SciForschen
Open HUB for Scientific Research
ISSN 2471-4968
International Journal of Molecular Genetics and Gene Therapy
Research Article
Volume: 2.1
Open Access

RNA-Dependent RNA Polymerase encoding Artifacts in Eukaryotic Transcriptomes

Stephen Winters-Hilt*

Computer Science Department, Connecticut College, New London, USA

***Corresponding author:** Stephen Winters-Hilt, Connecticut College, Computer Science Department, 270 Mohegan Ave, New London, CT 06320, USA, **E-mail:** swinters@conncoll.edu

Supplementary Data

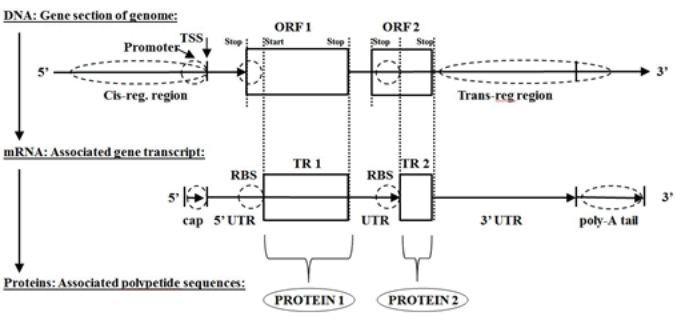
The possible viral origin of the spliceosome

The possible viral origin of the spliceosome, which is beginning to be indicated by results like those in [20], and described here, is relevant to two hypotheses: "Virus World" and "Introns Earliest". The Introns Earliest Hypothesis is discussed elsewhere [34], so will only be mentioned briefly in what follows. The Virus World Hypothesis suggests that the role of viruses could be significantly underestimated by the standard descriptions. In [35] it is even suggested that DNA and DNA replication may have first appeared in Virus World, from which three RNA cell lineages, each with its own symbiotic DNA virus lineage, could have then led to the cellular domains seen today [36].

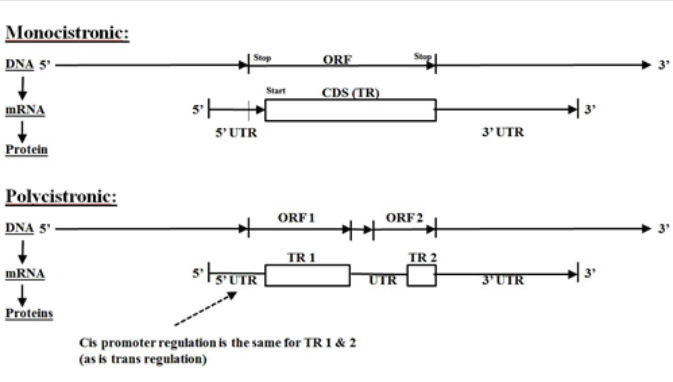
Pre-eukaryotes may have co-evolved an early form of RNAi (typical archaeons/prokaryotes have an unrelated form of RNA interference via CRISPR) with a mega Virus, where the battle for RNAi control of host, eventually gave rise to eukaryotic cell. Via modulation of RNAi, if nothing else, RdRp's have a pervasive role in eukaryotes in defense against foreign nucleic acids, developmental regulation, genome maintenance, and transcriptome modulation [37]. The RNAi transcriptional and post-transcriptional silencing in *Schizosaccharomycespombe*, for example, is accomplished with a single version of Dicer, RdRp, and Argonaute protein (Ago1) [38]. Some theories of viral eukaryogenesis indicate that the development of RNAi was a critical escalation in the battle for cellular control that led to the viral-nucleus dominated viral eukaryogenesis pathway [6]. Viral eukaryogenesis has been proposed in a variety of forms [6, 39-42], all consistent with the lack of RdRp in prokaryotes/archaea (as with RNAi), while RdRp is found in eukaryotes and viruses (as is RNAi). Viral eukaryogenesis is also indicated by a number of other recently discovered structural properties, including the ease of membrane vesicle bound transfer of genomic information [43], and an emerging picture of virus complexity and their co-evolving relation with cells [44-48]. So the important role of RdRp for RNAi purposes clearly explains why RdRp would have singular importance to eukaryotes, but is there a non-RNAi role for RdRp as well? In some RdRp RNA productions lengthy antisense strands are produced (as mentioned for *T. brucei* [19]), and this suggests a possible non-RNAi role, as will be discussed later.

