Published in final edited form as:

DNA Repair (Amst). 2016 August ; 44: 81–86. doi:10.1016/j.dnarep.2016.05.010.

Genome stability versus transcript diversity

Brian Magnuson^{1,2}, Karan Bedi¹, and Mats Ljungman^{1,2,*}

¹Department of Radiation Oncology, University of Michigan Comprehensive Cancer Center, and Translational Oncology Program, University of Michigan, Ann Arbor, USA

²Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, USA

Abstract

Our genome is protected from the introduction of mutations by high fidelity replication and an extensive network of DNA damage response and repair mechanisms. However, the expression of our genome, via RNA and protein synthesis, allows for more diversity in translating genetic information. In addition, the splicing process has become less stringent over evolutionary time allowing for a substantial increase in the diversity of transcripts generated. The result is a diverse transcriptome and proteome, which harbor selective advantages over a more tightly regulated system. Here, we describe mechanisms in place that both safeguard the genome and promote translational diversity, with emphasis on post-transcriptional RNA processing.

Keywords

RNA splicing; RNA transcription; protein translation; post-transcriptional RNA processing; DNA damage response; DNA repair; polymerase fidelity; proteomic diversity

1. Introduction

Heritable, molecular alterations of genomic information, or mutations, are the primary driver of evolutionary change among organisms and are responsible for the diversity of living things. Quality control in genome replication is paramount; yet the faithful expression of genetic material, via RNA and protein synthesis, must also be assured. These processes have an inherent error rate, determined by polymerase fidelity, error sensing mechanisms, and damage repair, all of which can feed back as selective pressure for genetic change. Between cell divisions DNA must be maintained, lest environmental factors, such as radiation and chemical mutagens, permanently alter the genetic information it carries.

*Correspondence and request for materials should be addressed to: ljungman@umich.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest
None.

Mutations can be advantageous, neutral, or deleterious, depending on their fitness or pathogenic effects (mutational rates and range of effects have been reviewed well in the past [1] but also updated recently [2]. Somatic cell mutations, along with non-mutational epigenetic changes, can cause disease such as cancer without passing along such genetic changes to progeny. Sources of mutation are environmental (radiation and chemicals), cell intrinsic (reactive oxygen species), and from the enzymatic errors of DNA polymerases. Perhaps counter intuitively, mutation rates are generally inversely proportional to genome size (mutation rate: viruses > unicellular microorganisms > multicellular eukaryotes), however mutation rates scale very well across organisms when taking into account the rate of base substitution per generation [3]. Additionally, in multicellular organisms, the mutation rates in somatic cells are much higher than in germ line cells and it is this high mutational burden and resulting selective pressure that has shaped the fidelity of DNA replication and repair, including DNA damage-sensitive cell cycle regulation, that we observe in nature.

The safeguarding of genomic information is thus well enforced; how true is this for processes responsible for the synthesis of gene products? The production of RNA and protein involves numerous steps under which various pressures (and lack of pressures) have shaped their ability to identify and correct mistakes. Compared to DNA replication and inheritance, these processes are quite noisy. New proteins are 5–6 orders of magnitude more likely to contain an misincorporated amino acid than by chance mutation of DNA and even post-translationally, proteins are subject to variability in folding, proteolytic cleavage, and other modifications [4]. Much of our understanding of error rates in the synthesis of RNA and protein comes from prokaryotes and single-celled eukaryotes. Multi-cellular eukaryotes differ substantially from these organisms in the realm of transcript processing and diversity. Although RNA splicing occurs in all domains of life, prokaryotic splicing mainly occurs in non-coding RNAs and is performed without the need for a spliceosome. The presence and length of introns in genes as well as the occurrence of alternatively spliced isoforms is strongly linked to an organism being multicellular. In recent years, it has become more evident that diversity associated with gene expression in humans and other multicellular eukaryotes has another critical layer—the production of multiple transcript isoforms from individual genes. The diversity of transcripts produced through alternative start and termination sites, alternative splicing, and selective RNA stability ultimately promote diverse proteomes. This diversity becomes an epigenetic tool for adaptation in organisms with long generation times and for specialization of somatic cells.

In this review, we will first discuss the high fidelity of replication and DNA repair, then the accuracy of RNA transcription and protein translation, and finally post-transcriptional RNA processing. Our post-transcriptional processing discussion will touch on concepts of co-transcriptional splicing, quality control, and the role of the RNA exosome in clearing out aberrantly spliced RNA.

