products and therefore relax the initial constraint (2).

B. Cellular stresses activate the expression of specific target genes while increasing the production of intracellular ROS. One strands of transcriptionally induced genes is exposed to ROS, promoting deamination of methyl cytosine, which gives rise to thymine. C>T mutations change Arg codons into Trp and Cys codons. Trp- and Cys-containing proteins may play a role in protecting cells from ROS, therefore relaxing the initial constraints (2).

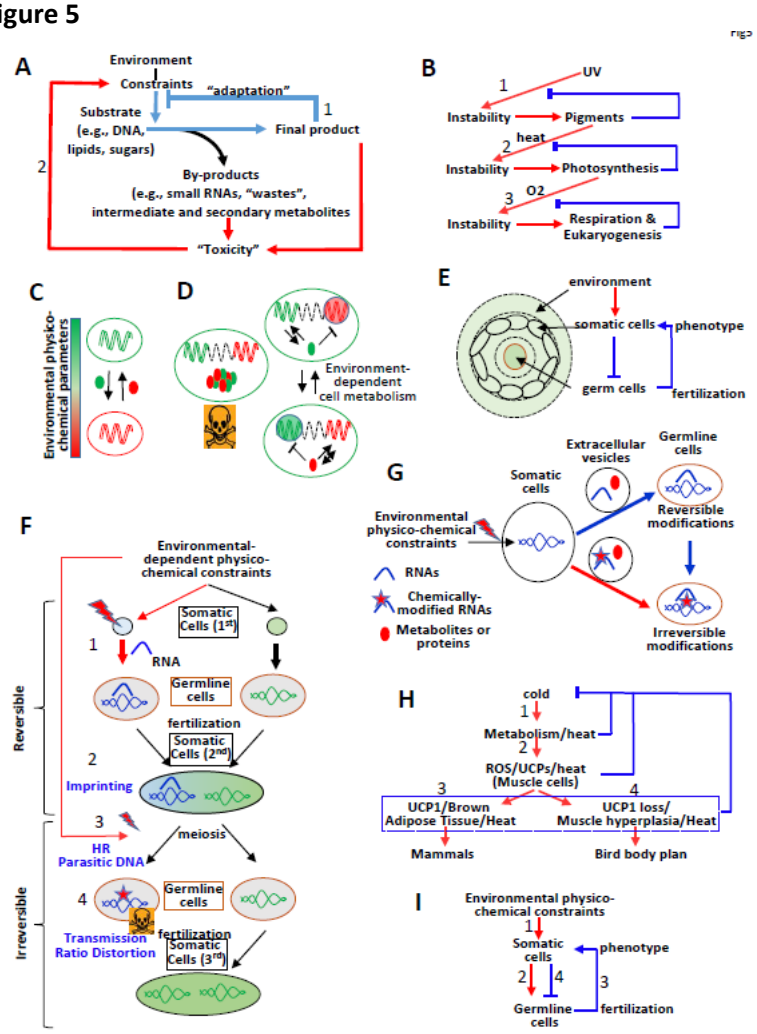
C. Transcriptional activation of genes induces neo-insertion of repeated sequences within transcription-dependent ssDNA. Neo-insertion of repeated sequences may facilitate co-regulation of two genes by

bringing them closer to each other in space, and it also promotes recombination. Both processes coordinate the production of gene products, contributing to relaxing the initial constraints.

D. Cellular stress induces chemical modifications of target genes, which affects chromatin organization and transcription fidelity ("epi-mutations"). Chemical modifications that induce biogenesis of new RNAs and proteins could allow survival of the cells in which these modifications took place. Chemical modifications in surviving cells lead to mutations that increase the survival probability of daughter cells.

E. Cellular stress induces physical constraints simultaneously on target genes and on proteins produced from these genes (1 & 2). By disrupting translation, for example by inducing nascent protein unfolding, a stress can induce translation stop and co-translational cleavage of mRNAs. RNA fragments generated during translation and/or transcription hybridize on the complementary DNA strand and locally generate DNA and/or chromatin modifications, thus increasing the probability of mutations to occur in the targeted regions (3). This process stops only when the gene and its products obtain physicochemical properties that relax the initial constraints (4).

Figure 5



A. Any biochemical reaction leads to the synthesis of a final product (1) and "waste", by-products, or secondary metabolites. These compounds are not necessarily essential to the cell's survival, but they are the result of vital cellular activities. Therefore, these compounds are not "random products" even though they can have cellular toxic effects by interacting with cellular polymers (2), which can lead to mutations.

B. UV radiations induce genomic instability favoring genomes that produce pigments that in turn protect them from the initial constraint (1). Likewise, UV radiation–absorbing pigments induce genomic instability favoring genomes that produce photosynthetic reaction centers that in turn protect them from the initial constraint (2). Likewise, O_2 production by photosynthesis induces genomic instability favoring

genomes that produce oxidases and cellular components that in turn protect them from the initial constraint (3).

C. Two unicellular organisms (green and red), which are adapted to different environment-dependent constraints, cooperate by exchanging various components in a fluctuating environment.

D. Not all enzymatic reactions generated from a genome can take place simultaneously (left panel). The selective compaction of different regions of a genome according to the cellular metabolic state (and therefore its environment) through chemical modifications of histones allow to protect some genome parts and repressing their potentially toxic expression while allowing the expression of genes whose products contribute to maintain the cellular homeostasis.

E. Somatic cells "buffer" environment-depend constraints, allowing to maintain the stability of the genome from the germ cells. When "protected" by somatic cells (i.e., when the phenotype of the organism is adapted to its environment), germ cells give rise to gametes that generate after fertilization the same somatic cells, i.e., phenotype.

F. Somatic cells are constrained by environmental fluctuations, and their activities can have consequences for germ cells, for example through the transfer of metabolites or small RNAs from somatic to germline cells (1). These compounds can change the activity of germ cells, affect the development of the body after fertilization, and cause mutations in germ cells (2 & 3). These mutations could lead to the emergence of somatic cells (3 & 4) whose activities maintain the integrity of the germ cell genome of the following generations.

G. Somatic cell genes that are constrained by environmental parameters produce extracellular vesicles containing RNAs that can either induce reversible epigenetic mutations or irreversible genetic mutations.

H. Cold induces physiological adaptation by activating cellular respiration that increases cellular heat production (1). However, an increase in cellular respiration increases ROS production, which can lead to activation of UCP proteins. This simultaneously balances the ROS genesis and increases heat production in muscle cells (2) or in brown adipose tissue (3). The loss of UCP1 in bird ancestor may have led to muscle hyperplasia for heat production with consequences on bird body plan (4).

I. Environment-dependent physicochemical constraints can challenge somatic cells (1) and therefore challenge the stability of germline DNA (red arrows), leading to germline DNA mutations (2). After fertilization, the modified germline DNA gives rise to somatic cells (the phenotype, 3) that protect the germline DNA from environment-dependent physicochemical constraints (4).

Glossary

Chemical modifications: Chemical modifications can change the physicochemical properties of monomers in nucleic and amino acid polymers (for example, a chemically modified nucleotide can "mimic" another nucleotide). Chemical modifications of nucleotides can therefore result in "transcription mutagenesis", RNA editing, recoding during translation, and genetic mutations during replication. Similarly, modified amino acids within a protein (i.e., post-translational modifications) change the physicochemical properties of the modified proteins. As chemical modifications depend on the cellular metabolic activities that change under environmental fluctuations, this implies that environment-dependent metabolic activities can change the activities and nature (i.e., sequence) of gene products. Therefore, the cellular activities depend on genomic sequences and their chemical modifications in a manner that depends on extracellular and intracellular physicochemical parameters.

Composition bias: The notion of composition bias relates to the fact that the global (or partial) composition of a polymer is different from that expected by chance. For example, the frequency of a mono-, di-, or trinucleotide may be higher than expected by chance. Nucleotide or amino acid composition biases correspond to specific physicochemical properties as each monomer has specific physicochemical properties. The notion of composition bias is used in this manuscript in tight relation with the resulting and associated physicochemical properties.

Damage and error: The word "error" is often used in biology. I do not use this term, as "error" refers to a deviation from a code. Codes are proposed by biologists to assemble a set of observations and to summarize the sum of knowledge at a time point. Therefore, deviation from a code can well be because an established code imperfectly describes the reality. I will therefore rather use the word "damage" that can be used to highlight phenomena that challenge the physical integrity of biological objects (see "Genome and genetic stability" and "Mutational bias and signature").

Emergence: Emergence describes a process of coming into existence or the fact that a biological object (e.g., polymers) made of different elements (e.g., nucleotides) has physicochemical and structural properties that are more complex than the sum of the properties of the elements that composed them.