The viral endosymbiosis could have been a very slow transitional process given the likely mutualism 'melting pot' of endosymbionts, including both proteobacteria (eventually leading to mitochondria, etc.) and viruses (including mega viruses) that could have existed at the LECA. This could have existed then since it still exists now, in present day *Acanthamoeba*, which is the host to the megaviruses, Mimivirus, Marseillesvirus, and other nucleocytoplasmic large DNA viruses (the NCLDVs), as well as a variety of bacteria [49], often hosting multiple virus and bacterial visitors at the same time. Furthermore, phylogenetic analysis shows that the NCLDV viruses have an ancient relation to eukaryotes, thus the LECA [50-53], placing them in the mutualism context at the time of the hypothesized viral eukaryogenesis. The NCLDV viruses possess their own mRNA capping protein (eIF4E), mRNA capping enzyme, and DdRp, to ensure preferential production of their mRNAs by the host, and eIF4E, at least, is thought to be unique to the virus (not obtained from the host *Acanthamoeba*) [54-56]. The mutualism 'melting pot' environment of the early eukaryote formation is further supported by viral mutualism in the form of favorable RdRp exchange from T7 to the mitochondrial host genome, which is thought to have occurred at the time of the LECA (or earlier) since it is present in all known eukaryotic mitochondria (with only one known exception [49]).

S.1) Classic central dogma of biology

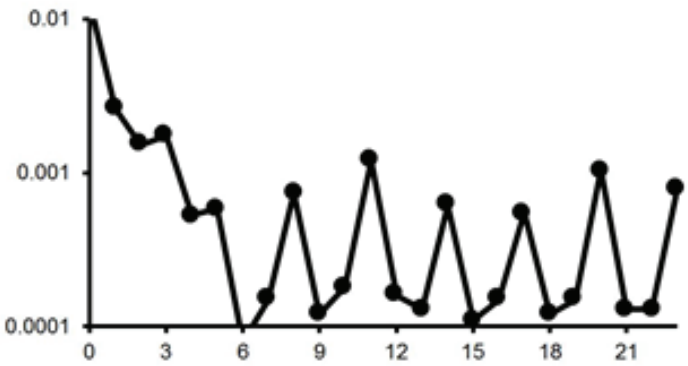


Suppl. Figure 1: Classic Central Dogma, operon (polycistronic): DNA→mRNA→Protein Set



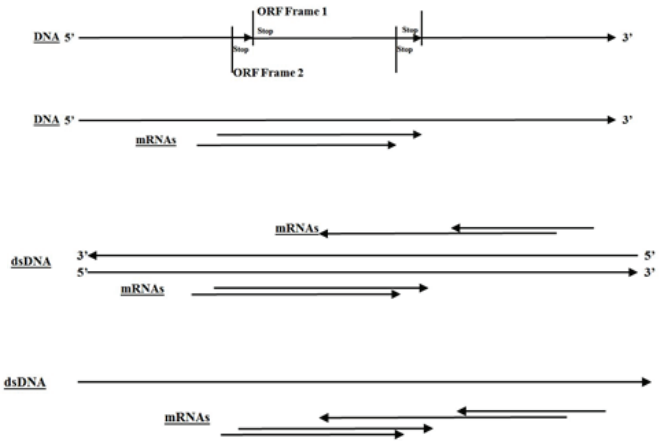
Suppl Figure 2: mRNA Transcriptome Topology Schematics: monocistronic and polycistronic

