The Reading Frame Surveillance Hypothesis and the Origins of RNAi

John T. Gray
Audentes Therapeutics, Inc.
600 California Street
17th Floor
San Francisco, California, 94108

Abstract

Pathways of gene regulation mediated by Argonaute proteins derive specificity from short, bound RNA molecules and are increasingly implicated in broad areas of biology, however a unifying principle for the origin of these pathways has been lacking. I here present a hypothetical Reading Frame Surveillance (RFS) model which proposes that primordial ribosomes utilized evenly spaced cleavage of a complementary RNA to monitor and preserve the reading frame during translation of a message. Furthermore, I show how published data from the fields of gene silencing, RNA processing, antigen presentation, development, and oncogenesis support the possibility that vestiges of this primordial pathway are extant and contribute to these processes. The model provides rational mechanistic interpretations of the data leading to multiple testable hypotheses in each of these areas. In particular, the argument is put forward that not only does an RNA dependent RNA polymerase likely exist today in vertebrates, as has been proposed previously, but that it also routinely copies transcribed RNAs in the nucleus, and is likely a ribozyme.

The Reading Frame Surveillance Model

The impetus for this model of primordial protein translation was a theoretical speculation as to how a ribosome might mechanistically prevent +1 frameshifting when translating an mRNA. Reading of mRNA by the ribosome entails direct binding of tRNA anti-codons to each codon of the message (Figure 1), and steric clashes between adjacent tRNAs should

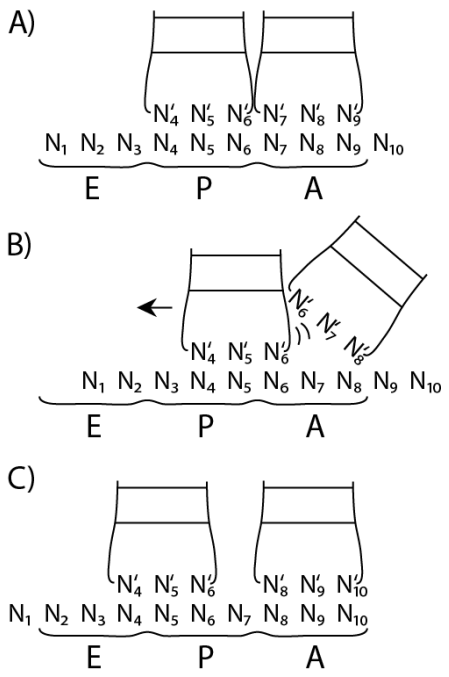

Figure 1. Steric Clashes Generated During -1 but Not +1 Frameshifting
Positioning of A and P site tRNA anticodon stems are shown with (A) proper positioning, (B) -1 frameshifting without the necessary sliding of the P-site tRNA (arrow), and (C) +1 frameshifting, with no steric clashes.

theoretically limit -1 frameshifting to some degree, as that event requires that two tRNAs be bound to the same base on the mRNA, or that the mRNA contain a short homopolymeric sequence that allows slippage of the P-site tRNA to make room for binding in the A-site after the frameshifting 2-nucleotide translocation event, as has been described for extant translating ribosomes[1]. Plus-one frameshifting, on the other hand, could occur without such steric inhibition whenever a one-nucleotide translocation allowed binding of the A-site tRNA in a +1 shifted position relative to the previous codon.

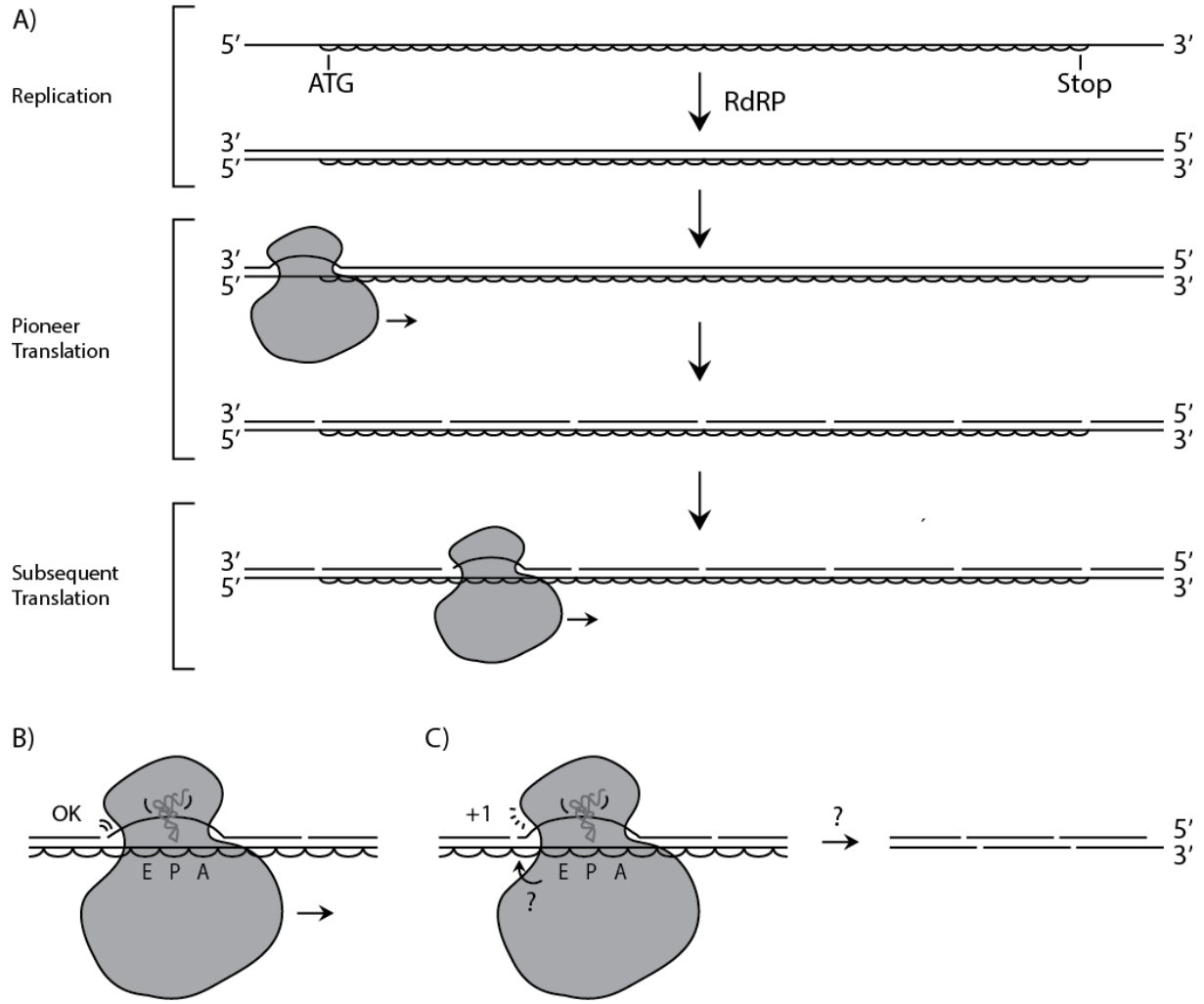

Figure 2. The Reading Frame Surveillance Model

A) The different phases of translation are illustrated as detailed in the text, with the coding RNA indicated by a line with downward bumps representing codons. Copying by RdRP and pioneer translation with a 7-codon phasing decorates the coding RNA segment with evenly spaced 21 nt scRNAs. Subsequent translation proceeds by sequential, transient dissociation of each scRNA from the coding RNA, allowing sensing of the frame at each scRNA terminus. B) Idealized schematic for a possible mechanism of sensing a properly positioned transcript based on scRNA termini. C) Sensing of frameshifted translation, including hypothetical cleavage of the template RNA (left), and possible downstream consequence of repeated cleavage events on transcripts bearing a deletion (right).

