Prensner John R. (Orcid ID: 0000-0002-7024-636X)

Shining a Light on the Dark Proteome: Non-Canonical Open Reading Frames and their Encoded MiniProteins as a New Frontier in Cancer Biology

Authors: Zoe Posner[1], Ian Yannuzzi[1], and John R. Prensner[2]

Affiliations: [1]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [2]Department of Pediatrics, University of Michigan, MI, USA.

Keywords: Miniprotein, Non-coding genome, Open reading frame, cancer

Word count: 7332

Address correspondence to:
John Prensner, MD PhD
Assistant Professor of Pediatrics, University of Michigan
Medical Science Research Building II, Room 2560B
1150 Medical Center Drive
Ann Arbor, MI, 48109
Email: prensner@umich.edu
Phone: 734-936-9814

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pro.4708

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

Abstract

In the decades following the discovery that genes encode proteins, scientists have tried to exhaustively and comprehensively characterize the human genome. Recent advances in computational methods along with transcriptomic and proteomic techniques have now shown that historically non-coding genomic regions may contain non-canonical open reading frames (ncORFs), which may encode functional miniproteins or otherwise exert regulatory activity through coding-independent functions. Increasingly, it is clear that these ncORFs may play critical roles in major human diseases such as cancer. In this review, we summarize the history and current progress of ncORF research and explore the known functions of ncORFs and the miniproteins they may encode. We particularly highlight the emerging body of evidence supporting a role for ncORFs and miniproteins contributions in cancer. Finally, we provide a blueprint for high-priority areas of future research for ncORFs in cancer, focusing on ncORF detection, functional characterization, and therapeutic intervention.

Word count: 146

Author Manuscript

Introduction

Gregor Mendel's proposal that heredity is transmitted in discrete units spurred a long-term interest to dissect, quantify, and characterize such units—known today as genes. Some 60 years after Mendel's death, a major breakthrough came when George Beadle and Edward Tatum performed their famed bread mold experiments to form the "one gene, one enzyme hypothesis" (Beadle & Tatum, 1941). Winning not only the 1958 Nobel Prize for their work, they would inspire a generation of leading researchers. By 1963, only 10 years from the advent of the double helix model of DNA, much about the central dogma was already appreciated: DNA is transcribed into mRNA, which encodes proteins via triplet codons with help from transfer RNAs and ribosomes (Nirenberg, 1963).

Which nucleotides, then, does the ribosome choose to translate? Seminal work from Joan Steitz defined that ribosomes engage RNAs in ~29 nucleotide fragments (Steitz, 1969; Steitz, Dube, & Rudland, 1970), and comprehensive efforts by Marilyn Kozak worked to compile the first catalog of 211 known messenger RNAs (Marilyn Kozak, 1984), leading to the scanning model of ribosome initiation (M. Kozak, 1989). Yet, it was the advent of DNA sequencing and the efforts of the Human Genome Project that resulted in the assembly of the first complete human genome, bringing the possibility of quantifying and characterizing all human genes nearer. Taking this a step further, the Encyclopedia of DNA Elements (ENCODE), formed shortly after the Human Genome Project's completion, sought to characterize all functional genomic elements in humans (ENCODE Project Consortium et al., 2007).

The Human Genome Project first estimated that the human genome contains approximately 30,000-40,000 protein-coding genes (Consortium & Mouse Genome Sequencing Consortium, 2002; Lander et al., 2001). Later scrutiny trimmed this list to ~25,000 (International Human

Author Manuscript

Genome Sequencing Consortium, 2004), and since the mid-2000s, this list has remained fairly static, settling at an estimation of 19,600 protein-coding genes (Harrow et al., 2012). By and large, identification of these genes in humans and other mammals relied on a number of parametric assumptions; namely, a protein size (>100 amino acids), an AUG start codon, monocistronic transcript requirements, and no overlap in open reading frames (Consortium & Mouse Genome Sequencing Consortium, 2002; Lander et al., 2001). In rare individual cases, these guidelines were modified, such as for the annotation of the *CDKN2A* gene, which was previously well-characterized to encode both the INK4a and p14-ARF proteins (Ouelle, Zindy, Ashmun, & Sherr, 1995), as well as for the transcript encoding the RPP14 and HTD2 proteins (Autio et al., 2008).

In this historical context, the two decades since the completion of the Human Genome Project have witnessed a tremendous expansion of research on RNA translation. Collectively, this work has spurred the nomination of thousands of human genetic elements that interact with the ribosome but are not annotated as protein coding in major reference databases. These elements are termed non-canonical open reading frames (ncORFs). ncORFs are often small (<100 amino acids), can have near-cognate start codons, may exist on a polycistronic transcript, and may overlap with other open reading frames (Mohaupt et al., 2022). Since these features defy many of the annotation guidelines set following the Human Genome Project, many ncORFs were missed during early gene discovery efforts. A subset of these nominated ncORFs likely encode stable small protein species termed miniproteins, which may have important implications for the human proteome. Representing a new aspect of human genome functionality, ncORFs promise to yield important insights into diseases like cancer. As a disease of dysregulated signaling and developmental pathways, ncORFs and miniproteins represent a class of yet unexplored potential cancer vulnerabilities and predictive biomarkers through intrinsic and regulatory functions.

Author Manuscript

In this review, we will provide an overview of the state of ncORFs research, describe the implications of these ncORFs in cancer, and provide a future outlook for ncORFs as new players in the landscape of cancer research.

2. Early Investigations of ncORFs in the Human Genome

2.1 Miniprotein Discovery Before the Human Genome Project

The origins of ncORF research date back to early observations of small open reading frames. In 1987, Marilyn Kozak explored "nonfunctional" upstream AUG codons in known vertebrate mRNAs (M. Kozak, 1987). Kozak interpreted these upstream open reading frames (uORFs) as a mechanism of translational regulation through ribosomal hogging, leading to decreased translation of the downstream CDS. While Kozak's work on uORFs suggested ribosome engagement at non-canonical sites, it did not address whether ncORF translation results in stable and bioactive miniproteins, despite interest in whether this was the case. Rather, the topic of discovery of protein-coding CDSs was largely based on early computational approaches, which struggled to predict real, functional miniproteins and differentiate them from those that might arise by chance among human DNA sequences. As such, early protein/miniprotein research relied on the use of cDNA libraries, ribosome binding models, and serial analysis of gene expression (SAGE) experiments to nominate candidate ncORFs (Hemm, Paul, Schneider, Storz, & Rudd, 2008; Kastenmayer et al., 2006; Kondo et al., 2007; V. E. Velculescu et al., 1997; Victor E. Velculescu, Zhang, Vogelstein, & Kinzler, 1995).

Despite these difficulties, the possibility of miniproteins persistently sparked attention from groups interested in genome characterization, particularly in non-mammal organisms such as *Drosophila*, *Escherichia coli*, and *Saccharomyces cerevisiae* (Hemm et al., 2008; Kastenmayer et

Author Manuscript

al., 2006; Kondo et al., 2007). Yeast was the first eukaryotic organism to gain general acceptance for the fact that ncORFs are widespread throughout its genome. Using early cDNA analyses of the yeast transcriptome along with homology-based inferences, one study identified 299 ncORFs within the *S. cerevisiae* genome, representing approximately ~5% of all *S. cerevisiae* genes (Kastenmayer et al., 2006). Many of these genes were essential for yeast viability (Kastenmayer et al., 2006). Additional work on ncORFs in unicellular organisms was also pursued in *E. coli*, where early investigations used sequence conservation and ribosome binding site models to identify potential ncORFs of 16-50 aa in the intergenic regions of the genome (Hemm et al., 2008). This work not only observed multiple candidate ncORFs, but also validated 20 previously identified and 18 newly predicted ncORFs using endogenous epitope tagging of the genomic site (Hemm et al., 2008).

Drosophila was the first multicellular organism to be extensively interrogated for ncORFs. Here, studies first identified candidate ncORFs through cDNA sequencing libraries that aimed to uncover non-coding RNA (ncRNA) transcripts (Inagaki et al., 2005; Rubin et al., 2000; Tupy et al., 2005). For example, initial analyses identified a putative ncRNA termed MRE29, which was nominated as a non-coding RNA due to the absence of a long coding sequence >100 amino acids. Yet, this transcript was later found to encode four embryonic lethal ncORFs involved in regulation of actin-based cell morphogenesis (Kondo et al., 2007). Likewise, the *tarsal-less* gene was characterized as a carefully regulated and locally expressed polycistronic mRNA encoding several ncORFs of 11 aa involved in development (Galindo, Pueyo, Fouix, Bishop, & Couso, 2007). As such, while appreciation for ncORFs in non-mammal genomes grew, there remained challenges in annotation and validation of ncORFs in the human genome.

Author Manuscript

2.2 Early Methods in Protein-Coding Gene Annotation

Initial efforts to identify the locations of protein-coding genes in mammals relied on genetic (or linkage) mapping methods (Bell & Haldane, 1937; Robson, 1988), which determined the location of genes on a chromosome through analyzing patterns of inheritance in families or by comparing the DNA of different, unrelated individuals. However, genetic mapping, particularly through classical pedigree analysis, is low-throughput and impractical for large-scale annotation efforts. With the advent of Sanger sequencing, large-scale projects became practical, prompting the development of methods to determine protein-coding regions of the genome.

The earliest of these methods involved early bioinformatic approaches, including neural network and information theory techniques to predict protein-coding regions from DNA (Farber, Lapedes, & Sirotkin, 1992; Lapedes, Barnes, Burks, & Farber, 1989). However, such work was dependent on training computational models on the limited numbers of previously annotated protein-coding DNA sequences, making model performance highly dependent on the length of the ORF to the exclusion of ORFs shorter than 90 codons (Farber et al., 1992). Therefore, cDNA libraries were used as an alternative method to identify functional miniproteins (Inagaki et al., 2005; Rubin et al., 2000; Stallmeyer, Drugeon, Reiss, Haenni, & Mendel, 1999; Tupy et al., 2005). These projects aimed to annotate cDNA clones containing the full-length ORF sequence for each gene (Rubin et al., 2000). These cDNA libraries became important resources for identifying bicistronic transcripts (Stallmeyer et al., 1999) and putative lncRNAs (Inagaki et al., 2005; Tupy et al., 2005)—many of which were later determined to encode miniproteins (Hartford & Lal, 2020).

Despite early successes, cDNA libraries have to overcome some significant limitations—such as leading to false positive and negative results, and only being representative of sequences

Author Manuscript

from mature mRNA. For example, the lncRNA, *Xist*, which plays a major role in X-chromosome inactivation, was originally thought to be a protein-coding gene from a cDNA library experiment (Borsani et al., 1991). Since start codons and ORFs are inferred based on cDNA sequence, it can be challenging to differentiate true lncRNAs from non-canonical ORFs. Moreover, cDNA libraries may over-represent the most abundant mRNA transcripts, thus underrepresenting or completely missing lowly expressed transcripts (P. Liang & Pardee, 1992). Since the entire human genome was not fully sequenced at this point, early cDNA libraries reconstructed transcripts through aligning cDNAs to each other and concatenating cDNA fragments. This method, however, became increasingly skeptical of small ORFs—which had a greater likelihood of representing individual portions of a single larger ORF once the cDNA fragments were concatenated. This challenge led to early methods establishing a 100 aa cutoff to decrease the chance of such artifacts, but this made false negative results for ncORFs more common (Burge and Karlin 1997). Ultimately, early sequence prediction approaches and use of cDNA libraries yielded results that were challenging to interpret, necessitating new and improved methods to accurately identify miniproteins.