S.2) ORFs and ORF overlap topology in prokaryotic dsDNA



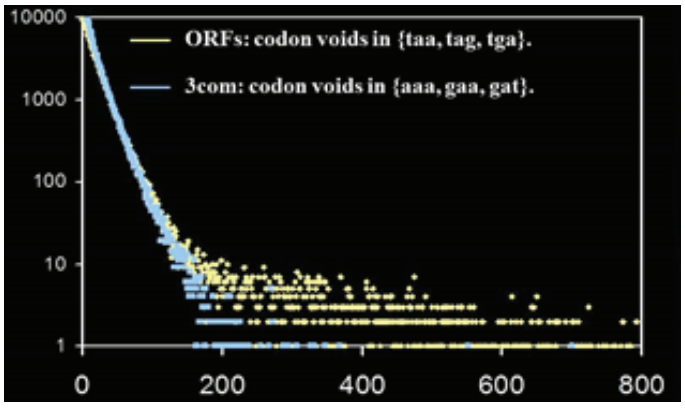
Suppl Figure 3: Codon structure is revealed in the *V. cholera* genome by mutual information between nucleotides in the genomic sequence. Reprinted with permission [58].

Once Codon grouping are revealed, a frequency analysis on codon can be done, and the stop codons are found to be rare. Focusing on the stop codons it is easily found that the gaps between stop codons can be quite anomalous compared to the gaps between other codons (Suppl. Figure 4). ORFs are "open reading frames", where the reference to what is open is lack of encounter with a stop codon {(uaa),(uag),(uga)} when traversing the genome with a particular codon framing , e.g., ORFs are regions devoid of stop codons when traversed with the codon framing choice of the ORF. When referring to ORFs in most of the analysis we refer to ORFs of length 300 bases or greater. The restriction to larger ORFs is due to their highly anomalous occurrences and likely biological encoding origin (Suppl. Figure 4), e.g., the long ORFs give a strong indication of containing the coding region of a gene. By restricting to transcripts with ORFs ≥ 300 in length, we have a resulting pool of transcripts that are mostly true coding transcripts.

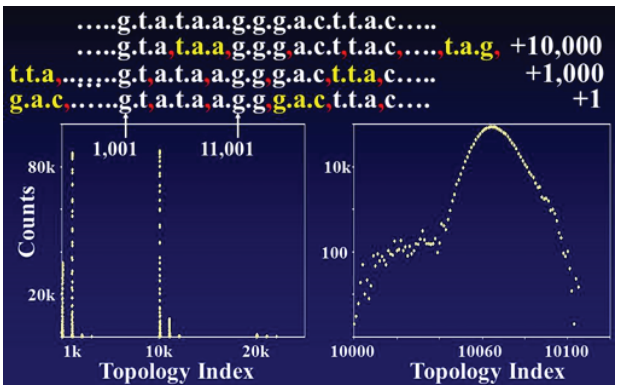


Suppl. Figure 5: Top. DNA ORF and mRNA Transcriptome Topology Schematics: Compact Notation (overlap reads have to be different frame, thus at most 3). Genome is typically dsDNA, so have reverse complement strand with its own gene encodings, also possibly overlapping with frame shift. **Middle dsDNA and mRNA Transcriptome Topology Schematics: Compact Notation with forward and reverse encodings.** Prokaryotes have significant overlap encoding from both frameshift and reverse complement reads (for opposite strand). Prokaryotes also often have operon structure. **Bottom. dsDNA and mRNA Transcriptome Topology Schematics: Single-line dsDNA Compact Notation with forward and reverse encodings.**

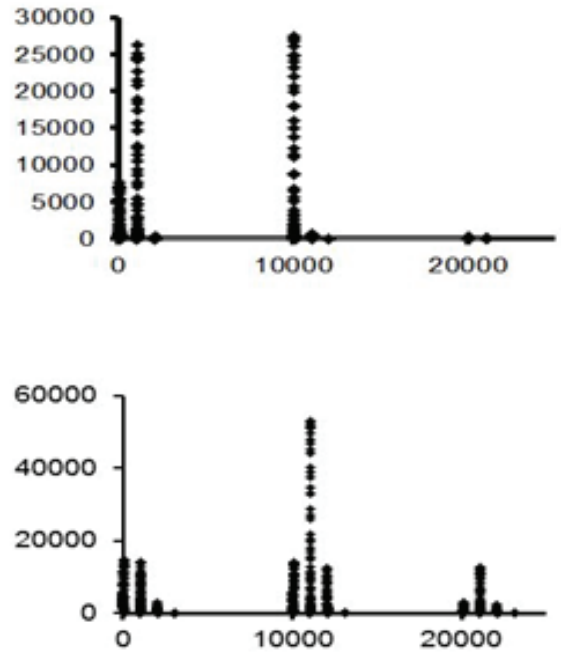
A long ORF can overlap with another ORF that has a different framing (since the codon has length three bases, there are three possible framings). When working with dsDNA genomes both strands encode, so with reference to one strand listed in genbank, say, one needs to perform three frame passes in the normal 'forward' direction, then three frame passes on the reverse complement of the reference strand, in order to identify all ORFs ≥ 300 bases. Once this is done it is found that many of the ORFs overlap. An accounting of this overlap 'topology' is indexed and plotted as shown in Suppl. Figure 6 and 7 (topology is a study of connectedness, and overlap extent is a form of connectedness in this context).