Genome and genetic stability: Stability can refer to the physical integrity of components (e.g., genome stability) but also to maintenance over time of a particular sequence (e.g., genetic stability). Both notions are related, as the physical instability of a genome can generate sequence variations and sequence variations are the product of physicochemical modifications of DNA."

Genome and phenotype: A genome is a template that gives rise directly or indirectly to polymers (DNAs, RNAs, and proteins) in response to environment-dependent physicochemical parameters. Some of these polymers (proteins as gene products) can modify molecules supplied by the cellular environment and give rise to metabolites ("gene product products"). The phenotype corresponding to all the cellular or organism activities allows its originating template to be protected from environmental fluctuations and its integrity maintained. If not, the template changes (genome mutations) and gives rise to another phenotype, until it protects its originating genome against environmental fluctuations.

Mutational bias and signature: Mutations depend on physicochemical laws. For example, a mutating agent has specific physicochemical properties and therefore reacts with nucleotides and sequences which also have specific physicochemical properties. These reactions lead therefore to specific modifications of specific sequences. This corresponds to mutational bias and signature. Mutations are not "errors" since they are the consequences of physicochemical processes.

Box 1: Genome physical organization: from gene expression regulation to gene product functions.

A DNA molecule is a template producing polymers (i.e., RNAs) in a manner that depends on its physical and structural properties at multiple scales. At the scale of a few nucleotides, DNA structure is determined by the arrangement of bases, the nature of which determines the depth and hydropathy of the DNA grooves, as well as, local electrostatic and electronic properties, which collectively determine interactions between DNA and proteins such as transcription factors recruiting RNA polymerases[27]. At the scale of a dozen to hundreds of nucleotides, the nucleotide composition determines DNA thermodynamic- and mechanical-properties, such as resistance and response to topological stresses with consequences on transcription initiation and elongation[26,32,240]. Nucleotide composition-dependent mechanical properties, such as DNA bendability and rigidity, play a role in nucleosome positioning. The nucleosome frequency and the physical properties of the sequences between nucleosomes determines i) the state of chromatin compaction, and ii) DNA/chromatin folding (e.g., the 3D organization)[241,242]. Chromatin organization and DNA folding play a major role in transcription regulation by allowing coordinated regulation of genes, as co-regulated genes in eukaryotes are parts of 3D structural units called topologically-associated domains (TADs) that correspond to regions of several hundred thousand of Kbps that fold and are nearby in the 3D nuclear space. An emerging model is that genes hosted by the same TAD share the same chromatin organization and depend on the same transcriptional regulators that form biomolecular condensates, allowing the simultaneous and efficient transcription of co-regulated genes[133-137]. As DNA folding relies on nucleotide composition-depend physical properties that also determine the sets of interacting regulatory proteins, and as these features play a role in gene expression regulation, there is an overlap between the 1D (i.e., nucleotide composition bias) and 3D genome organization. Co-regulated genes, or genes within the same TADs, have similar nucleotide composition biases[243-246]. Of note, co-regulated genes co-evolved toward the same nucleotide composition bias[245-247].

Interestingly, coding exons can share the same nucleotide composition bias than their flanking introns and their hosting genes[112,248]. As a consequence, mRNAs produced from co-regulated genes share the same nucleotide composition bias and are likely regulated by the same set of RNAs binding proteins, that interact with nucleotide composition biased sequences, in agreement with the concept or RNA regulons[51,246,249 250]. Furthermore, since the genetic code is not random, coding sequences sharing the same nucleotide composition bias encode for proteins having the same amino acid composition bias or the same physicochemical properties (see main text). As amino acid composition determines in turn the protein physicochemical properties, co-localized proteins and/or proteins contributing to the same cellular processes have similar amino acid composition biases[251,252].

In conclusion, composition bias determines on the one hand the location of genes in the 3D nuclear space and on the other hand, gene product functions. Accordingly, genes contributing to the same cellular processes are in linear proximity in prokaryotes, and in 3D proximity in the nuclear space in eukaryotes[145-147,247,253-255]. Composition biases driving physicochemical properties of polymers therefore establish a straightforward link between production and function of gene products.

Box 2: The genetic code is not a cipher

While the genetic code is often described as a "frozen accident" and a cipher, early studies in the 1980s point out some physicochemical properties shared by amino acids and their cognate codons or anticodons. For example, the hydropathy of amino acids correlates with the hydropathy of the principal dinucleotides (excluding the wobble position) of their corresponding anticodons[45,57,58]. Along the same line, some protein physicochemical properties can be shared by their cognate mRNA. For example, the average hydrophobicity of a protein correlates with the average hydrophobicity of its cognate mRNA[249]. The physicochemical similarities between proteins and their cognate mRNAs may have played a role in the co-adaptation of these polymers to the same physicochemical parameters (see main text) but may also have been important for the physical interactions between proteins and their cognate RNAs[45,256,257]. Indeed, the ability of a protein to preferentially interact with its cognate RNA could have limited the free diffusion of proteins (see main text) and played a role in nascent protein folding[258]. Supporting a role for protein-RNA interaction in the evolution of the genetic code, i) some amino acids preferentially interact with their corresponding codons or anti-codons; ii) some RNA- and DNA-binding proteins bind preferentially to sequences owing to amino acid binding to their cognate codons; iii) mRNAs enriched in a particular nucleobase (e.g., Gs) tend to encode proteins that interact with mRNAs made of the same nucleotide; and iv) many proteins, such as ribosomal proteins, bind to their own mRNAs[257-259 256,260,261].

The physical parameters of molecular dynamics regarding nucleic and amino acid polymerization could explain the triplet-based nature of the genetic code. Indeed, it is well established that three-base codon structure of the genetic code contributes to translation efficiency and fidelity and molecular dynamics modelling suggests that charged particles (e.g., ribosomes) interacting with a polymer (e.g., an RNA) via electrostatic forces moves dynamically along the polymer in steps of three monomers[262]. Quite remarkably, there is now increasing evidence that triplets also play a major role in nucleic acid polymer properties and biogenesis: i) triplets correspond to the width of the minor groove in a double-stranded nucleic acid polymer, and backbone atoms that are in proximity across the minor grove are separated by three nucleotides on the complementary strand[263,264]; ii) a 3-base periodicity has been observed outside coding sequences and provides, for example, specificity for the positioning of the transcription preinitiation complex[265]; iii) codon bias affects transcription by affecting RNA folding, which favors transcription elongation by reducing pausing and RNA polymerase backtracking[70,266-268]; iv) intra- and inter-trinucleotide stacking interactions contribute to stabilizing base pairing during the translation process but may have also played a role in replication early in evolution[75]. Collectively, these observations suggest that the 3-base genetic code may have been constrained by physical parameters, allowing the simultaneous enhancement of RNA and protein biogenesis[262-264,269]. This would have been particularly important if both processes were physically coupled (see main text). In conclusion, physicochemical rules and parameters constrained the evolution of the genetic code that cannot be considered as a "frozen accident" but as an evolutionary process constrained by physicochemical laws.

References

1 Koonin, E. V. The Origin at 150: is a new evolutionary synthesis in sight? *Trends Genet* **25**, 473-475, doi:10.1016/j.tig.2009.09.007 (2009).

2 Noble, D., Jablonka, E., Joyner, M. J., Muller, G. B. & Omholt, S. W. Evolution evolves: physiology returns to centre stage. *J Physiol* **592**, 2237-2244, doi:10.1113/jphysiol.2014.273151 (2014).

3 Laland, K. *et al.* Does evolutionary theory need a rethink? *Nature* **514**, 161-164, doi:10.1038/514161a (2014).

4 Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**, 704-714, doi:10.1038/nrg.2016.104 (2016).

5 Tomkova, M. & Schuster-Bockler, B. DNA Modifications: Naturally More Error Prone? *Trends Genet* **34**, 627-638, doi:10.1016/j.tig.2018.04.005 (2018).

6 Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**, 213-223, doi:10.1038/nrg3890 (2015).

7 Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Bockler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol* **19**, 129, doi:10.1186/s13059-018-1509-y (2018).

8 Boulikas, T. Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J Mol Evol* **35**, 156-180, doi:10.1007/bf00183227 (1992).

9 Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393-395, doi:10.1038/ng.363 (2009).

10 Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48**, 126-133, doi:10.1038/ng.3469 (2016).

11 Danchin, E. & Pocheville, A. Inheritance is where physiology meets evolution. *J Physiol* **592**, 2307-2317, doi:10.1113/jphysiol.2014.272096 (2014).

12 Yona, A. H., Frumkin, I. & Pilpel, Y. A relay race on the evolutionary adaptation spectrum. *Cell* **163**, 549-559, doi:10.1016/j.cell.2015.10.005 (2015).

13 Noble, R. & Noble, D. Was the Watchmaker Blind? Or Was She One-Eyed? *Biology (Basel)* **6**, doi:10.3390/biology6040047 (2017).