2.3. Miniprotein Discovery After the Human Genome Project

Sequence-Based Computational Miniprotein Prediction

Following the Human Genome Project, miniprotein discovery and protein-coding annotation relied heavily on computational methods (**Figure 1**) (Z. Wang, Chen, & Li, 2004), particularly evolutionary-based alignment techniques (M. F. Lin, Jungreis, & Kellis, 2011; Mudge et al., 2019; L. Wang et al., 2013). Pressure to develop these methods increased with the rise in lncRNA discovery through the invention of RNA sequencing (Cabili et al., 2011; Guttman et al., 2009; Trapnell et al., 2010). From work in bacterial genomes, the CRITICA (Coding Region

Author Manuscript

Identification Tool Invoking Comparative Analysis) program was developed as one of the earliest ncORF annotation tools (Badger & Olsen, 1999; Frith et al., 2006). CRITICA combines a comparative sequence homology alignment analysis with a statistical analysis of hexanucleotide frequencies in predicted coding and non-coding frames. Applying this method to large mouse cDNA collections identified more than 3,000 potential ORFs < 100 aa (Frith et al., 2006). Despite these advances, computational prediction methods still suffered from high false-positive rates, casting doubt on the true biological relevance of ncORFs.

Methods involving conservation-based comparative analyses significantly improved bioinformatic protein prediction and include conservation-analysis at both the nucleotide and amino-acid levels (Lin et al. 2011; Trapnell et al. 2010; Mudge et al. 2019; Mackowiak et al. 2015). To illustrate, the program PhyloCSF uses a multispecies nucleotide alignment to calculate the likelihood that a particular sequence represents a conserved protein-coding region (M. F. Lin et al., 2011). PhyloCSF has been applied to whole-genome sequencing data for multiple model organisms, facilitating the identification of some novel protein-coding genes (Mudge et al., 2019). However, ncORFs are generally less well conserved than canonical protein-coding ORFs and these models still rely on previously identified ncORFs for training and validation; both caveats necessitate continued optimization of these methods for ncORFs (Dinger, Pang, Mercer, & Mattick, 2008).

Outside of strict conservation-based methods of miniprotein prediction, other algorithms include AUGUSTUS, sOFR finder, and Coding-Potential Assessment Tool (CPAT). AUGUSTUS establishes a generalized Hidden Markov Model and combines ab initio gene prediction with a comparative genomics approach to accurately identify gene structures (Stanke et al., 2006). AUGUSTUS uses a species-specific training set of high-quality gene annotations to develop

Author Manuscript

parameters to predict new genes by first identifying potential transcription start sites and promoter elements. Furthermore, AUGUSTUS uses alignments to closely related species to detect regions of conservation and provide further statistical power for accurate gene prediction. Unlike AUGUSTUS, sORF finder was developed specifically for the prediction of small ORFs (Hanada et al. 2010). sORF finder uses a high-quality nucleotide sequence as input—scanning the input sequence in all reading frames and searching for ORFs that meet certain criteria, such as minimum length, absence of internal stop codons, sequence conservation, codon usage bias, and secondary structure predictions. Lastly, the CPAT program combines machine learning algorithms with sequence features to differentiate between coding and noncoding sequences (Wang et al. 2013). It utilizes a training set of known coding and non-coding sequences to establish positive and negative training examples. CPAT extracts various sequence features from the training sequences, such as nucleotide composition, ORF length, ORF coverage, hexamer usage bias, and conservation scores, to capture discriminatory patterns between coding and non-coding regions. Using the extracted features, CPAT trains a machine learning model, typically a support vector machine (SVM) algorithm. Once the machine learning model is trained, CPAT can be applied to assess the coding potential of novel transcript sequences. For additional information on these and other computational tools, there are more comprehensive reviews suggested here (Wang et al. 2004; Do and Choi 2006; Alioto 2012; Goel et al. 2013; Klasberg et al. 2016).

Mass Spectrometry for Miniprotein Discovery

To address limitations of phylogenetic solely computational-based approaches, advances in mass spectrometry (MS) provided crucial experimental evidence of miniprotein translation (Oyama et al., 2004; Svensson, Sköld, Svenningsson, & Andren, 2003). Mass spectrometry, which

Author Manuscript

can directly identify amino acid sequences from miniproteins, helped to facilitate the first high-throughput experimental evidence of miniproteins (Oyama et al., 2004, 2007; Slavoff et al., 2012). Importantly,, shotgun MS techniques are biased against miniproteins due to fewer tryptic cleavage sites, greater protein instability, and lack of comparative databases for analysis (Yewdell, 2022). Optimization of MS techniques through biochemical enrichment, less stringent cutoffs, and other prior isolation steps addresses these limitations (Orr, Mao, Storz, & Qian, 2020). For example, Slavoff *et al.* 2012 combined MS with RNA sequencing, enriching for ncORFs by inhibiting degradation and using electrostatic repulsion hydrophilic interaction chromatography to separate peptides prior to MS. Leveraging a custom RNA-seq database, Slavoff and colleagues identified evidence of protein translation for 86 novel ncORFs (Slavoff et al., 2012). Further optimization of the MS fractionation step has facilitated discovery of additional novel ncORFs (J. Ma et al., 2016, 2014).

Slavoff and colleagues' use of a custom RNA-seq database highlights the broader point that curating protein sequence databases with improved genomic and transcriptomic data increases the ability to identify ncORF-encoded miniproteins through MS. (Nesvizhskii 2014). A variety of computational programs, including customProDB, Galaxy Integrated Omics, and QUILTs can aid in the design of these custom databases, and are comprehensively reviewed elsewhere (Ruggles et al. 2017). As newer proteomics techniques improve in sensitivity (Brunner et al., 2022), these methods will become increasingly important for characterizing the landscape of ncORFs and clarifying intrinsic versus regulatory functions.

Ribosome Profiling

Author Manuscript

Ribosome profiling, or Ribo-seq, has been the most successful method for identifying ncORFs. This technique uses deep sequencing of ribosome footprints, or the fragment of mRNA that is protected from nucleolytic digestion by the ribosome, resulting in an indirect measurement of active translation (Ingolia, Ghaemmaghami, Newman, & Weissman, 2009). Therefore, a wide variety of distinct classes of ncORFs have been described with Ribo-seq (**Figure 2**). While this technique can provide high resolution mapping of ORFs, the quality of ribosome profiling data and fluctuations in periodicity can alter the ncORFs detectable through this approach (Hsu et al., 2016). As Ribo-seq has grown in popularity, a large number of computational tools have been developed to analyze profiling data, particularly for ORF identification. Many of these tools, such as RiboTools and RiboSeqR are notable for ease of use given limited computational background. For greater review of computational tools and approaches for Ribo-seq analysis, we direct our readers to recent comprehensive reviews (Calviello and Ohler 2017; Wang et al. 2019; Kiniry et al. 2020).

Ribo-seq is also unlimited in its ability to detect non-AUG start codons, both for ncORFs as well as annotated CDSs (Andreev et al., 2022; Fedorova, Kiniry, Andreev, Mudge, & Baranov, 2022; Ingolia, Lareau, & Weissman, 2011). Application of drugs, such as homoharringtonine or lactimidomycin, that stall ribosomes at translation initiation sites enable further resolution on the use of non-AUG start codons (Lee et al., 2012). Notably, Ribo-seq has the potential to describe very tiny ncORFs with putative functional miniproteins (Sandmann et al., 2023), whereas MS may be unable to resolve very small ORFs due to challenges with uniquely aligning peptides less than 8 amino acids. Given the utility of Ribo-seq, it has been used as the basis for genomic annotation of ncORFs in standardized databases (Mudge et al., 2022). Finally, efforts to expand ncORF

databases are ongoing. Especially notable are recent efforts to define the landscape of human small ORF (smORF) translation in primary cells and human tissues (Chothani et al. 2022).

3. ncORFs in Cancer

With the advances in Ribo-seq and MS enabling identification of a growing number of ncORFs, cancer biologists have increasingly taken note of whether these ncORFs—and the miniproteins they encode—may expand knowledge of cancer pathogenesis and treatment options. With the possibility of both protein-coding or coding-independent activities, ncORFs may act as crucial regulators of transcription, translation, and post-translational protein function (Y. Liu, Zeng, & Wu, 2022), and characterization of ncORFs and miniproteins has revealed that many are tightly associated with cancer and frequently dysregulated early in oncogenesis (Erady et al., 2021; Zou et al., 2019). Extending this further, CRISPR/Cas9 loss of function screens targeting ncORFs have identified multiple miniproteins essential for cancer cell viability (J. Chen et al., 2020; Prensner et al., 2021). Here, we review the major coding functions of ncORFs and their regulatory interests, highlighting their contributions to cancer development and progression (**Figure 3**).

3.1. ncORF-Intrinsic Interests

An intriguing conclusion from ncORF discovery efforts is that some ncORFs indeed expand the known proteome by encoding functional miniproteins (Anderson et al., 2015; Pauli et al., 2014). These miniproteins, as a consequence of their small size, tend to contain few functional domains. Yet, despite this relative lack of domain architecture, they frequently interact with protein complexes (J. Chen et al., 2020) and can act as allosteric regulators and peptide hormones. Miniproteins also localize to specific cellular compartments (Na et al., 2022) and often reside in

Author Manuscript

membranes (Pang et al., 2020; Senís et al., 2021). In this section, we review different mechanisms underlying miniprotein activity, while highlighting cancer-relevant miniproteins and ones with known involvement in tumor signaling pathways.

Author Manuscript

Miniprotein Interactions with Protein Complexes

Miniprotein characterization studies have elucidated that some modulate the activity of protein complexes. Intriguingly, complex-associated miniproteins are mitochondrially-enriched and may comprise over 28% of proteins involved in the electron transport chain (ETC) (C. Liang et al., 2022). For example, the ncORF-encoded BRAWNIN protein functions as an assembly factor, essential for Complex III formation and stability (S. Zhang et al., 2020). BRAWNIN may further operate with two other ncORF-encoded miniproteins to regulate the biogenesis of a critical protein subunit of Complex III (C. Liang et al., 2022). LYRM2 is another mitochondrial miniprotein and associates with Complex I in the ETC to promote oxidative phosphorylation (Huang et al., 2019). In colorectal cancer, oxidative phosphorylation is overactive and LYRM2 is consistently upregulated. Correspondingly, miniproteins may contribute to metabolic reprogramming in cancer, a process known to impact tumor cell proliferation, metastasis, and drug adaptation (Nayak, Kapur, Barroilhet, & Patankar, 2018).