Primitive ribosomes have been argued to have been more error prone[2], increasing the challenge of controlling mRNA positioning and translocation to limit +1 frameshifting. Combined with the suggestion that genomic RNA replication would cause primordial translation templates to be double stranded[3], I propose that a primordial ribosome could have used the complementary RNA strand (hereafter referred to as 'cRNA') to measure the template RNA during translation to facilitate recognition and response to frameshifting events. This Reading Frame Surveillance (RFS) hypothesis postulates that measured cleavage of an annealed cRNA copy of the mRNA during a first round of 'scanning' translation could create a molecular ruler that in subsequent rounds of translation allowed continuous monitoring of the progression of the mRNA through the ribosome with respect to the frame being read. An idealized schematic outlining this model translation process is presented in Figure 2a, which shows the postulated primordial translation of a 42 codon mRNA, first in the pioneer round of translation, and later when that same message is translated repetitively. A key element of this model is that during the pioneer round, translation or ribosomal scanning of a uniform set of codons is associated with cleavage of the cRNA template at a set distance from the A-site, which results in the mRNA being decorated with short complementary RNA (hereafter generally referred to as 'scRNA') oligonucleotides. The 7-codon length of the scRNA diagrammed in this figure is somewhat arbitrarily chosen for simplicity, and different lengths of scRNA are consistent with similar mechanistic processes. It is predicted that the length of the scRNA is determined by the ribosome, and derives from a pulsatile process that reads a defined set of codons before pausing and cleaving the cRNA, and then resets the ribosome to translate another segment. Thus during the pioneer round, which by the model must have a reduced error rate with regard to frame-shifting, ribosomal scanning of 7-codon segments would generate 21 nucleotide scRNAs, 8-codon segments 24, etc. The extent to which actual peptidyl bond formation on a growing polypeptide chain would be required for such a mechanism to occur is not clear.

When a previously measured mRNA is subsequently translated, the RFS model postulates that the translational reading frame is monitored by sensing of scRNA termini with respect to the position of the tRNAs bound in the ribosome active sites. Figure 2b crudely presents one of several possible mechanisms for this sensing event, with the ribosome sensing the 5' end of an scRNA after mRNA translocation restores complete scRNA pairing with its complementary sequence on the mRNA, and confirming that translation of the previous segment occurred without frameshifting. Figure 2c presents a schematic of a +1 frameshifted complex, which increases the distance between the ribosome active site and the scRNA 5' end, which could be sensed by the ribosome to trigger alteration or abortion of the translation of that mRNA.

An additional enhancement to the RFS model postulates that scRNA molecules can dissociate from the mRNA template that was translated during their biogenesis and bind to daughter RNA molecules of the same polarity to facilitate RFS on those templates. In this way scRNA could

function as a form of epigenetic memory of previous translations, and facilitate the detection of errors occurring during genome replication. If, for example, a single base pair deletion is created during the replication of an RNA molecule, (which is essentially a +1 frame-shifting event during RNA replication), and scRNAs derived from translation of an intact RNA decorate the new, mutated RNA, then translation downstream of the deletion would be sensed as +1 frameshifted at each scRNA every time it is read by the ribosome, perhaps triggering a more severe response such as cleavage of the RNA message. If such a 'replication proof-reading' cleavage by ribosomes was positioned properly it might generate double stranded RNA molecules with terminal 2 bp 3' overhangs (Fig. 2c, right), as have been found to facilitate loading of RNA fragments onto the RNA Induced Silencing Complex (RISC) in extant eukaryotic cells[4].

RNAi

The RISC is one of a many complexes that participate in the regulatory pathways that are frequently referred to as RNA interference, or RNAi[5]. Over the past two decades, a growing body of research has illuminated the mechanisms whereby 18-30 nucleotide (nt) RNAs provide sequence specificity to the conserved Argonaute family of proteins, which bind the short RNAs and target their regulatory functions using the specificity derived from base pairing of the short RNAs with target RNA or DNA in cells (reviewed in [6-8]). The mechanism by which short RNAs are generated and loaded onto Argonaute proteins is dependent on the pathway in question, but typically require that the RNA initially be paired with a complementary strand and processed by endonucleolytic cleavage to generate the short, single stranded form loaded into the Argonaute active site. In some species, an RNA-dependent RNA Polymerase (RdRP) copies RNA transcripts to generate dsRNA that is subsequently processed to produce the short RNA molecules that are then loaded onto Argonaute proteins. Complexes containing Argonaute domain proteins have been shown to regulate a wide variety of processes in cells, including mRNA translation and stability in the cytoplasm[9,10], suppression of invasive genetic elements[5], and the regulation of mRNA transcription in eukaryotic nuclei[11]. Argonaute family members have been identified in all domains of life[7], suggesting that small RNA regulated pathways are a fundamental element of the molecular biology of living systems.

If an RFS mechanism did exist in the primordial world, then it is natural to wonder whether and how it might have given rise to the extant short regulatory RNA pathways summarized above. It is also tempting to speculate that extant protein translation might in some organisms or contexts still yet implement some mechanistic aspects of the RFS model. Notably, the human Argonaute family member Ago2 was originally characterized as a translation initiation factor (EIF2C)[12], suggesting that an Argonaute domain protein can contribute to translation initiation, which in some systems might include the first round of translation, when the RFS model predicts that cRNA would be cleaved. Although a vast quantity of data on the mechanisms of ribosomal translation (reviewed in [13]) have not explicitly established anything like an RFS mechanism, studies of protein translation are typically limited to complex mixtures of either cellular extracts or less efficient

purified components, and it remains possible that RFS has either not been tested directly in those systems or that RFS is not essential for ribosomal translation to occur in all contexts. Likewise, sequencing of short RNA molecules in cells has not provided evidence that mRNA open-reading frames are paired with short complementary RNA molecules, but for occasional piwi RNAs such as those described in worms[14], and it is not clear whether non-canonical nucleotide structures, such as modified bases[15] or a lack of free terminal hydroxyl residues, or general inefficiencies in the recovery of short RNA molecules, have led to significant under sampling of RFS-derived scRNA-type molecules.

Although many species possess RdRP enzymes, they have been argued to be lacking in all vertebrates and other species which are known to possess robust RNAi type pathways[7], suggesting that the RdRP generation of a cRNA could not happen in vertebrates and that a mechanism similar to RFS could only exist in lower organisms. Yet there is clear evidence that in some vertebrates mRNA is copied into a cRNA in the cell, for example with globin mRNA in developing erythroblasts[15], and exogenously introduced RNA in Xenopus eggs[16]. Also, deep sequencing efforts have revealed antisense RNA molecules in human cells with a structure suggesting they result from copying poly-adenylated mRNA transcripts[17], and which in some cases have the portion of the transcript complementary to the coding sequence trimmed away, as might be expected to result from an RFS process (e.g., the 3' end of the FAU locus in Figure 1, ibid). Lastly, studies of the replication of the hepatitis delta virus convincingly implicate a host encoded RNA dependent RNA polymerase activity in mammalian cells that is necessary for the replication of the viral RNA genome[18]. A reasonable interpretation of this collection of data is that an RdRP enzyme exists in mammalian cells but has eluded specific identification.