14 Booker, T. R., Jackson, B. C. & Keightley, P. D. Detecting positive selection in the genome. *BMC Biol* **15**, 98, doi:10.1186/s12915-017-0434-y (2017).

15 Wideman, J. G., Novick, A., Munoz-Gomez, S. A. & Doolittle, W. F. Neutral evolution of cellular phenotypes. *Curr Opin Genet Dev* **58-59**, 87-94, doi:10.1016/j.gde.2019.09.004 (2019).

16 Lanfear, R., Kokko, H. & Eyre-Walker, A. Population size and the rate of evolution. *Trends Ecol Evol* **29**, 33-41, doi:10.1016/j.tree.2013.09.009 (2014).

17 Barrett, R. D. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* **12**, 767-780, doi:10.1038/nrg3015 (2011).

18 Darwin, C. On the Origin of Species. *Murray* (1859).

19 Koonin, E. V. Why the Central Dogma: on the nature of the great biological exclusion principle. *Biol Direct* **10**, 52, doi:10.1186/s13062-015-0084-3 (2015).

20 Echave, J. & Wilke, C. O. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annu Rev Biophys* **46**, 85-103, doi:10.1146/annurev-biophys-070816-033819 (2017).

21 Reed, C. J., Lewis, H., Trejo, E., Winston, V. & Evilia, C. Protein adaptations in archaeal extremophiles. *Archaea* **2013**, 373275, doi:10.1155/2013/373275 (2013).

22 Harmel, R. & Fiedler, D. Features and regulation of non-enzymatic post-translational modifications. *Nat Chem Biol* **14**, 244-252, doi:10.1038/nchembio.2575 (2018).

23 Panda, A. & Ghosh, T. C. Prevalent structural disorder carries signature of prokaryotic adaptation to oxic atmosphere. *Gene* **548**, 134-141, doi:10.1016/j.gene.2014.07.002 (2014).

24 Elser, J. J., Acquisti, C. & Kumar, S. Stoichiogenomics: the evolutionary ecology of macromolecular elemental composition. *Trends Ecol Evol* **26**, 38-44, doi:10.1016/j.tree.2010.10.006 (2011).

25 Bragg, J. G., Thomas, D. & Baudouin-Cornu, P. Variation among species in proteomic sulphur content is related to environmental conditions. *Proc Biol Sci* **273**, 1293-1300, doi:10.1098/rspb.2005.3441 (2006).

26 Vinogradov, A. E. DNA helix: the importance of being GC-rich. *Nucleic Acids Res* **31**, 1838-1844, doi:10.1093/nar/gkg296 (2003).

27 Harteis, S. & Schneider, S. Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J Mol Sci* **15**, 12335-12363, doi:10.3390/ijms150712335 (2014).

28 Dans, P. D. *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* **42**, 11304-11320, doi:10.1093/nar/gku809 (2014).

29 Goncearenco, A., Ma, B. G. & Berezovsky, I. N. Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. *Nucleic Acids Res* **42**, 2879-2892, doi:10.1093/nar/gkt1336 (2014).

30 Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**, 1998-2004, doi:10.1101/gr.1649303 (2003).

31 Urrutia, A. O. & Hurst, L. D. The signature of selection mediated by expression on human genes. *Genome Res* **13**, 2260-2264, doi:10.1101/gr.641103 (2003).

32 Reymer, A., Zakrzewska, K. & Lavery, R. Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. *Nucleic Acids Res* **46**, 1684-1694, doi:10.1093/nar/gkx1270 (2018).

33 Chen, K., Zhao, B. S. & He, C. Nucleic Acid Modifications in Regulation of Gene Expression. *Cell Chem Biol* **23**, 74-85, doi:10.1016/j.chembiol.2015.11.007 (2016).

34 Olinski, R., Gackowski, D. & Cooke, M. S. Endogenously generated DNA nucleobase modifications source, and significance as possible biomarkers of malignant transformation risk, and role in anticancer therapy. *Biochim Biophys Acta Rev Cancer* **1869**, 29-41, doi:10.1016/j.bbcan.2017.11.002 (2018).

35 Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644-656, doi:10.1016/j.cell.2017.01.002 (2017).

36 Reichenberger, E. R., Rosen, G., Hershberg, U. & Hershberg, R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol Evol* **7**, 1380-1389, doi:10.1093/gbe/evv063 (2015).

37 Seward, E. A. & Kelly, S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol* **17**, 226, doi:10.1186/s13059-016-1087-9 (2016).

38 Kelly, S. The Amount of Nitrogen Used for Photosynthesis Modulates Molecular Evolution in Plants. *Mol Biol Evol* **35**, 1616-1625, doi:10.1093/molbev/msy043 (2018).

39 Smarda, P. *et al.* Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytol* **200**, 911-921, doi:10.1111/nph.12399 (2013).

40 Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E. & Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet* **30**, 308-321, doi:10.1016/j.tig.2014.04.006 (2014).

41 Quintales, L., Soriano, I., Vazquez, E., Segurado, M. & Antequera, F. A species-specific nucleosomal signature defines a periodic distribution of amino acids in proteins. *Open Biol* **5**, 140218, doi:10.1098/rsob.140218 (2015).

42 Babbitt, G. A., Alawad, M. A., Schulze, K. V. & Hudson, A. O. Synonymous codon bias and functional constraint on GC3-related DNA backbone dynamics in the prokaryotic nucleoid. *Nucleic Acids Res* **42**, 10915-10926, doi:10.1093/nar/gku811 (2014).

43 Bailey, S. F., Hinz, A. & Kassen, R. Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nat Commun* **5**, 4076, doi:10.1038/ncomms5076 (2014).

44 Taylor, F. J. & Coates, D. The code within the codons. *Biosystems* **22**, 177-187 (1989).

45 Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the genetic code. *Proc Natl Acad Sci U S A* **55**, 966-974, doi:10.1073/pnas.55.4.966 (1966).

46 Prilusky, J. & Bibi, E. Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. *Proc Natl Acad Sci U S A* **106**, 6662-6666, doi:10.1073/pnas.0902029106 (2009).

47 Panda, A., Podder, S., Chakraborty, S. & Ghosh, T. C. GC-made protein disorder sheds new light on vertebrate evolution. *Genomics* **104**, 530-537, doi:10.1016/j.ygeno.2014.09.003 (2014).

48 Brbic, M., Warnecke, T., Krisko, A. & Supek, F. Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled. *Genome Biol Evol* **7**, 1519-1532, doi:10.1093/gbe/evv088 (2015).

49 Goncearenco, A. & Berezovsky, I. N. The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. *Biol Direct* **9**, 29, doi:10.1186/s13062-014-0029-2 (2014).

50 Warnecke, T., Weber, C. C. & Hurst, L. D. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem Soc Trans* **37**, 756-761, doi:10.1042/BST0370756 (2009).

51 Fontrodona, N. *et al.* Interplay between coding and exonic splicing regulatory sequences. *Genome Res* **29**, 711-722, doi:10.1101/gr.241315.118 (2019).

52 Faure, G., Ogurtsov, A. Y., Shabalina, S. A. & Koonin, E. V. Adaptation of mRNA structure to control protein folding. *RNA Biol* **14**, 1649-1654, doi:10.1080/15476286.2017.1349047 (2017).

53 Brunak, S. & Engelbrecht, J. Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. *Proteins* **25**, 237-252, doi:10.1002/(SICI)1097-0134(199606)25:2<237::AID-PROT9>3.0.CO;2-E (1996).

54 Trifonov, E. N., Volkovich, Z. & Frenkel, Z. M. Multiple levels of meaning in DNA sequences, and one more. *Ann N Y Acad Sci* **1267**, 35-38, doi:10.1111/j.1749-6632.2012.06589.x (2012).

55 Ponce de Leon, M., de Miranda, A. B., Alvarez-Valin, F. & Carels, N. The Purine Bias of Coding Sequences is Determined by Physicochemical Constraints on Proteins. *Bioinform Biol Insights* **8**, 93-108, doi:10.4137/BBI.S13161 (2014).

56 Tagami, S., Attwater, J. & Holliger, P. Simple peptides derived from the ribosomal core potentiate RNA polymerase ribozyme function. *Nat Chem* **9**, 325-332, doi:10.1038/nchem.2739 (2017).

57 Kun, A. & Radvanyi, A. The evolution of the genetic code: Impasses and challenges. *Biosystems* **164**, 217-225, doi:10.1016/j.biosystems.2017.10.006 (2018).

58 Koonin, E. V. & Novozhilov, A. S. Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet* **51**, 45-62, doi:10.1146/annurev-genet-120116-024713 (2017).

59 Gulik, P. T. On the Origin of Sequence. *Life (Basel)* **5**, 1629-1637, doi:10.3390/life5041629 (2015).

60 Grosjean, H. & Westhof, E. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res* **44**, 8020-8040, doi:10.1093/nar/gkw608 (2016).