Nuclear-localized miniproteins may act in protein complexes to regulate DNA transcription and repair. For instance, the miniprotein SEHBP tightly binds histone H2B and interacts with scaffolding proteins that modulate chromatin accessibility, indicating a role in transcriptional regulation (Koh et al., 2021). Tumor-suppressive transcriptional regulators have also been identified. Zhang and coworkers investigated an lncRNA-derived circRNA that generates the functional miniprotein PINT87aa. This miniprotein interacts with the PAF1 complex

to modulate promoter binding and potentially stall RNA Pol II, thereby inhibiting oncogene elongation (M. Zhang et al., 2018). While not involved in transcriptional regulation, pTINCR is a ubiquitin-like and primarily nuclearly-localized miniprotein that stabilizes SUMO1 and SUMO2/3 (Boix et al., 2022). Consistent with this function, overexpression of pTINCR results in SUMOYlation of CDC42, promoting its activation to mediate maintenance of epithelial cell state (ibid). As such, pTINCR likely acts as a tumor-suppressor by suppressing epithelial to mesenchymal transition.

Membrane associated miniproteins frequently regulate protein complexes. For instance, the 59aa SMIM30 miniprotein mediates the SRC/YES1 complex anchoring to the cytoplasmic surface of the plasma membrane, necessary for downstream MAPK signaling (Pang et al., 2020). SMIM30 is regulated by c-MYC and overexpressed in HCC, potentially explaining the mechanism underlying MAPK signaling dysregulation in HCCs. Thus, SMIM30 provides an example of oncogene hijacking of ncORF expression and illustrates the subsequent pro-tumorigenic effects that ncORFs can orchestrate. In other cases, membrane-associated miniproteins are downregulated in cancers, suggesting potential anti-tumor roles. pTUNAR is an lncRNA-encoded transmembrane protein that localizes to the endoplasmic reticulum and maintains calcium homeostasis through activation of the SERCA2 complex, which transports calcium into the ER (Senís et al., 2021). Downregulation of TUNAR lncRNA in glioblastoma likely contributes to calcium dysregulation in these cancers.

Allosteric Regulation by Miniproteins

Miniproteins can also regulate cell function by acting as allosteric activators or inhibitors. For instance, Wang *et al* characterized an lncRNA-encoded miniprotein, termed ASRPS, that binds

Author Manuscript

Author Manuscript

the coil-coil domain of STAT3, inhibiting its phosphorylation and subsequent transcriptional activity (Y. Wang et al., 2020). Given the role of STAT3 in promoting metastasis, ASRPS acts as a tumor-suppressing miniprotein and is downregulated in triple negative breast cancer (Y. Wang et al., 2020). Another lncRNA encoded miniprotein, CIP2A-BP, was also found to act as a tumor-suppressor in triple-negative breast cancer. CIP2A-BP acts by competing with PP2A for CIP2A binding (Guo et al., 2020). PP2A binding stabilizes the CIP2A protein, allowing for its activation and subsequent signaling in AKT/NFkB pathways (J. Wang et al., 2017). Thus, CIP2A-BP, by inhibiting downstream CIP2A signaling, reduces pro-tumorigenic signaling in triple-negative breast cancer. Interrogating binding sites of both oncogenic and anti-tumor miniproteins offers therapeutic opportunities to develop small-molecule binders.

Miniproteins as Peptide Hormones

Due to their small size, miniproteins are logical candidates to act as peptide hormones to mediate both short and long range intercellular communication. Despite this expectation, thus far, few peptide-hormone miniproteins are known. ELABELA is a 54aa peptide hormone that activates PI3K/AKT signaling via the apelin receptor to stimulate hESC growth and maintain self-renewal through inhibition of apoptotic pathways (Ho et al., 2015). Interestingly, ELABELA also has coding-independent functions—its lncRNA transcript can sequester the p53 inhibitor hnRNPL (Ho et al., 2015). A second miniprotein, apelin, acts as an ELABELA agonist, also binding the apelin receptor through a two-site binding mechanism (Y. Ma et al., 2017). Intriguingly, Gao *et al.* have suggested that a circular form of E-cadherin produces a distinct, secreted miniprotein that activates EGFR in glioblastoma (Gao et al., 2021). While it has been suggested that up to 80 additional miniproteins may be secreted, whether they act as signaling peptides is unknown (F. Hu, Lu,

Munoz, Saveliev, & Turner, 2022). Finally, ncORFs, particularly those associated with lncRNAs and miRNAs, can act in hormone signaling through coding-independent roles via RNA-based regulation, as summarized elsewhere (Pardini & Calin, 2019).

Author Manuscript

Molecular Mimicry

Some ncORFs may share sequence similarities with annotated ORFs, and sufficient sequence homology may permit a non-canonical peptide to bind its cognate protein's target or regulator. In either instance, the binding mode determines whether this interaction reinforces or competes with the canonical protein's function. As an example, STORM, encoded by *linc00689*, is upregulated by stress-induced eIF4E phosphorylation. STORM mimics SRP19 to inhibit SRP complex assembly (Min et al., 2017). In addition, a formerly uncharacterized gene, Cxorf67 was found to encode EZHIP, a protein whose overexpression is a biomarker in posterior fossa ependymoma and midline gliomas H3-WT (Antin et al., 2020). EZHIP inhibits a gene-silencing protein complex, PRC2, by mimicking the sequence of the K27M onco-histone (Antin et al., 2020; Jain et al., 2019).

Translational Regulation by Miniproteins

A major mode by which miniproteins regulate translation is by modulating mRNA stability. The NoBody miniprotein interacts with mRNA decapping proteins to remove the 5' cap from mRNA transcripts, triggering nonsense mediated decay (D'Lima et al., 2017). Other miniproteins enhance mRNA stability. For instance, some miniproteins can sponge miRNA. This was demonstrated by Xu and coworkers who identified the HOXB-AS3 miniprotein which sponges the tumor-suppressive miR-378a-3p to upregulate lactate dehydrogenase-A (LDHA)

Author Manuscript

expression (Xu et al., 2021). Further, miniproteins can interact with m6A readers to strengthen target recognition, in turn bolstering mRNA stability and target translation. The onco-peptide IGF2BP1 is a cancer-relevant miniprotein that strengthens m6A recognition of c-Myc, contributing to its overexpression (Zhu et al., 2020).

Miniproteins can also drive alternative splicing of mRNA transcripts. The SRSP miniprotein interacts with the splicing factor SRSF3 to promote alternate splicing of target transcripts, including SP4 and likely others (Meng et al., 2020). SRSP upregulation in colorectal cancer lends support to the potential role for transcript splicing in this disease.

It is worth noting that, while uORFs have well defined coding-independent roles in translational regulation (Johnstone, Bazzini, & Giraldez, 2016), the impact of miniproteins on mRNA translation remains more controversial. One challenge has been the lack of scalable high-throughput technologies to address this question comprehensively, and single-gene studies remain time-consuming to complete.

3.2. Mechanisms of Regulation by ncORFs

ncORFs are increasingly recognized as major translational regulators. Here, we focus on the coding-independent regulatory interests of ncORFs, review how the intricate regulation of ncORFs allows them to finely tune translation, and summarize how the regulatory properties of ncORFs are hijacked during oncogenesis to promote cancer.

Regulation of Main CDS Translation by uORFs and dORFs

The most widely known model of translational regulation by a ncORF is through uORFs, which may negatively regulate the translation of their downstream (cognate) ORF (Johnstone et

Author Manuscript

al., 2016). uORFs exert their suppressive effect by causing ribosome stalling and dissociation, and have the capacity to reduce main CDS translation substantially in some cases (H.-H. Chen & Tarn, 2019). When uORFs are present, translation of the main CDS requires that ribosomes re-initiate at the cognate CDS or otherwise scan past the uORFs through a process termed "leaky scanning" (Wright, Yi, Weissman, & Chen, 2022). This process itself is influenced by sequence features of the uORF that govern either a weaker or stronger ability to bind the ribosomal pre-initiation complex (H.-H. Chen & Tarn, 2019). Other crucial uORF properties that influence its impact on the cognate ORF include the length of the uORF and intercistronic sequences, uORF secondary structure, and the position of the uORF termination codon (Barbosa, Peixeiro, & Romão, 2013; Y. Lin et al., 2019).

Consequently, modifications of uORFs may modulate their regulatory activities. For example, m6A methylation sites may modulate ribosome scanning and start codon selection to favor ribosome binding to the canonical start site (Vasudevan et al., 2020), and Zhou *et al* found that during the integrated stress response (ISR), m6A methylation of the 5'UTR of ATF4 promoted translation of the cognate CDS. There is additional evidence that, when uORF translation is highly favored, specialized translation factors are necessary to promote cognate CDS translation (Weber et al., 2022). Specifically, DAP5 interacts with eIF4A to modulate ribosome scanning and shift ribosomes towards the main CDS. Significantly, DAP5 transcriptional targets are enriched for proteins involved in cell migration, proliferation, and development, including multiple proto-oncogenes (Liberman, Marash, & Kimchi, 2009). Moreover, uORFs can regulate mRNA stability by promoting ribosome reinitiation at a downstream AUG rather than an annotated start codon, resulting in an erroneously truncated protein isoform and consequent NMD (Arribere & Gilbert, 2013).

Author Manuscript

In addition, there is also evidence that uORFs can promote the translation and stability of the main CDS, particularly as an adaptive response to stress conditions. In yeast, translation of the GCN4 transcription factor is largely controlled by four uORFs located in the GCN4 mRNA (Hinnebusch 1997). Under non-starved states, the uORFs downregulate GCN4 translation. However, under starved conditions, phosphorylation of eukaryotic translation initiation factor 2 (eIF2) causes a reduction in the eIF2·GTP·Met-tRNAiMet ternary complex formation. This reduces the ability of the ribosome to bind to uORFs. Resulting in the ribosome scanning through the uORFs and initiating translation at the main GCN4 reading frame instead (Hinnebusch 1997). Work by Andreev and colleagues also indicates that uORFs overcome eIF2 phosphorylation-mediated translational downregulation in other contexts as well (Andreev et al. 2015). Here, the authors induced the eIF2 stress response and found that mRNAs resistant to eIF2 inhibition, with only one exception, contain a uORF that is highly translated under normal conditions (Andreev et al. 2015). These results suggest that uORFs play an important role in enhancing the translation of the main CDS upon eIF2 phosphorylation in the stress response.

Since ribosomes terminate at stop codons in the CDS, downstream ORFs (dORFs) are less prevalent in mammalian transcripts (Mudge et al., 2022) and less studied. In some cases, dORFs may enhance main CDS translation. Using Ribo-seq to identify dORFs, Wu and coworkers found that all mRNAs containing dORFs had higher translation efficiency, and that this effect was not influenced by uORF length or sequence (Wu et al., 2020). The mechanism underlying this enhancement of the main CDS translation is unresolved. The prevailing hypothesis is that ribosome recruitment through cap-independent mechanisms (including IRES) drives dORF translational activity (Ruiz Cuevas et al., 2021), although further work is needed.

Author Manuscript

3.3. ncORF Dysregulation and Mutations in Cancer

ncORF dysregulation is both a response to and driver of oncogenesis

Cancer is a disease of rapid and unchecked cell proliferation resulting from gene dysregulation via genetic and epigenetic mechanisms. These mechanisms may impact ncORFs by altering their transcription, translation, or both. Typically, in instances where ncORFs lie within a coding gene, transcription of the ncORF and main CDS are regulated by the same elements. Thus, if transcription of a given gene is aberrantly expressed during oncogenesis, the associated ncORF is as well. Following transcription, modifications to ncRNA transcripts can further regulate their ultimate expression, often independently from the main CDS.

