Biochimica et Biophysica Acta 1819 (2012) 566–577

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

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

Review

Nuclear export as a key arbiter of "mRNA identity" in eukaryotes ☆

Alexander F. Palazzo *, Abdalla Akef

Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada M5S 1A8

ARTICLE INFO

Article history:
Received 4 October 2011
Received in revised form 23 December 2011
Accepted 29 December 2011
Available online 9 January 2012

Keywords:
mRNA export
mRNA identity
Coupling
TREX
SSCR
ALREX

ABSTRACT

Over the past decade, various studies have indicated that most of the eukaryotic genome is transcribed at some level. The pervasiveness of transcription might seem surprising when one considers that only a quarter of the human genome comprises genes (including exons and introns) and less than 2% codes for protein. This conundrum is partially explained by the unique evolutionary pressures that are imposed on species with small population sizes, such as eukaryotes. These conditions promote the expansion of introns and non-functional intergenic DNA, and the accumulation of cryptic transcriptional start sites. As a result, the eukaryotic gene expression machinery must effectively evaluate whether or not a transcript has all the hallmarks of a protein-coding mRNA. If a transcript contains these features, then positive feedback loops are activated to further stimulate its transcription, processing, nuclear export and ultimately, translation. However if a transcript lacks features associated with "mRNA identity", then the RNA is degraded and/or used to inhibit further transcription and translation of the gene. Here we discuss how mRNA identity is assessed by the nuclear export machinery in order to extract meaningful information from the eukaryotic genome. In the process, we provide an explanation of why certain sequences that are enriched in protein-coding genes, such as the signal sequence coding region, promote mRNA nuclear export in vertebrates. This article is part of a Special Issue entitled: Nuclear Transport and RNA Processing.

1. Introduction

1.1. The nucleus and the expansion of non-coding sequences: two features that distinguish eukaryotes from prokaryotes

The nucleus is the defining feature of the eukaryotic cell. It compartmentalizes the cellular space into two distinct regions: the nucleoplasm, where RNA is synthesized, processed and packaged, and the cytoplasm, where mature mRNA is translated into proteins. This is in striking contrast to prokaryotes, in which transcription and translation occur concurrently in the same compartment. Another important difference between these two groups is the percentage of their genomes that encode protein. In multicellular eukaryotes, protein-coding sequences account for a small fraction of the genome, varying from 1.5 to 36% [1–3], while in prokaryotes, the majority of the genome encodes protein [4]. Although these two properties seem to be only marginally related, we will explore how the compartmentalization of the cell becomes vital when only a small proportion of the genome encodes proteins.

1.2. Transcriptional expansion: an adaptive feature of eukaryotes or a byproduct of small population size?

Why are eukaryotic genomes so much larger than those of prokaryotes? One reason for the difference is that a substantial proportion of the eukaryotic genome is made up of transposons and other self-replicating entities which are not likely to have a functional role [1]. A second factor is the presence of introns, which in humans accounts for almost a quarter of the genome [1]. However, it is clear that even if these entities are omitted, eukaryotes have experienced a vast expansion in genomic sequence that does not code for protein. It has been assumed by many that this increase was a consequence of natural selection acting to expand the amount of functional information and organismal complexity [5,6], which could have taken the form of an amplification in 1) functional non-coding transcriptional products, 2) DNA regulatory sequences that direct RNA transcription and chromosome architecture and/or 3) RNA regulatory sequences that impact alternative splicing and other RNA processing events.

In support of the expansion of functional non-coding transcripts, several large-scale analyses have indicated that most of the eukaryotic genome is transcribed, albeit at some low level [7,8]. However, evidence has been mounting that many of these non-coding transcripts do not have any specific purpose, but are simply the result of widespread non-specific RNA polymerase activity. First, most of the non-coding

Abbreviations: ALREX, (Alternative mRNA Export); CBC, (Cap Binding Complex); CTD, (Carboxy-Terminal Domain); EJC, (Exon Junction Complex); mRNP, (Messenger Ribonucleoprotein); MSCR, (Mitochondrial Targeting Sequence Coding Region); NPC, (Nuclear Pore Complex); Pol II, (RNA Polymerase II); TREX, (Transcription Export); SR, (Serine/Arginine-Rich); SSCR, (Signal Sequence Coding Region); UAS, (Upstream Activating Sequence)

☆ This article is part of a Special Issue entitled: Nuclear Transport and RNA Processing.

* Corresponding author at: Department of Biochemistry, University of Toronto Medical Sciences Building, Room 5242 Toronto, Ontario, Canada M5S 1A8. Tel.: +1 416 978 7234; fax: +1 416 978 8548.
E-mail address: alex.palazzo@utoronto.ca (A.F. Palazzo).

1874-9399/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbagrm.2011.12.012

transcribed regions are poorly conserved [8–10]. Second, it is likely that the transcriptional products from these regions are rapidly degraded [11–18]. Third, although initial reports indicated that these RNAs are quite abundant [8], it now appears that the steady state level of these transcripts is very low [19,20]. Of course, many examples have been found that contradict these general features. Non-coding RNAs have been found to regulate transcription [21], mRNA translation and stability [22], histone modification [23], DNA methylation [24], DNA recombination [25], and even cross-regulate other non-coding RNAs [26]. However, even if the number of functional non-coding RNAs has dramatically increased, as exemplified by recent studies (for example [27,28]), it appears that all of these novel functional transcripts are derived from only a small percentage of the total genome [10,29].