61 Szostak, J. W. On the origin of life. *Medicina (B Aires)* **76**, 199-203 (2016).

62 Usui, K., Ichihashi, N. & Yomo, T. A design principle for a single-stranded RNA genome that replicates with less double-strand formation. *Nucleic Acids Res* **43**, 8033-8043, doi:10.1093/nar/gkv742 (2015).

63 Bansho, Y. *et al.* Importance of parasite RNA species repression for prolonged translation-coupled RNA self-replication. *Chem Biol* **19**, 478-487, doi:10.1016/j.chembiol.2012.01.019 (2012).

64 Eigen, M., Biebricher, C. K., Gebinoga, M. & Gardiner, W. C. The hypercycle. Coupling of RNA and protein biosynthesis in the infection cycle of an RNA bacteriophage. *Biochemistry* **30**, 11005-11018, doi:10.1021/bi00110a001 (1991).

65 Carter, C. W., Jr. & Wills, P. R. Interdependence, Reflexivity, Fidelity, Impedance Matching, and the Evolution of Genetic Coding. *Mol Biol Evol* **35**, 269-286, doi:10.1093/molbev/msx265 (2018).

66 Saad, N. A ribonucleopeptide world at the origin of life: Co-evolution of RNA. *Journal of Systematics and Evolution*, doi: 10.1111/jse.12287 (2018).

67 Francis, B. R. The Hypothesis that the Genetic Code Originated in Coupled Synthesis of Proteins and the Evolutionary Predecessors of Nucleic Acids in Primitive Cells. *Life (Basel)* **5**, 467-505, doi:10.3390/life5010467 (2015).

68 Gounaris, Y. An evolutionary theory based on a protein-mRNA co-synthesis hypothesis. *Journal of Biological Research-Thessaloniki* **15**, 3-16 (2011).

69 Gordon, K. H. Were RNA replication and translation directly coupled in the RNA (+protein?) World? *J Theor Biol* **173**, 179-193, doi:10.1006/jtbi.1995.0054 (1995).

70 Zamft, B., Bintu, L., Ishibashi, T. & Bustamante, C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc Natl Acad Sci U S A* **109**, 8948-8953, doi:10.1073/pnas.1205063109 (2012).

71 Chen, X., Yang, J. R. & Zhang, J. Nascent RNA folding mitigates transcription-associated mutagenesis. *Genome Res* **26**, 50-59, doi:10.1101/gr.195164.115 (2016).

72 Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**, 504-508, doi:10.1126/science.1184939 (2010).

73 Auboeuf, D. Alternative mRNA processing sites decrease genetic variability while increasing functional diversity. *Transcription* **9**, 75-87, doi:10.1080/21541264.2017.1373891 (2018).

74 Morgens, D. W. The protein invasion: a broad review on the origin of the translational system. *J Mol Evol* **77**, 185-196, doi:10.1007/s00239-013-9592-x (2013).

75 Attwater, J., Raguram, A., Morgunov, A. S., Gianni, E. & Holliger, P. Ribozyme-catalysed RNA synthesis using triplet building blocks. *Elife* **7**, doi:10.7554/eLife.35255 (2018).

76 Maizels, N. & Weiner, A. M. Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc Natl Acad Sci U S A* **91**, 6729-6734, doi:10.1073/pnas.91.15.6729 (1994).

77 Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* **3**, e5, doi:10.1371/journal.pcbi.0030005 (2007).

78 Granold, M., Hajieva, P., Tosa, M. I., Irimie, F. D. & Moosmann, B. Modern diversification of the amino acid repertoire driven by oxygen. *Proc Natl Acad Sci U S A* **115**, 41-46, doi:10.1073/pnas.1717100115 (2018).

79 Chowdhury, K. *et al.* Presence of a consensus DNA motif at nearby DNA sequence of the mutation susceptible CG nucleotides. *Gene* **639**, 85-95, doi:10.1016/j.gene.2017.10.001 (2018).

80 Szpiech, Z. A. *et al.* Prominent features of the amino acid mutation landscape in cancer. *PLoS One* **12**, e0183273, doi:10.1371/journal.pone.0183273 (2017).

81 Tsuber, V., Kadamov, Y., Brautigam, L., Berglund, U. W. & Helleday, T. Mutations in Cancer Cause Gain of Cysteine, Histidine, and Tryptophan at the Expense of a Net Loss of Arginine on the Proteome Level. *Biomolecules* **7**, doi:10.3390/biom7030049 (2017).

82 Son, H., Kang, H., Kim, H. S. & Kim, S. Somatic mutation driven codon transition bias in human cancer. *Sci Rep* **7**, 14204, doi:10.1038/s41598-017-14543-1 (2017).

83 Tan, H., Bao, J. & Zhou, X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Sci Rep* **5**, 12566, doi:10.1038/srep12566 (2015).

84 Suarez-Villagran, M. Y., Azevedo, R. B. R. & Miller, J. H., Jr. Influence of Electron-Holes on DNA Sequence-Specific Mutation Rates. *Genome Biol Evol* **10**, 1039-1047, doi:10.1093/gbe/evy060 (2018).

85 Bender, A., Hajieva, P. & Moosmann, B. Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria. *Proc Natl Acad Sci U S A* **105**, 16496-16501, doi:10.1073/pnas.0802779105 (2008).

86 Vetsigian, K., Woese, C. & Goldenfeld, N. Collective evolution and the genetic code. *Proc Natl Acad Sci U S A* **103**, 10696-10701, doi:10.1073/pnas.0603780103 (2006).

87 Wong, J. T., Ng, S. K., Mat, W. K., Hu, T. & Xue, H. Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life. *Life (Basel)* **6**, doi:10.3390/life6010012 (2016).

88 Di Giulio, M. The aminoacyl-tRNA synthetases had only a marginal role in the origin of the organization of the genetic code: Evidence in favor of the coevolution theory. *J Theor Biol* **432**, 14-24, doi:10.1016/j.jtbi.2017.08.005 (2017).

89 Copley, S. D., Smith, E. & Morowitz, H. J. A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proc Natl Acad Sci U S A* **102**, 4442-4447, doi:10.1073/pnas.0501049102 (2005).

90 van der Knaap, J. A. & Verrijzer, C. P. Undercover: gene control by metabolites and metabolic enzymes. *Genes Dev* **30**, 2345-2369, doi:10.1101/gad.289140.116 (2016).

91 Schvartzman, J. M., Thompson, C. B. & Finley, L. W. S. Metabolic regulation of chromatin modifications and gene expression. *J Cell Biol* **217**, 2247-2259, doi:10.1083/jcb.201803061 (2018).

92 Reid, M. A., Dai, Z. & Locasale, J. W. The impact of cellular metabolism on chromatin dynamics and epigenetics. *Nat Cell Biol* **19**, 1298-1306, doi:10.1038/ncb3629 (2017).

93 Cairns, J., Overbaugh, J. & Miller, S. The origin of mutants. *Nature* **335**, 142-145, doi:10.1038/335142a0 (1988).

94 Wright, B. E. A biochemical mechanism for nonrandom mutations and evolution. *J Bacteriol* **182**, 2993-3001, doi:10.1128/jb.182.11.2993-3001.2000 (2000).

95 Correa, R., Thornton, P. C., Rosenberg, S. M. & Hastings, P. J. Oxygen and RNA in stress-induced mutation. *Curr Genet* **64**, 769-776, doi:10.1007/s00294-017-0801-9 (2018).

96 Sebastian, R. & Oberdoerffer, P. Transcription-associated events affecting genomic integrity. *Philos Trans R Soc Lond B Biol Sci* **372**, doi:10.1098/rstb.2016.0288 (2017).

97 Wang, G. & Vasquez, K. M. Effects of Replication and Transcription on DNA Structure-Related Genetic Instability. *Genes (Basel)* **8**, doi:10.3390/genes8010017 (2017).

98 Merrikh, H. Spatial and Temporal Control of Evolution through Replication-Transcription Conflicts. *Trends Microbiol* **25**, 515-521, doi:10.1016/j.tim.2017.01.008 (2017).

99 Chen, Y. H. *et al.* Transcription shapes DNA replication initiation and termination in human cells. *Nat Struct Mol Biol* **26**, 67-77, doi:10.1038/s41594-018-0171-0 (2019).

100 Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285-311, doi:10.1146/annurev-genom-082908-150001 (2009).

101 Long, H. *et al.* Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol* **2**, 237-240, doi:10.1038/s41559-017-0425-y (2018).

102 Pozzoli, U. *et al.* Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol Biol* **8**, 99, doi:10.1186/1471-2148-8-99 (2008).