The N6-methyladenosine (m6A) modification occurs on ncORFs, lncRNAs, and circRNAs (M. Ma et al., 2022). This modification functions in cap-independent translation during the integrated stress response (ISR). In this context, EIF3 binds m6A residues to recruit the 43S ribosomal complex to initiate translation at these sites (Meyer et al., 2015). m6A modifications are prevalent in ncORFs, potentially explaining why their expression persists during ISR despite global reductions in protein translation. Given that ISR is frequently induced during tumorigenesis (Meyer, 2019; Tian et al., 2021), ncORF translation is plausibly enhanced in this context, though further work is needed in this area.

In addition to promoting the translation and activity of oncogenes while inhibiting tumor suppressors in cancer, ncORFs themselves are also dysregulated in cancer settings. Indeed, Erady and coworkers performed a pan-cancer differential expression analysis of ncORF transcriptional levels using RNA-sequencing data from The Cancer Genome Atlas (TCGA) and Genotype-Expression (GTex), and found that commonly expressed ncORF transcripts are often dysregulated across cancer types (Erady et al., 2021). This dysregulation of ncORF expression also had

Author Manuscript

prognostic value, hinting at their potential as biomarkers. Finally, characterization of differentially expressed ncORFs revealed that many are enriched in sites amenable to post-translational modifications, suggesting a major potential mechanism for their (dys)regulation. Aberrant ncORF expression in cancer is corroborated by other studies. For instance, Zhang and coworkers demonstrated that the Ras oncoprotein upregulates *Orilnc1* in BRAF-mutant cancers to further drive RAS/RAF activation (D. Zhang et al., 2017).

Importantly, the relationship between ncORF dysregulation and cancer is bidirectional—cancers can drive dysregulation of ncORFs and ncORFs can act during tumor initiation to upregulate pro-tumorigenic translation. This striking finding was demonstrated by Sendoel and coworkers using an SOX2-inducible mouse model (Sendoel et al., 2017). SOX2 induction drove significant shifts in mRNA translation efficiency and ribosome occupancy, with enhanced occupancy at 5'UTRs. Further, they observed eIF2 repression and eIF2A de-repression. EIF2A preferentially directs initiator-tRNA to uORFs, driving increased uORF expression. Finally, examination of the uORFs upregulated upon SOX2 induction showed enhancement for uORFs with downstream ORFs with cancer-associated functions. Taken together, these findings strongly support a role for ncORFs in early stages of tumorigenesis (Sendoel et al., 2017).

Mutations of ncORFs and their Relevance in Cancer

The elucidation of functional miniproteins has motivated efforts to uncover cancer-associated ncORF mutations. Conceptually, such mutations fall into two categories; namely, cancer variants that impact existing ncORFs and those that generate novel ncORFs.

Mutations occurring in endogenous ncORFs can alter miniprotein translation or translation of the main CDS. As a general proof of concept, Whiffin and colleagues demonstrated that uORF

Author Manuscript

point mutations that either generate start codons or premature termination codons are under strong negative selection, implying that uORF disruption by way of mutation can be deleterious and potentially disease causing (Whiffin et al., 2020). In a more cancer-specific context, Occhi and coworkers identified a 4bp deletion in an endogenous and highly conserved uORF within the 5'UTR of *CDKN1B*, a known tumor-suppressor gene. The deletion results in an elongated miniprotein and shortens the intercistronic space, blocking ribosome reinitiation at the cognate CDS and ultimately suppressing *CDKN1B* expression (Occhi et al., 2013)

Cancer mutations can also produce novel ncORFs that disrupt main CDS expression. For instance, Liu and coworkers characterized a 5'UTR point mutation that generates a novel AUG-start codon to cause a melanoma-predisposing *CDKN2A* loss of function mutation (L. Liu et al., 1999). Likewise, a mutation in the NF2 5'UTR that produces a novel upstream overlapping ORF was shown to produce NF2 loss of function to drive neurofibromatosis (Whiffin et al., 2020). These examples notwithstanding, there are few known cancer-driving mutations in ncORFs. Myriad factors account for this under-discovery. Principally, cancer-variants are typically identified using exome sequencing. Inherently, the majority of ncORF variants—which reside in flanking regions or outside exons—are missed by this approach (Bailey et al., 2020). Capturing these ncORFs instead requires the use of whole-genome sequencing.

The International Cancer Genome Consortium (ICGC) has undertaken a massive effort to identify unique cancer driver mutations using whole-genome sequencing, integrating their own sequencing data with that from the Pan-Cancer Analysis of Whole Genomes (PCAWG) and The Cancer Genome Atlas (TCGA) (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Leveraging this combined dataset, Rheinbay and colleagues analyzed whole-genome sequencing data from 2,658 samples to define cancer driver mutations in non-coding

genomic regions (Rheinbay et al., 2020). This analysis identified non-coding driver mutations affecting TP53, NFKBIZ, TOB1, and BRD4 (among other protein-coding genes). The authors also found that cancer-driving mutations in non-coding regions are far rarer than those in coding regions. Nonetheless, since the non-coding genome vastly exceeds the coding genome, the total cancer-burden of non-coding mutations is likely still significant (Schipper & Posthuma, 2022). Crucially, the Rheinbay study interrogated variants in nearby transcriptional start sites, which is not well suited to detect ncORF variants. In a later attempt to better annotate cancer-drivers in non-coding regions, Dietlein and coworkers developed a computational method that stratifies variants based on their genomic region. This accounts for the varied likelihood that a mutation is cancer-driving based on genomic context (Dietlein et al., 2022). Through this novel approach, they identified significantly more non-coding driver mutations relative to findings from other studies. This suggests that as the community continues to develop improved computational approaches, more ncORF driver mutations will be detected.

Even as methods to detect ncORF variants improve, evaluating the functional effects of ncORF mutations adds additional challenge. Most methods to analyze the functional effects of a mutation utilize conservation of the affected amino acid residue. Specifically, variant annotation methods assess whether a mutation occurs at a conserved residue such that it would drastically impact protein structure (e.g. by altering key side chains, changing the isoelectric point, etc). For young or *de novo* ORFs, evolutionary conservation analysis cannot well predict the effects of an ncORF variant on its coding or regulatory function (Vakirlis, Vance, Duggan, & McLysaght, 2022). Given that ncORFs are generally non conserved, new approaches are needed to better scrutinize effects of variants.

Author Manuscript

Author Manuscript

There are further difficulties in computationally mining for cancer-associated ncORF mutations. Statistical models identify cancer-driving mutations based on recurrence in coding-regions only (Cibulskis et al., 2013). As a result, these models cannot be applied for the classification of somatic mutations in ncORFs or non-coding regions. Efforts to develop new analytical approaches to identify somatic mutations outside of coding-regions are ongoing (Dietlein et al., 2022).

Given the multi-faceted role that ncORFs play in translational regulation and cell signaling, it is likely that there are cancer-relevant mutations in ncORFs. However, their elucidation is complicated by a lack of computational approaches to facilitate their detection and analyze their functional relevance. The development of annotation methods that do not rely on evolutionary conservation will be crucial to these efforts.

Cancer-specific ncORF Annotations

An overarching challenge for ncORFs remains their inconsistent annotation in human genome databases. To this end, a recent international collaborative effort is engaged in efforts to standardize the reference annotation database for ncORFs (Mudge et al., 2022). Yet, reference gene annotations are, by definition, designed for the reference human genome, rather than disease-specific genomes. Beyond employing copy number variation and translocations that may physically disrupt the genome, cancers also deregulate normally quiescent regions of the genome. Cancer-specific transcriptomes are therefore well-established (W. Hu et al., 2022); many RNA transcripts are induced uniquely in certain cancer states and those transcripts are not annotated in the reference human genome. Likewise, cancer-specific ncORF translations are an emerging topic

(Vibert et al., 2022) and disease-specific ncORFs are likely to reflect the activity of oncogenic drivers.

Ongoing research into ncORFs uniquely expressed in cancer promises to elucidate new mechanisms underlying oncogenesis and open therapeutic avenues. Molecular studies of cancer-unique ncORFs and their potential functional products may uncover new signaling and regulatory events that drive tumor progression. Clinically, these ncORFs have promise as biomarkers or as tumor specific antigens.

3.4. Miniproteins and the Immunopeptidome

Tumor immunosurveillance relies heavily on CD8+ T-cell recognition of tumor cells, largely mediated by the MHC-I system, wherein MHC-I molecules present antigens of cytosolic origin to CD8+ cells. Antigen presentation occurs when endogenous proteins are degraded into small fragments (8-10aa) and subsequently bind to the peptide-binding site of MHC-I (Hewitt, 2003). Cancer cells, due to the expression of proteins bearing non-synonymous mutations, often present aberrant peptide fragments, or tumor-associated neoantigens. Recognition of these neoantigens by CD8+ cells facilitates the T-cell mediated antitumor response (Smith et al., 2019).

The "immunopeptidome" describes the set of peptides presented by MHC molecules (Yewdell, 2022). Intriguingly, the immunopeptidome is not representative of the proteome, meaning that the abundance of an antigenic peptide does not necessarily correlate with its protein abundance (Dersh, Hollý, & Yewdell, 2021). In fact, miniproteins—despite comprising only a small fraction of translated proteins—make up a relatively higher proportion of the MHC-I immunopeptidome, approximately 7.5% (Yewdell, 2022). Consistent with this finding, non-

Author Manuscript

Author Manuscript

coding ORFs generate MHC-I associated peptides (MAPs) at a 5-fold higher efficiency compared to canonical transcripts (Ruiz Cuevas et al., 2021).

Since cancer-associated genetic alterations can drive upregulated expression of miniproteins (Ruiz Cuevas et al., 2021), tumor cells may display antigens deriving from miniproteins at a higher rate than non-tumor cells. Indeed, Ouspenskaia and coworkers identified cancer-enriched ncORF MAPs in glioblastoma and melanomas (Ouspenskaia et al., 2022). Relatedly, Chong et al identified an immunogenic tumor-specific antigen derived from the dORF of the ABCB5 gene, which is pro-oncogenic and differentially in melanoma cell lines (Chong et al. 2020). Their finding suggests that upregulated cancer-promoting genes containing ncORFs may also drive increased tumor antigen production. Importantly, the function and stability of a miniprotein is irrelevant to its efficiency in generating MAPs. As a whole, ncORFs represent an understudied source of tumor antigens that could be leveraged in immunotherapeutic applications, including for the development of cancer vaccines.

Conclusion

Studies of ncORFs are poised to reveal new dimensions in cancer biology. Collectively, ncORFs may reveal new aspects of cancer pathogenesis, both through gene regulation as well as through the production of proteins or peptide products. To deliver on this promise, the set of computational and experimental tools to find and functionally validate ncORFs and their miniprotein products requires ongoing expansion and improvement. While current identification of ncORFs mainly relies on Ribo-seq and mass spectrometry, new computational methods (Nabi, Dilekoglu, Adebali, & Tastan, 2023) potentiate higher-confidence and higher-throughput discovery. For instance, the TIS Transformer program, which uses artificial intelligence to map

the human genome, was able to detect ncORFs and predict those that encode miniproteins with high performance (Clauwaert, Gupta, McVey, & Menschaert, 2022).