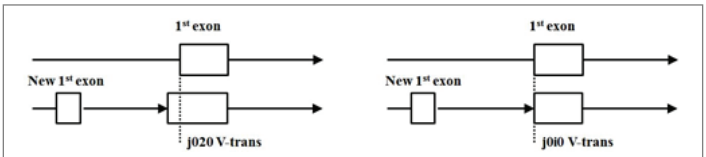
Suppl Figure 4: ORF encoding structure is revealed in the V. cholera genome by gaps between stop codons in the genomic sequence. X-axis shows the size of the gap in codon count between reference codons (stops for conventional ORFs, or 3com set for comparisons in table), Y-axis show the counts. Reprinted with permission [58].



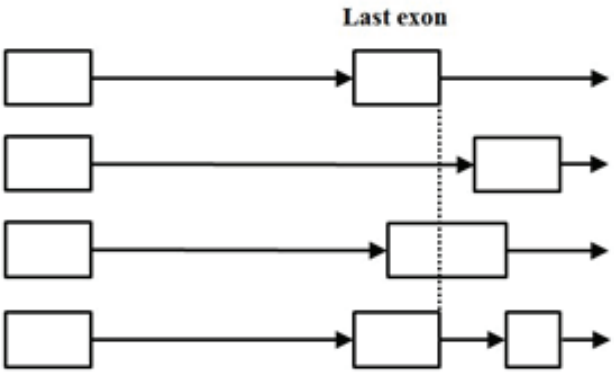
Suppl Figure 6: Quadratic peak in log space indicates Gaussian distribution, an expected result on the sum over smORF distributions due to law of large numbers. The topological index shows a shoulder up to 10040, with double counting on 'non-void' codons due to reverse reads pooling on counts, indicates that there are 20 codons in "strong use". Reprinted with permission [58].



Suppl Figure 7: Topology-Index histograms-*Chlamydia trachomatis* genomein Top panel, and *Deinococcus radiodurans* genome in Bottom panel. *C. trachomatis*, like *V. cholerae*, shows very little overlapping gene structure. *D. radiodurans*, on the other hand, is dominated by genes that overlap other genes.Reprinted with permission [58].

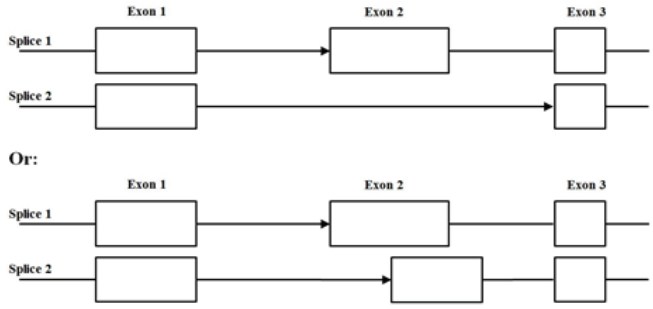


Suppl. Figure 9: Alt-Splice Encoding-Alternative Splicing that results in different start exons. Reprinted with permission [64].



Suppl. Figure 10: Alt-Splice Encoding-Alternative Splicing that results in different end exons. Reprinted with permission [64]

S.3 Alternative Splicing in Eukaryotes



Suppl. Figure 8: Alt-Splice Encoding-Alternative Splicing that results in different internal exons. Reprinted with permission [64]