2. Replication fidelity

The DNA replication error rate when including mismatch correction has been widely reported as $\sim 10^{-10}$, that is, 1 error in 10 billion bases [5]. For the human genome (3.2×10^9 bp in length) 0.32 bp are mutated on average per replication cycle, an incredible level of

accuracy. In humans, it is estimated that 400 cell divisions occur before the first sperm cell is produced and 30 cell divisions before the first egg cell. Roughly, this gives us $(400 \times 0.32) + (30 \times 0.32) = 138$ mutations per generation in humans [1] and [6]. Of course, mutation rates vary among species, within species, at different times, and at different genomic loci, thus modeling is complicated and measurement approaches are varied [7] and [2]. Mutation rates per generation have been measured with whole genome sequencing, both for *de novo* mutations on short time scales and, in combination with fossil evidence, phylogenetic mutations. For humans, *de novo* mutations occur at a rate of $\sim 1.2 \times 10^{-8}$ / bp / generation [7].

The core DNA replication polymerases (family B; α , δ , and ϵ) exhibit a combined error rate of 10^{-7} – 10^{-8} indicating that most of the replication fidelity is due to nucleotide selectivity and proofreading within the replisome (Pol δ and ϵ). The remaining orders of magnitude are captured by DNA mismatch repair enzymes and general DNA damage response pathways operating independently of replication or after lesion-induced stalling of replicating polymerases. In cells, the overall mutational load is carried mostly by DNA repair rather than DNA replication synthesis. Pol β , for example, is a DNA repair polymerase involved in base excision repair (BER) and is several orders of magnitude more error-prone than DNA replication (10^{-6} vs. 10^{-10}). The Y family polymerases, involved in translesion synthesis, have even higher error rates, suggesting that inaccuracy is favored over more drastic repair mechanisms that could cause chromosomal breakage and trigger cell death. DNA repair polymerases synthesize only short stretches of DNA relative to replicative polymerases, thus reducing their cumulative contribution to genomic mutation. Measured error rates of individual polymerases in both human and yeast have previously been compiled [5].

3. The DNA damage response

The cellular DNA damage response (DDR) is highly conserved among eukaryotes both in terms of its general mechanism and the specific proteins involved. The two main arms of this response are, 1) the repair process itself and 2) cell cycle checkpoint activation and/or apoptosis. Cell cycle checkpoints provide a window of time in which the cell may attempt DNA repair while apoptosis cleanses the tissue of damaged cells in the interest of genomic quality control. As a kinase cascade with the damage-sensing ATM and ATR at its apex, the DDR network modifies over one thousand proteins and the various signaling and repair factors that translocate to DNA lesions create large “repair foci” that are visible under light microscopy [8]. The DDR is sensitive, coordinated, and comprehensive; the sheer scale of its mobilization underscores the importance of safeguarding the genome. Protecting the information content of the genome is clearly of critical importance to all organisms, however when genomic information is transmitted into a work order, such exquisite attention to detail drops substantially.

4. Fidelity of RNA synthesis

Like DNA polymerases in replication, accuracy of RNA polymerases is determined by nucleotide selection and proofreading. RNA polymerase II (RNAPII), responsible for transcribing the bulk of protein-coding and non-coding genes, has been reported to have an error rate between 10^{-6} and 10^{-5} [9] and [10], at least one order of magnitude higher than

DNA replicative polymerases (10^{-8} - 10^{-7}). The accuracy of RNA polymerase is mostly due to nucleotide selectivity, but derives about one order of magnitude increased fidelity from proofreading. A number of structure-function studies have determined RNA polymerase proofreading to be a three-step process: mismatch identification, backtracking, and cleavage [11].

Most methods for measuring RNA polymerase error rates utilize exogenous reporters (e.g. [12] or *in vitro* assays, both of which may suffer from internal bias, but more recent studies have used next generation sequencing approaches to measure endogenous RNA errors *in vivo*. One such study used a modified RNA-seq approach to measure transcriptional errors in *C. elegans*, reporting a 4×10^{-6} error rate [13], suggesting fidelity of transcription to be much higher than previously thought. Another study described a method that analyzes standard RNA-seq data (rather than using a specialized library preparation) to estimate RNA transcriptional error rates and found that the mammalian error rate agreed with previous estimates ($\sim 10^{-5}$) [14]. Interestingly, the nucleotide sequences of protein-coding DNA may have been refined during evolution to mitigate the potential mutagenic effects of transcription. A logical extension to this observation is that maintaining a certain level of transcriptional error is beneficial or at least balanced by the energy cost that would be needed to improving the fidelity of transcription.

5. Diversity generated by RNA splicing

In eukaryotes, primary transcripts generated by RNA polymerase II undergo several processing steps, including capping, splicing and 3' end processing [15]. The number of genes containing introns varies dramatically among eukaryotes. In brewer's yeast, with its small, compact genome, only 5% of its 6000 genes contain introns. Most of these genes contain only one intron and the introns are no more than 1000 bp long [16] and [17]. In stark contrast, the human genome is comprised of about 26,000 genes, 94% of which contain introns, about 7 apiece, on average [18] and [19].