Experimental functional validation efforts will benefit from new approaches such as CRISPR/Cas9 based screening (Prensner et al., 2021), bio-orthogonal non-canonical amino acid tagging (BONCAT) (Cao, Chen, Khitun, & Slavoff, 2023), proximity biotinylation to determine miniprotein subcellular localization (Na et al., 2022), and use of AlphaFold for structural predictions (Perrakis & Sixma, 2021). Recent technologies that adapt DNA-sequencing approaches for high-throughput protein analysis (Layton, McMahon, & Greenleaf, 2019; Yu et al., 2023) may also accelerate miniprotein characterization.

While specific ncORF-directed therapeutics remain a future goal at this time, active work in this area lends credibility to the hope for new disease-relevant therapeutic insights through the study of ncORFs (**Figure 4**). Cancer or tissue specific ncORFs are attractive therapeutic targets to disrupt downstream translation of undruggable oncoproteins with greater selectivity than drugging the relevant transcription factors. Miniproteins themselves may be also utilized as biomarkers or in drug design. For instance, miniproteins that scaffold protein complexes could be used in molecular glues. Miniprotein-based drugs are not simply theoretical—the miniprotein Myc inhibitor, Omomyc, recently advanced through Phase I clinical trials (Llombart & Mansour, 2022; Peptomyc, n.d.). Finally, the immuno-oncology field is actively exploring the use of miniproteins as tumor-specific neoantigens. While these therapeutic applications will require additional investigation, they highlight the exciting translational potential of ncORFs.

Meanwhile, the fundamental question of which ncORFs operate as cancer drivers compared to bystanders in oncogenesis remains unanswered, and will require substantial efforts by the cancer research community. Other essential aspects of ncORF biology also require further

Author Manuscript

elucidation—for instance, it remains unknown whether specific ribosome binding proteins are associated with ncORFs and their regulation. Similarly, greater mechanistic understanding of the translation initiation of ncORFs is still needed. Much of the dark proteome remains shrouded in mystery; like all great unknowns, unraveling ncORF and miniprotein biology offers to enrich our understanding of the complexities underlying cell signaling and regulation while also engendering new therapeutic strategies and targets.

Author Manuscript

Author Manuscript

Acknowledgements

We would like to acknowledge members of the Golub lab at the Broad Institute of MIT and Harvard for helpful conversations and discussions. We acknowledge www.biorender.com for the generation of figures.

Funding

J.R.P. acknowledges funding from the National Institutes of Health/National Cancer Institute (K08-CA263552-01A1), the Alex's Lemonade Stand Foundation Young Investigator Award (#21-23983), the St. Baldrick's Foundation Scholar Award (#931638), The DIPG/DMG Research Funding Alliance, and a Collaborative Pediatric Cancer Research Awards Program/Kids Join the Fight award (#22FN23).

Author contributions

Zoe Posner: Conceptualization; Data Curation; Methodology; Visualization; Writing – Original Draft Preparation; Writing – Review & Editing

Ian Yannuzzi: Conceptualization; Data Curation; Methodology; Visualization; Writing – Original Draft Preparation; Writing – Review & Editing

John R. Prensner: Conceptualization; Funding Acquisition; Supervision; Writing – Review & Editing

Declaration of interests

The authors declare no competing interests.

ORCID

Zoe Posner: https://orcid.org/0000-0003-0043-0281

Ian Yannuzzi: https://orcid.org/0000-0002-0216-4805

John R. Prensner: https://orcid.org/0000-0002-7024-636X

Author Manuscript

Author Manuscript

References

Alioto, T. (2012). Gene prediction. *Methods in Molecular Biology* , *855*, 175–201.

Anderson, D. M., Anderson, K. M., Chang, C.-L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., … Olson, E. N. (2015). A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell*, Vol. 160, pp. 595–606. doi: 10.1016/j.cell.2015.01.009

Andreev, D. E., Loughran, G., Fedorova, A. D., Mikhaylova, M. S., Shatsky, I. N., & Baranov, P. V. (2022). Non-AUG translation initiation in mammals. *Genome Biology*, *23*(1), 111.

Andreev, D. E., O'Connor, P. B. F., Fahey, C., Kenny, E. M., Terenin, I. M., Dmitriev, S. E., … Baranov,

P. V. (2015). Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife*, *4*, e03971.

Antin, C., Tauziède-Espariat, A., Debily, M.-A., Castel, D., Grill, J., Pagès, M., … Varlet, P. (2020). EZHIP is a specific diagnostic biomarker for posterior fossa ependymomas, group PFA and diffuse midline gliomas H3-WT with EZHIP overexpression. *Acta Neuropathologica Communications*, Vol. 8. doi: 10.1186/s40478-020-01056-8

Arribere, J. A., & Gilbert, W. V. (2013). Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Research*, *23*(6), 977–987.

Autio, K. J., Kastaniotis, A. J., Pospiech, H., Miinalainen, I. J., Schonauer, M. S., Dieckmann, C. L., & Hiltunen, J. K. (2008). An ancient genetic link between vertebrate mitochondrial fatty acid synthesis and RNA processing. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *22*(2), 569–578.

Badger, J. H., & Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, *16*(4), 512–524.

Bailey, M. H., Meyerson, W. U., Dursi, L. J., Wang, L.-B., Dong, G., Liang, W.-W., … PCAWG Consortium. (2020). Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. *Nature Communications*, *11*(1), 4748.

Barbosa, C., Peixeiro, I., & Romão, L. (2013). Gene expression regulation by upstream open reading frames and human disease. *PLoS Genetics*, *9*(8), e1003529.

Beadle, G. W., & Tatum, E. L. (1941). Genetic Control of Biochemical Reactions in Neurospora.

Author Manuscript

Proceedings of the National Academy of Sciences, Vol. 27, pp. 499–506. doi: 10.1073/pnas.27.11.499

Bell, J., & Haldane, J. B. S. (1937). The linkage between the genes for colour-blindness and haemophilia in man. *Proceedings. Biological Sciences / The Royal Society*, *123*(831), 119–150.

Boix, O., Martinez, M., Vidal, S., Giménez-Alejandre, M., Palenzuela, L., Lorenzo-Sanz, L., … Abad, M. (2022). pTINCR microprotein promotes epithelial differentiation and suppresses tumor growth through CDC42 SUMOylation and activation. *Nature Communications*, *13*(1), 6840.

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., … Ballabio, A. (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, *351*(6324), 325–329.

Brunner, A.-D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., … Mann, M. (2022). Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Molecular Systems Biology*, *18*(3), e10798.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–1927.

Calviello, L., & Ohler, U. (2017). Beyond Read-Counts: Ribo-seq Data Analysis to Understand the

Functions of the Transcriptome. *Trends in Genetics: TIG*, *33*(10), 728–744.

Cao, X., Chen, Y., Khitun, A., & Slavoff, S. A. (2023). BONCAT-based Profiling of Nascent Small and Alternative Open Reading Frame-encoded Proteins. *Bio-Protocol*, *13*(1), e4585.

Chen, H.-H., & Tarn, W.-Y. (2019). uORF-mediated translational control: recently elucidated mechanisms and implications in cancer. *RNA Biology*, *16*(10), 1327–1338.

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., … Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science*, *367*(6482), 1140–1146.

Chong, Chloe, Markus Müller, Huisong Pak, Dermot Harnett, Florian Huber, Delphine Grun, Marion Leleu, et al. 2020. “Integrated Proteogenomic Deep Sequencing and Analytics Accurately Identify Non-Canonical Peptides in Tumor Immunopeptidomes.” Nature

Author Manuscript

Communications 11 (1): 1293.

Chothani, Sonia P., Eleonora Adami, Anissa A. Widjaja, Sarah R. Langley, Sivakumar Viswanathan, Chee Jian Pua, Nevin Tham Zhihao, et al. 2022. "A High-Resolution Map of Human RNA Translation." Molecular Cell 82 (15): 2885–99.e8.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219.

Clauwaert, J., Gupta, R., McVey, Z., & Menschaert, G. (2022). TIS Transformer: Remapping the Human Proteome Using Deep Learning (p. 2021.11.18.468957). doi: 10.1101/2021.11.18.468957

Consortium, M. G. S., & Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, Vol. 420, pp. 520–562. doi: 10.1038/nature01262

Crosetto, N., Bienko, M., & van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond.

Nature Reviews. Genetics, *16*(1), 57–66.

Dersh, D., Hollý, J., & Yewdell, J. W. (2021). A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion. *Nature Reviews. Immunology*, *21*(2), 116–128.

Dietlein, F., Wang, A. B., Fagre, C., Tang, A., Besselink, N. J. M., Cuppen, E., … Van Allen, E. M. (2022). Genome-wide analysis of somatic noncoding mutation patterns in cancer. *Science*, *376*(6589), eabg5601.

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Computational Biology*, *4*(11), e1000176.

D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., … Slavoff, S. A. (2017). A human microprotein that interacts with the mRNA decapping complex. *Nature Chemical Biology*, *13*(2), 174–180.

Do, J. H., & Choi, D.-K. (2006). Computational approaches to gene prediction. *Journal of Microbiology* ,

44(2), 137–144.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R.,

Author Manuscript

Gingeras, T. R., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816.

Erady, C., Boxall, A., Puntambekar, S., Suhas Jagannathan, N., Chauhan, R., Chong, D., … Prabakaran, S. (2021). Pan-cancer analysis of transcripts encoding novel open-reading frames (nORFs) and their potential biological functions. *NPJ Genomic Medicine*, *6*(1), 4.

Farber, R., Lapedes, A., & Sirotkin, K. (1992). Determination of eukaryotic protein coding regions using neural networks and information theory. *Journal of Molecular Biology*, *226*(2), 471–479.

Fedorova, A. D., Kiniry, S. J., Andreev, D. E., Mudge, J. M., & Baranov, P. V. (2022). Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. *Nature Communications*, *13*(1), 7910.

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., … Grimmond, S. M. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genetics*, *2*(4), e52.

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., & Couso, J. P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biology*, *5*(5), e106.

Gao, X., Xia, X., Li, F., Zhang, M., Zhou, H., Wu, X., … Zhang, N. (2021). Circular RNA-encoded oncogenic E-cadherin variant promotes glioblastoma tumorigenicity through activation of EGFR–STAT3 signalling. *Nature Cell Biology*, *23*(3), 278–291.

Goel, N., Singh, S., & Aseri, T. C. (2013). A comparative analysis of soft computing techniques for gene

prediction. *Analytical Biochemistry*, *438*(1), 14–21.

Guo, B., Wu, S., Zhu, X., Zhang, L., Deng, J., Li, F., … Zhou, Y. (2020). Micropeptide CIP 2A- BP encoded by LINC 00665 inhibits triple- negative breast cancer progression. *The EMBO Journal*, Vol. 39. doi: 10.15252/embj.2019102190

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., … Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, *458*(7235), 223–227.