Eukaryotic Reading Frame Surveillance

Figure 3 was therefore constructed to explore how an RdRP activity in eukaryotes might support an RFS pathway and influence eukaryotic mRNA transcription and processing, and guide the creation of testable hypotheses to explore the relevance of the RFS model to eukaryotic cell biology. Eukaryotes uniquely possess nuclei wherein the primary transcripts of genes are post-transcriptionally processed to remove introns and splice the remaining exons together, generating mature mRNAs competent for translation in the cytoplasm with contiguous, uninterrupted, open reading frames. The model presented in Figure 3 shows RFS occurring in concert with cRNA synthesis and a process of scanning translation that decorates exonic regions of the primary transcript with scRNAs. Splicing joins these decorated exonic regions, juxtaposing the bordering scRNAs and demarcating their junction with a signaling complex that can adjust the register of RFS monitoring during translation downstream of the junction. After all introns have been removed, and the message contains an acceptable contiguous open reading frame, it is 'licensed' for export to the cytoplasm and robust translation.

Although this model leaves many details to be established, such as how the scanning translation process is initiated and terminated to define which of the many short open reading frames in primary transcripts are decorated with scRNAs, it is

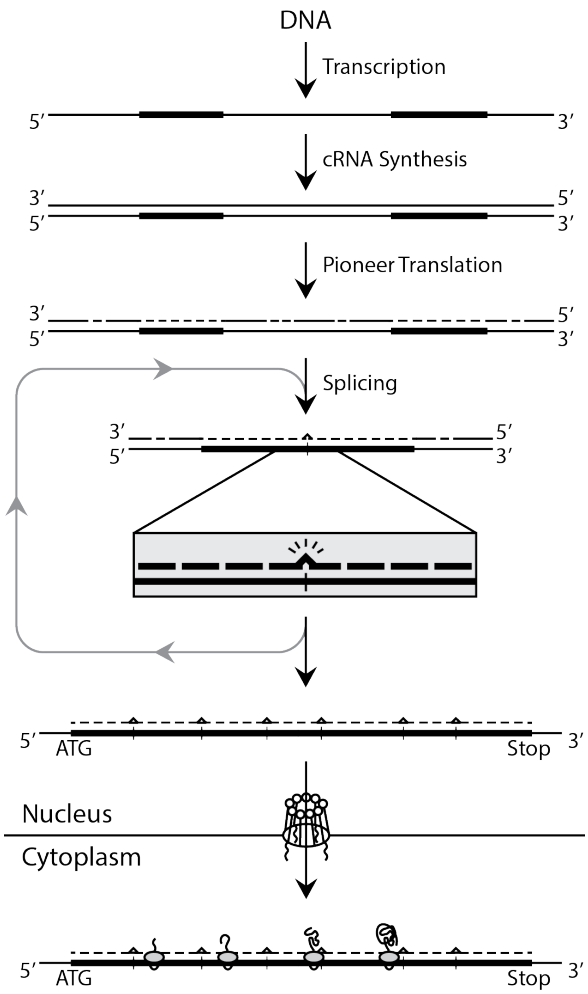

Figure 3. A Model for Eukaryotic RFS
The segregation of exon definition in the nucleus from translation in the cytoplasm is shown schematically. Two internal exons (thicker segments) are shown with RFS pioneer translation leading to decoration of the exonic regions with a series of regularly spaced scRNAs in the nucleus. Discrimination of a minimum length of properly spaced decorations is postulated to define exonic regions relative to introns, which likely also contain some scRNAs. Splicing creates an exon junction complex (EJC), schematically highlighted in the inset. After all exons have been spliced together, the transcript is exported from the nucleus.

consistent with current perspectives on the mechanisms by which exons are selected for inclusion into mRNAs (reviewed in [19]). For example, splice site recognition has long been known to involve complementary pairing of short spliceosomal RNAs to splice site sequences, and although this pairing could occur coordinately with the generation of scRNAs on a message, the simple recognition of splice site consensuses is not always enough to define all splice sites. When introns are long, as frequently occurs in higher eukaryotes, splice site selection is increasingly influenced by the more nebulous mechanism of exon recognition[19]. In support of RFS contributing to this process, studies have implicated the translatability of an exon as contributing to its inclusion into an mRNA[20,21], and recent evidence supports the translation of primary RNAs in the nucleus[22], including portions of those segments destined to be removed as introns[23]. Although clearly not definitive evidence for RFS, it is adequate to at least consider whether the general mechanism of RFS might provide insights into some key open questions of eukaryotic molecular biology, specifically where published data already may support it, and where testable hypotheses can guide future experimentation in extant eukaryotic cells to validate or refine the model.

Codon Usage

In eukaryotes, exons generally have elevated GC content relative to adjacent introns, and in higher eukaryotes this feature is more pronounced in those regions of the genome with lower overall GC content and longer introns[24,25]. This property of coding exons is frequently revealed in analyses of codon usage bias, in particular the GC content of the third ‘wobble’ position of codons. Codon usage bias is often attributed to selective pressure

to maintain optimal translation rates in the context of charged tRNA abundances[26], but there are many possible mechanistic causes for altered use of synonymous codons (reviewed in [27,28]), and recent descriptions of physiologic alteration of tRNA profiles in cells[29] raise the possibility that codon usage bias in transcribed genes is a cause, as opposed to a consequence, of tRNA abundances. In the RFS model, GC-rich coding exons would be predicted to bind to scRNAs with increased thermodynamic stability, which could enhance translational fidelity and processivity. Thus codon usage in eukaryotes could be influenced by reading frame surveillance, by providing selective pressure to maintain high GC content in the wobble position. In support of this hypothesis, the severity of disease causing mutations in the Factor IX gene has been observed to correlate with the change in pairing free energy caused by each coding sequence mutation[30], an observation difficult to explain without invoking a need for pairing of the coding sequences with a complementary strand in some form.

Nonsense Mediated Decay

Mechanistic studies of mRNA splicing have also illustrated a requirement for Exon Junction Complexes (EJC, markers of the location of exon-exon junctions in mRNA), for the Nonsense Mediated Decay pathway of mRNA surveillance (NMD, reviewed in [31]), a pathway that down-modulates levels of mRNA when those RNAs contain nonsense mutations. In eukaryotic cells, pre-termination codons (PTCs) most robustly initiate NMD when positioned upstream of an EJC, which prompted the original hypothesis that a pioneer round of translation is necessary for NMD to occur prior to stabilization and translation of the spliced mRNA in the cytoplasm[32]. The RFS model presented in Figure 3 provides a plausible mechanistic basis for why an EJC might be required for translation in the cytoplasm (to reset the register of RFS at the beginning of each exon) and suggests that scanning of a message for PTCs may occur sequentially and concurrently with RFS across each exon as a spliced message is assembled in the nucleus. Future studies of alternative splicing in vertebrate cells should take into consideration the possibility that exon selection could be influenced by scRNA decorations on primary transcripts resulting from pioneer translation events in the nucleus, a process long hypothesized to be crucial for PTC recognition in the NMD pathway.

Additionally, the RFS model provides a mechanism for the generation of small RNAs that can direct additional regulatory RNA pathways in cells in response to PTCs. For example, the transcriptional silencing of PTC containing genes in the nucleus (Nonsense Mediated Transcriptional Gene Silencing, or NMTGS), which has been shown to be dependent on translation of the mutated RNA in HeLa cells[33], might be mediated by RFS derived scRNAs generated during pioneer translation of PTC containing transcripts and loaded onto a RISC-type complex in response to the abortion of translation by the PTC. In support of this hypothesis, the SMG family proteins, central mediators of the NMD pathway[31], have structural homology to Chp1, a protein shown in yeast to interact with the Argonaute protein Ago1 and mediate heterochromatic silencing of centromeric repeats[34], a well characterized example of RNAi mediated transcriptional silencing. Further characterization of the structure of the EJC may reveal bound, short RNA molecules that can be characterized to

establish whether RFS mechanisms might play a role in the above pathways, or whether other NMD factors, such as UPF1[35], might recognize scRNAs paired to mRNAs and modulate their activities based on scRNA spacing or density.

