OPINION

# Error prevention and mitigation as forces in the evolution of genes and genomes

*Tobias Warnecke and Laurence D. Hurst*

Abstract | Why are short introns rarely a multiple of three nucleotides long? Why do essential genes cluster? Why are genes in operons often lined up in the order in which they are needed in the encoded pathway? In this Opinion article, we argue that these and many other — ostensibly disparate — observations are all pieces of an emerging picture in which multiple aspects of gene anatomy and genome architecture have evolved in response to error-prone gene expression.

Faithful transmission of information is crucial for life. This is perhaps most evident when genetic instructions are passed on from parents to their offspring. If the information is corrupted at this stage, the result — depending on the sequence affected — may be lethal. A complex suite of mechanisms is in place to avoid detrimental effects of this kind, starting from the inbuilt capacity of DNA polymerases to backtrack and proof-read their own output. However, the fidelity of information transmission is not only crucial during replication, but also during the day-to-day running of the cell. Information that is encoded in the DNA must be read-off and processed to yield biological effector molecules. This is typically a multi-step process, which provides ample opportunities for errors to be made along the way. A gene can be transcribed at the wrong time or at levels that are too high or too low to achieve optimal functioning; it can also be mistranscribed, mis-spliced or mistranslated. Once translated, the protein can misfold, localize incorrectly or be activated or degraded too soon or too late.

To safeguard the integrity of biological information, the cellular machines that decode that information (such as the ribosome and the spliceosome) operate with intrinsically high fidelity[1], and there are many quality control pathways that detect and eliminate erroneous gene products. Exactly how many errors occur and are caught after the act, however, depends on the gene that is being expressed. Some transcripts are particularly susceptible to accidental frameshifts during translation, others habitually escape quality control so that errors may go unnoticed.

> “multiple facets of gene anatomy and genome architecture may be adaptations to error-prone gene expression”

In this Opinion article, we highlight how both gene anatomy (that is, the composition and structure of genes and their products) and genome architecture (that is, the arrangement of genes in the genome) have evolved to reduce the rate at which errors occur (error prevention) or to limit deleterious effects if an error has already been made (error mitigation) (FIG. 1). Drawing on case studies from, among other systems, *Escherichia coli*, humans, *Saccharomyces cerevisiae* and *Paramecium tetraurelia*, we first discuss how selection has moulded gene anatomies to facilitate high-fidelity information transmission or to enable faulty products to be intercepted by quality control mechanisms. Thereafter, we argue that non-random genome architecture — from the composition of local gene neighbourhoods to the differential distribution of genes across chromosomes — may also frequently reflect selection against erroneous gene expression. We focus specifically on how cells prevent transcript levels from falling below a critical threshold in the face of stochastic gene expression. Our aim is not to present a comprehensive inventory of genomic adaptations to erroneous gene expression. For the most part, we do not discuss adaptations relating to protein stability and misfolding, which have been well reviewed recently[2]. Neither do we address the creative potential of error-prone gene expression, in which a lack of fidelity in gene expression generates natural variation that can result in increased fitness in novel environments[2]. Rather, we focus on a few examples that illustrate the diversity of molecular signatures that are associated with error management, while highlighting current progress and areas for future research.

## The role of gene anatomy

***Preventing faulty gene products.*** Arguably the most extensive support for a role of error prevention in the evolution of gene anatomy comes from the study of synonymous codon usage. Codons — both individually and in the context of their neighbours — differ in their propensity to be mistranslated or to induce frameshifts. Consistent with selection to reduce errors during translation, functionally important and structurally sensitive sites are enriched for less error-prone codons in a taxonomically diverse range of species (reviewed in REF. 2). Beyond individual codons, certain codon combinations also seem to be avoided. Notably, protein-coding sequence in *S. cerevisiae, E. coli,* and *Caenorhabditis elegans* is depleted for mononucleotide repeats[3]. This signature is absent in introns, supporting selective avoidance, rather than mutational bias, as the causative mechanism. As mononucleotide repeats are prone to slippage during transcription[4] and translation[5], the most

