BioSystems 111 (2013) 156–174

Contents lists available at SciVerse ScienceDirect

BioSystems

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

Polymerization of non-complementary RNA: Systematic symmetric nucleotide exchanges mainly involving uracil produce mitochondrial RNA transcripts coding for cryptic overlapping genes

Hervé Seligmann[a,b,*]

[a] National Natural History Museum Collections, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel
[b] Department of Life Sciences, Ben Gurion University, 84105 Beer Sheva, Israel

ARTICLE INFO

Article history:
Received 24 October 2012
Received in revised form 24 January 2013
Accepted 29 January 2013

Keywords:
Expressed sequence tags
Nucleotide misinsertion
Human DNA polymerase gamma
Genome compression
Antitermination tRNA
Termination codon

ABSTRACT

Usual DNA→RNA transcription exchanges T→U. Assuming different systematic symmetric nucleotide exchanges during translation, some GenBank RNAs match exactly human mitochondrial sequences (exchange rules listed in decreasing transcript frequencies): C↔U, A↔U, A↔U+C↔G (two nucleotide pairs exchanged), G↔U, A↔G, C↔G, none for A↔C, A↔G+C↔U, and A↔C+G↔U. Most unusual transcripts involve exchanging uracil. Independent measures of rates of rare replicational enzymatic DNA nucleotide misinsertions predict frequencies of RNA transcripts systematically exchanging the corresponding misinserted nucleotides. Exchange transcripts self-hybridize less than other gene regions, self-hybridization increases with length, suggesting endoribonuclease-limited elongation. Blast detects stop codon depleted putative protein coding overlapping genes within exchange-transcribed mitochondrial genes. These align with existing GenBank proteins (mainly metazoan origins, prokaryotic and viral origins underrepresented). These GenBank proteins frequently interact with RNA/DNA, are membrane transporters, or are typical of mitochondrial metabolism. Nucleotide exchange transcript frequencies increase with overlapping gene densities and stop densities, indicating finely tuned counterbalancing regulation of expression of systematic symmetric nucleotide exchange-encrypted proteins. Such expression necessitates combined activities of suppressor tRNAs matching stops, and nucleotide exchange transcription. Two independent properties confirm predicted exchanged overlap coding genes: discrepancy of third codon nucleotide contents from replicational deamination gradients, and codon usage according to circular code predictions. Predictions from both properties converge, especially for frequent nucleotide exchange types. Nucleotide exchanging transcription apparently increases coding densities of protein coding genes without lengthening genomes, revealing unsuspected functional DNA coding potential.

1. Introduction

The question 'why are there several stop codons?' (Krizek and Krizek, 2012) has an apparently satisfying answer: off frame, protein coding genes include numerous stops (Seligmann and Pollock, 2004a,b; Singh and Pardasani, 2009; Tse et al., 2010) which decrease protein synthesis costs due to unprogrammed ribosomal slippage (Seligmann, 2007, 2010a; Warnecke and Hurst, 2011). In addition, the genetic code's codon–amino acid assignments maximize off frame stop numbers (Itzkovitz and Alon, 2007), and third codon positions that are part of off frame stops tend to mutate less than comparable positions (Seligmann, 2012a). However, this explanation hides a further function that stop codons play in off frame sequences: it seems that when antitermination (suppressor) tRNAs are active in translation, the regular genetic code is *de facto* transformed into another, stopless genetic code (Seligmann, 2010b). Translating sequences into proteins according to that overlapping code reveals numerous previously undetected genes and proteins, their number coevolving with capacities of antitermination tRNAs (tRNAs with anticodons matching stops) to translate the stops they include (Faure et al., 2011; Seligmann, 2011a, 2012a,b). Inclusion of stop codons in the regular genetic code enables a double coding system, based on the same sequences, and whose expression is efficiently regulated by the presence or absence of suppressor (antitermination) tRNAs. That way, numbers of coded proteins can be high while keeping a relatively short genome, by switching from the regular genetic code to a stopless code.

Genome length is an important factor limiting replication and cellular multiplication rates, apparently affecting also developmental rates of metazoan organisms (Sessions and Larson, 1987; Gregory and Hebert, 1999; Chipman et al., 2001). Ample data

* Correspondence address: National Natural History Museum Collections, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel.
E-mail address: varanuseremius@gmail.com

0303-2647/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biosystems.2013.01.011

suggest that even at the level of single amino acids, protein sequences minimize metabolic synthesis costs (Akashi and Gojobori, 2002; Seligmann, 2003; Barton et al., 2010), notably of cognate amino acids (Perlstein et al., 2007; Alves and Savageau, 2005; Seligmann, 2012b). Protein length reduction apparently follows similar principles (Brocchieri and Karlin, 2005; Warringer and Blomberg, 2006; Seligmann, 2012b). Considering this, it is very probable that similar forces decrease genome length. Accordingly, there would be a strong advantage for being able to code for more proteins, while keeping the genome short, a phenomenon that increases coding density by coding compression, such as overlapping genes, including those induced by antitermination tRNA activity (Seligmann, 2011a, 2012c,f; Faure et al., 2011). Recent analyses suggest that mitochondrial genomes include several overlapping genes coded in the 3′-to-5′ direction of regular protein coding genes, apparently expressed upon putative 'invertase' activity, which would invert the sequence polymerized into RNA in the 3′-to-5′ direction (Seligmann, 2012d). A further mechanism apparently increasing coding density is that of protein coding genes based on tetracodons, quadruplet codons recognized by (among others) tRNAs with expanded anticodons (Seligmann, 2012e). Mitochondrial genes for ribosomal RNAs seem also to include overlapping protein coding genes (Seligmann, 2013)

It is in this context that a group of phenomena called RNA recoding is considered here. These imply typically changing frames (Namy et al., 2005) and various phenomena of exon/intron reshuffling (i.e., Jin et al., 2007; Lev-Maor et al., 2007). In some cases, recoding alters the nucleotides used, such as adenosine-to-inosine RNA editing (Reenan, 2005; Paz et al., 2007; Daniel et al., 2011).