Epigenetic Silencing

One of the best studied examples of RNAi mediated transcriptional silencing occurs in the *C. elegans* germline, where the worm specific Piwi-clade Argonaute proteins PRG-1 and CSR-1 have been shown to carry epigenetic memories of previously expressed genes via short RNAs complementary to mRNA sequences. PRG-1 mediates an epigenetic silencing pathway termed 'RNAe', which suppresses transcription when piwi RNA (piRNA) transcripts are cleaved and loaded onto PRG-1, which in turn recruits RdRP to mRNA transcripts to propagate and amplify a signal that silences expression of invasive DNA elements[36]. An opposing pathway termed 'RNAa' counteracts RNAe via CSR-1, which also binds RNA complementary to self mRNA transcripts yet mediates the transcriptional activation of genes[37]. Transgenes that are introduced into the worm germline become either silenced or 'licensed' after multiple generations depending on which mode eventually predominates (RNAe vs. RNAa, respectively)[14]. This system, along with other Piwi RNA based pathways in eukaryotes, have been described as genetic immune systems, because they are capable of both licensing genes that are naturally expressed, and using that awareness of self to recognize and silence foreign elements to defend the genome, similar to the way mammalian adaptive immune systems utilize clonal populations of antigen specific lymphocytes to recognize foreign pathogens and proliferate to mount a robust immune response against them.

Cellular immune systems also, however, play a vital role in identifying and controlling the negative consequences of genetic mutation, (for example, cancer[38]), and it is intriguing to speculate that the piRNA system described above for C. elegans might not also promote genome surveillance by enabling the silencing of mutated host encoded alleles to prevent expression of defective proteins. For example, if the first transgene specific RNAe and RNAa molecules generated in the C. elegans germline were actually scRNAs derived from nuclear RFS translation of the newly transcribed RNAs, then those new scRNAs could be routed to the RNAa or RNAe pathways depending upon whether the transgene transcript interacts with the existing cohort of licensing RNAa and RNAe without being identified as a mutated copy of an existing gene transcript. It would therefore be interesting to test whether transgenes possessing in frame and +1 frameshifted identity to previously licensed germline genes trigger RNAa or RNAe differentially, and whether similar studies in animals might reveal a contribution of RFS pathways towards the silencing of gene transfer vectors.

Adaptive Immunity

In another intriguing parallel with the adaptive immune system, antigen presentation into major histocompatibility complex (MHC) Class I molecules, a key step in the immune surveillance of tumors in mammals, has been noted to favor unstable, and by inference, mutated, proteins (the DRiP hypothesis[39,40]), and has recently been demonstrated to occur as a result of nuclear translation[22,23,41]. MHC Class I molecules optimally bind peptides 8-

10 amino acids in length, a size consistent with the translation of 24-30 nts, which is also the size range for many RNAi pathways, suggesting that the pulsatile nuclear translation of RNA at the core of the RFS model could be simultaneously generating both small RNAs and MHC class I peptides in the nucleus. Further characterization of the cis-acting signals that guide the selection of RNAs for nuclear translation may provide validation of the RFS model and expand our understanding of the molecular processes that define adaptive immunity in higher vertebrates.

In addition, it has been observed that immune cytokines can enhance silencing of vector transgenes after gene delivery[42], but how these two molecular pathways interact is unknown. Several innate immune pattern recognition molecules are activated by specific forms of nucleic acids in the cytoplasm[43], a feature which is typically ascribed to the need to detect viral infections, but are increasingly understood to also be triggered by endogenous nucleic acids[44]. Protein Kinase R (PKR), a central component of antiviral defense pathways, is only activated by dsRNA 30 nts or longer[45], and it is not clear what response an mRNA decorated with scRNAs would stimulate. An RFS mechanism of protein translation that simultaneously seeds both Argonaute- and lymphocyte receptor-mediated pathways could be regulated by cellular stress responses to facilitate gene specific responses to mutations that cause stress. Experiments delivering vector transgenes that do or do not encode stress inducing mutant versions of endogenous genes to animals may reveal connections between cellular stress and gene silencing pathways, and how such events could influence not only innate and adaptive host responses to pathogen infection, but also aberrant cellular differentiation, as when for example IL6 facilitates the reversion of differentiated cells into pluripotent states[46].

Development

The transition of a quiescent progenitor cell to a terminally differentiated cell could be argued to require the greatest increase in protein synthesis of any developmental stage, as most differentiated tissues are composed of large cells that contain or secrete high quantities of the proteins that define their structure and function. Although much study has revealed how mutations in genes necessary for regulating the expansion of stem and progenitor cell populations can lead to increased protein translation, unregulated growth, and cancer[47], questions still remain around the linkage between cellular growth and division, and a key question in this research is how cells activate growth at terminal differentiation distinctly from growth to achieve cell expansion. Notably, recent research has shown that protein translation in both hematopoietic[48] and skin[49] stem cells is less robust than it is after those cells differentiate into committed progenitors. The formation of functionally specialized tissue is the ultimate purpose of stem and progenitor cell expansion, so it is reasonable to hypothesize that in many cells the RFS pathway is regulated to become increasingly stringent during terminal differentiation, when tissue specific genes are expressed and at peak demand for protein synthesis, concurrent with blockade of cell cycle progression. Such a system might allow cells to most robustly initiate and monitor the progress of cellular differentiation to ensure the healthy development of tissues and organs.

In healthy mammalian cells, many growth signals are communicated to the

protein translation apparatus via mechanistic/mammalian Target of Rapamycin (mTOR), which is a serine/threonine kinase that stimulates global protein synthesis via phosphorylation of a number of targets (reviewed in [47,50]). Two of the most well studied targets of mTOR are the eIF4F complex, a 3-subunit ribosomal cofactor that regulates cap-dependent translation initiation, and the ribosomal protein S6 Kinase (S6K), a kinase that activates both translation initiation and elongation via phosphorylation of multiple targets. The eIF4F complex, consisting of eIF4A, eIF4E, and eIF4G, is assembled in response to phosphorylation of the eIF4E-binding proteins (4E-BPs) by mTOR, which then release eIF4E and allow it to bind the 7-methyl-guanosine cap structure at the 5' end of mRNAs. This event initiates the nucleation of the other EIF4F components, including eIF4A, a helicase that enhances translation of mRNAs containing double stranded regions in their 5' untranslated regions (UTRs)[50]. S6K activation can also stimulate eIF4F activity via other targets, but was first characterized as the kinase that phosphorylates ribosomal protein S6 (rpS6), a signaling event that affects translation by still unknown mechanisms[51].

The RFS model is a protein translation mechanism, and so if it functions in mammalian cells it would likely be regulated by these established pathways, perhaps as a stringent form of cap-dependent translation that is induced upon cellular differentiation. Consistent with a developmental regulation of RFS stringency, protein-coding transcripts highly expressed in cells at an early developmental stage appear to prefer codons with A or T in the wobble position[52], which as described above would not bind scRNAs as tightly as those containing the more common GC-rich codons characteristic of most mammalian genes. Additionally, cap-dependent translation is initiated by the entire eIF4F complex, which as stated earlier obligately contains an RNA helicase that is capable of melting double stranded RNA structures in the 5' UTRs of many mRNAs. Although this helicase activity most significantly enhances the translation of those mRNAs containing such structures, it remains possible that eIF4A facilitates the translation of all GC-rich mRNAs, when the RFS pathway is active, by melting scRNAs from mRNA templates during cytoplasmic translation. Such a developmentally regulated RFS mechanism could thereby provide an additional layer of complexity to the many characterized mechanisms of translational regulation via 5' UTRs[53]. Experimental confirmation of an RFS pathway influence on mTOR regulated translation, however, might require careful selection of the developmental state of the cells being tested.