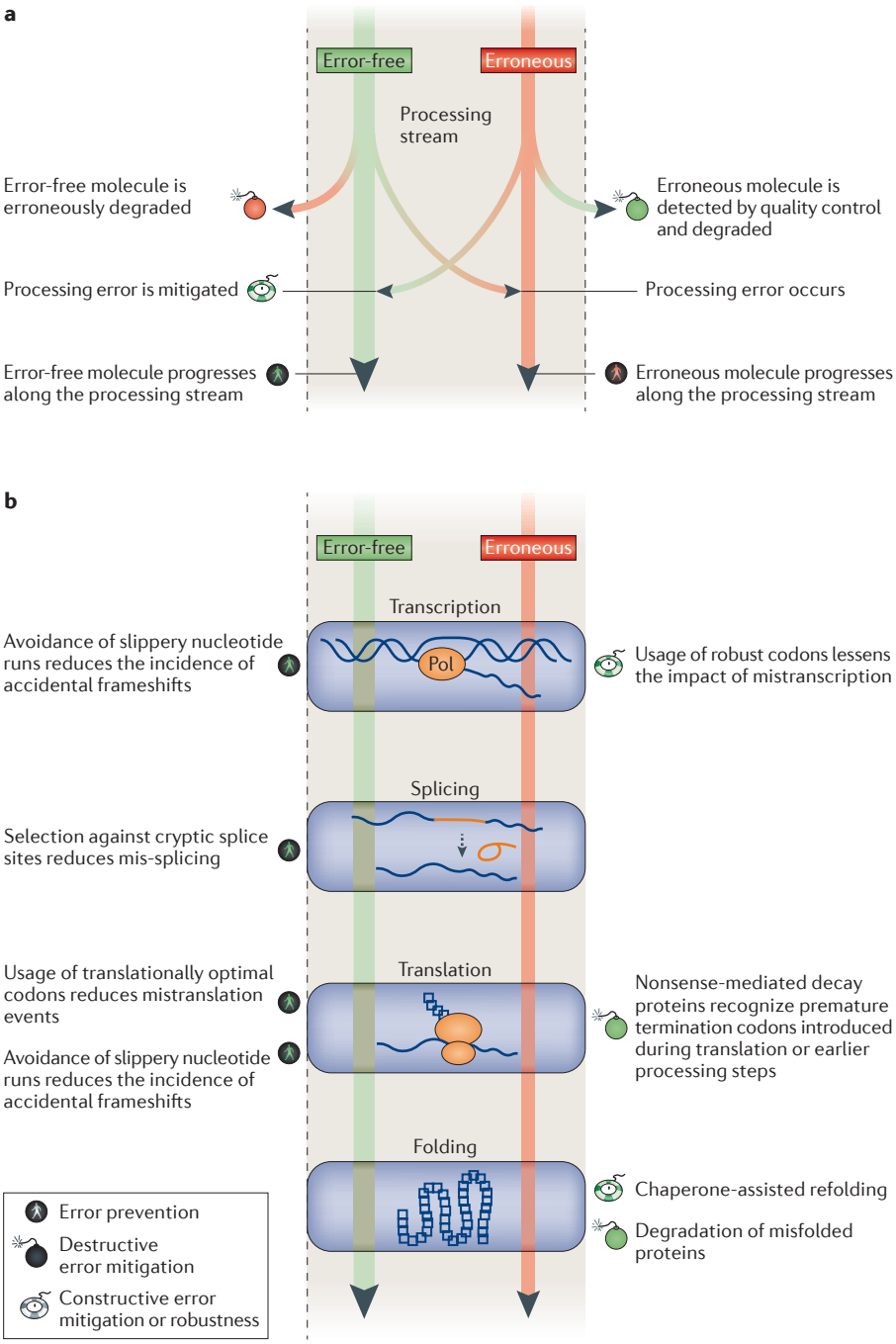


Figure 1 | **Error prevention and mitigation, from transcription to protein folding. a** | At any point during gene expression, the relevant expression product is either error-free or has accumulated one or more errors. Both error-free and erroneous intermediates can progress further along the same processing stream or be degraded. In addition, new errors can be acquired and previous errors mitigated. For example, this can occur when chaperones unfold or disaggregate proteins that had initially failed to fold correctly, so that error-free gene products can become erroneous and vice versa. These alternative processing fates have different consequences for fitness, with presumably beneficial and detrimental fates shown by green and red symbols, respectively. Features of the gene that promote error-free processing constitute adaptations for error prevention. Conversely, we can speak of error mitigation when an error has already occurred, but that error is either corrected outright (termed constructive mitigation), as in the chaperone example above, or its impact is reduced, such as through targeting an erroneous transcript for degradation (termed destructive mitigation). **b** | Some key steps in the expression of protein-coding genes, showing examples of error prevention, as well as constructive and destructive mitigation (see the main text for details). Pol, RNA polymerase.

parsimonious explanation is selection against error-prone nucleotide composition. However, whether selection principally operates to reduce transcription or translation errors remains unclear.

***Dealing with faulty gene products.*** Even if an error fails to be prevented, the fitness consequences that ensue may be minimal. Famously, neighbouring triplets in the genetic code tend to specify biochemically similar amino acids, so that single-nucleotide substitutions rarely lead to radical amino acid replacements[6]. This property — which may reflect past selection for an error-mitigation capacity or may constitute a by-product of genetic code evolution[7] — makes the code robust not only to genetic mutations but also to transcriptional and translational errors[8], for which misreading often involves a single base.

However, not all single-nucleotide changes have a minor impact on functionality. Notably, errors that create premature termination codons (PTCs; also known as nonsense mutations) are unlikely to yield functional protein products. Translation of PTC-containing transcripts can have serious repercussions for organismal fitness by wasting translational resources and generating potentially toxic truncated peptides[9]. In eukaryotes, deleterious effects of PTC-containing transcripts are mitigated by the intervention of a dedicated quality control system. During the pioneer round of translation, the nonsense-mediated decay (NMD) machinery recognizes that the PTC occurs too early (relative to specific downstream features that differ across species) and triggers degradation of the transcript, thus precluding the production of truncated and potentially deleterious protein products[10].