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., & Shiu, S.-H. (2010). sORF finder: a

Author Manuscript

program package to identify small open reading frames with high coding potential. *Bioinformatics* , *26*(3), 399–400.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774.

Hartford, C. C. R., & Lal, A. (2020). When Long Noncoding Becomes Protein Coding. *Molecular and Cellular Biology*, *40*(6). doi: 10.1128/MCB.00528-19

Hemm, M. R., Paul, B. J., Schneider, T. D., Storz, G., & Rudd, K. E. (2008). Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology*, *70*(6), 1487–1501.

Hewitt, E. W. (2003). The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, *110*(2), 163–169.

Hinnebusch, A. G. (1997). Translational regulation of yeast GCN4. A window on factors that control

initiator-trna binding to the ribosome. *The Journal of Biological Chemistry*, *272*(35), 21661–21664.

Ho, L., Tan, S. Y. X., Wee, S., Wu, Y., Tan, S. J. C., Ramakrishna, N. B., … Reversade, B. (2015). ELABELA Is an Endogenous Growth Factor that Sustains hESC Self-Renewal via the PI3K/AKT Pathway. *Cell Stem Cell*, *17*(4), 435–447.

Hope, I. A., & Struhl, K. (1985). GCN4 protein, synthesize in vitro, binds HIS3 regulatory sequences:

Implications for general control of amino acid biosynthetic genes in yeast. *Cell*, *43*(1), 177–188.

Hsu, P. Y., Calviello, L., Wu, H.-Y. L., Li, F.-W., Rothfels, C. J., Ohler, U., & Benfey, P. N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(45), E7126–E7135.

Huang, Q., Chen, Z., Cheng, P., Jiang, Z., Wang, Z., Huang, Y., … Huang, J. (2019). LYRM2 directly regulates complex I activity to support tumor growth in colorectal cancer by oxidative phosphorylation. *Cancer Letters*, *455*, 36–47.

Hu, F., Lu, J., Munoz, M. D., Saveliev, A., & Turner, M. (2022). ORFLine: a bioinformatic

Author Manuscript

pipeline to prioritise small open reading frames identifies candidate secreted small proteins from lymphocytes. *Bioinformatics* , *38*(9), 2673–2673.

Hu, W., Wu, Y., Shi, Q., Wu, J., Kong, D., Wu, X., … Li, S. (2022). Systematic characterization of cancer transcriptome at transcript resolution. *Nature Communications*, *13*(1), 6803.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93.

Inagaki, S., Numata, K., Kondo, T., Tomita, M., Yasuda, K., Kanai, A., & Kageyama, Y. (2005). Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila. *Genes to Cells: Devoted to Molecular & Cellular Mechanisms*, *10*(12), 1163–1173.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, *324*(5924), 218–223.

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945.

Jain, S. U., Do, T. J., Lund, P. J., Rashoff, A. Q., Diehl, K. L., Cieslik, M., … Lewis, P. W. (2019). PFA ependymoma-associated protein EZHIP inhibits PRC2 activity through a H3 K27M-like mechanism. *Nature Communications*, *10*(1), 2146.

Johnstone, T. G., Bazzini, A. A., & Giraldez, A. J. (2016). Upstream ORF s are prevalent translational repressors in vertebrates. *The EMBO Journal*, Vol. 35, pp. 706–723. doi: 10.15252/embj.201592759

Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W.-C., Yang, H., … Basrai, M. A. (2006). Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Research*, *16*(3), 365–373.

Klasberg, S., Bitard-Feildel, T., & Mallet, L. (2016). Computational Identification of Novel Genes: Current and Future Perspectives. *Bioinformatics and Biology Insights*, *10*, 121–131.

Kiniry, Stephen J., Audrey M. Michel, and Pavel V. Baranov. 2020. "Computational Methods for Ribosome Profiling Data Analysis." *Wiley Interdisciplinary Reviews. RNA* 11 (3):

Author Manuscript

e1577.

Koh, M., Ahmad, I., Ko, Y., Zhang, Y., Martinez, T. F., Diedrich, J. K., … Bollong, M. J. (2021). A short ORF-encoded transcriptional regulator. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(4). doi: 10.1073/pnas.2021943118

Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., & Kageyama, Y. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology*, *9*(6), 660–665.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Research*, Vol. 12, pp. 857–872. doi: 10.1093/nar/12.2.857

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, *15*(20), 8125–8148.

Kozak, M. (1989). The scanning model for translation: an update. *Journal of Cell Biology*, Vol. 108, pp. 229–241. doi: 10.1083/jcb.108.2.229

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Lapedes, A., Barnes, C., Burks, C., & Farber, R. (1989). Application of neural networks and other machine learning algorithms to DNA sequence analysis. *ACM SIGCAS Computers and Society*. doi: 10.4324/9780429501463-15/application-neural-networks-machine-learning-algorithms-dna-sequence-analysis-lapedes-barnes-burks-farber-sirotkin

Layton, C. J., McMahon, P. L., & Greenleaf, W. J. (2019). Large-Scale, Quantitative Protein Assays on a High-Throughput DNA Sequencing Chip. *Molecular Cell*, *73*(5), 1075–1082.e4.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., & Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), E2424–E2432.

Liang, C., Zhang, S., Robinson, D., Ploeg, M. V., Wilson, R., Nah, J., … Ho, L. (2022). Mitochondrial microproteins link metabolic cues to respiratory chain biogenesis. *Cell*

Author Manuscript

Reports, *40*(7), 111204.

Liang, P., & Pardee, A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, *257*(5072), 967–971.

Liberman, N., Marash, L., & Kimchi, A. (2009). The translation initiation factor DAP5 is a regulator of cell survival during mitosis. *Cell Cycle*, Vol. 8, pp. 204–209. doi: 10.4161/cc.8.2.7384

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* , *27*(13), i275–i282.

Lin, Y., May, G. E., Kready, H., Nazzaro, L., Mao, M., Spealman, P., … McManus, C. J. (2019). Impacts of uORF codon identity and position on translation regulation. *Nucleic Acids Research*, *47*(17), 9358–9367.

Liu, L., Dilworth, D., Gao, L., Monzon, J., Summers, A., Lassam, N., & Hogg, D. (1999). Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nature Genetics*, *21*(1), 128–132.

Liu, Y., Zeng, S., & Wu, M. (2022). Novel insights into noncanonical open reading frames in cancer. *Biochimica et Biophysica Acta, Reviews on Cancer*, *1877*(4), 188755.

Llombart, V., & Mansour, M. R. (2022). Therapeutic targeting of "undruggable" MYC. *EBioMedicine*, *75*, 103756.

Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., … Saghatelian, A. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Analytical Chemistry*, *88*(7), 3967–3975.

Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., … Saghatelian, A. (2014). Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of Proteome Research*, Vol. 13, pp. 1757–1765. doi: 10.1021/pr401280w

Ma, M., Ye, T., Wang, J., Zhao, H., Zhang, S., Li, P., & Zhao, G. (2022). N6-methyladenosine Modification of Noncoding RNAs: Mechanisms and Clinical Applications in Cancer. *Diagnostics (Basel, Switzerland)*, *12*(12). doi: 10.3390/diagnostics12122996

Mackowiak, Sebastian D., Henrik Zauber, Chris Bielow, Denise Thiel, Kamila Kutz, Lorenzo Calviello, Guido Mastrobuoni, et al. 2015. "Extensive Identification and Analysis of Conserved Small ORFs in Animals." Genome Biology 16 (September): 179.

Matsumoto, A., Clohessy, J. G., & Pandolfi, P. P. (2017). SPAR, a lncRNA encoded mTORC1

Author Manuscript

inhibitor. *Cell Cycle*, Vol. 16, pp. 815–816. doi: 10.1080/15384101.2017.1304735

Ma, Y., Yue, Y., Ma, Y., Zhang, Q., Zhou, Q., Song, Y., … Xu, F. (2017). Structural Basis for Apelin Control of the Human Apelin Receptor. *Structure* , *25*(6), 858–866.e4.

Meng, N., Chen, M., Chen, D., Chen, X.-H., Wang, J.-Z., Zhu, S., … Yan, G.-R. (2020). Small Protein Hidden in lncRNA LOC90024 Promotes "Cancerous" RNA Splicing and Tumorigenesis. *Advancement of Science*, *7*(10), 1903233.

Meyer, K. D. (2019). m6A-mediated translation regulation. *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms*, *1862*(3), 301–309.

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., … Jaffrey, S. R. (2015). 5' UTR m(6)A Promotes Cap-Independent Translation. *Cell*, *163*(4), 999–1010.

Min, K.-W., Davila, S., Zealy, R. W., Lloyd, L. T., Lee, I. Y., Lee, R., … Yoon, J.-H. (2017). eIF4E phosphorylation by MST1 reduces translation of a subset of mRNAs, but increases lncRNA translation. *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms*, *1860*(7), 761–772.

Mohaupt, P., Roucou, X., Delaby, C., Vialaret, J., Lehmann, S., & Hirtz, C. (2022). The alternative proteome in neurobiology. *Frontiers in Cellular Neuroscience*, *16*, 1019680.

Mudge, J. M., Jungreis, I., Hunt, T., Gonzalez, J. M., Wright, J. C., Kay, M., … Kellis, M. (2019). Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Research*, *29*(12), 2073–2087.

Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., … van Heesch, S. (2022). Standardized annotation of translated open reading frames. *Nature Biotechnology*, *40*(7), 994–999.

Nabi, A., Dilekoglu, B., Adebali, O., & Tastan, O. (2023). Discovering misannotated lncRNAs using deep learning training dynamics. *Bioinformatics* , *39*(1). doi: 10.1093/bioinformatics/btac821

Nayak, A., Kapur, A., Barroilhet, L., & Patankar, M. (2018). Oxidative Phosphorylation: A Target for Novel Therapeutic Strategies Against Ovarian Cancer. *Cancers*, Vol. 10, p. 337. doi: 10.3390/cancers10090337

Na, Z., Dai, X., Zheng, S.-J., Bryant, C. J., Loh, K. H., Su, H., … Slavoff, S. A. (2022). Mapping subcellular localizations of unannotated microproteins and alternative proteins with MicroID. *Molecular Cell*, *82*(15), 2900–2911.e7.

Author Manuscript

Nesvizhskii, Alexey I. 2014. "Proteogenomics: Concepts, Applications and Computational Strategies." Nature Methods 11 (11): 1114–25.

Nirenberg, M. W. (1963). The Genetic Code: II. *Scientific American*, Vol. 208, pp. 80–95. doi: 10.1038/scientificamerican0363-80

Occhi, G., Regazzo, D., Trivellin, G., Boaretto, F., Ciato, D., Bobisse, S., … Scaroni, C. (2013). A novel mutation in the upstream open reading frame of the CDKN1B gene causes a MEN4 phenotype. *PLoS Genetics*, *9*(3), e1003350.

Orr, M. W., Mao, Y., Storz, G., & Qian, S.-B. (2020). Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Research*, *48*(3), 1029–1042.