One particular system where a developmental onset of the RFS pathway is consistent with published data is in the enigmatic observations of 'revertant' muscle fibers in Duchenne's Muscular Dystrophy (DMD) patients and mouse models of that disease[54]. The *DMD* gene encoding dystrophin is unusually long, with more than 70 exons, and DMD patients often possess mutations that disrupt the open reading frame to generate a protein devoid of the C-terminal domains, which destabilizes the muscle fiber leading to destructive cycles of degeneration and repair leading to profound muscle weakness and often early death. Altered splicing of mutated *DMD* gene transcripts, either induced by exogenous oligonucleotides or occurring spontaneously in rare revertant fibers, can exclude

mutated exons and restore the translational reading frame such that partially functional protein is generated[55]. When visualized via dystrophin domain-specific immunohistochemistry, natural revertant fibers reside in what appear to be clonal clusters that grow by the addition of similarly spliced adjacent fibers after each cycle of degeneration and repair, prompting the hypothesis that tissue resident satellite stem cells pick up epigenetic cues from revertant fibers and use them during differentiation to replicate the specific alternative splicing pattern established in the original revertant fiber[55,56]. These carefully documented observations are strikingly consistent with scRNA decoration of *DMD* primary transcripts during RFS translation in the nucleus and utilization of those scRNAs to establish an RNA splicing isoform-specific epigenetic signature of gene expression that can be communicated either between adjacent cells or from progenitors to differentiated progeny to facilitate homogeneous tissue development and repair. Experimental confirmation of the RFS model could thus not only broadly influence our understanding of the regulation of alternative splicing during mammalian development, but also provide a mechanistic explanation for this established *DMD* gene behavior and stimulate the development of novel strategies for the design of genetic therapies to treat this and other diseases.

Oncogenesis

In mammalian organisms, the accumulation of mutations in somatic cellular genomes causes cancer, in part by abrogating the normal control of proliferative signaling to allow continuous growth, robust protein synthesis, and the disabling of tumor suppression pathways that detect and respond to the mutational events that accumulate as individuals age[38]. The *CDKN2A* tumor suppressor locus (reviewed in [57]), is one of the most frequently mutated loci in tumors, and bears a unique genetic architecture in that two independent promoters generate mRNA transcripts that share an exon translated in two different reading frames (see diagram in Figure 4), prompting consideration of how the RFS pathway might affect its expression. Transcription initiating at exon 1α generates the INK4A transcript and protein, one of a redundant family of 4 tumor suppressors (INK4A, B, C, and D) that regulate the retinoblastoma gene product by binding and inhibiting the cyclin dependent kinases 4 and 6 (CDK4/6) in response to mitogenic stimuli. Transcription products initiating at exon 1β upstream of 1α are translated to produce the ARF protein (p14ARF in humans and p19ARF in mice), which acts via MDM2, (a nuclearly localized protein associated with ribosomes), to activate p53, a key sensor of DNA damage and inducer of senescence and apoptosis. Exon 2 is present in both transcripts, but in the ARF transcript is translated in the -1 frame relative to that of INK4A. The ARF exon 1β promoter is silenced in most stem cells, becomes modestly active during development and with advancing age, and is most strongly activated under conditions of oncogenic stress (e.g., expression of oncogenic Ras). The activation of p53 by the ARF protein requires only amino acids encoded by the first exon (1β), and yet there appears to be strong selective pressure to maintain the absence of in frame stop codons in the second exon shared with the INK4A transcript[58]. The function served by this unique overlapping genetic arrangement has puzzled oncologists for over a decade.

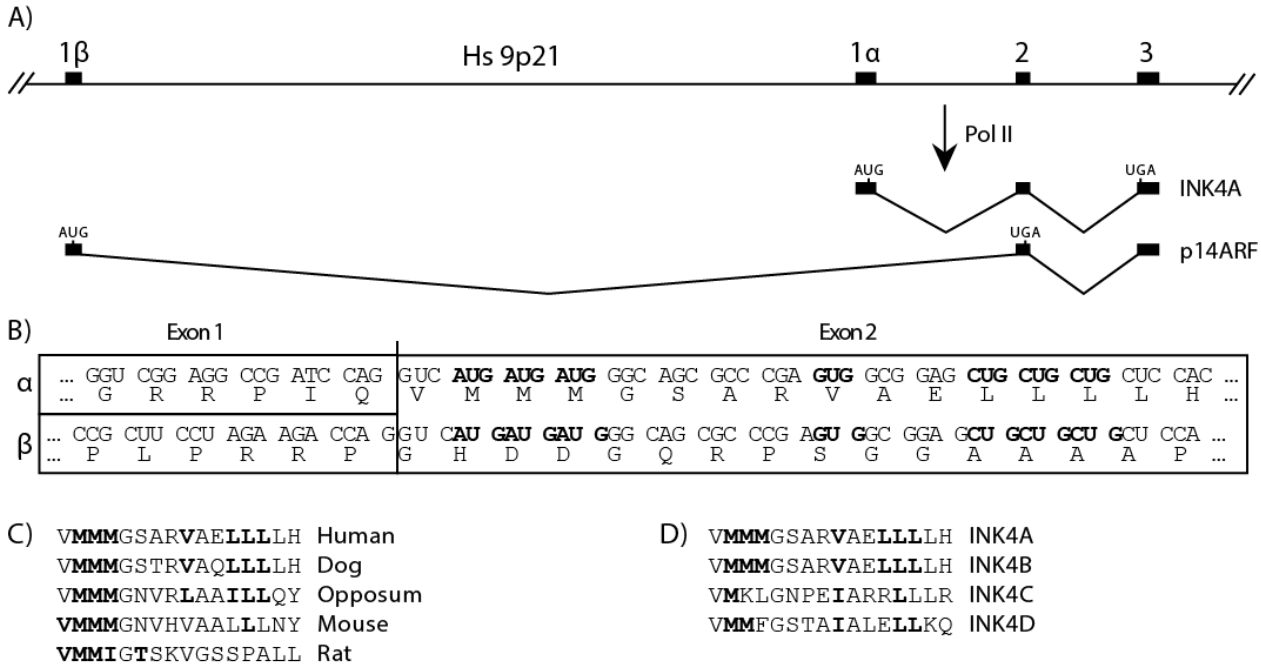

Figure 4. The Human *CDKN2A* Locus
A) To scale drawing of the DNA chromosome (top) and the two RNA transcripts derived from it (below). Each exon is labeled as in [57], and the location of start (AUG) and stop (UGA) codons are shown on the resulting mRNA. B) Detailed sequence view of partial sequences of each transcript, initiating either with exon 1α or 1β. The nucleotide sequence, which is the same for both transcripts in Exon 2, is shown above the translated amino acid sequence, which differs between the two. Potential initiation codons (with 2 of 3 nucleotides matching AUG) are indicated in bold. C) Amino acid sequence alignment for the INK4A transcript from several mammals, again with possible initiation codons in bold. D) Alignment of the corresponding amino acids from all 4 human INK4 paralogs, again with possible initiation codons in bold.