Intriguingly, the presence of NMD seems to have systematically affected the evolution of gene anatomies. The most striking evidence for this comes from the ciliate *P. tetraurelia*. Most introns in its genome (>96%) are very short (<34 nucleotides), and transcriptome analysis has revealed high rates (~1%) of intron retention[11]. Curiously, fewer than expected of these short introns have a length that is divisible by three nucleotides (*3n*), and those that do are more likely to harbour in-frame stop codons[11]. Why would this be? Introns that are not *3n* lead to frameshifts when accidentally retained in the mRNA; their retention is likely to generate a PTC in the new reading frame[12] and thus render these erroneous transcripts subject to

NMD. By contrast, introns that neither contain in-frame PTCs nor cause PTCs by inducing a frameshift escape detection by NMD and may be repeatedly translated into erroneous proteins. The scarcity of *3n* introns lacking stop codons — observed not only in *P. tetraurelia* but also in species as diverse as humans, *Arabidopsis thaliana* and the fungus *Yarrowia lipolytica*[11,13] — strongly suggests that intron length and composition have been shaped by selection to ensure that mis-spliced transcripts are recognized by NMD.

Some researchers additionally suggest that selection has favoured the retention of nucleotide triplets that encode out-of-frame stop codons (known as 'ambush codons') and therefore terminate translation when a frameshift occurs upstream[14]. However, unless GC content is very high[15,16], translation is usually terminated shortly after an accidental frameshift, even in the absence of such triplets. This is because, at least in the human genome, common codons are often those that generate a partial stop codon when a 1 base pair frameshift occurs in the 5′ or 3′ direction (FIG. 2). In humans, after a frameshift, only ~15 further amino acids are translated on average before the ribosome encounters a stop codon in the new reading frame[12]. As the cost savings through dedicated ambush codons would therefore typically be very small, it remains questionable how many, if any, out-of-frame termination triplets are actually maintained by selection to provide ambush functionality.

Some classes of transcripts are unable to trigger NMD when a PTC is introduced and seem to have evolved alternative features to reduce the fitness cost of faulty gene products. In mammals, to be recognized by the major NMD pathway the PTC must be located some distance upstream of the last exon-junction complex, which is deposited during splicing. Therefore, intronless transcripts have a problem: when a PTC is generated (for example, during transcription) these transcripts cannot trigger NMD. Do these genes therefore have to bear a higher error load? Recent evidence suggests not. Intronless genes in mammals instead use fewer codons for which the introduction of a single incorrect nucleotide during transcription would result in a stop codon[17]. Thus, when errors cannot be mitigated by one method, their impact may be alleviated through a complementary route, or selection may operate to reduce their incidence by promoting the fixation of less error-prone states.

| | | | |
|---|---|---|---|
| TTT (Phe) | TCT (Ser) | TAT (Tyr) | TGT (Cys) |
| TTC (Phe) | TCC (Ser) | TAC (Tyr) | TGC (Cys) |
| TTA (Leu) +1 | TCA (Ser) | TAA (Stop) | TGA (Stop) |
| TTG (Leu) +1 | TCG (Ser) | TAG (Stop) | TGG (Trp) |
| CTT (Leu) | CCT (Pro) | CAT (His) | CGT (Arg) |
| CTC (Leu) | CCC (Pro) | CAC (His) | CGC (Arg) |
| CTA (Leu) +1 | CCA (Pro) | CAA (Gln) | CGA (Arg) |
| CTG (Leu) +1 | CCG (Pro) | CAG (Gln) | CGG (Arg) |
| ATT (Ile) | ACT (Thr) | AAT (Asn) −1 | AGT (Ser) −1 |
| ATC (Ile) | ACC (Thr) | AAC (Asn) −1 | AGC (Ser) −1 |
| ATA (Ile) +1 | ACA (Thr) | AAA (Lys) −1 | AGA (Arg) −1 |
| ATG (Met) +1 | ACG (Thr) | AAG (Lys) −1 | AGG (Arg) −1 |
| GTT (Val) | GCT (Ala) | GAT (Asp) −1 | GGT (Gly) |
| GTC (Val) | GCC (Ala) | GAC (Asp) −1 | GGC (Gly) |
| GTA (Val) +1 | GCA (Ala) | GAA (Glu) −1 | GGA (Gly) |
| GTG (Val) +1 | GCG (Ala) | GAG (Glu) −1 | GGG (Gly) |

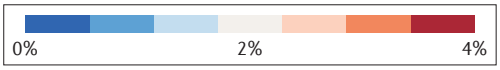


Figure 2 | **Common codons encode partial termination signals.** A heat map illustrating codon usage frequencies in the human genome. When a 1 base pair frame-shift occurs in either the upstream (-1) or downstream (+1) direction, many codons can form part of a stop codon in the new reading frame, which terminates translation. These codons are significantly more abundant on average than those that do not introduce a partial stop codon ($P = 0.025$, logistic regression). Codon usage frequencies are for human nuclear genes and are taken from the Kazusa Codon Usage Database.