In parallel to these studies, the development of population genetics has uncovered many of the evolutionary forces that shape genomic content. One important principle derived from these analyses is that the ability for natural selection to weed out mildly deleterious mutations, such as the insertion of non-functional DNA sequences (i.e., introns and intergenic sequences), increases with the number of breeding individuals [2–4]. As a consequence, species that have a low population size, as is the case with most eukaryotes, cannot effectively select out these genetic alterations. Interestingly, the widespread conservation of intronic positions within orthologous genes across diverse branches of the eukaryotic tree strongly suggests that all eukaryotes descended from a common ancestor that was intron-rich [30–34]. Since the elimination of introns is promoted by a large effective population size, it is also likely that this last common eukaryotic ancestor (and all intermediates between this organism and metazoans) had a small number of breeding individuals [34] and a fairly substantial amount of non-functional intergenic sequence in its genome. Indeed, several studies have indicated that both introns and intergenic regions are being eliminated from certain eukaryotic lineages that have probably experienced a relatively recent increase in their effective population number (one example being *Saccharomyces cerevisiae*, which has experienced a recent loss of introns [30–33] and likely intergenic sequences [35,36]). However, it is likely that these lineages are the exception rather than the rule, as the level of intergenic sequence in most unicellular eukaryotic genomes is about 50% [3].

Other mildly deleterious mutations that accumulate in organisms with low population sizes are short DNA elements that promote some cellular activity in non-functional genomic regions. These elements include cryptic transcriptional start sites, whose sequence tends to be highly degenerate [37–40]. This would explain why both the mouse and human genomes contain about an order of magnitude more promoter regions as compared to protein-coding genes [8,41]. It is also worth noting that although RNA polymerase II (Pol II) does not efficiently initiate transcription at non-promoter sites, the proliferation of non-functional DNA may also increase the frequency of spurious transcription initiation by increasing the amount of non-specific substrate. Indeed, spurious initiation of Pol II-driven transcription has been observed at nucleosome-free sites *in vivo* [42,43]. As a result of all these forces, we can begin to understand why a large fraction of active Pol II is associated with intergenic regions throughout the yeast and human genomes [8,44]. In fact, it has been estimated that in *S. cerevisiae*, only 10% of DNA-associated Pol II is bound to protein-coding genes, while the rest is associated with these intergenic regions [45].

1.3. Sorting transcripts that contain mRNA identity features from spurious transcription, a key role of the nucleus

From this vantage point the real question that we should be asking is how functional information (i.e., conserved protein-coding transcripts) is extracted from all the transcriptional noise found within the eukaryotic genome. The answer appears to be that transcripts bearing hallmarks of protein-coding genes are identified by an extensive network of feedback and feed-forward loops between various machineries present at different steps of the gene expression pathway (for example, Fig. 1). Thus at each step, features associated with mRNA identity are acted on by one set of factors, and these directly promote the activity of other machines responsible for subsequent and previous steps. This phenomenon, generally known as "coupling" [46–49], was previously viewed as a method for either enhancing the efficiency of gene expression or distinguishing properly processed from unprocessed mRNAs. However, through this extensive coupling network, a system for identifying mRNAs from spurious transcription also emerges. It should be noted that the concept of "mRNA identity" was originally used to describe how protein-

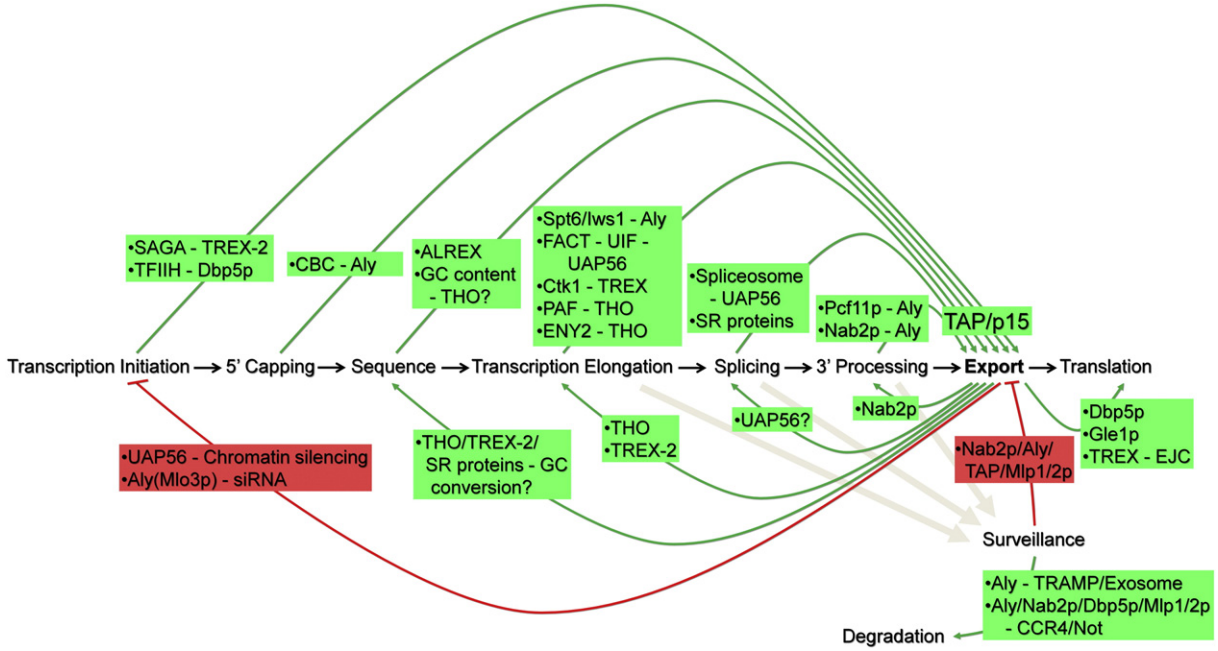

Fig. 1. Coupling between nuclear mRNA export and various steps of gene expression. Green arrows represent a positive feed-forward or feedback regulation while red lines represent a negative feedback relationship.