In the RFS model, pioneer translation would decorate both the ARF and INK4A transcripts with scRNAs to demarcate the reading frames translated, and these scRNAs could be subsequently monitored by the ribosome to maintain the fidelity of translation of each message in the proper frame (assuming the scRNAs remain bound to the original mRNA only). On the ARF transcript, the INK4A coding sequence exists in the +1 frame, and the beginning of the shared exon 2 contains a striking array of methionine and other potential initiator codons in the INK4A frame (Figure 4B). This unique cluster of initiation codons is positioned perfectly to enhance the frequency of +1 frameshifted translation on the ARF transcript, which would trigger the proposed RFS pathway and perhaps generally inhibit translation of the ARF protein. The methionines are relatively well conserved in INK4A (Figure 4C), yet X-ray crystallographic structures of the INK4 proteins in complex with CDK4/6 suggest that the methionine side chains are not involved in binding[59], and the INK4C gene contains only one methionine in this position (Figure 4D), even though it is

capable of complementing INK4A and B function[57], and all 4 paralogs reportedly have similar biochemical properties[60]. Although this slightly stronger conservation of initiating codons within the *INK4A* gene relative to the other *INK4* paralogs is not definitive evidence for an RFS mechanism, the lack of perfect conservation of the INK4A exon 2 methionines may reflect a species difference in the degree of RFS influence on ARF translation, and the unusual repetition of initiator codons at the beginning of exon 2 is consistent with a need to maintain the potential for frameshifted translation for proper ARF function.

Promotion of +1 frameshifted translation on the ARF transcript may thus allow the activation of p53 to be governed by not only the activity of the exon 1β promoter, but also the ability of cellular ribosomes to either prevent or sense frameshifting events and respond to them. Prolonged activation of mTOR by oncogenic Ras has been shown to increase the translation of ARF independently from the previously characterized transcriptional activation mediated by Dmp1[61], and although the mechanism whereby this activation occurs is unknown, it is consistent with an RFS mechanism and linkage of that pathway to the down regulation of translation on the ARF transcript. This raises the possibility that the RFS pathway naturally suppresses alternative translation on all transcripts with increasing stringency as development progresses, and initiates specific tumor suppression pathways when this becomes more difficult. Given that tumor antigens can be expressed from alternative reading frames of normal genes[62], it will be valuable to understand whether disabling of RFS is an obligate step for carcinogenesis and cellular immortalization, and how such events might impact the immune profile of tumors. More fundamentally, further exploration of the mechanism of ARF translational activation, in particular in primary cells prior to replicative senescence, should provide insights into how p53 senses and responds to DNA damage, and how the hypothetical RFS pathway might contribute to it and tumor suppression in general.

The RNA Dependent RNA Polymerase

This paper presents a new model for the primordial origin of short regulatory RNA molecules that is not only congruent with several observed properties of modern RNAi systems, but also can be adopted to guide experimentation into the molecular mechanisms of eukaryotic gene regulation and silencing, RNA processing, immunology, development, and oncogenesis. The model is as of yet, however, entirely speculative, and while numerous experiments have been suggested, confirmation of the model most critically awaits confirmation that in extant cells, transcribed RNAs are used as a template to generate complementary RNAs, which are then cleaved at regularly spaced intervals to somehow facilitate the translation of the original template. The published experimental systems described in this work as being suitable for obtaining experimental validation of the RFS model are almost exclusively from higher vertebrate species, which challenges the carefully researched data supporting the absence of an RNA dependent RNA Polymerase in those species[63], an issue which has been debated[16,18].

The requirement for an RdRP at the initiation phase of protein translation might suggest that we should look for it to be associated with the small subunit of the ribosome, which is known to first bind

independently to mRNA templates prior to translation. Older experiments did characterize an RdRP activity specifically associated with ribosomes[64], but that activity was disputed[65] and never identified. The previously mentioned ribosomal protein S6 is conserved throughout eukaryotes, is a well-known target of mTOR via S6K, and might be hypothesized to play a role in regulating such an activity. RpS6 heterozygotes die during gastrulation in a p53 dependent manner[66]. Defective phosphorylation of rpS6, either by deletion of S6K ($S6K^{-/-}$) or by knock-in of phosphorylation site mutations ($rpS6^{P-/-}$), leads to mice with small cells as a result of reduced cell growth, as opposed to increased cell division, and yet global protein translation is mildly up regulated[51]. Recently, it has also been observed that $rpS6^{P-/-}$ murine embryonic fibroblasts show decreased translational accuracy[67], consistent with a model whereby phosphorylation of rpS6 enhances reading frame surveillance to provide the increased translational precision necessary for generation of terminally differentiated cells. S6K phosphorylates rpS6 at its extreme C-terminus, separated by a long alpha-helical segment from the bulk of the protein,

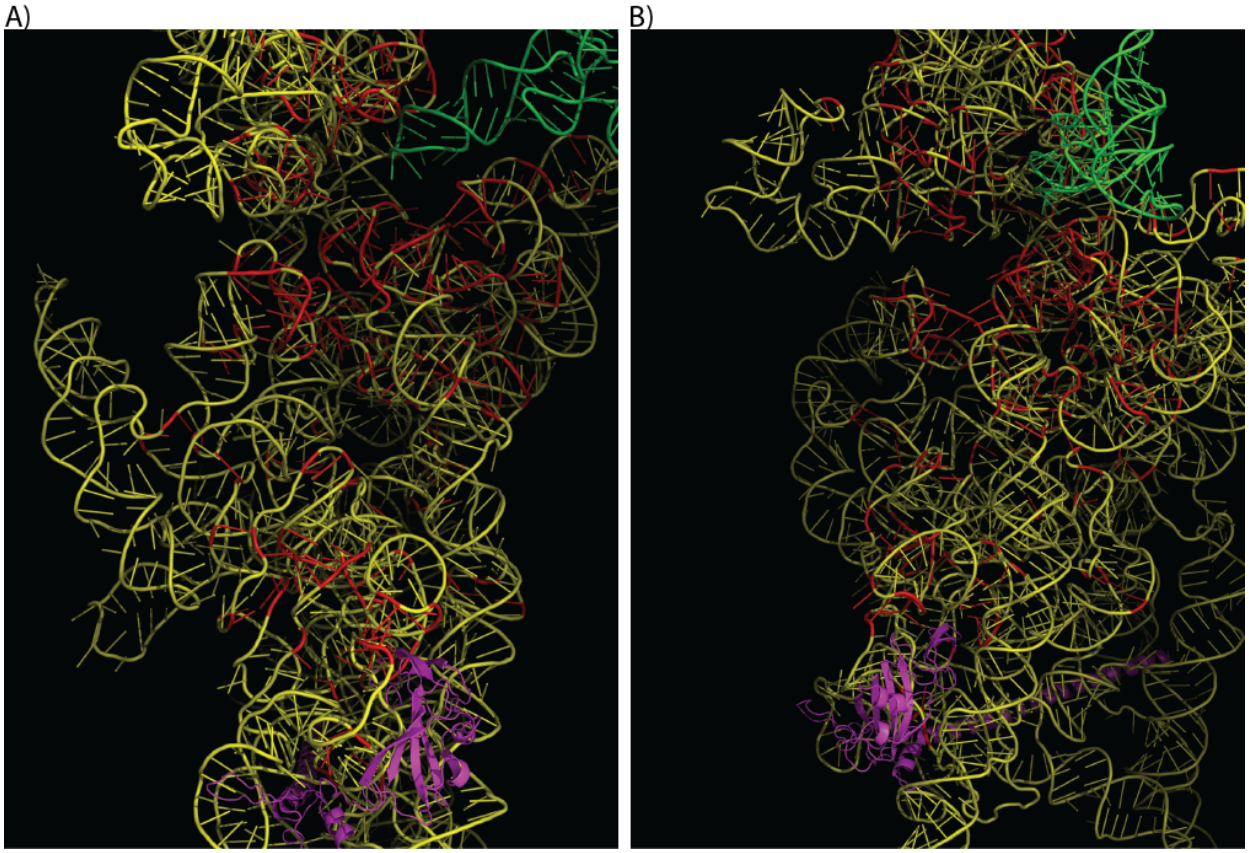

Figure 5. Small Subunit Ribosomal RNA Structure and Sequence Conservation
A) Molecular representation of the human small subunit ribosomal RNA as found in Protein Data Bank record 4V6X, with ribosomal protein S6 in magenta, and E-site tRNA in green. The SSU rRNA is shown in yellow, with residues 100% conserved in all domains of life in red [68]. Side view, with head domain at the top, and the interface with the large subunit on the right. B) View of the interface, with colors as in A.

which is integrated with several small subunit (SSU) RNA helices in the lower part of the SSU body. This region of the ribosome, (sometimes referred to as the 'ratchet' because of its motion relative to the large subunit during translation[69]), is distant (>70 Angstroms) from the decoding center, (the active site for tRNA binding), and yet possesses many residues conserved in all 3 kingdoms of life (Figure 5). It could be postulated that RdRP activity might be associated with this region of the ribosome during translation initiation, when the subunits are separated, and possibly that in eukaryotes phosphorylation of rpS6 activates the RdRP, although how or whether this actually occurs is impossible to determine without first identifying the RdRP enzyme.