Following the same logic, *E. coli* genes differ in their propensity to use translationally optimal codons depending on whether or not their protein products are clients of the chaperonin GroEL; genes encoding obligate GroEL clients are relatively depleted of optimal codons[18]. This is consistent with obligate GroEL substrates experiencing selective relief because the chaperone can mitigate at least some of the deleterious effects of mistranslation-induced protein misfolding.

***Unravelling error-adapted gene anatomy.*** Dissecting differential interactions between individual genes and quality control machineries arguably constitutes the most informative route to understanding adaptive gene anatomy because it can reveal subsets of substrates that behave unlike others and can form the cornerstone for critical tests. Importantly, the need for studying interactions goes beyond considering the cellular machineries that carry out processing and quality control. Ancillary interactions (those that are not directly involved in generating the error) nonetheless shape error propensities and hence the need for adaptive solutions. Consider transcripts that are populated by RNA-binding proteins, such as those deposited during splicing. Some of these transcripts may — based on sequence information alone — seem to be liable to erroneous interactions but the transcripts are actually protected from errors by their binding partners. It seems, for example, that many cryptic polyadenylation sites in human transcripts can persist without deleterious consequences because they are rendered unusable by the nearby binding of the U1 small nuclear ribonucleoprotein (snRNP), which prevents cleavage[19]. Integrating increasingly comprehensive interaction data sets to capture the complex context of gene expression will therefore be essential for understanding adaptive gene anatomy.

***The known unknowns.*** In both humans and *P. tetraurelia* the proportion of transcripts that are estimated to contain splicing errors is comparatively high (with lower-bound estimates in the region of 1% of transcripts[11,20]). Similarly high or higher error rates have been reported for other steps of transcript processing; these findings

Box 1 | **Distinguishing functional and aberrant isoforms — the major challenge**

With deep-sequencing platforms providing ample raw data for analysing erroneous gene expression, a major challenge is to discriminate aberrant from functional transcript isoforms. Three criteria are frequently used to assess likely functionality: rarity, evolutionary conservation and characteristic sequence features. Each of these criteria has drawbacks, particularly when they are used individually.

**Rarity**

Very rare transcript isoforms are sometimes assumed to be erroneously produced[20]. Although this can be a useful approximation, there are obvious pitfalls to equating erroneous with rare isoforms. Many transcripts with crucial biological functions, including many of those that encode transcription factors, are present at low levels, and sometimes very low levels. Conversely, not everything that is common is necessarily functional. If error mitigation is metabolically cheap and errors are not immediately deleterious (that is, cells can tolerate the continued presence of the erroneous product), erroneous isoforms might be more ubiquitous than is commonly assumed.

**Conservation**

Conservation of transcripts across species has also been used to categorize isoforms according to likely functionality. Many splicing isoforms are poorly conserved, strengthening the argument that mis-splicing is widespread[46,47]. Inevitably, this approach yields some false positives, that is, isoforms that look like errors but are not; these isoforms, being species-specific, might be particularly interesting for understanding the phenotypic variation between species. There will also be false negatives; these are isoforms that do not look like errors but are. If error-prone states are maintained by pleiotropy (BOX 3), errors might be frequent, systematic and systematically conserved. However, the rate of such false negatives is completely unknown.

**Characteristic sequence features**

For some expression processes there are characteristic sequence features that are thought to indicate that the isoform was produced in error. The presence of premature termination codons (PTCs) is regarded as a strong indicator that something has gone awry. Yet even PTC-containing isoforms cannot be automatically classified as errors. For example, nonsense-mediated decay (NMD)-targeted isoforms of serine- and arginine-rich (SR) protein-encoding genes are highly conserved across mammals, and their products participate in auto-regulatory feedback loops[48] making their production functional rather than erroneous. For primary transcripts, comparing their sequences to the DNA template can reveal the presence of transcription errors. However, observed discrepancies might largely be technical (that is, sequencing errors), and post-transcriptional modifications (such as RNA editing) need to be ruled out.

might be indicative of efficient downstream error mitigation. In yeast, the fraction of transcripts that are polyadenylated prematurely, and therefore lack a stop codon, may be as high as 10%[21]. In the absence of a stop codon, attempts by the ribosome to translate the poly(A) tail lead to mRNA degradation and translational repression. This suggests that poly(A) tails function as part of an error-control system[22,23]. However, most incorrectly polyadenylated transcripts never reach the ribosome because they are degraded by the nuclear exosome at the site of transcription[24]. Despite the high apparent error rate, we know little about the sequence features that are involved in this mitigation process. Targeted knockdown of exosome components in conjunction with high-throughput sequencing might shed light on this issue.