103 Kudla, G., Helwak, A. & Lipinski, L. Gene conversion and GC-content evolution in mammalian Hsp70. *Mol Biol Evol* **21**, 1438-1444, doi:10.1093/molbev/msh146 (2004).

104 Yin, Y. *et al.* Dynamics of spontaneous flipping of a mismatched base in DNA duplex. *Proc Natl Acad Sci U S A* **111**, 8043-8048, doi:10.1073/pnas.1400667111 (2014).

105 Sankar, T. S., Wastuwidyaningtyas, B. D., Dong, Y., Lewis, S. A. & Wang, J. D. The nature of mutations induced by replication-transcription collisions. *Nature* **535**, 178-181, doi:10.1038/nature18316 (2016).

106 Sueoka, N. Wide intra-genomic G+C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. *Gene* **300**, 141-154, doi:10.1016/s0378-1119(02)01046-6 (2002).

107 Chen, C. L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**, 447-457, doi:10.1101/gr.098947.109 (2010).

108 Kenigsberg, E. *et al.* The mutation spectrum in genomic late replication domains shapes mammalian GC content. *Nucleic Acids Res* **44**, 4222-4232, doi:10.1093/nar/gkw268 (2016).

109 Gul, I. S. *et al.* GC Content of Early Metazoan Genes and Its Impact on Gene Expression Levels in Mammalian Cell Lines. *Genome Biol Evol* **10**, 909-917, doi:10.1093/gbe/evy040 (2018).

110 Rao, Y. S. *et al.* Selection for the compactness of highly expressed genes in Gallus gallus. *Biol Direct* **5**, 35, doi:10.1186/1745-6150-5-35 (2010).

111 Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**, 1616-1625, doi:10.1101/gr.134445.111 (2012).

112 Lemaire, S. *et al.* Interplay between gene nucleotide composition bias and splicing. *bioRxiv 605832* (2019).

113 Quandt, E. M., Traverse, C. C. & Ochman, H. Local genic base composition impacts protein production and cellular fitness. *PeerJ* **6**, e4286, doi:10.7717/peerj.4286 (2018).

114 Gorochowski, T. E., Ignatova, Z., Bovenberg, R. A. & Roubos, J. A. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res* **43**, 3022-3032, doi:10.1093/nar/gkv199 (2015).
115 Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255-258, doi:10.1126/science.1170160 (2009).
116 Rao, Y., Wang, Z., Chai, X., Nie, Q. & Zhang, X. Hydrophobicity and aromaticity are primary factors shaping variation in amino acid usage of chicken proteome. *PLoS One* **9**, e110381, doi:10.1371/journal.pone.0110381 (2014).
117 Du, M. Z. *et al.* The GC Content as a Main Factor Shaping the Amino Acid Usage During Bacterial Evolution Process. *Front Microbiol* **9**, 2948, doi:10.3389/fmicb.2018.02948 (2018).
118 Gao, N., Lu, G., Lercher, M. J. & Chen, W. H. Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. *Sci Rep* **7**, 10572, doi:10.1038/s41598-017-11159-3 (2017).
119 Hull, R. M., Cruz, C., Jack, C. V. & Houseley, J. Environmental change drives accelerated adaptation through stimulated copy number variation. *PLoS Biol* **15**, e2001333, doi:10.1371/journal.pbio.2001333 (2017).
120 Ambriz-Avina, V., Yasbin, R. E., Robleto, E. A. & Pedraza-Reyes, M. Role of Base Excision Repair (BER) in Transcription-associated Mutagenesis of Nutritionally Stressed Nongrowing Bacillus subtilis Cell Subpopulations. *Curr Microbiol* **73**, 721-726, doi:10.1007/s00284-016-1122-9 (2016).
121 Shewaramani, S. *et al.* Anaerobically Grown Escherichia coli Has an Enhanced Mutation Rate and Distinct Mutational Spectra. *PLoS Genet* **13**, e1006570, doi:10.1371/journal.pgen.1006570 (2017).
122 Liu, H. & Zhang, J. Yeast Spontaneous Mutation Rate and Spectrum Vary with Environment. *Curr Biol* **29**, 1584-1591 e1583, doi:10.1016/j.cub.2019.03.054 (2019).
123 Maharjan, R. P. & Ferenci, T. A shifting mutational landscape in 6 nutritional states: Stress-induced mutagenesis as a series of distinct stress input-mutation output relationships. *PLoS Biol* **15**, e2001477, doi:10.1371/journal.pbio.2001477 (2017).
124 Chu, X. L. *et al.* Temperature responses of mutation rate and mutational spectrum in an Escherichia coli strain and the correlation with metabolic rate. *BMC Evol Biol* **18**, 126, doi:10.1186/s12862-018-1252-8 (2018).
125 Matsuba, C., Ostrow, D. G., Salomon, M. P., Tolani, A. & Baer, C. F. Temperature, stress and spontaneous mutation in Caenorhabditis briggsae and Caenorhabditis elegans. *Biol Lett* **9**, 20120334, doi:10.1098/rsbl.2012.0334 (2013).
126 Rogozin, I. B. *et al.* Mutational signatures and mutable motifs in cancer genomes. *Brief Bioinform* **19**, 1085-1101, doi:10.1093/bib/bbx049 (2018).
127 White, K. A. *et al.* Cancer-associated arginine-to-histidine mutations confer a gain in pH sensing to mutant proteins. *Sci Signal* **10**, doi:10.1126/scisignal.aam9931 (2017).
128 Saier, M. H., Jr., Kukita, C. & Zhang, Z. Transposon-mediated directed mutation in bacteria and eukaryotes. *Front Biosci (Landmark Ed)* **22**, 1458-1468 (2017).
129 Newman, A. G., Bessa, P., Tarabykin, V. & Singh, P. B. Activity-DEPendent Transposition. *EMBO Rep* **18**, 346-348, doi:10.15252/embr.201643797 (2017).
130 Vandecraen, J. *et al.* Zinc-Induced Transposition of Insertion Sequence Elements Contributes to Increased Adaptability of Cupriavidus metallidurans. *Front Microbiol* **7**, 359, doi:10.3389/fmicb.2016.00359 (2016).
131 Grandbastien, M. A. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim Biophys Acta* **1849**, 403-416, doi:10.1016/j.bbagrm.2014.07.017 (2015).
132 Miousse, I. R. *et al.* Response of transposable elements to environmental stressors. *Mutat Res Rev Mutat Res* **765**, 19-39, doi:10.1016/j.mrrev.2015.05.003 (2015).
133 Rada-Iglesias, A., Grosveld, F. G. & Papantonis, A. Forces driving the three-dimensional folding of eukaryotic genomes. *Mol Syst Biol* **14**, e8214, doi:10.15252/msb.20188214 (2018).
134 Meyer, S., Reverchon, S., Nasser, W. & Muskhelishvili, G. Chromosomal organization of transcription: in a nutshell. *Curr Genet* **64**, 555-565, doi:10.1007/s00294-017-0785-5 (2018).
135 Lin, Y. H., Forman-Kay, J. D. & Chan, H. S. Theories for Sequence-Dependent Phase Behaviors of Biomolecular Condensates. *Biochemistry* **57**, 2499-2508, doi:10.1021/acs.biochem.8b00058 (2018).