The origin of the ribosome has been argued to be one of the seminal events in the history of life on earth, as it allowed primordial self-replicating RNA based systems to synthesize proteins as all modern cells must do to survive[70]. Although we do not know how this transition from an "RNA World" to proteinaceous life occurred, the discovery of RNAs with catalytic activity[71] promoted an awareness that even in extant cells, RNAs can be more than simple information strings or structural scaffolds. Most notably for this discussion, the large subunit ribosomal RNA was itself subsequently determined to be a ribozyme that performs the peptidyl transferase reaction at the heart of protein synthesis[72]. Given that ribozymes can be engineered to possess template dependent polymerase activity[73], and that an RdRP activity is postulated by the RFS model to be integral to primordial ribosomal translation, it seems logical to suggest that the hypothesized RNA dependent RNA polymerase might be a ribozyme, as this would explain why bioinformatic searches for protein based RdRP activity in vertebrates have failed[63], even though there is strong evidence that one exists[18]. It is, perhaps, somewhat radical to propose that the enzyme postulated more than forty years ago to be the molecular ancestor of all life on earth[74,75] still exists, in us, today, but a rigorous vetting of this hypothesis requires that we consider it. Once this step has been taken, is it too bold to propose that it might be the small subunit ribosomal RNA itself? An RdRP ribozyme must have been present at the beginning, when the first peptidyl transferase ribozyme was formed, and there is a harmonious symmetry to the idea that these two primordial ribozymes are each reflected in extant small and large ribosomal subunit RNAs, arguably the two most essential and highly conserved genetic sequences of all life on earth. It is hoped that this sharing of the Reading Frame Surveillance Hypothesis will enable the scientific community to confirm whether or not the idea has merit.

Acknowledgements

I thank Richard Lawn for critical reading of the manuscript. The author has been generously supported by Audentes Therapeutics, Inc., the American Lebanese Syrian Associated Charities, and the Assisi Foundation of Memphis. The author is currently a salaried employee of Audentes Therapeutics, Inc., owns equity interest in that company, and receives royalty payments resulting from licensing of gene transfer technology.

References

1. Chamorro M, Parkin N, Varmus HE. An RNA pseudoknot and an optimal heptameric shift site are required for

highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc Natl Acad Sci U S A.* 1992;89(2):713-717.
2. Woese CR. On the evolution of the genetic code. *Proc Natl Acad Sci U S A.* 1965;54(6):1546-1552.
3. Zenkin N. Hypothesis: emergence of translation as a result of RNA helicase evolution. *Journal of molecular evolution.* 2012;74(5-6):249-256.
4. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* 2001;15(2):188-200.
5. Tabara H, Sarkissian M, Kelly WG, et al. The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell.* 1999;99(2):123-132.
6. Joshua-Tor L, Hannon GJ. Ancestral roles of small RNAs: an Ago-centric perspective. *Cold Spring Harb Perspect Biol.* 2011;3(10):a003772.
7. Burroughs AM, Ando Y, Aravind L. New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. *Wiley interdisciplinary reviews. RNA.* 2014;5(2):141-181.
8. Grewal SI. RNAi-dependent formation of heterochromatin and its diverse functions. *Curr Opin Genet Dev.* 2010;20(2):134-141.
9. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell.* 2009;136(2):215-233.
10. Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet.* 2015;16(7):421-433.
11. Yamanaka S, Mehta S, Reyes-Turcu FE, et al. RNAi triggered by specialized machinery silences developmental genes and retrotransposons. *Nature.* 2013;493(7433):557-560.
12. Koesters R, Adams V, Betts D, et al. Human eukaryotic initiation factor EIF2C1 gene: cDNA sequence, genomic organization, localization to chromosomal bands 1p34-p35, and expression. *Genomics.* 1999;61(2):210-218.
13. Hershey JW, Sonenberg N, Mathews MB. Principles of translational control: an overview. *Cold Spring Harb Perspect Biol.* 2012;4(12).
14. Seth M, Shirayama M, Gu W, Ishidate T, Conte D, Jr., Mello CC. The C. elegans CSR-1 argonaute pathway counteracts epigenetic silencing to promote germline gene expression. *Dev Cell.* 2013;27(6):656-663.
15. Volloch V, Schweitzer B, Rits S. Antisense globin RNA in mouse erythroid tissues: structure, origin, and possible function. *Proc Natl Acad Sci U S A.* 1996;93(6):2476-2481.
16. Pelczar H, Woisard A, Lemaitre JM, Chachou M, Andeol Y. Evidence for an RNA polymerization activity in axolotl and Xenopus egg extracts. *PLoS One.* 2010;5(12):e14411.
17. Kapranov P, Ozsolak F, Kim SW, et al. New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature.* 2010;466(7306):642-646.
18. Tseng CH, Lai MM. Hepatitis delta virus RNA replication. *Viruses.* 2009;1(3):818-831.
19. Hertel KJ. Combinatorial control of exon recognition. *J Biol Chem.* 2008;283(3):1211-1215.
20. Miriami E, Sperling R, Sperling J, Motro U. Regulation of splicing: the importance of being translatable. *RNA.* 2004;10(1):1-4.
21. Jaillon O, Bouhouche K, Gout JF, et al. Translational control of intron splicing in

eukaryotes. *Nature.* 2008;451(7176):359-362.
22. Dolan BP, Knowlton JJ, David A, Bennink JR, Yewdell JW. RNA polymerase II inhibitors dissociate antigenic peptide generation from normal viral protein synthesis: a role for nuclear translation in defective ribosomal product synthesis? *J Immunol.* 2010;185(11):6728-6733.
23. Apcher S, Millot G, Daskalogianni C, Scherl A, Manoury B, Fahraeus R. Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. *Proc Natl Acad Sci U S A.* 2013;110(44):17951-17956.
24. Louie E, Ott J, Majewski J. Nucleotide frequency variation across human genes. *Genome Res.* 2003;13(12):2594-2601.
25. Amit M, Donyo M, Hollander D, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell reports.* 2012;1(5):543-556.
26. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol.Biol.Evol.* 1985;2(1):13-34.
27. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet.* 2011;12(1):32-42.
28. Hershberg R, Petrov DA. Selection on Codon Bias. *Annual Review of Genetics.* 2008;42:287-299.
29. Pavon-Eternod M, Gomes S, Geslain R, Dai Q, Rosner MR, Pan T. tRNA over-expression in breast cancer and functional consequences. *Nucleic Acids Res.* 2009;37(21):7268-7280.
30. Hamasaki-Katagiri N, Salari R, Simhadri VL, et al. Analysis of F9 point mutations and their correlation to severity of haemophilia B disease. *Haemophilia : the official journal of the World Federation of Hemophilia.* 2012;18(6):933-940.
31. Popp MW, Maquat LE. The dharma of nonsense-mediated mRNA decay in mammalian cells. *Mol Cells.* 2014;37(1):1-8.
32. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol.* 1998;18(9):5272-5283.
33. Stalder L, Muhlemann O. Transcriptional silencing of nonsense codon-containing immunoglobulin micro genes requires translation of its mRNA. *J.Biol.Chem.* 2007;282(22):16079-16085.
34. Schalch T, Job G, Shanker S, Partridge JF, Joshua-Tor L. The Chp1-Tas3 core is a multifunctional platform critical for gene silencing by RITS. *Nat Struct Mol Biol.* 2011;18(12):1351-1357.
35. Hogg JR, Goff SP. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell.* 2010;143(3):379-389.
36. Shirayama M, Seth M, Lee HC, et al. piRNAs initiate an epigenetic memory of nonself RNA in the C. elegans germline. *Cell.* 2012;150(1):65-77.
37. Conine CC, Moresco JJ, Gu W, et al. Argonautes promote male fertility and provide a paternal memory of germline gene expression in C. elegans. *Cell.* 2013;155(7):1532-1544.
38. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* 2011;144(5):646-674.
39. Yewdell JW, Anton LC, Bennink JR. Defective ribosomal products (DRiPs): a major source of antigenic peptides for