For yet other stages of transcript processing, even basic error estimates are lacking. For example, how many phosphorylation or dephosphorylation events happen off-target or at the wrong time? Importantly, this is not simply a problem of quantification. The major challenge is distinguishing erroneous from functional isoforms (BOX 1), especially when (ostensible) telltale signs like PTCs are absent or unknown.

## The role of genome architecture

Genome architecture (that is, the order, spacing and orientation of genes in the genome) can be highly non-random[25]. In part, this reflects the action of selection. Genes, through recombination, retrotransposition or similar processes, repeatedly sample different genomic locations and, over evolutionary time, come to reside in locations that confer high fitness. In bacteria, non-random gene order is primarily due to the fact that gene expression is organized into polycistronic transcripts, with genes participating in the same biochemical pathway or protein complex often being colocated in the same operon. However, pathway-based clustering of genes cannot explain every aspect of non-random genome architecture. For example, in *E. coli* and other bacteria essential genes cluster around the origin of replication and preferentially reside on the leading strand[26]. Proximity to the origin may be adaptive because it enhances the expression of core genes during multiple concurrent rounds of replication. More pertinently, preferential location on the leading strand is considered beneficial because it prevents the transcribing RNA polymerase from colliding head-on with the DNA polymerase during replication.

***Controlling stochasticity.*** Ensuring that the production of transcripts does not suddenly cease or fluctuate violently is not only a challenge during cell division when the replication machinery competes for access to the DNA. Cells must regularly adjust the expression of some genes without disturbing the expression of others, especially those that are sensitive to changes in dosage. In addition, molecular binding dynamics — such as between transcription factors (or chromatin remodelling complexes) and DNA — are intrinsically stochastic[27]. As a result, gene product levels may fall below (or rise above) a critical threshold. However, the degree of stochasticity (noise) in the expression levels of individual genes can vary dramatically[28], thus highlighting the possibility that noise is an evolvable trait[27,29]. Stochasticity can be reduced by making use of specific promoter architectures (prominently, an absence of TATA motifs is associated with low-noise genes[30,31]), by raising the overall expression level[28], by increasing gene copy number[27,32] or by altering genetic network wiring to include noise-abating feedback loops[33,34]. In the remainder of this section we argue that, in addition, genome architecture has been moulded by selection at several scales of organization to dampen noise.

***Noise-abating genome architecture.*** Genes with similar noise tolerance are not randomly scattered across chromosomes but instead form clusters. Notably, in yeast, noise-sensitive genes (both essential and non-essential) cluster together[35]. These clusters are located in domains of open chromatin, suggesting that the noisiness of individual genes is, at least in part, determined by regional chromatin states. This is consistent with observations that, in both mammals and yeast, neighbouring genes have correlated bursting kinetics, and that transgenes adopt the bursting kinetics of their new host domain[32,36]. Bursting is the pulse-like, non-continuous mode of transcript production in which periods of active transcription are

interspersed by periods of inactivity, and these kinetics may be a general feature of transcriptional activity. Indeed, a model in which genes are allowed to recombine into domains of differential noise recreates the observed clustering of essential genes in 'quiet' domains[35].

Selection to dampen noise may also influence gene order and orientation at even finer levels. Pairs of genes that are head-to-head in orientation — that is, those that are transcribed divergently from the same promoter but from different strands — show reduced noise[30,37]. This is probably because the use of a shared bidirectional promoter generates a mutually reinforcing chromatin micro-environment that leads to reduced stochastic fluctuations[37]. Similarly, leaky transcription can be modulated by the expression of antisense transcripts, which often share their promoter with a downstream sense gene[37,38]. Consistent with expectations, this type of organization is more common for genes that are expected to be more sensitive to noise, such as essential genes and those that encode proteins that participate in complexes[37].

The fine-scale organization of some bacterial operons is also consistent with selection for noise abatement. Lovdok *et al.*[39] compared the structures of operons across bacteria. They found that some local gene arrangements in operons encoding the chemotaxis pathway are much more highly conserved than others. This is surprising: genes residing in the same operon are transcriptionally coupled, so why should there be selection to maintain a particular order? Intriguingly, the affected genes show strong coupling at the level of translation in *E. coli*, which the authors show reduces noise in the output that is generated by the pathway[39]. It therefore seems likely that these 'neighbourhoods' are preferentially conserved because they buffer pathway output against fluctuations in the concentration of individual proteins.

Noise also seems to have influenced gene order in some metabolic operons where, curiously, genes often lie in the order in which they are required in the corresponding biochemical pathway (a phenomenon termed colinearity[40]). Again, this is surprising given that the genes are transcriptionally coupled. However, it is consistent with a model in which, at low transcription rates, pathways occasionally collapse (because, owing to stochastic effects, a crucial component of the pathway is completely absent), but these pathways are more easily restarted when gene order is colinear[40]. Colinearity is indeed exclusive to operons that are expressed at low levels[40].

### Box 2 | The evolutionary context of error-proofing

In order to understand why some genes have error-adaptive features yet others do not, we need to take into account the population genetic context in which genes and genomes have evolved.