Ouelle, D. E., Zindy, F., Ashmun, R. A., & Sherr, C. J. (1995). Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell*, Vol. 83, pp. 993–1000. doi: 10.1016/0092-8674(95)90214-7

Ouspenskaia, T., Law, T., Clauser, K. R., Klaeger, S., Sarkizova, S., Aguet, F., … Regev, A. (2022). Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nature Biotechnology*, *40*(2), 209–217.

Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., … Sugano, S. (2004). Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Research*, *14*(10B), 2048–2052.

Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T., & Sugano, S. (2007). Diversity of translation start sites may define increased complexity of the human short ORFeome. *Molecular & Cellular Proteomics: MCP*, *6*(6), 1000–1006.

Pang, Y., Liu, Z., Han, H., Wang, B., Li, W., Mao, C., & Liu, S. (2020). Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. *Journal of Hepatology*, *73*(5), 1155–1169.

Pardini, B., & Calin, G. A. (2019). MicroRNAs and Long Non-Coding RNAs and Their Hormone-Like Activities in Cancer. *Cancers*, *11*(3). doi: 10.3390/cancers11030378

Pauli, A., Norris, M. L., Valen, E., Chew, G.-L., Gagnon, J. A., Zimmerman, S., … Schier, A. F. (2014). Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science*, *343*(6172), 1248636.

Peptomyc, S. L. (n.d.). A Phase 1/2 Study to Evaluate the Safety, Pharmacokinetics, and Anti-Tumour Activity of the MYC Inhibitor OMO-103 Administered Intravenously in Patients

Author Manuscript

With Advanced Solid Tumours (Clinical Trial Registration No. NCT04808362).

Perrakis, A., & Sixma, T. K. (2021). AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Reports*, *22*(11), e54046.

Prensner, J. R., Enache, O. M., Luria, V., Krug, K., Clauser, K. R., Dempster, J. M., … Golub, T. R. (2021). Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nature Biotechnology*, *39*, 697–704.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., … Getz, G. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, *578*(7793), 102–111.

Robson, E. B. (1988). The human gene map. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *319*(1194), 229–237.

Rubin, G. M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., & Harvey, D. A. (2000). A Drosophila complementary DNA resource. *Science*, *287*(5461), 2222–2224.

Ruggles, Kelly V., Karsten Krug, Xiaojing Wang, Karl R. Clauser, Jing Wang, Samuel H. Payne, David Fenyö, Bing Zhang, and D. R. Mani. 2017. "Methods, Tools and Current Perspectives in Proteogenomics*." Molecular & Cellular Proteomics: MCP 16 (6): 959–81.

Ruiz Cuevas, M. V., Hardy, M.-P., Hollý, J., Bonneil, É., Durette, C., Courcelles, M., … Yewdell, J. W. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Reports*, *34*(10), 108815.

Sandmann, C.-L., Schulz, J. F., Ruiz-Orera, J., Kirchner, M., Ziehm, M., Adami, E., … Hubner, N. (2023). Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Molecular Cell*. doi: 10.1016/j.molcel.2023.01.023

Schipper, M., & Posthuma, D. (2022). Demystifying non-coding GWAS variants: an overview of computational tools and methods. *Human Molecular Genetics*, *31*(R1), R73–R83.

Sendoel, A., Dunn, J. G., Rodriguez, E. H., Naik, S., Gomez, N. C., Hurwitz, B., … Fuchs, E. (2017). Translation from unconventional 5' start sites drives tumour initiation. *Nature*, *541*(7638), 494–499.

Senís, E., Esgleas, M., Najas, S., Jiménez-Sábado, V., Bertani, C., Giménez-Alejandre, M., … Abad, M. (2021). TUNAR lncRNA Encodes a Microprotein that Regulates Neural Differentiation and Neurite Formation by Modulating Calcium Dynamics. *Frontiers in Cell*

Author Manuscript

and Developmental Biology, *9*, 747667.

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., … Saghatelian, A. (2012). Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology*, *9*(1), 59–64.

Smith, C. C., Selitsky, S. R., Chai, S., Armistead, P. M., Vincent, B. G., & Serody, J. S. (2019). Alternative tumour-specific antigens. *Nature Reviews. Cancer*, *19*(8), 465–478.

Stallmeyer, B., Drugeon, G., Reiss, J., Haenni, A. L., & Mendel, R. R. (1999). Human molybdopterin synthase gene: identification of a bicistronic transcript with overlapping reading frames. *American Journal of Human Genetics*, *64*(3), 698–705.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435–W439.

Steitz, J. A. (1969). Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. *Nature*, Vol. 224, pp. 957–964. doi: 10.1038/224957a0

Steitz, J. A., Dube, S. K., & Rudland, P. S. (1970). Control of Translation by T4 Phage : Altered Ribosome Binding at R17 Initiation Sites. *Nature*, Vol. 226, pp. 824–827. doi: 10.1038/226824a0

Svensson, M., Sköld, K., Svenningsson, P., & Andren, P. E. (2003). Peptidomics-based discovery of novel neuropeptides. *Journal of Proteome Research*, *2*(2), 213–219.

Tian, X., Zhang, S., Zhou, L., Seyhan, A. A., Hernandez Borrero, L., Zhang, Y., & El-Deiry, W. S. (2021). Targeting the Integrated Stress Response in Cancer Therapy. *Frontiers in Pharmacology*, *12*, 747837.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient

template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159.

Author Manuscript

Tupy, J. L., Bailey, A. M., Dailey, G., Evans-Holm, M., Siebel, C. W., Misra, S., … Rubin, G. M. (2005). Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5495–5500.

Vakirlis, N., Vance, Z., Duggan, K. M., & McLysaght, A. (2022). De novo birth of functional microproteins in the human lineage. *Cell Reports*, *41*(12), 111808.

Vasudevan, D., Neuman, S. D., Yang, A., Lough, L., Brown, B., Bashirullah, A., … Ryoo, H. D. (2020). Translational induction of ATF4 during integrated stress response requires noncanonical initiation factors eIF2D and DENR. *Nature Communications*, *11*(1), 4677.

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial Analysis of Gene Expression. *Science*, Vol. 270, pp. 484–487. doi: 10.1126/science.270.5235.484

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr, … Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, *88*(2), 243–251.

Vibert, J., Saulnier, O., Collin, C., Petit, F., Borgman, K. J. E., Vigneau, J., … Delattre, O. (2022). Oncogenic chimeric transcription factors drive tumor-specific transcription, processing, and translation of silent genomic regions. *Molecular Cell*, *82*(13), 2458–2471.e9.

Wang, Hongwei, Yan Wang, and Zhi Xie. 2019. "Computational Resources for Ribosome Profiling: From Database to Web Server and Software." Briefings in Bioinformatics 20 (1): 144–55.

Wang, J., Okkeri, J., Pavic, K., Wang, Z., Kauko, O., Halonen, T., … Westermarck, J. (2017). Oncoprotein CIP 2A is stabilized via interaction with tumor suppressor PP 2A/B56. *EMBO Reports*, Vol. 18, pp. 437–450. doi: 10.15252/embr.201642788

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*(6), e74.

Wang, Y., Wu, S., Zhu, X., Zhang, L., Deng, J., Li, F., … Zhou, Y. (2020). LncRNA-encoded polypeptide ASRPS inhibits triple-negative breast cancer angiogenesis. *The Journal of Experimental Medicine*, *217*(3). doi: 10.1084/jem.20190950

Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics*, *2*(4), 216–221.

Author Manuscript

Weber, R., Kleemann, L., Hirschberg, I., Chung, M.-Y., Valkov, E., & Igreja, C. (2022). DAP5 enables main ORF translation on mRNAs with structured and uORF-containing 5' leaders. *Nature Communications*, *13*(1), 7510.

Whiffin, N., Karczewski, K. J., Zhang, X., Chothani, S., Smith, M. J., Evans, D. G., … Ware, J. S. (2020). Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nature Communications*, *11*(1), 2523.

Wright, B. W., Yi, Z., Weissman, J. S., & Chen, J. (2022). The dark proteome: translation from noncanonical open reading frames. *Trends in Cell Biology*, *32*(3), 243–258.

Wu, Q., Wright, M., Gogol, M. M., Bradford, W. D., Zhang, N., & Bazzini, A. A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. *The EMBO Journal*, *39*(17), e104763.

Xia, X., Li, X., Li, F., Wu, X., Zhang, M., Zhou, H., … Zhang, N. (2019). A novel tumor suppressor protein encoded by circular AKT3 RNA inhibits glioblastoma tumorigenicity by competing with active phosphoinositide-dependent Kinase-1. *Molecular Cancer*, *18*(1), 131.

Xu, S., Jia, G., Zhang, H., Wang, L., Cong, Y., Lv, M., … Wang, Y. (2021). LncRNA HOXB-AS3 promotes growth, invasion and migration of epithelial ovarian cancer by altering glycolysis. *Life Sciences*, *264*, 118636.

Yeh, C.-H., Bellon, M., & Nicot, C. (2018). FBXW7: a critical tumor suppressor of human cancers. *Molecular Cancer*, *17*(1), 115.

Yewdell, J. W. (2022). MHC Class I Immunopeptidome: Past, Present, and Future. *Molecular & Cellular Proteomics: MCP*, *21*(7), 100230.

Yu, L., Kang, X., Li, F., Mehrafrooz, B., Makhamreh, A., Fallahi, A., … Wanunu, M. (2023). Unidirectional single-file transport of full-length proteins through a nanopore. *Nature Biotechnology*. doi: 10.1038/s41587-022-01598-3

Zhang, D., Zhang, G., Hu, X., Wu, L., Feng, Y., He, S., … Zhang, L. (2017). Oncogenic RAS Regulates Long Noncoding RNA Orilnc1 in Human Cancer. *Cancer Research*, *77*(14), 3745–3757.

Zhang, M., Zhao, K., Xu, X., Yang, Y., Yan, S., Wei, P., … Zhang, N. (2018). A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nature Communications*, *9*(1), 4475.

Zhang, S., Reljić, B., Liang, C., Kerouanton, B., Francisco, J. C., Peh, J. H., … Ho, L. (2020). Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nature Communications*, *11*(1), 1312.

Zhu, S., Wang, J.-Z., Chen, D., He, Y.-T., Meng, N., Chen, M., … Yan, G.-R. (2020). An oncopeptide regulates m6A recognition by the m6A reader IGF2BP1 and tumorigenesis. *Nature Communications*, *11*(1), 1685.

Zou, Q., Xiao, Z., Huang, R., Wang, X., Wang, X., Zhao, H., & Yang, X. (2019). Survey of the translation shifts in hepatocellular carcinoma with ribosome profiling. *Theranostics*, *9*(14), 4141–4155.

Author Manuscript

Author Manuscript

Figure Legends

Figure 1. Methods for the discovery or validation of non-canonical open reading frames (ncORFs).

Non-canonical open reading frames are putatively identified through computational methods, ribosome profiling, or mass spectrometry approaches. These computational approaches include sequence based approaches that predict the coding potential of a non-canonical open reading frame based on sequence conservation or by analyzing a nucleotide sequence for codon bias and coding potential. Other computational approaches assess ribosome-profiling data to identify the non-canonical open reading frame that gave rise to a given protected (translated) RNA transcript. Finally, translated miniproteins can be studied through mass spectrometry and other functional approaches, including CRISPR-based functional screens and protein structure-function studies, including novel approaches like BONCAT. Ultimately, the successful identification and validation of a non-canonical ORF will require the simultaneous use of multiple methodological approaches. Figure created with biorender.com.