MHC class I molecules? *J Immunol.* 1996;157(5):1823-1826.
40. Dolan BP, Bennink JR, Yewdell JW. Translating DRiPs: progress in understanding viral and cellular sources of MHC class I peptide ligands. *Cell Mol Life Sci.* 2011;68(9):1481-1489.
41. Apcher S, Daskalogianni C, Lejeune F, et al. Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *Proc Natl Acad Sci U S A.* 2011;108(28):11572-11577.
42. Suzuki M, Bertin TK, Rogers GL, et al. Differential type I interferon-dependent transgene silencing of helper-dependent adenoviral vs. adeno-associated viral vectors in vivo. *Mol Ther.* 2013;21(4):796-805.
43. O'Neill LA, Bowie AG. Sensing and signaling in antiviral innate immunity. *Curr Biol.* 2010;20(7):R328-333.
44. Roers A, Hiller B, Hornung V. Recognition of Endogenous Nucleic Acids by the Innate Immune System. *Immunity.* 2016;44(4):739-754.
45. Lemaire PA, Anderson E, Lary J, Cole JL. Mechanism of PKR Activation by dsRNA. *J Mol Biol.* 2008;381(2):351-360.
46. Brady JJ, Li M, Suthram S, Jiang H, Wong WH, Blau HM. Early role for IL-6 signalling during generation of induced pluripotent stem cells revealed by heterokaryon RNA-Seq. *Nat Cell Biol.* 2013;15(10):1244-1252.
47. Ruggero D. Translational control in cancer etiology. *Cold Spring Harb Perspect Biol.* 2013;5(2).
48. Signer RA, Magee JA, Salic A, Morrison SJ. Haematopoietic stem cells require a highly regulated protein synthesis rate. *Nature.* 2014;509(7498):49-54.
49. Blanco S, Bandiera R, Popis M, et al. Stem cell function and stress response are controlled by protein synthesis. *Nature.* 2016;534(7607):335-340.
50. Roux PP, Topisirovic I. Regulation of mRNA translation by signaling pathways. *Cold Spring Harb Perspect Biol.* 2012;4(11).
51. Meyuhas O. Ribosomal Protein S6 Phosphorylation: Four Decades of Research. *Int Rev Cell Mol Biol.* 2015;320:41-73.
52. Ren L, Gao G, Zhao D, Ding M, Luo J, Deng H. Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. *Genome biology.* 2007;8(3):R35.
53. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science.* 2016;352(6292):1413-1416.
54. Hoffman EP, Morgan JE, Watkins SC, Partridge TA. Somatic reversion/suppression of the mouse mdx phenotype in vivo. *J Neurol Sci.* 1990;99(1):9-25.
55. Lu QL, Morris GE, Wilton SD, et al. Massive idiosyncratic exon skipping corrects the nonsense mutation in dystrophic mouse muscle and produces functional revertant fibers by clonal expansion. *J Cell Biol.* 2000;148(5):985-996.
56. Yokota T, Lu QL, Morgan JE, et al. Expansion of revertant fibers in dystrophic mdx muscles reflects activity of muscle precursor cells and serves as an index of muscle regeneration. *Journal of cell science.* 2006;119(Pt 13):2679-2687.
57. Sherr CJ. Ink4-Arf locus in cancer and aging. *Wiley Interdiscip Rev Dev Biol.* 2012;1(5):731-741.

58. Szklarczyk R, Heringa J, Pond SK, Nekrutenko A. Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. *Proc Natl Acad Sci U S A.* 2007;104(31):12807-12812.
59. Russo AA, Tong L, Lee JO, Jeffrey PD, Pavletich NP. Structural basis for inhibition of the cyclin-dependent kinase Cdk6 by the tumour suppressor p16INK4a. *Nature.* 1998;395(6699):237-243.
60. Roussel MF. The INK4 family of cell cycle inhibitors in cancer. *Oncogene.* 1999;18(38):5311-5317.
61. Miceli AP, Saporita AJ, Weber JD. Hypergrowth mTORC1 signals translationally activate the ARF tumor suppressor checkpoint. *Mol Cell Biol.* 2012;32(2):348-364.
62. Wang RF, Parkhurst MR, Kawakami Y, Robbins PF, Rosenberg SA. Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J Exp Med.* 1996;183(3):1131-1140.
63. Zong J, Yao X, Yin J, Zhang D, Ma H. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene.* 2009;447(1):29-39.
64. Downey KM, Byrnes JJ, Jurmark BS, So AG. Reticulocyte RNA-dependent RNA polymerase. *Proc Natl Acad Sci U S A.* 1973;70(12):3400-3404.
65. Boyd CD, Fitschen W. Ribosome bound RNA-dependent RNA polymerase: changes in total and specific activity during maturation of avian erythrocytes. *Biochem Biophys Res Commun.* 1975;67(3):1210-1218.
66. Panic L, Tamarut S, Sticker-Jantscheff M, et al. Ribosomal protein S6 gene haploinsufficiency is associated with activation of a p53-dependent checkpoint during gastrulation. *Mol Cell Biol.* 2006;26(23):8880-8891.
67. Wittenberg AD, Azar S, Klochendler A, et al. Phosphorylated Ribosomal Protein S6 Is Required for Akt-Driven Hyperplasia and Malignant Transformation, but Not for Hypertrophy, Aneuploidy and Hyperfunction of Pancreatic beta-Cells. *PLoS One.* 2016;11(2):e0149995.
68. Petrov AS, Bernier CR, Gulen B, et al. Secondary structures of rRNAs from all three domains of life. *PLoS One.* 2014;9(2):e88222.
69. Frank J, Agrawal RK. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature.* 2000;406(6793):318-322.
70. Woese CR. Translation: in retrospect and prospect. *Rna.* 2001;7(8):1055-1067.
71. Zaug AJ, Cech TR. The intervening sequence RNA of Tetrahymena is an enzyme. *Science.* 1986;231(4737):470-475.
72. Noller HF, Hoffarth V, Zimniak L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science.* 1992;256(5062):1416-1419.
73. Bartel DP, Doudna JA, Usman N, Szostak JW. Template-directed primer extension catalyzed by the Tetrahymena ribozyme. *Mol Cell Biol.* 1991;11(6):3390-3394.
74. Crick FH. The origin of the genetic code. *J Mol Biol.* 1968;38(3):367-379.
75. Orgel LE. Evolution of the genetic apparatus. *J Mol Biol.* 1968;38(3):381-393.