RNA splicing in eukaryotes is performed by the spliceosome, a large, complex machine, consisting of more than 200 proteins and of 5 small nuclear RNAs [20]. Splicing and polyadenylation are thought to take place while the RNA is still engaged with the chromatin (co-transcriptionally) [21], [22], [23], [24], [25], [26] and [27]; and the spliceosome attempts to recognize *bona fide* splice sites and presumably keep pace with RNAPII, which elongates at greater than 1000 bp/min [28], [29], [30] and [31]. Two hypotheses attempt to explain the mechanism of co-transcriptional splicing. The "recruitment model" for splicing states that factors involved in splicing and other processing events are recruited to the elongating RNAPII via the C-terminal domain (CTD) of the polymerase [26], [32] and [27]. The "kinetic model" for splicing suggests that the RNAPII elongation rate influences the efficiency of splicing such that slower elongation rates provide more time for splice junction recognition and spliceosome assembly, thus favoring efficient splicing [33], [25], [26] and [27]. Conversely, splicing may regulate the rate of transcription elongation through an "elongation checkpoint" that presumably prevents transcript release from the chromatin in the event of incomplete splicing [34] and [35]. Nucleosomes are strongly phased over exons; as transcription "speed bumps", they slow down transcription elongation and increase the

chances of productive splicing [36], [37] and [38]. Indeed, we and others have found that exon density in the path of RNAPII correlates with slower elongation rate [29] and [30].

Analysis of post-transcriptional RNA (mRNA-based gene expression studies) have revealed that splicing is noisy. Though alternative splicing (AS) was first described many decades ago [39] and [40], we have more recently learned that it occurs much more frequently and in more cell types than was previously thought. Next-generation sequencing technology has revealed that the mammalian transcriptome is generously infused with splice variants; some are conserved but many are species-specific. The splicing error rate in humans (per intron) has been estimated to be 7×10^{-3} and most errors fall into two categories: splice site recognition or exon recognition [41]. AS events are encoded in the genome via splicing-related sequences and epigenetic mechanisms and it has become apparent that AS events are commonplace, indicate a propensity of noise in the splicing of pre-mRNA [41]. Although splicing decisions are directed mainly by sequences within introns, codon usage near splice junctions can influence splicing efficiency as well and thus elicit selective pressure independent of the protein that they encode [41]. Like polypeptides with translation errors, splice variants can evade the degradation pathway. Of all the steps in the production line from DNA to proteins, RNA splicing is by far the most important in terms of generating diversity. In fact, it is thought that AS has been selected during evolution to promote increased complexity through degeneration of splicing site consensus sequences [20]. Taken together, proteome diversity in multicellular eukaryotes is driven, in large part, by transcriptome diversity due to generation of AS transcripts [42].

6. Post-transcriptional quality control

RNA degradation is carried out by the RNA exosome, a machine located in both the nucleus and the cytoplasm [43]. The exosome is a two-layered, cylindrical ring consisting of nine proteins; six bottom ring subunits, and four in the top ring, or cap. Bound at either end of this cylindrical core are two 3'-5' exoribonucleases. The exosome is activated by and receives its substrates from the TRAMP and NEXT complexes in the nucleus and the Ski complex in the cytoplasm [44] and [45]. A variety of RNA classes are sensed and removed by the nuclear exosome: pre-mRNA, -tRNA, and -rRNA, including their by-products, such as introns and post-3' cleavage site RNA [44] and [46]; various types of non-coding and non-functional RNAs [47] [48], [49] and [50]; and ribosome-associated mRNA, including those targeted by nonsense-mediated and non-stop decay mechanisms [44]. It has become increasingly apparent that an important function of the RNA exosome is to dispose of aberrantly spliced pre-mRNA [51], [52], [53] and [54]. In fact, it has been estimated that more than 50% of the nascent pre-mRNA produced in yeast cells are rapidly degraded before having a chance of becoming spliced [54]. If splicing is as error-prone as evidence suggests, a substantial amount of rapid, post-transcriptional cleanup, ostensibly via the exosome, is predicted. In humans and other multicellular eukaryotes where splicing is especially complex, the extent of nascent RNA degradation is even more of a mystery.

Some amount of mRNA is formed well enough to escape this first step in quality control, but can still carry with it variability/noise from the splicing process, or alternative splicing events, as described above. A number of annotated human exons yield premature

termination codons and are likely targeted by nonsense-mediated decay (NMD) (2–4% of all exons according to [55]). These exons have low inclusion rates, are less conserved, and exist in genes with a greater-than-average number of splice variants, though given that they are fast-evolving, they were not determined to be under relaxed selection [55]. NMD, non-stop decay, and other decay mechanisms use the RNA exosome to clear out transcripts that are detrimental to translation. These degradation pathways represent a point of contraction that prevents aberrant transcripts from becoming proteins.