In particular, the leverage of selection can vary substantially within and between genomes. Within genomes, with other factors being equal, selection is stronger for more highly expressed genes; this accounts for the strong link between the expression level and the degree of optimal codon usage and the higher splicing fidelity of these genes[2,20].

Between genomes, differences in effective population sizes ($N_e$) will affect the leverage of selection. For example, the scarcity of optimal codons in obligate endosymbionts is commonly attributed to stronger drift due to small population sizes[49]. Similarly, recent evidence suggests that the average stability of proteins is reduced in small populations[50]. The short-term effect of this lower protein stability could be a higher error load owing to elevated rates of protein unfolding and undesirable interactions following the exposure of hydrophobic surfaces. However, in the long term, increased interactivity might facilitate the evolution of a more complex and versatile protein interactome[50].

Does a reduced $N_e$ inevitably compromise the capacity to mitigate errors? Not necessarily. Although selection may become too weak to promote the adaptation of individual genes, selection on system-wide mitigation mechanisms, such as chaperones and nonsense-mediated decay (NMD) proteins, may actually become stronger because of the elevated workload from multiple increasingly poorly adapted substrates. Indeed, the prevalence, in small populations, of global versus local solutions to error mitigation was recently predicted by evolutionary modelling[51] and is consistent with the overexpression of GroEL in *Buchnera* and other endosymbionts (see REF. 49 and references therein).

Beyond individual genes, the evolution of broad trends in genome architecture, such as genome size and the number of genes in the genome, has also been attributed to differences in effective population size[52]. However, to what extent adaptive genome architecture, such as the clustering of noise-sensitive genes, breaks down under reduced $N_e$ remains largely unexplored.

## Does X-treme noise foster relocation?

In *C. elegans*, in which genome-wide RNA interference (RNAi) knockdown data are available, only 5.6% of X-linked genes are essential compared with 12.8% of autosomal genes[41]. X chromosomes in mammals and *C. elegans* are also depleted for genes that are haploinsufficient in yeast, although the same is not true for *Drosophila melanogaster*[42]. Why might essential genes avoid X chromosomes? The answer may, as above, in part be related to noise.

Genes that are haploid regarding the number of chromosomal copies from which they are expressed (haploid-expressed genes) are expected to be high-noise genes[43]. This is because stochastic fluctuations in transcript production are more effectively dampened if a second target for transcription is present. As predicted, haploid-expressed human autosomal genes seems to be especially noisy[44]. In mammals, both sexes are effectively haploid for the X chromosome, either by virtue of being male or following the inactivation of one copy in females. As essential and haploinsufficient genes tend to be low-noise genes[35], a simple rationalization is thus that essential genes are selected to avoid the high noise context of the haploid-expressed X chromosome.

This hypothesis is very much speculation. More stringent support is needed to reinforce the hypothesis that noise has contributed to inter-chromosomal differences in gene content and that higher ploidy, as predicted by theory[29], confers fitness benefits by reducing noise. One useful experiment would be to induce ploidy differences *de novo* in a suitable organism and to subsequently assay noise. If the above hypothesis is correct, we would expect polyploidization to be associated with reduced variability in the expression of individual genes. The observation that in plants of the genus *Senecio*, polyploidization of artificial hybrids globally reduces variance between different genes[45] is intriguing in this regard.

## Conclusions

We have argued that multiple facets of gene anatomy and genome architecture may be adaptations to error-prone gene expression. Concerning genome architecture, the evolution of non-random gene neighbourhoods might often reflect selection to prevent detrimental stochastic fluctuations in transcript levels. Current evidence remains largely limited to a few model species; however, in order to assess the relative role of noise in shaping genome architecture more broadly, it will be imperative to obtain comparative measures of noise and dosage sensitivity in a wider range of organisms. In addition, we

Box 3 | **Pleiotropy and the maintenance of error**

Minimal error rates do not necessarily equate to optimal fitness because reducing the incidence of errors usually comes at a cost. The textbook example of such cases is translation, in which there is an intrinsic speed–accuracy trade-off that governs the interactions between ribosomes and mRNAs[53,54]. Higher ribosomal accuracy is easily evolved but often selected against, because the resulting slower protein production has a net negative effect on fitness[55].

A second type of trade-off concerns the coding potential of the information carrier (such as DNA or mRNA). Notably, protein-coding sequences not only specify amino acid content but also encode additional information about translational speed, RNA secondary structure and regulatory binding sites[56,57]. Error-prone sites may be maintained because a more accurate alternative causes a net fitness loss by compromising information that is unrelated to protein-coding (for example, by abrogating an exonic splicing enhancer[58]).