136 Erdel, F. & Rippe, K. Formation of Chromatin Subcompartments by Phase Separation. *Biophys J* **114**, 2262-2270, doi:10.1016/j.bpj.2018.03.011 (2018).
137 Rieder, D., Trajanoski, Z. & McNally, J. G. Transcription factories. *Front Genet* **3**, 221, doi:10.3389/fgene.2012.00221 (2012).
138 Gu, Z. *et al.* Enrichment analysis of Alu elements with different spatial chromatin proximity in the human genome. *Protein Cell* **7**, 250-266, doi:10.1007/s13238-015-0240-7 (2016).
139 Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**, 1963-1976, doi:10.1101/gr.168872.113 (2014).
140 Puc, J., Aggarwal, A. K. & Rosenfeld, M. G. Physiological functions of programmed DNA breaks in signal-induced transcription. *Nat Rev Mol Cell Biol* **18**, 471-476, doi:10.1038/nrm.2017.43 (2017).
141 Kaiser, V. B. & Semple, C. A. Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline. *Genome Biol* **19**, 101, doi:10.1186/s13059-018-1483-4 (2018).
142 Schwer, B. *et al.* Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc Natl Acad Sci U S A* **113**, 2258-2263, doi:10.1073/pnas.1525564113 (2016).
143 Roychowdhury, T. & Abyzov, A. Chromatin organization modulates the origin of heritable structural variations in human genome. *Nucleic Acids Res* **47**, 2766-2777, doi:10.1093/nar/gkz103 (2019).
144 Mellor, J., Woloszczuk, R. & Howe, F. S. The Interleaved Genome. *Trends Genet* **32**, 57-71, doi:10.1016/j.tig.2015.10.006 (2016).
145 Hurst, L. D., Pal, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**, 299-310, doi:10.1038/nrg1319 (2004).
146 Yin, Y., Zhang, H., Olman, V. & Xu, Y. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. *Proc Natl Acad Sci U S A* **107**, 6310-6315, doi:10.1073/pnas.0911237107 (2010).
147 Nutzmann, H. W., Huang, A. & Osbourn, A. Plant metabolic clusters - from genetics to genomics. *New Phytol* **211**, 771-789, doi:10.1111/nph.13981 (2016).
148 Gordon, A. J., Satory, D., Halliday, J. A. & Herman, C. Lost in transcription: transient errors in information transfer. *Curr Opin Microbiol* **24**, 80-87, doi:10.1016/j.mib.2015.01.010 (2015).
149 Bradley, C. C., Gordon, A. J. E., Halliday, J. A. & Herman, C. Transcription fidelity: New paradigms in epigenetic inheritance, genome instability and disease. *DNA Repair (Amst)*, 102652, doi:10.1016/j.dnarep.2019.102652 (2019).
150 Reid-Bayliss, K. S. & Loeb, L. A. Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations. *Proc Natl Acad Sci U S A* **114**, 9415-9420, doi:10.1073/pnas.1709166114 (2017).
151 Xu, L. *et al.* RNA polymerase II transcriptional fidelity control and its functional interplay with DNA modifications. *Crit Rev Biochem Mol Biol* **50**, 503-519, doi:10.3109/10409238.2015.1087960 (2015).
152 Morreall, J. *et al.* Evidence for Retromutagenesis as a Mechanism for Adaptive Mutation in Escherichia coli. *PLoS Genet* **11**, e1005477, doi:10.1371/journal.pgen.1005477 (2015).
153 Sekowska, A., Wendel, S., Fischer, E. C., Norholm, M. H. H. & Danchin, A. Generation of mutation hotspots in ageing bacterial colonies. *Sci Rep* **6**, 2, doi:10.1038/s41598-016-0005-4 (2016).
154 Williamson, A. K., Zhu, Z. & Yuan, Z. M. Epigenetic mechanisms behind cellular sensitivity to DNA damage. *Cell Stress* **2**, 176-180, doi:10.15698/cst2018.07.145 (2018).
155 Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101-114, doi:10.1016/j.cell.2019.02.051 (2019).
156 Elfman, J. & Li, H. Chimeric RNA in Cancer and Stem Cell Differentiation. *Stem Cells Int* **2018**, 3178789, doi:10.1155/2018/3178789 (2018).
157 Keskin, H. *et al.* Transcript-RNA-templated DNA recombination and repair. *Nature* **515**, 436-439, doi:10.1038/nature13682 (2014).
158 Yang, Y. G. & Qi, Y. RNA-directed repair of DNA double-strand breaks. *DNA Repair (Amst)* **32**, 82-85, doi:10.1016/j.dnarep.2015.04.017 (2015).

159 Shapiro, J. A. Living Organisms Author Their Read-Write Genomes in Evolution. *Biology (Basel)* **6**, doi:10.3390/biology6040042 (2017).
160 Khanduja, J. S., Calvo, I. A., Joh, R. I., Hill, I. T. & Motamedi, M. Nuclear Noncoding RNAs and Genome Stability. *Mol Cell* **63**, 7-20, doi:10.1016/j.molcel.2016.06.011 (2016).
161 Auboeuf, D. Putative RNA-Directed Adaptive Mutations in Cancer Evolution. *Transcription*, 0, doi:10.1080/21541264.2016.1221491 (2016).
162 Auboeuf, D. Genome evolution is driven by gene expression-generated biophysical constraints through RNA-directed genetic variation: A hypothesis. *Bioessays* **39**, doi:10.1002/bies.201700069 (2017).
163 Ibrahim, F., Maragkakis, M., Alexiou, P. & Mourelatos, Z. Ribothrypsis, a novel process of canonical mRNA decay, mediates ribosome-phased mRNA endonucleolysis. *Nat Struct Mol Biol* **25**, 302-310, doi:10.1038/s41594-018-0042-8 (2018).
164 Ikeuchi, K., Izawa, T. & Inada, T. Recent Progress on the Molecular Mechanism of Quality Controls Induced by Ribosome Stalling. *Front Genet* **9**, 743, doi:10.3389/fgene.2018.00743 (2018).
165 Linster, C. L., Van Schaftingen, E. & Hanson, A. D. Metabolite damage and its repair or pre-emption. *Nat Chem Biol* **9**, 72-80, doi:10.1038/nchembio.1141 (2013).
166 Mulkidjanian AY, J. W. On the origin of photosynthesis as inferred from sequence analysis. *Photosynthesis Research* **51**, 27-42 (1997).
167 Wolstencroft RD, R. J. Photosynthesis: Likelihood of Occurrence and Possibility of Detection on Earth-like Planets. *Icarus* **157**, 535-548 (2002).
168 Michaelian, K. & Simeonov, A. Fundamental molecules of life are pigments which arose and co-evolved as a response to the thermodynamic imperative of dissipating the prevailing solar spectrum *Biogeosciences* **12**, 4913–4937 (2015).
169 Degli Esposti, M., Mentel, M., Martin, W. & Sousa, F. L. Oxygen Reductases in Alphaproteobacterial Genomes: Physiological Evolution From Low to High Oxygen Environments. *Front Microbiol* **10**, 499, doi:10.3389/fmicb.2019.00499 (2019).
170 Muller, M. *et al.* Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* **76**, 444-495, doi:10.1128/MMBR.05024-11 (2012).
171 Forte, E. *et al.* The Terminal Oxidase Cytochrome bd Promotes Sulfide-resistant Bacterial Respiration and Growth. *Sci Rep* **6**, 23788, doi:10.1038/srep23788 (2016).
172 Margulis, L., Chapman, M., Guerrero, R. & Hall, J. The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. *Proc Natl Acad Sci U S A* **103**, 13080-13085, doi:10.1073/pnas.0604985103 (2006).
173 Kurland, C. G. & Andersson, S. G. Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev* **64**, 786-820, doi:10.1128/mmbr.64.4.786-820.2000 (2000).
174 Speijer, D. Alternating terminal electron-acceptors at the basis of symbiogenesis: How oxygen ignited eukaryotic evolution. *Bioessays* **39**, doi:10.1002/bies.201600174 (2017).
175 Raymond, J. & Segre, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764-1767, doi:10.1126/science.1118439 (2006).
176 Jiang, Y. Y. *et al.* The impact of oxygen on metabolic evolution: a chemoinformatic investigation. *PLoS Comput Biol* **8**, e1002426, doi:10.1371/journal.pcbi.1002426 (2012).
177 Desmond, E. & Gribaldo, S. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biol Evol* **1**, 364-381, doi:10.1093/gbe/evp036 (2009).
178 Zhang, X., Barraza, K. M. & Beauchamp, J. L. Cholesterol provides nonsacrificial protection of membrane lipids from chemical damage at air-water interface. *Proc Natl Acad Sci U S A* **115**, 3255-3260, doi:10.1073/pnas.1722323115 (2018).
179 Galea, A. M. & Brown, A. J. Special relationship between sterols and oxygen: were sterols an adaptation to aerobic life? *Free Radic Biol Med* **47**, 880-889, doi:10.1016/j.freeradbiomed.2009.06.027 (2009).
180 Deng, Y. & Almsherqi, Z. A. Evolution of cubic membranes as antioxidant defence system. *Interface Focus* **5**, 20150012, doi:10.1098/rsfs.2015.0012 (2015).

181 Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3**, MDNA3-0050-2014, doi:10.1128/microbiolspec.MDNA3-0050-2014 (2015).

182 de Lange, T. A loopy view of telomere evolution. *Front Genet* **6**, 321, doi:10.3389/fgene.2015.00321 (2015).

183 Coros, C. J., Piazza, C. L., Chalamcharla, V. R. & Belfort, M. A mutant screen reveals RNase E as a silencer of group II intron retromobility in Escherichia coli. *RNA* **14**, 2634-2644, doi:10.1261/rna.1247608 (2008).

184 Belfort, M. Mobile self-splicing introns and inteins as environmental sensors. *Curr Opin Microbiol* **38**, 51-58, doi:10.1016/j.mib.2017.04.003 (2017).

185 Friedman, K. & Heller, A. On the Non-Uniform Distribution of Guanine in Introns of Human Genes: Possible Protection of Exons against Oxidation by Proximal Intron Poly-G Sequences. *J. Phys. Chem. B* **105**, 11859-11865 (2001).