Hanahan and Weinberg's "hallmarks of cancer" [56] and [57] did not emphasize the role of alternative splicing in tumorigenesis, though others believe it should be on this seminal list [58]. There are more than 300 splicing factors and nearly all of them have been found mutated in cancer [59]. Altered expression levels of splicing factors in cancer also contributes to high levels of aberrantly spliced transcripts [59]. In a large study of over 800 cancer patients and 16 different tumor types, elevated levels of intron retention in mRNA were found, suggesting pervasive splicing defects are a general cancer phenomenon [60]. Mutation and dysregulation of splicing factors broadens the transcriptional landscape, promoting cancer progression and chemotherapy resistance. Given that cancer cells are more prone to splicing errors, the RNA exosome is under extra pressure to purge the pool of aberrantly spliced transcripts and prevent their translation. Numerous cancer types, therefore, may be highly sensitive to small molecules that inhibit RNA cleanup. The mechanism of action of the chemotherapeutic agent, 5-fluorouracil (5-FU), has been linked to the RNA exosome and new drugs that target RNA degradation may have even stronger anti-cancer effects [61], [62], [63] and [64] yet the RNA exosome remains a largely untapped therapeutic target and could represent an exciting new avenue for cancer intervention.

7. Fidelity of protein synthesis

The requirements for accurate translation of RNA to functional polypeptides are transfer RNA charging (aminoacyl-tRNA synthetases), ribosome assembly, codon reading and tRNA selection, and protein folding capacity. Researchers have been occupied with measuring protein translation errors for decades; earlier and more recent studies found the error rate to vary among different organisms (though differences are also likely driven by methods)— 6×10^{-4} - 5×10^{-3} in Gram-negative *E. coli* [65] and [66], 10^{-2} in Gram-positive *B. subtilis* [67], 10^{-5} in *S. cerevisiae* [68], and 3×10^{-4} in mammals [69]. However in all cases it is clear that translation is less accurate than transcription ($\sim 10^{-5}$) and excessively error-prone relative to DNA synthesis (10^{-10}).

The ribosome employs an induced fit mechanism, reminiscent of that used by DNA and RNA polymerases [70]. Closely cognate tRNAs are more likely to be erroneously incorporated than others and the chances of this occurring is tRNA concentration-dependent. Additionally, tRNA charging has an inherent error rate of about 10^{-4} , relying on an editing mechanism to maintain this level of fidelity. Methods have also distinguished the rate of misincorporation at the ribosome due to selectivity and proofreading [71] and [72].

The general experimental framework for understanding mistranslation has been determined for prokaryotes and single-celled eukaryotes [73] and [74]. In addition, a significant effort has been made to model the protein translation system, incorporating codon misreading at the ribosome, charged tRNA concentration, tRNA mischarging, etc. [75], [76], [77], [78] and [79]. Recent evidence also demonstrates how tRNA synthetase fidelity is critical for mammalian fitness. A mutation in the editing domain of the mouse alanine tRNA synthetase was found to be responsible for mistranslation, leading to the accumulation of misfolded proteins, induction of the chronic unfolded protein response, and ultimately neuronal cell loss and ataxia [80]. Importantly, tRNA synthetase editing errors in general have deleterious effects on mammalian cells, rather than an isolated case specific to the alanine tRNA synthetase [81].

The primary driving force in selection for translational fidelity has been the effect on protein misfolding. Evolutionary selection against translation errors are more likely driven by their effects on protein folding than by other factors and results in two properties of mistranslation. First, errors are tolerated where they do not significantly alter protein folding, and second, errors affecting surface/solvent-exposed amino acids are much better tolerated than those in buried/hydrophobic amino acids that are necessary for efficient folding. Additionally, genome-wide studies have shown that genes encoding highly expressed proteins are under greater selective pressure to remain unchanged and that codons in such genes selectively exhibit lower amino acid misincorporation rates [82] and [83]. Ribosomes are also closely linked to the protein-degrading machinery (the ubiquitin-proteasome system) to relieve the cell of misfolded proteins. Indeed, a significant portion of nascent protein is ubiquitylated and somewhere between 6% and 30% is destined for degradation, indicating that these quality control mechanisms are in place to correct for the high rate of mistranslation and misfolding naturally occurring during protein synthesis. Substantial selective pressure on the rate of translation elongation is kinetic, given the co-translational nature of protein folding. In essence, slower dwell times mean improved tRNA-codon matching, but subtle variances in speed, pausing, etc. can have noticeable effects on the final product, even between synonymous codons [84]. Though a long-standing notion is that selection directs codon usage to be more efficient, there is evidence that less-optimal codons may be selected for [85], [86] and [87].

