DNA Repair xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

DNA Repair

journal homepage: www.elsevier.com/locate/dnarepair

Genome stability versus transcript diversity

Brian Magnuson[a,b], Karan Bedi[a], Mats Ljungman[a,b,*]

[a] Department of Radiation Oncology, University of Michigan Comprehensive Cancer Center, and Translational Oncology Program, University of Michigan, Ann Arbor, USA
[b] Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
RNA splicing
RNA transcription
Protein translation
Post-transcriptional RNA processing
DNA damage response
DNA repair
Polymerase fidelity
Proteomic diversity

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 that 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.

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 inherent rates of error, 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.

Mutations can be advantageous, neutral, or deleterious, depending on their fitness or pathogenic effects (mutational rates and range of effects have been well-reviewed 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 misincorporated amino acids 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

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).

* Corresponding author at: Department of Radiation Oncology, University of Michigan Comprehensive Cancer Center, and Translational Oncology Program, University of Michigan, Ann Arbor, USA.
E-mail address: ljungman@umich.edu (M. Ljungman).

http://dx.doi.org/10.1016/j.dnarep.2016.05.010
1568-7864/

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. A model representing the flow of errors in these processes is provided in Fig. 1a.

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 cells divisions before the first egg cell. Roughly, this gives us $(400 \times 0.32) + (30 \times 0.32) = 138$ mutations per generation in humans [1,6]. Of course, mutation rates vary among species, within species, at different times, and at different genomic loci, making modeling complicated and measurement approaches varied [7,2]. 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 (Pols 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} [9,10], 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 [10].

Most methods for measuring RNA polymerase error rates utilize exogenous reporters, e.g. Ref. [11], or *in vitro* assays, both of which may suffer from internal bias. 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 [12], 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 ($\sim 10^{-5}$) [13]. 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 [14]. The number of genes containing introns varies dramatically among eukaryotes. In brewer's yeast, with it's 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 [15,16]. In stark contrast, the human genome is comprised of about 26,000 genes, 94% of which contain introns, about 7 apiece, on average [17,18].

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 [19]. Splicing and polyadenylation are thought to take place while the RNA is still engaged with the chromatin (co-transcriptionally) [20–26] and the spliceosome attempts to recognize *bona fide* splice sites and presumably keeps pace with RNAPII, which elongates at greater than 1000 bp/min [27–30]. Two hypotheses attempt to explain the mechanism of co-transcriptional

Please cite this article in press as: B. Magnuson, et al., Genome stability versus transcript diversity, DNA Repair (2016), http://dx.doi.org/10.1016/j.dnarep.2016.05.010