1.1. Nucleotide exchanges as a working hypothesis for cryptic overlapping genes

The systematic 'recoding' of T (thymidine) to U (uracil) in transcription from DNA to RNA is also a type of recoding, by DNA→RNA polymerases that systematically exchange T by U, and U by T for reverse transcriptases. This suggests the hypothesis that coding density might be increased by other types of systematic nucleotide exchanges, i.e. A by C and C by A (or any other symmetric exchange of this type). The fact that during regular DNA replication, ribonucleotides are frequently inserted instead of deoxynucleotides by the mitochondrial DNA polymerase gamma (Kasiviswanathan and Copeland, 2011) indicates that polymerases have some flexibility in that respect. Misinsertion of non-complementary nucleotides is also a basic property of polymerase (mis)function (Lee and Johnson, 2006). The possibility of polymerase activity implying systematic misinsertions, producing non-complementary DNA and/or RNA strands, cannot be excluded.

Such recoded RNA, based on the template of regular DNA sequence, could code for additional protein coding gene(s). Interestingly, if this occurs at DNA level, this could be a mechanism for producing new genes, but in this case the assumed mechanism of transcription exchanging between nucleotides implies that genes code according to 'direct' (non-exchanging) and exchange transcription. In some ways, the former can be seen as explicit, and the latter as implicit coding, nevertheless, both levels would be inherent simultaneously to the gene's primary structure.

Hence if such nucleotide exchanging activity exists, by some kind of unknown or modified DNA→RNA polymerases during RNA polymerization or editing, inducing such activity might unleash a very large coding potential, enabling to code for proteins without increasing genome size. In addition, this system implies very simple regulation, as each set of genes associated with a given type of nucleotide exchange would be induced by the expression of its specific 'nucleotide exchanger' polymerase/editing activity.

Table 1
The nine different RNA sequences produced from transcription of a single DNA sequence (ACGT) according to the nine types of symmetric nucleotide exchange rules. The amino acid coded by the three first nucleotides according to the vertebrate mitochondrial genetic code is also indicated, as well as the percentage of nucleotides that remain identical after that type of exchange transcription.

Exchange rule	Initial DNA 5′-ACGT-3′	Codon for Thr	Similarity to initial DNA sequence
A↔C	5′-CAGU-3′	Gln	50%
A↔G	5′-GCAU-3′	Ala	50%
A↔U	5′-UCGA-3′	Ser	50%
C↔G	5′-AGCU-3′	Ser	50%
C↔U	5′-AUGC-3′	Met	50%
G↔U	5′-ACUG-3′	Thr	50%
A↔C and G↔U	5′-CAUG-3′	His	0%
A↔G and C↔U	5′-GUAC-3′	Val	0%
A↔U and C↔G	5′-UGCA-3′	Cys	0%

In total, considering only the four usual nucleotides, nine symmetric nucleotide exchanges are possible, multiplying by nine the coding potential of any single sequence. Six of these involve only two types of nucleotides (A↔C, A↔G, A↔U, C↔G, C↔U, G↔U) and three all four types of nucleotides, implying two symmetric exchanges (A↔C+G↔U, A↔G+C↔U, and A↔U+C↔G). Table 1 shows the different RNA sequences produced by each of these rules from a single, given initial DNA sequence. Note that this procedure alters at least 50% of the nucleotides in the initial sequence used in Table 1, and that the amino acid coded by the three first nucleotides in that sequence is changed in almost all cases after systematic symmetric nucleotide exchange.

Along the same lines, asymmetric nucleotide recodings are also possible (such as an exchange rule including three nucleotide exchanges, i.e., A→C, C→G and G→A, in total 14 asymmetric exchange possibilities exist (including also rules with four asymmetric nucleotide exchanges). For practical reasons, I explore here only symmetric exchanges Separating symmetric from asymmetric exchanges is also justified by the possibility that symmetric and asymmetric nucleotide exchanges may depend upon different types of polymerization (or editing) mechanisms.

First, I explore GenBank's EST (expressed sequence tags) RNA databank for sequences matching the 'exchanged' human mitochondrial genome according to each of the nine symmetric exchange rules and report the results for the various types of exchanges. Then Blast alignment analyses explore whether RNA recoded by each of these exchanges could be coding for proteins, using various bioinformatics methods to indicate whether the detected putative overlapping genes seem functional or not. A meta-analysis of the data shows that frequencies of RNAs associated with the different types of symmetric exchanges are proportional to the bioinformatics estimations of overlap protein coding gene functionalities, indicating that coding compression through RNA exchange/editing occurs, and this at different frequencies for different types of nucleotide exchanges. Most notably, DNA nucleotide misinsertion rates during replication predict rates of nucleotide exchanging RNA transcription.

2. Materials and methods

2.1. Sequence manipulations and alignments with existing RNA transcripts

All analyses are done for GenBank's reference complete human mitochondrial genome (NC_012920). Its entire sequence is copy pasted from GenBank into a blank Microsoft Word file. In 'Word', the sequence of the genome was altered by using the software's 'Replace' function, mimicking a putative systematic nucleotide exchange. For example, for the symmetric exchange rule A↔C, the function 'Replace' was used to replace all 'A's in the genome by 'X', then all 'C's by 'A', and then all 'X's by 'C'. The intermediate stage using 'X' (or any other arbitrary symbol differing from the four letters used to symbolize the four nucleotides) is necessary to avoid that 'A's changed into 'C's at the first step are changed back into 'A' at the second step. The resulting