In contrast to DNA mutation, amino acid misincorporation is quite well tolerated, suggesting that translation diversity can provide a selective benefit. The genetic code is neither universal nor optimal for some species. The CTG clade of fungi, whose CUG codon was reassigned from leucine to serine, harbors a high error rate due to this codon's ambiguity (3% leucine misincorporation). *Candida albicans*, a CTG fungus, was found to tolerate induced leucine misincorporation of over 25% and result in diverse cellular forms and properties. Through CUG ambiguity and frequent use of this codon, *C. albicans* can substantially enhance its overall protein diversity to degrees of magnitude beyond what its genome provides [88]. Moreover, CUG usage versus other serine codons is related to gene function. For example, CUG is rare in ribosomal protein coding genes but high in genes coding for cell surface proteins, suggesting its use is related to adaptive responses. Mammalian cells were found to exhibit a low, but regular misacylation of non-Met-tRNAs with methionine (~1% of all methionine was misincorporated in this manner) and this behavior was significantly

enhanced by oxidative stress. The peppering of proteins with methionine may be deleterious to their functions, but the redox-active residue may provide a general buffer against reactive oxygen species [89].

8. Conclusions

Protein diversity, and the resulting phenotypic diversity, represents an invention space for new evolutionary possibilities. Unlike safeguarding the genome, which aims to prevent the degradation of information, the systems that encourage invention, such as alternative splicing and mistranslation, are *maintained* or even *encouraged* rather than *improved* upon through natural selection [90]. For prokaryotes and simple eukaryotes, the majority of diversity beyond that specifically encoded in the genome is realized through variability in protein synthesis. In multicellular eukaryotes, on the other hand, RNA processing, specifically splicing, is arguably a greater driver of diversity in gene expression that amplify the unique output of their genomes (see Fig. 1b).

The hypothetical translation mechanism of the distant evolutionary past was riddled with an ambiguous coding system and rudimentary form of the genetic code resulting in a highly variable proteome (Woese's "statistical proteins") [91]. Primitive genes, therefore, encoded at best a general protein structure rather than a specific sequence of amino acids. Due to lax constraints on coding, changing of codon assignments had little deleterious impact on the final product; thus the system was ideal for development of the genetic code. The fidelity of protein synthesis in modern species is very high relative to that primitive system, so clearly improvements in the fidelity of translation has been evolutionarily beneficial. However, optimization of splicing and protein synthesis lagged behind DNA replication [92]. In multicellular eukaryotes, especially metazoans, RNA transcription represents another layer in gene expression diversity, as alternative initiation, termination, and splicing together expand the roughly 26,000 gene catalog (in humans) by at least 5 orders of magnitude. The high levels of errors in splicing and translation have provided layers of molecular diversity important for the adaptation and survival of organisms through evolutionary time.

Acknowledgments

This work was supported by National Institute of Environmental Health Sciences (1R21ES020946), the National Human Genome Research Institute (1R01HG006786) and the University of Michigan Pancreatic Cancer Center. BM is supported by the University of Michigan School of Public Health Environmental Toxicology and Epidemiology Program, National Institute of Environmental Health Sciences (T32ES007062).

Abbreviations

Pol	polymerase
BER	base-excision repair
DDR	DNA damage response
ATM	ataxia telangiectasia mutated
ATR	ATM-related

RNAPII	RNA polymerase II
CTD	C-terminal domain
AS	alternative splicing
TRAMP	Trf4/Air2/Mtr4p polyadenylation complex
NEXT	nuclear exosome targeting complex
NMD	nonsense-mediated decay
5-FU	5-fluorouracil
UBS	ubiquitin-proteasome system
QC	quality control

References

1. Crow JF. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* 2000; 1:40–47. [PubMed: 11262873]
2. Shendure J, Akey JM. The origins, determinants, and consequences of human mutations. *Science.* 2015; 349:1478–1483. [PubMed: 26404824]
3. Lynch M. Evolution of the mutation rate. *Trends Genet.* 2010; 26:345–352. [PubMed: 20594608]
4. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* 2009; 10:715–724. [PubMed: 19763154]
5. McCulloch SD, Kunkel TA. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* 2008; 18:148–161. [PubMed: 18166979]
6. Vogel F, Rathenberg R. Spontaneous mutation in man. *Adv. Hum. Genet.* 1975; 5:223–318. [PubMed: 805513]
7. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* 2012; 13:745–753. [PubMed: 22965354]
8. Dellaire G, Bazett-Jones DP. Beyond repair foci: subnuclear domains and the cellular response to DNA damage. *Cell Cycle.* 2007; 6:1864–1872. [PubMed: 17660715]
9. Liu X, Bushnell DA, Kornberg RD. RNA Polymerase II Transcription: Structure and Mechanism. *Biochim. Biophys. Acta.* 2013; 1829:2. [PubMed: 23000482]
10. Sydow JF; Cramer. [accessed February 10, 2016] RNA polymerase fidelity and transcriptional proofreading. - PubMed - NCBI. (n.d.). <http://www.ncbi.nlm.nih.gov/pubmed/19914059>
11. Sydow JF, Cramer P. RNA polymerase fidelity and transcriptional proofreading. *Curr. Opin. Struct. Biol.* 2009; 19:732–739. [PubMed: 19914059]
12. Shaw RJ, Bonawitz ND, Reines D. Use of an in vivo reporter assay to test for transcriptional and translational fidelity in yeast. *J. Biol. Chem.* 2002; 277:24420–24426. [PubMed: 12006589]
13. Gout J-F, Thomas WK, Smith Z, Okamoto K, Lynch M. Large-scale detection of in vivo transcription errors. *Proc. Natl. Acad. Sci. U. S. A.* 2013; 110:18584–18589. [PubMed: 24167253]
14. Carey LB. RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *Elife.* 2015; 4
15. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature.* 2002; 416:499–506. [PubMed: 11932736]
16. Spingola M, Grate L, Haussler D, Ares M Jr. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA.* 1999; 5:221–234. [PubMed: 10024174]
17. Ast G. How did alternative splicing evolve? *Nat. Rev. Genet.* 2004; 5:773–782. [PubMed: 15510168]