186 Amit, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* **1**, 543-556, doi:10.1016/j.celrep.2012.03.013 (2012).

187 Enright, H., Miller, W. J., Hays, R., Floyd, R. A. & Hebbel, R. P. Preferential targeting of oxidative base damage to internucleosomal DNA. *Carcinogenesis* **17**, 1175-1177, doi:10.1093/carcin/17.5.1175 (1996).

188 Colangeli, R. *et al.* The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates. *Proc Natl Acad Sci U S A* **106**, 4414-4418, doi:10.1073/pnas.0810126106 (2009).

189 Speijer, D. Birth of the eukaryotes by a set of reactive innovations: New insights force us to relinquish gradual models. *Bioessays* **37**, 1268-1276, doi:10.1002/bies.201500107 (2015).

190 Ljungman, M. & Hanawalt, P. C. Efficient protection against oxidative DNA damage in chromatin. *Mol Carcinog* **5**, 264-269 (1992).

191 Cannan, W. J., Tsang, B. P., Wallace, S. S. & Pederson, D. S. Nucleosomes suppress the formation of double-strand DNA breaks during attempted base excision repair of clustered oxidative damages. *J Biol Chem* **289**, 19881-19893, doi:10.1074/jbc.M114.571588 (2014).

192 D'Souza, G. *et al.* Ecology and evolution of metabolic cross-feeding interactions in bacteria. *Nat Prod Rep* **35**, 455-488, doi:10.1039/c8np00009c (2018).

193 Jeltsch, A. Oxygen, epigenetic signaling, and the evolution of early life. *Trends Biochem Sci* **38**, 172-176, doi:10.1016/j.tibs.2013.02.001 (2013).

194 Drinnenberg, I. A. *et al.* EvoChromo: towards a synthesis of chromatin biology and evolution. *Development* **146**, doi:10.1242/dev.178962 (2019).

195 Aravind, L., Burroughs, A. M., Zhang, D. & Iyer, L. M. Protein and DNA modifications: evolutionary imprints of bacterial biochemical diversification and geochemistry on the provenance of eukaryotic epigenetics. *Cold Spring Harb Perspect Biol* **6**, a016063, doi:10.1101/cshperspect.a016063 (2014).

196 Rokas, A. The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet* **42**, 235-251, doi:10.1146/annurev.genet.42.110807.091513 (2008).

197 Michod, R. E. Evolution of individuality during the transition from unicellular to multicellular life. *Proc Natl Acad Sci U S A* **104 Suppl 1**, 8613-8618, doi:10.1073/pnas.0701489104 (2007).

198 Kaiser, D. Building a multicellular organism. *Annu Rev Genet* **35**, 103-123, doi:10.1146/annurev.genet.35.102401.090145 (2001).

199 Kupiec, J. J. A Darwinian theory for the origin of cellular differentiation. *Mol Gen Genet* **255**, 201-208, doi:10.1007/s004380050490 (1997).

200 Bernstein, H., Bernstein, C. & Richard E. Michod. Meiosis as an Evolutionary Adaptation for DNA Repair. *DNA Repair, Inna Kruman, IntechOpen*, doi:DOI: 10.5772/25117. (2011).

201 Horandl, E. & Speijer, D. How oxygen gave rise to eukaryotic sex. *Proc Biol Sci* **285**, doi:10.1098/rspb.2017.2706 (2018).

202 Poljsak, B., Milisav, I., Lampe, T. & Ostan, I. Reproductive benefit of oxidative damage: an oxidative stress "malevolence"? *Oxid Med Cell Longev* **2011**, 760978, doi:10.1155/2011/760978 (2011).

203 Immler, S. & Otto, S. P. The Evolutionary Consequences of Selection at the Haploid Gametic Stage. *Am Nat* **192**, 241-249, doi:10.1086/698483 (2018).

204 Wahl, M. E. & Murray, A. W. Multicellularity makes somatic differentiation evolutionarily stable. *Proc Natl Acad Sci U S A* **113**, 8362-8367, doi:10.1073/pnas.1608278113 (2016).
205 Goldsby, H. J., Knoester, D. B., Ofria, C. & Kerr, B. The evolutionary origin of somatic cells under the dirty work hypothesis. *PLoS Biol* **12**, e1001858, doi:10.1371/journal.pbio.1001858 (2014).
206 Karimi, J., Goodarzi, M. T., Tavilani, H., Khodadadi, I. & Amiri, I. Increased receptor for advanced glycation end products in spermatozoa of diabetic men and its association with sperm nuclear DNA fragmentation. *Andrologia* **44 Suppl 1**, 280-286, doi:10.1111/j.1439-0272.2011.01178.x (2012).
207 Sies, H. Strategies of antioxidant defense. *Eur J Biochem* **215**, 213-219, doi:10.1111/j.1432-1033.1993.tb18025.x (1993).
208 Sales, V. M., Ferguson-Smith, A. C. & Patti, M. E. Epigenetic Mechanisms of Transmission of Metabolic Disease across Generations. *Cell Metab* **25**, 559-571, doi:10.1016/j.cmet.2017.02.016 (2017).
209 Vanhees, K., Vonhogen, I. G., van Schooten, F. J. & Godschalk, R. W. You are what you eat, and so are your children: the impact of micronutrients on the epigenetic programming of offspring. *Cell Mol Life Sci* **71**, 271-285, doi:10.1007/s00018-013-1427-9 (2014).
210 da Silveira, J. C. *et al.* Cell-secreted vesicles containing microRNAs as regulators of gamete maturation. *J Endocrinol* **236**, R15-R27, doi:10.1530/JOE-17-0200 (2018).
211 Chen, Q. *et al.* Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science* **351**, 397-400, doi:10.1126/science.aad7977 (2016).
212 Chen, Q., Yan, W. & Duan, E. Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. *Nat Rev Genet* **17**, 733-743, doi:10.1038/nrg.2016.106 (2016).
213 Danchin, E., Pocheville, A. & Huneman, P. Early in life effects and heredity: reconciling neo-Darwinism with neo-Lamarckism under the banner of the inclusive evolutionary synthesis. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180113, doi:10.1098/rstb.2018.0113 (2019).
214 Klosin, A. & Lehner, B. Mechanisms, timescales and principles of trans-generational epigenetic inheritance in animals. *Curr Opin Genet Dev* **36**, 41-49, doi:10.1016/j.gde.2016.04.001 (2016).
215 Horsthemke, B. A critical view on transgenerational epigenetic inheritance in humans. *Nat Commun* **9**, 2973, doi:10.1038/s41467-018-05445-5 (2018).
216 Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat Rev Genet* **20**, 235-248, doi:10.1038/s41576-018-0092-0 (2019).
217 Zhang, Y. *et al.* Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. *Nat Cell Biol* **20**, 535-540, doi:10.1038/s41556-018-0087-2 (2018).
218 Hollick, J. B. Paramutation and related phenomena in diverse species. *Nat Rev Genet* **18**, 5-23, doi:10.1038/nrg.2016.115 (2017).
219 Fishman, L. & McIntosh, M. Standard Deviations: The Biological Bases of Transmission Ratio Distortion. *Annu Rev Genet*, doi:10.1146/annurev-genet-112618-043905 (2019).
220 Tock, A. J. & Henderson, I. R. Hotspots for Initiation of Meiotic Recombination. *Front Genet* **9**, 521, doi:10.3389/fgene.2018.00521 (2018).
221 Brachet, E., Sommermeyer, V. & Borde, V. Interplay between modifications of chromatin and meiotic recombination hotspots. *Biol Cell* **104**, 51-69, doi:10.1111/boc.201100113 (2012).
222 Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol* **17**, 241, doi:10.1186/s13059-016-1110-1 (2016).
223 Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* **48**, 935-939, doi:10.1038/ng.3597 (2016).
224 Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-1442, doi:10.1016/j.cell.2012.11.019 (2012).
225 Skinner, M. K., Guerrero-Bosagna, C. & Haque, M. M. Environmentally induced epigenetic transgenerational inheritance of sperm epimutations promote genetic mutations. *Epigenetics* **10**, 762-771, doi:10.1080/15592294.2015.1062207 (2015).