Figure 2. Major categories of ncORFs.

Non-canonical open reading frames are genomic regions bookended by start and stop codons, which confer the potential for translation of these genomic regions. ncORFs can exist within a variety of genomic regions including: upstream of a protein-coding sequence (CDS) in the 5'UTR, downstream of a CDS in the 3' UTR, contained internally within a canonical CDS, overlapping by spanning from a 5' UTR into the CDS or from the CDS into the 3' UTR. Outside of annotated protein-coding genes, ncORFs can also exist within long non-coding RNA (lncRNA), within

retroviral genes, and within pseudogenes. Finally, non-canonical back splicing of pre-mRNA transcripts can give rise to circular RNAs that encode unique translational products. ncORF transcription does not guarantee translation into miniproteins. However, ncORF functionality can be a product of coding-independent activity, coding-dependent activity, or both. Finally, cancer-associated ncORFs, where applicable, are highlighted herein according to their classification. Figure created with biorender.com.

Figure 3. Miniprotein involvement in major cancer signaling pathways.

Miniproteins can suppress or promote pro-oncogenic signaling pathways in diverse manners. At the plasma membrane, AKT3-174aa—encoded by a circular RNA isoform of AKT3—competes with p-PDK-1 to inhibit AKT activation. Consistent with this tumor-suppressive function, circAKT3 is downregulated in glioblastoma (GBM) samples (Xia et al., 2019). Also overexpressed in GBM, C-E-Cad is a secreted miniprotein encoded by a circular isoform of E-cadherin that binds the CR2 domain of EGFR to activate it (Gao et al., 2021). The 59aa SMIM30 miniprotein mediates the SRC/YES1 complex anchoring to the cytoplasmic surface of the plasma membrane. SRC/YES1 membrane-anchoring enables phosphorylation and subsequent activation of the complex, which induces downstream MAPK signaling (Pang et al., 2020). SMIM30 is regulated by c-MYC and overexpressed in hepatocellular carcinoma. In the cytoplasm, the tumor-suppressive FBXW7-185aa promotes c-myc degradation by suppressing USP28 inhibition of the canonical FBXW7 protein (Yeh, Bellon, & Nicot, 2018). Further, CIP2A-BP competes with PP2A for binding to CIP2A (Guo et al., 2020) which prevents CIP2A stabilization, thereby inhibiting downstream AKT/NFkB pathway activation. Localized to the lysosome, the lncRNA-encoded

Author Manuscript

Author Manuscript

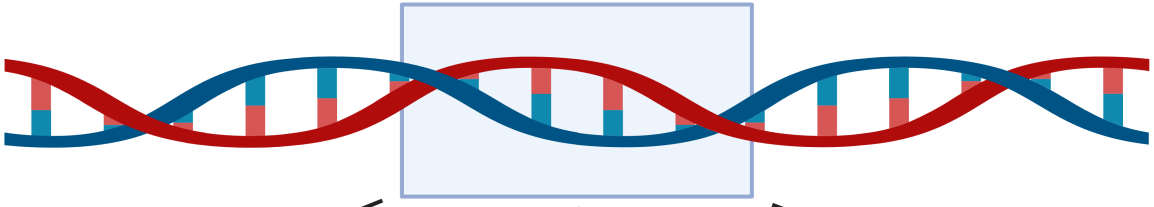
SPAR miniprotein interacts with v-ATPase to prevent mTORC1 recruitment to the complex, attenuating growth-factor-independent mTORC1 activation (Matsumoto, Clohessy, & Pandolfi, 2017). Nuclear-localized miniproteins are implicated in transcriptional regulation. For instance, SEHBP tightly binds histone H2B and interacts with scaffolding proteins that modulate chromatin accessibility, strongly suggesting a role in transcriptional regulation (Koh et al., 2021). Figure created with biorender.com.

Figure 4. Clinical applications of miniproteins, ranging from drug development to biomarkers.

Cancer cells are enriched for ncORF MHC-I associated peptides (MAPs), suggesting that they could serve as a source of tumor antigens for immunotherapeutic applications. Additionally, pharmacological targeting of cancer-specific ncRNAs to silence pro-oncogenic miniprotein translation could enable downstream inhibition of undruggable targets. Miniproteins can also be used to inform drug design. Miniprotein characterization could be used to identify protein-protein interaction hotspots or inhibitor binding sites. Miniprotein conjugation to other drugs could be explored for potential application in molecular glues or heterobifunctional compounds. Miniproteins are also potentially suitable as non-antibody binding scaffolds. Finally, differential expression of ncORFs and miniproteins in cancer potentiates their use as potential biomarkers, especially in cases of differentially expressed secreted miniproteins. Figure created with biorender.com.

Author Manuscript

Potential Coding Region



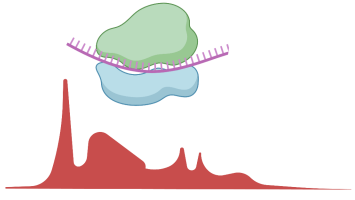
Sequence-Based Approaches

PhyloCSF
sORF Finder
PhastCons
CRITICA
CPAT



Transcriptomics-Based Approaches

Ribo-seq



Ribosome-profiling interpretation

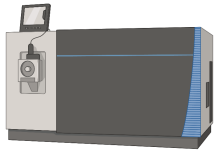
ORFquant
ORFscore
PROTEOFORMER

RNA-seq



Proteomics-Based Approaches

Mass spectrometry



Functional validation

CRISPR-based functional screens

Knockdown analysis

Protein analysis (structure-function studies)

Author Manuscript

ncORF =

coding sequence =

Categories of non-canonical ORFs and Cancer Relevance

upstream ORFs

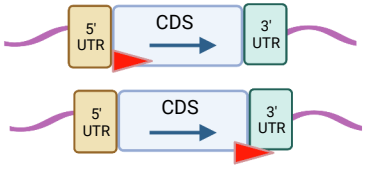


SEHBP
ASNSD1
CPNE1

downstream ORFs

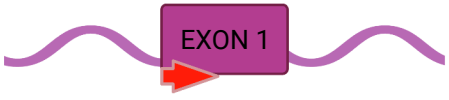


overlapping uORF/dORFs



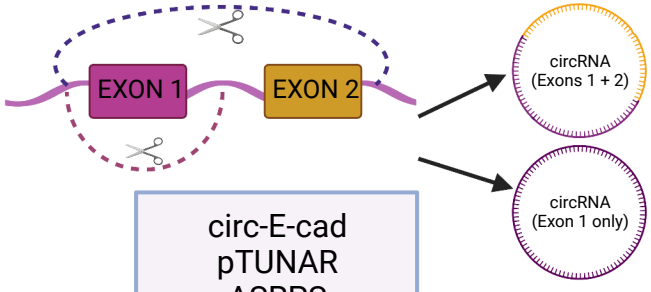
alt-RPL36
POLG

long non-coding RNA



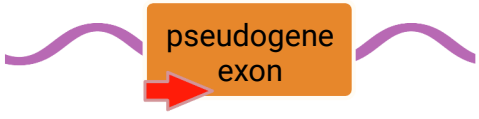
PINT87aa
pTINCR
SMIM30
CIP2A-BP
NoBody
HOXB-AS3
Orilnc1

circular RNA ORFs



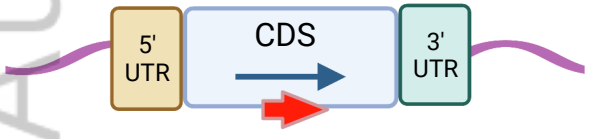
circ-E-cad
pTUNAR
ASRPS
circRasGEF1B
PINT87aa

pseudogenic ORFs

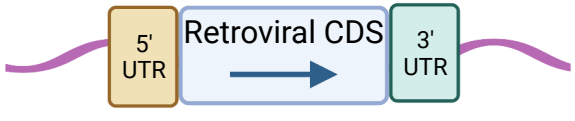


OCT4 pseudogenes
pseudo-BRAF

internal out-of-frame ORFs



retroviral ORFs



Manuscript

Author

C-E-Cad

PIP3

PDK-1

AKT3-174aa

PIP3

AKT

P

P

AKT signaling cascade

P

EGFR

STAT3

P

SRC

SMIM30

YES1

MAPK activation

P

Ub

Ub

Ub

Ub

c-myc

c-myc

proteasomal degradation

FBXW7

USP28

FBXW7-185aa

v-ATPase

mTORC1OFF

SPAR

Ragulator Complex

Lysosome

PI3K

CIP2A

PP2A

AKT

CIP2A BP

= miniprotein

SEHBP

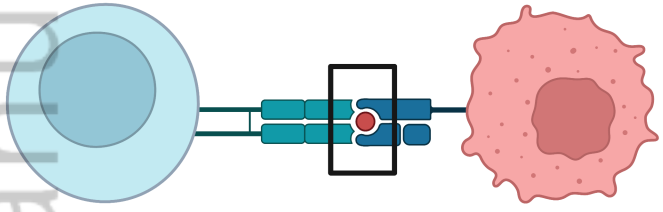
HMGN 1/3

Transcriptional regulation by SEHBP binding to H2B

Author Manuscript

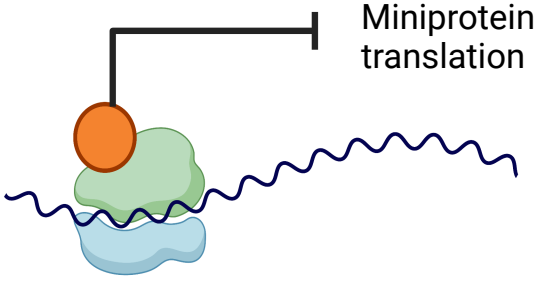
Therapeutic and Diagnostic Opportunities

Immuno-Oncology: New cancer-specific epitope targets

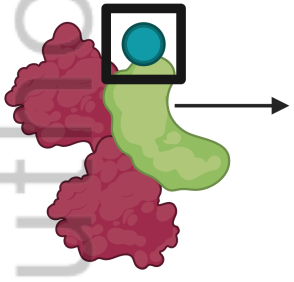


Miniproteins generate tumor-specific HLA-peptides

Pharmacological targeting at the RNA level



Miniprotein-informed Drug Design

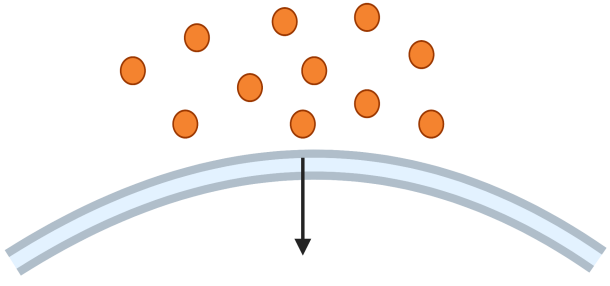


Identification of PPI hotspots

Rational design of PPI inhibitors

Conjugate miniproteins to other drugs

Miniproteins as non-antibody binding scaffolds



Biomarkers based on differential ncORF or miniprotein expression