18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001; 291:1304–1351. [PubMed: 11181995]
20. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* 2015; 84:291–323. [PubMed: 25784052]
21. Ameer A, Zaghlool A, Halvardson J, Wetterbom A, Gyllenstein U, Cavelier L, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* 2011; 18:1435–1440. [PubMed: 22056773]
22. Bhatt DM, Pandya-Jones A, Tong A-J, Barozzi I, Lissner MM, Natoli G, et al. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*. 2012; 150:279–290. [PubMed: 22817891]
23. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. *Nature*. 2012; 489:101–108. [PubMed: 22955620]
24. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, 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.* 2012; 22:1616–1625. [PubMed: 22955974]
25. Darnell JE Jr. Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. *RNA*. 2013; 19:443–460. [PubMed: 23440351]
26. Merkhofer EC, Hu P, Johnson TL. Introduction to cotranscriptional RNA splicing. *Methods Mol. Biol.* 2014; 1126:83–96. [PubMed: 24549657]
27. Naftelberg S, Shiran N, Schor IE, Gil A, Kornblihtt AR. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* 2015; 84:165–198. [PubMed: 26034889]
28. Singh J, Padgett RA. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* 2009; 16:1128–1133. [PubMed: 19820712]
29. Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, et al. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* 2014; 24:896–905. [PubMed: 24714810]
30. Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*. 2014; 3:e02407. [PubMed: 24843027]
31. Fuchs G, Voichek Y, Benjamin S, Gilad S, Amit I, Oren M. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* 2014; 15:R69. [PubMed: 24887486]
32. Bentley DL. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* 2014; 15:163–175. [PubMed: 24514444]
33. Braberg H, Jin H, Moehle EA, Chan YA, Wang S, Shales M, et al. From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell*. 2013; 154:775–788. [PubMed: 23932120]
34. Martins SB, Rino J, Carvalho T, Carvalho C, Yoshida M, Klose JM, et al. Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3' end of human genes. *Nat. Struct. Mol. Biol.* 2011; 18:1115–1123. [PubMed: 21892168]
35. Chathoth KT, Barrass JD, Webb S, Beggs JD. A splicing-dependent transcriptional checkpoint associated with pre-spliceosome formation. *Mol. Cell*. 2014; 53:779–790. [PubMed: 24560925]
36. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, et al. Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* 2009; 16:996–1001. [PubMed: 19684599]
37. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* 2009; 16:990–995. [PubMed: 19684600]
38. Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, et al. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*. 2015; 161:526–540. [PubMed: 25910207]
39. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* 1977; 74:3171–3175. [PubMed: 269380]