226 Guerrero-Bosagna, C. *et al.* Transgenerational epigenetic inheritance in birds. *Environ Epigenet* **4**, dvy008, doi:10.1093/eep/dvy008 (2018).
227 Wurdinger, T. *et al.* Extracellular vesicles and their convergence with viral pathways. *Adv Virol* **2012**, 767694, doi:10.1155/2012/767694 (2012).
228 Koonin, E. V. Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol Direct* **12**, 5, doi:10.1186/s13062-017-0177-2 (2017).
229 Durdevic, Z. & Schaefer, M. Dnmt2 methyltransferases and immunity: an ancient overlooked connection between nucleotide modification and host defense? *Bioessays* **35**, 1044-1049, doi:10.1002/bies.201300088 (2013).
230 Rechavi, O. Guest list or black list: heritable small RNAs as immunogenic memories. *Trends Cell Biol* **24**, 212-220, doi:10.1016/j.tcb.2013.10.003 (2014).
231 Zheng, Y., Lorenzo, C. & Beal, P. A. DNA editing in DNA/RNA hybrids by adenosine deaminases that act on RNA. *Nucleic Acids Res* **45**, 3369-3377, doi:10.1093/nar/gkx050 (2017).
232 Zhang, X., Cozen, A. E., Liu, Y., Chen, Q. & Lowe, T. M. Small RNA Modifications: Integral to Function and Disease. *Trends Mol Med* **22**, 1025-1034, doi:10.1016/j.molmed.2016.10.009 (2016).
233 Blagojevic, D. P., Grubor-Lajsic, G. N. & Spasic, M. B. Cold defence responses: the role of oxidative stress. *Front Biosci (Schol Ed)* **3**, 416-427 (2011).
234 Speijer, D. Being right on Q: shaping eukaryotic evolution. *Biochem J* **473**, 4103-4127, doi:10.1042/BCJ20160647 (2016).
235 Oelkrug, R., Goetze, N., Meyer, C. W. & Jastroch, M. Antioxidant properties of UCP1 are evolutionarily conserved in mammals and buffer mitochondrial reactive oxygen species. *Free Radic Biol Med* **77**, 210-216, doi:10.1016/j.freeradbiomed.2014.09.004 (2014).
236 Rowland, L. A., Bal, N. C. & Periasamy, M. The role of skeletal-muscle-based thermogenic mechanisms in vertebrate endothermy. *Biol Rev Camb Philos Soc* **90**, 1279-1297, doi:10.1111/brv.12157 (2015).
237 Nowack, J., Giroud, S., Arnold, W. & Ruf, T. Muscle Non-shivering Thermogenesis and Its Role in the Evolution of Endothermy. *Front Physiol* **8**, 889, doi:10.3389/fphys.2017.00889 (2017).
238 Newman, S. A. Form and function remixed: developmental physiology in the evolution of vertebrate body plans. *J Physiol* **592**, 2403-2412, doi:10.1113/jphysiol.2014.271437 (2014).
239 Newman, S. A., Mezentseva, N. V. & Badyaev, A. V. Gene loss, thermogenesis, and the origin of birds. *Ann N Y Acad Sci* **1289**, 36-47, doi:10.1111/nyas.12090 (2013).
240 Il'icheva, I. A. *et al.* Structural features of DNA that determine RNA polymerase II core promoter. *BMC Genomics* **17**, 973, doi:10.1186/s12864-016-3292-z (2016).
241 Todolli, S., Perez, P. J., Clauvelin, N. & Olson, W. K. Contributions of Sequence to the Higher-Order Structures of DNA. *Biophys J* **112**, 416-426, doi:10.1016/j.bpj.2016.11.017 (2017).
242 Travers, A. & Muskhelishvili, G. DNA structure and function. *FEBS J* **282**, 2279-2295, doi:10.1111/febs.13307 (2015).
243 Ramirez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**, 189, doi:10.1038/s41467-017-02525-w (2018).
244 Jabbari, K. & Bernardi, G. An Isochore Framework Underlies Chromatin Architecture. *PLoS One* **12**, e0168023, doi:10.1371/journal.pone.0168023 (2017).
245 Lian, S. *et al.* Intrachromosomal colocalization strengthens co-expression, co-modification and evolutionary conservation of neighboring genes. *BMC Genomics* **19**, 455, doi:10.1186/s12864-018-4844-1 (2018).
246 Bessiere, C. *et al.* Probing instructions for expression regulation in gene nucleotide compositions. *PLoS Comput Biol* **14**, e1005921, doi:10.1371/journal.pcbi.1005921 (2018).
247 Yin, H., Wang, G., Ma, L., Yi, S. V. & Zhang, Z. What Signatures Dominantly Associate with Gene Age? *Genome Biol Evol* **8**, 3083-3089, doi:10.1093/gbe/evw216 (2016).
248 Fuertes, M. A., Rodrigo, J. R. & Alonso, C. Do Intron and Coding Sequences of Some Human-Mouse Orthologs Evolve as a Single Unit? *J Mol Evol* **82**, 247-250, doi:10.1007/s00239-016-9746-8 (2016).

249 Polyansky, A. A., Hlevnjak, M. & Zagrovic, B. Analogue encoding of physicochemical properties of proteins in their cognate messenger RNAs. *Nat Commun* **4**, 2784, doi:10.1038/ncomms3784 (2013).
250 Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-543, doi:10.1038/nrg2111 (2007).
251 Cascarina, S. M. & Ross, E. D. Proteome-scale relationships between local amino acid composition and protein fates and functions. *PLoS Comput Biol* **14**, e1006256, doi:10.1371/journal.pcbi.1006256 (2018).
252 Wang, T. & Tang, H. The physical characteristics of human proteins in different biological functions. *PLoS One* **12**, e0176234, doi:10.1371/journal.pone.0176234 (2017).
253 Karathia, H., Kingsford, C., Girvan, M. & Hannenhalli, S. A pathway-centric view of spatial proximity in the 3D nucleome across cell lines. *Sci Rep* **6**, 39279, doi:10.1038/srep39279 (2016).
254 Paz, A., Frenkel, S., Snir, S., Kirzhner, V. & Korol, A. B. Implications of human genome structural heterogeneity: functionally related genes tend to reside in organizationally similar genomic regions. *BMC Genomics* **15**, 252, doi:10.1186/1471-2164-15-252 (2014).
255 Tsochatzidou, M., Malliarou, M., Papanikolaou, N., Roca, J. & Nikolaou, C. Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of Saccharomyces cerevisiae. *Nucleic Acids Res* **45**, 5818-5828, doi:10.1093/nar/gkx198 (2017).
256 Hlevnjak, M. & Zagrovic, B. Malleable nature of mRNA-protein compositional complementarity and its functional significance. *Nucleic Acids Res* **43**, 3012-3021, doi:10.1093/nar/gkv166 (2015).
257 Nahalka, J. Protein-RNA recognition: cracking the code. *J Theor Biol* **343**, 9-15, doi:10.1016/j.jtbi.2013.11.006 (2014).
258 Biro, J. C. Coding nucleic acids are chaperons for protein folding: a novel theory of protein folding. *Gene* **515**, 249-257, doi:10.1016/j.gene.2012.12.048 (2013).
259 Yarus, M. The Genetic Code and RNA-Amino Acid Affinities. *Life (Basel)* **7**, doi:10.3390/life7020013 (2017).
260 de Ruiter, A. & Zagrovic, B. Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments. *Nucleic Acids Res* **43**, 708-718, doi:10.1093/nar/gku1344 (2015).
261 Root-Bernstein, R. & Root-Bernstein, M. The ribosome as a missing link in prebiotic evolution II: Ribosomes encode ribosomal proteins that bind to common regions of their own mRNAs and rRNAs. *J Theor Biol* **397**, 115-127, doi:10.1016/j.jtbi.2016.02.030 (2016).
262 Aldana-Gonzalez, M., Cocho, G., Larralde, H. & Martinez-Mekler, G. Translocation properties of primitive molecular machines and their relevance to the structure of the genetic code. *J Theor Biol* **220**, 27-45, doi:10.1006/jtbi.2003.3108 (2003).
263 Babbitt, G. A. *et al.* Triplet-Based Codon Organization Optimizes the Impact of Synonymous Mutation on Nucleic Acid Molecular Dynamics. *J Mol Evol* **86**, 91-102, doi:10.1007/s00239-018-9828-x (2018).
264 Taghavi, A., van der Schoot, P. & Berryman, J. T. DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase. *Q Rev Biophys* **50**, e15, doi:10.1017/S0033583517000130 (2017).
265 Goldshtein, M. & Lukatsky, D. B. Specificity-Determining DNA Triplet Code for Positioning of Human Preinitiation Complex. *Biophys J* **112**, 2047-2050, doi:10.1016/j.bpj.2017.04.023 (2017).
266 Lukacisin, M., Landon, M. & Jajoo, R. Sequence-specific thermodynamic properties of nucleic acids influence both transcriptional pausing and backtracking in yeast. *PLoS One* **12**, e0174066, doi:10.1371/journal.pone.0174066 (2017).
267 Trotta, E. Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Res* **41**, 9382-9395, doi:10.1093/nar/gkt740 (2013).
268 Dai, Z. & Dai, X. Gene expression divergence is coupled to evolution of DNA structure in coding regions. *PLoS Comput Biol* **7**, e1002275, doi:10.1371/journal.pcbi.1002275 (2011).
269 Bosaeus, N. *et al.* A stretched conformation of DNA with a biological role? *Q Rev Biophys* **50**, e11, doi:10.1017/S0033583517000099 (2017).