The nature and severity of these trade-offs remains largely uncharacterized but should vary substantially across species depending, for example, on the growth strategy of the organism. The principal implication here is that high error levels may frequently be optimal and generate a persistent error reservoir even at large effective population sizes. We therefore speculate that, regardless of the capacity of selection to purge weakly deleterious mutations at individual sites (as discussed in REF. 51), error mitigation might often be favoured during evolution. In turn, efficient error mitigation can dramatically lower the cost of gene expression errors and thereby alleviate the severity of pleiotropic trade-offs. This may have been an important factor in the evolution of regulatory complexity, which — founded on combinatorial control — typically comes at a cost of making occasional errors[59,60]. The presence of error mitigation can also permit rapid sampling of phenotypic space in processes such as V(D)J recombination, which generates a substantial number of non-functional and potentially harmful isoforms.

anticipate that the study of genome organization in three-dimensional space, powered by high-resolution chromatin capture techniques, will provide crucial insights into adaptive interactions between genome architecture and expression processes. In particular, it will be interesting to explore whether the spatial segregation of transcription foci in the nucleus limits erroneous interactions (for example, by confining promiscuously binding transcription factors to a defined nuclear domain).

In relation to gene anatomy, we highlighted several cases in which there is strong evidence that selection has acted on transcript structure and composition to reduce the fitness burden of erroneous gene products. For many stages of gene expression, however, we remain ignorant about whether there are sequence-level adaptations to facilitate error prevention or mitigation. To advance our understanding in this regard, it will be crucial to combine system-level molecular interaction data with knowledge about the evolutionary regime that governs fixation probabilities (BOX 2) and the structure of pleiotropy in the system (BOX 3).

Finally, as modern sequencing technologies continue to unearth increasing numbers of rare isoforms, understanding the molecular signatures of error adaptation may yield valuable clues for understanding what is functionally relevant diversity and what is not.

*Tobias Warnecke is at the Centre for Genomic Regulation (CRG) and UPF, Carrer Dr. Aiguader 88, 08003 Barcelona, Spain.*

*Laurence D. Hurst is at the Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK.*

*e-mails:* tobias.warnecke@crg.eu; bssldh@bath.ac.uk

doi:10.1038/nrg3092

1. Fox-Walsh, K. L. & Hertel, K. J. Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl Acad. Sci. USA* **106**, 1766–1771 (2009).
2. Drummond, D. A. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nature Rev. Genet.* **10**, 715–724 (2009).
3. Ackermann, M. & Chao, L. DNA Sequences shaped by selection for stability. *PLoS Genet.* **2**, e22 (2006).
4. Wagner, L. A., Weiss, R. B., Driscoll, R., Dunn, D. S. & Gesteland, R. F. Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli*. *Nucleic Acids Res.* **18**, 3529–3535 (1990).
5. Weiss, R. B., Dunn, D. M., Atkins, J. F. & Gesteland, R. F. Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 687–693 (1987).
6. Woese, C. R. On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 1546–1552 (1965).
7. Massey, S. E. A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* **67**, 510–516 (2008).
8. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
9. Khajavi, M., Inoue, K. & Lupski, J. R. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur. J. Hum. Genet.* **14**, 1074–1081 (2006).
10. Maquat, L. E. & Carmichael, G. G. Quality control of mRNA function. *Cell* **104**, 173–176 (2001).
11. Jaillon, O. *et al.* Translational control of intron splicing in eukaryotes. *Nature* **451**, 359–362 (2008).
12. Itzkovitz, S. & Alon, U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* **17**, 405–412 (2007).
13. Mekouar, M. *et al.* Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biol.* **11**, R65 (2010).
14. Seligmann, H. & Pollock, D. D. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* **23**, 701–705 (2004).
15. Warnecke, T., Huang, Y., Przytycka, T. M. & Hurst, L. D. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. *Genome Biol. Evol.* **2**, 636–645 (2010).
16. Clarke, C. H. The consequences of base-pair substitution mutations in AT- and GC-rich bacteria. *J. Theor. Biol.* **105**, 117–131 (1983).
17. Cusack, B. P., Arndt, P. F., Duret, L. & Crollius, H. R. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet.* **7**, e1002276 (2011).
18. Warnecke, T. & Hurst, L. D. GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. *Mol. Syst. Biol.* **6**, 340 (2010).
19. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).
20. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* **6**, e1001236 (2010).
21. Frischmeyer, P. A. *et al.* An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science* **295**, 2258–2261 (2002).
22. Ito-Harashima, S., Kuroha, K., Tatematsu, T. & Inada, T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* **21**, 519–524 (2007).
23. Ito, K. *et al.* RUNX3, a novel tumor suppressor, is frequently inactivated in gastric cancer by protein mislocalization. *Cancer Res.* **65**, 7743–7750 (2005).
24. Hilleren, P., McCarthy, T., Rosbash, M., Parker, R. & Jensen, T. H. Quality control of mRNA 3′-end processing is linked to the nuclear exosome. *Nature* **413**, 538–542 (2001).
25. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nature Rev. Genet.* **5**, 299–310 (2004).
26. Rocha, E. P. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genet.* **34**, 377–378 (2003).
27. Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–2013 (2005).
28. Newman, J. R. *et al.* Single-cell proteomic analysis of, *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
29. Wang, Z. & Zhang, J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl Acad. Sci. USA* **108**, E67–E76 (2011).
30. Woo, Y. H. & Li, W. H. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proc. Natl Acad. Sci. USA* **108**, 3306–3311 (2011).
31. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084–1091 (2008).
32. Becskei, A., Kaufmann, B. B. & van Oudenaarden, A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genet.* **37**, 937–944 (2005).
33. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
34. Kollmann, M., Løvdok, L., Bartholomé, K., Timmer, J. & Sourjik, V. Design principles of a bacterial signalling network. *Nature* **438**, 504–507 (2005).
35. Batada, N. N. & Hurst, L. D. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genet.* **39**, 945–949 (2007).
36. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
37. Wang, G. Z., Lercher, M. J. & Hurst, L. D. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol. Evol.* **3**, 320–331 (2011).
38. Xu, Z. *et al.* Antisense expression increases gene expression variability and locus interdependency. *Mol. Syst. Biol.* **7**, 468 (2011).