40. Chow LC, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. 1977. *Rev. Med. Virol.* 2000; 10:362–371. discussion 355–6. [PubMed: 11187030]
41. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 2010; 6:e1001236. [PubMed: 21151575]
42. Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat. Genet.* 2004; 36:1073–1078. [PubMed: 15448691]
43. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'→5' exoribonucleases. *Cell.* 1997; 91:457–466. [PubMed: 9390555]
44. Houseley J, Jonathan H, John L, David T. RNA-quality control by the exosome. *Nat. Rev. Mol. Cell Biol.* 2006; 7:529–539. [PubMed: 16829983]
45. Mitchell P M Phil. Exosome substrate targeting: the long and short of it. *Biochem. Soc. Trans.* 2014; 42:1129–1134. [PubMed: 25110014]
46. Lemay J-F, Larochelle M, Marguerat S, Atkinson S, Bähler J, Bachand F. The RNA exosome promotes transcription termination of backtracked RNA polymerase II. *Nat. Struct. Mol. Biol.* 2014; 21:919–926. [PubMed: 25240800]
47. Colin J, Libri D, Porrua O. Cryptic transcription and early termination in the control of gene expression. *Genet. Res. Int.* 2011; 2011:653494. [PubMed: 22567365]
48. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, et al. RNA exosome depletion reveals transcription upstream of active human promoters. *Science.* 2008; 322:1851–1854. [PubMed: 19056938]
49. Andersson R, Robin A, Claudia G, Irene M-E, Ilka H, Jette B, et al. An atlas of active enhancers across human cell types and tissues. *Nature.* 2014; 507:455–461. [PubMed: 24670763]
50. Pefanis E, Evangelos P, Jiguang W, Gerson R, Junghyun L, David K, et al. RNA Exosome-Regulated Long Non-Coding RNA Transcription Controls Super-Enhancer Activity. *Cell.* 2015; 161:774–789. [PubMed: 25957685]
51. West S, Gromak N, Norbury CJ, Proudfoot NJ. Adenylation and exosome-mediated degradation of cotranscriptionally cleaved pre-messenger RNA in human cells. *Mol. Cell.* 2006; 21:437–443. [PubMed: 16455498]
52. Bousquet-Antonelli C, Presutti C, Tollervey D. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell.* 2000; 102:765–775. [PubMed: 11030620]
53. Schneider C, Kudla G, Wlotzka W, Tuck A, Tollervey D. Transcriptome-wide analysis of exosome targets. *Mol. Cell.* 2012; 48:422–433. [PubMed: 23000172]
54. Gudipati RK, Xu Z, Lebreton A, Séraphin B, Steinmetz LM, Jacquier A, et al. Extensive degradation of RNA precursors by the exosome in wild-type cells. *Mol. Cell.* 2012; 48:409–421. [PubMed: 23000176]
55. Zhang Z, Xin D, Wang P, Zhou L, Hu L, Kong X, et al. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol.* 2009; 7:23. [PubMed: 19442261]
56. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000; 100:57–70. [PubMed: 10647931]
57. Hanahan D, Douglas H, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell.* 2011; 144:646–674. [PubMed: 21376230]
58. Ladomery M, Michael L. Aberrant Alternative Splicing Is Another Hallmark of Cancer. *Int. J. Cell Biol.* 2013; 2013:1–6.
59. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene.* 2015
60. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* 2015; 7:45. [PubMed: 26113877]
61. Lum PY, Armour CD, Stephanians SB, Cavet G, Wolf MK, Butler JS, et al. Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell.* 2004; 116:121–137. [PubMed: 14718172]

62. Fang F, Hoskins J, Butler JS. 5-fluorouracil enhances exosome-dependent accumulation of polyadenylated rRNAs. *Mol. Cell. Biol.* 2004; 24:10766–10776. [PubMed: 15572680]
63. Kammler S, Lykke-Andersen S, Jensen TH. The RNA Exosome Component hRrp6 Is a Target for 5-Fluorouracil in Human Cells. *Mol. Cancer Res.* 2008; 6:990–995. [PubMed: 18567802]
64. Silverstein RA, de Valdivia EG, Visa N. The Incorporation of 5-Fluorouracil into RNA Affects the Ribonucleolytic Activity of the Exosome Subunit Rrp6. *Mol. Cancer Res.* 2011; 9:332–340. [PubMed: 21289297]
65. Edelman P, Gallant J. Mistranslation in *E. coli*. *Cell.* 1977; 10:131–137. [PubMed: 138485]
66. Parker J, Friesen JD. Two out of three” codon reading leading to mistranslation in vivo. *Mol. Gen. Genet.* 1980; 177:439–445. [PubMed: 6768967]
67. Meyerovich M, Mamou G, Ben-Yehuda S. Visualizing high error levels during gene expression in living bacterial cells. *Proc. Natl. Acad. Sci. U. S. A.* 2010; 107:11543–11548. [PubMed: 20534550]
68. Stansfield I, Jones KM, Herbert P, Lewendon A, Shaw WV, Tuite MF. Missense translation errors in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 1998; 282:13–24. [PubMed: 9733638]
69. Loftfield RB, Vanderjagt D. The frequency of errors in protein biosynthesis. *Biochem. J.* 1972; 128:1353–1356. [PubMed: 4643706]
70. Rodnina MV; Wintermeyer. [accessed February 10, 2016] Ribosome fidelity: tRNA discrimination, proofreading and induced fit. - PubMed - NCBI. (n.d.). <http://www.ncbi.nlm.nih.gov/pubmed/11166571>
71. Kramer EB, Farabaugh PJ. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA.* 2007; 13:87–96. [PubMed: 17095544]
72. Zaher HS, Green R. Quality control by the ribosome following peptide bond formation. *Nature.* 2009; 457:161–166. [PubMed: 19092806]
73. Kramer EB, Vallabhaneni H, Mayer LM, Farabaugh PJ. A comprehensive analysis of translational missense errors in the yeast *Saccharomyces cerevisiae*. *RNA.* 2010; 16:1797–1808. [PubMed: 20651030]
74. Rodnina MV. Quality control of mRNA decoding on the bacterial ribosome. *Adv. Protein Chem. Struct. Biol.* 2012; 86:95–128. [PubMed: 22243582]
75. Fluitt A, Pienaar E, Viljoen H. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.* 2007; 31:335–346. [PubMed: 17897886]
76. Zouridis H, Hatzimanikatis V. Effects of codon distributions and tRNA competition on protein translation. *Biophys. J.* 2008; 95:1018–1033. [PubMed: 18359800]
77. Shah P, Gilchrist MA. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet.* 2010; 6:e1001128. [PubMed: 20862306]
78. Dutta A, Chowdhury D. A model for mis-sense error in protein synthesis: mischarged cognate tRNA versus mis-reading of codon, arXiv [physics.bio-Ph]. 2015 <http://arxiv.org/abs/1512.01790>.
79. Rudolf S, Lipowsky R. Protein Synthesis in *E. coli*: Dependence of Codon-Specific Elongation on tRNA Concentration and Codon Usage. *PLoS One.* 2015; 10:e0134994. [PubMed: 26270805]
80. Lee JW, Beebe K, Nangle LA, Jang J, Longo-Guess CM, Cook SA, et al. Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature.* 2006; 443:50–55. [PubMed: 16906134]
81. Nangle LA, Motta CM, Schimmel P. Global effects of mistranslation from an editing defect in mammalian cells. *Chem. Biol.* 2006; 13:1091–1100. [PubMed: 17052613]
82. DA, D.; CO, Wilke. [accessed February 10, 2016] Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. - PubMed - NCBI. (n.d.). <http://www.ncbi.nlm.nih.gov/pubmed/18662548>
83. al Yang JR E. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. - PubMed - NCBI. [accessed February 10, 2016] (n.d.). <http://www.ncbi.nlm.nih.gov/pubmed/20959819/>.
84. Marin M. Folding at the rhythm of the rare codon beat. *Biotechnol. J.* 2008; 3:1047–1057. [PubMed: 18624343]

85. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and quantitative assessment of the ubiquitin modified proteome. *Mol. Cell.* 2011; 44:325. [PubMed: 21906983]
86. Wang F, Durfee LA, Huibregtse JM. A cotranslational ubiquitination pathway for quality control of misfolded proteins. *Mol. Cell.* 2013; 50:368–378. [PubMed: 23583076]
87. Xu Y, Ma P, Shah P, Rokas A, Liu Y, Johnson CH. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature.* 2013; 495:116–120. [PubMed: 23417065]
88. Santos MAS, Gomes AC, Santos MC, Carreto LC, Moura GR. The genetic code of the fungal CTG clade. *C. R. Biol.* 2011; 334:607–611. [PubMed: 21819941]
89. Netzer N, Goodenbour JM, David A, Dittmar KA, Jones RB, Schneider JR, et al. Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature.* 2009; 462:522–526. [PubMed: 19940929]
90. Moura GR, Carreto LC, Santos MAS. Genetic code ambiguity: an unexpected source of proteome innovation and phenotypic diversity. *Curr. Opin. Microbiol.* 2009; 12:631–637. [PubMed: 19853500]
91. Woese CR. Evolution of the genetic code. *Naturwissenschaften.* 1973; 60:447–459. [PubMed: 4588588]
92. Kurland CG, Ehrenberg M. Optimization of translation accuracy. *Prog. Nucleic Acid Res. Mol. Biol.* 1984; 31:191–219. [PubMed: 6397771]

Figure 1. Decoding the genome leads to diversity

(a) At each step along the transfer of genetic information, errors are produced. In all domains of life, replication and repair of DNA are overall rigid, high-fidelity systems for safeguarding the genome. Transcribing and decoding this information, however, is more error-prone, leading to diverse transcriptomes and proteomes. Represented here is the flow and magnitude of errors (as black arrows) along a route from DNA to RNA to proteins. Some processes are outlets (brown text), whereby malformed products are removed, such as with apoptosis (removal of cells), RNA degradation (removal by the RNA exosome), and

protein degradation (removal by the ubiquitin proteasome system). Perturbations, such as cell stress, can increase error rates of many processes (indicated by red circles). Each step has an associated error rate, which may have been measured experimentally, and indicated in green where available. Green question marks represent areas where error rates (or rates or error disposal) have not been conclusively obtained. (b) For all organisms, erroneous protein translation plays a significant role in generating evolutionary diversity beyond what is contained in the genome. However, in multicellular eukaryotes, complex transcriptional and posttranscriptional mechanisms (alternative transcription initiation, splicing, stability, etc.) have greatly expanded the potential proteome. Here, the realized accumulation of diversity is represented by a widening arrow from genome to gene product. The bubbles on the sides represent expanding diversity due to process errors and contraction due to quality control measures (see points indicated by “errors” and “QC”). Contraction is akin to outlets as described in (a).