39. Lovdok, L. *et al.* Role of translational coupling in robustness of bacterial chemotaxis pathway. *PLoS Biol.* **7**, e1000171 (2009).
40. Kovacs, K., Hurst, L. D. & Papp, B. Stochasticity in protein levels drives colinearity of gene order in metabolic operons of *Escherichia coli*. *PLoS Biol.* **7**, e1000115 (2009).
41. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
42. de Clare, M., Pir, P. & Oliver, S. G. Haploinsufficiency and the sex chromosomes from yeasts to humans. *BMC Biol.* **9**, 15 (2011).
43. Cook, D. L., Gerber, A. N. & Tapscott, S. J. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc. Natl Acad. Sci. USA* **95**, 15641–15646 (1998).
44. Yin, S. Y. *et al.* Dosage compensation on the active X chromosome minimizes transcriptional noise of X-linked genes in mammals. *Genome Biol.***10**, R74 (2009).
45. Hegarty, M. J. *et al.* Transcriptome shock after interspecific hybridization in *senecio* is ameliorated by genome duplication. *Curr. Biol.* **16**, 1652–1659 (2006).
46. Melamud, E. & Moult, J. Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37**, 4873–4886 (2009).
47. Tress, M. L. *et al.* The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA* **104**, 5495–5500 (2007).
48. Lareau, L., Brooks, A., Soergel, D., Meng, Q. & Brenner, S. in *Alternative Splicing in the Postgenomic Era* (eds Blencowe, B. & Graveley, B.) 191–212 (Landes Biosciences, Austin, Texas, 2007).
49. Wernegreen, J. J. & Moran, N. A. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol. Biol. Evol.* **16**, 83–97 (1999).
50. Fernández, A. & Lynch, M. Non-adaptive origins of interactome complexity. *Nature* **474**, 502–505 (2011).
51. Rajon, E. & Masel, J. Evolution of molecular error rates and the consequences for evolvability. *Proc. Natl Acad. Sci. USA* **108**, 1082–1087 (2011).
52. Lynch, M. *The origins of genome architecture* (Sinauer Associates, Sunderland, Massachusetts, 2007).
53. Thompson, R. C. & Karim, A. M. The accuracy of protein biosynthesis is limited by its speed: high fidelity selection by ribosomes of aminoacyl-tRNA ternary complexes containing GTP[γS]. *Proc. Natl Acad. Sci. USA* **79**, 4922–4926 (1982).
54. Wohlgemuth, I., Pohl, C. & Rodnina, M. V. Optimization of speed and accuracy of decoding in translation. *EMBO J.* **29**, 3701–3709 (2010).
55. Ruusala, T., Andersson, D., Ehrenberg, M. & Kurland, C. G. Hyper-accurate ribosomes inhibit growth. *EMBO J.* **3**, 2575–2580 (1984).
56. Itzkovitz, S., Hodis, E. & Segal, E. Overlapping codes within protein-coding sequences. *Genome Res.* **20**, 1582–1589 (2010).
57. 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 (2009).
58. Warnecke, T. & Hurst, L. D. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol. Biol. Evol.* **24**, 2755–2762 (2007).
59. Boue, S., Letunic, I. & Bork, P. Alternative splicing and evolution. *Bioessays* **25**, 1031–1034 (2003).
60. Doma, M. K. & Parker, R. RNA quality control in eukaryotes. *Cell* **131**, 660–668 (2007).



**Acknowledgements**
T.W. is a recipient of a European Molecular Biology Organization (EMBO) Long-term Fellowship. L.D.H. is a Royal Society Wolfson Research Merit Award Holder.


**Competing interests statement**
The authors declare no competing financial interests.

**FURTHER INFORMATION**
**Tobias Warnecke's homepage:** http://big.crg.cat/people/twarnecke
**Laurence D. Hurst's homepage:** http://people.bath.ac.uk/bssldh/LaurenceDHurst/Home.html
**Kazusa Codon Usage Database:** http://www.kazusa.or.jp/codon

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**