Prediction of novel bioactive micropeptides in the immune system

Fengyuan Hu

The Babraham Institute

St Edmund's College, University of Cambridge

June 2020

Submitted for the degree of Doctor of Philosophy at the University of Cambridge

Declaration of Originality

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Fengyuan Hu

Table of Contents

Acknowledgements

The work presented in this thesis would not have been possible without the contribution of many other people. Firstly, I would like to express my sincere gratitude to my supervisor Dr Martin Turner for continuous support of my PhD study, for his patience, motivation, and immense knowledge, for bring forth the idea to search for novel functional micropeptides in the immune system, and also for providing the guidance that kept me on course. I have benefitted and learned greatly from his guidance during the four years of my PhD study. I would like to thank my Personal Committee, they are my Mentors Dr Alexander Saveliev and Dr Manuel D. Diaz-Muñoz, my Assessor Dr Mikhail Spivakov and my University Supervisor Professor Chris Smith, and they have shared valuable knowledge, insight and encouragement to me at different stages of my study. I would also like to thank Dr Louise Matheson, David Turner and Dr Jia Lu for sharing their comments and results.

I would also like to say big thank you to the rest of the Turner lab, past and present: Elisa Monzon-Casanova, Sarah Bell, Kirsty Bates, Fiamma Salerno, Beatriz Saenz, Vanessa D'Angeli, Oezge Gizlenci, Michael Screen, Twm Mitchell, Georg Petkau, Helena Ahlfors, Rebecca Newman, Alison Galloway, Ramkumar Venigalla, Krishnendu Chakraborty and Richard Coulson for being friends and advisors, for being supportive, providing discussion and helping with my general wellbeing. You made my time in the lab enjoyable.

My work at Babraham would never have been completed without the support from the staff at the institute. Thanks to Simon Andrews, Felix Krueger, Christel Krueger, Steven Wingett, Anne Segonds-Pichon and Laura Biggins from the Bioinformatics team for their help with bioinformatics and statistics. Thanks to Katarzyna Wojdyla, David Oxley and Judith Webster from Proteomics team for the discussion about mass spectrometry experiments for micropeptide dectection. I would also like to thank people from the University of Cambridge, thanks to Juan Mata and Caia Duncan from the Department of Biochemistry for sharing their experience with micropeptide discovery in the fission yeast *Schizosaccharomyces pombe.* Thanks to Daniel Hodson, Jie Gao and Joanna Krupka from the Department of Haematology for sharing their data and knowledge about human lymphoma biology.

I would like to thank my viva examiners Dr Julie Aspden and Dr Juan Mata for all the insightful discussions and valuable suggestions, this is truly a memorable experience.

I gratefully acknowledge my financial support from BBSRC for my studentship.

My acknowledgement would be incomplete without thanking the biggest source of my strength, my family. Thanks to my parents and my in-laws for being understanding and helpful. Thanks to my lovely babies Ray and Laurene for bringing me fun and keeping me going. And finally, thank you Wenting, for your love.

,

,

Abstract

Open reading frames (ORFs) are the genomic DNA sequences that have the potential to be translated. Genome annotation pipelines dismiss translation products of small ORFs (smORFs) of 100 codons or fewer (⩽ 300 nucleotides) as being unlikely to have a biological function. Recently, a number of micropeptides with diverse functions have been discovered in different organisms. The smORFs that encode them were identified from sequences originally annotated as non-coding regions of the genome including untranslated regions of an mRNA or a non-coding RNA (ncRNA). Newly discovered micropeptides have been shown to influence various biological processes and diseases. These new discoveries complement already characterised peptides and small proteins known to be important biological regulators. Within the immune system the best characterised of these include host defense anti-microbial peptides (~12-50 AA), chemokines (~90-100 AA) and cytokines (101-200 AA) that are known to play essential roles in normal and pathological immune reactions. Questions remain as to how widespread micropeptides are in the immune system and what their functional roles might be.

Here we describe a new analytical pipeline that performs a comprehensive and systematic analyses of RNA-Seq and Ribosome profiling to identify actively translated smORFs. In comparison to previously published pipelines, our pipeline is more stringent at smORF prediction. We have applied our pipeline to mouse B and T cells and discovered 5744 actively translated smORFs and their predicted translation products. smORFs were classified, and for each class, we performed analyses to look at their conservation, translation efficiency, and the biological processes linked to them. It has been shown in the UniProt database that a small subset of chemokines and majority of cytokines are between 101 and 200 AA long. With this in mind, we extended our analysis to candidate proteins of up to 200 AA in length and found evidence for translation of 945 such polypeptides. We further investigate whether the predicted micropeptides possess features of signal peptides which have a potential to be secreted and could act as immune regulators. Furthermore, verifying their existence and identifying their functions will be essential and potentially lead to useful applications.

Abbreviations

AA	Amino acid
ORF	Open reading frame
smORF	Translated small open reading frames
SEP	smORF-encoded peptide
HDP	Host defense peptide
AMP	Antimicrobial peptide
CDS	Coding DNA sequence
RPF	Ribosome protected fragment
TIS	Translation initiation site
TE	Translation efficiency
MS	Mass spectrometry
UTR	Untranslated region
TUF	Transcript with unknown function
LPS	Lipopolysaccharide
LTR	Long terminal repeat

Chapter One

Introduction

1.1 Roles of bioactive peptides and small proteins in the immune system

Biologically active peptides and small proteins belong to a class of molecules that play essential regulatory roles in diverse biological processes (Boonen et al, 2009; Andrews and Rothnagel, 2014; Cabrera-Quio et al., 2016). Neuropeptides and hormones are considered the best examples of extensively studied small proteins, they are derived from larger precursor proteins and contain N-terminal signal sequences (Fricker, 2005; Cunha et al., 2008). Within the immune system the best characterised of these include host defense antimicrobial peptides, hormones and cytokines that are known to have important functions in normal and pathological immune reactions.

Host defence peptides (HDPs), also known as antimicrobial peptides (AMPs) are produced by various cells and tissues in all classes of complex organisms, such as amphibians, birds, insects, mammals, and plants. Their lengths vary between 12 and 50 amino acids (AA). For instance, neutrophils in the human innate immune system produce alpha-defensins (Ganz et al., 1985), and another HDP called dermcidin is secreted by human sweat glands onto the skin (Schittek et al., 2001). The general defence mechanism is that HDPs interact with membranes of microbes and disrupt them. Some of the HPDs may form pores or holes in the membrane and others may change the membrane structure by poking into it in many places. Apart from having a role against microorganisms, they are also involved in activities in wound healing and in the maintenance of the microbiota (Hancock et al., 2016).

Peptide hormones or protein hormones are secreted from animal and plant’s cells. Plant’s hormones are able to influence plant’s growth and development including embryogenesis, the regulation of organ size, pathogen defense, stress tolerance and reproductive development (Shigenaga and Argueso, 2016; Pierre-Jerome et al., 2018; Bürger and Chory, 2019). Animal hormones play a role in regulating animal's growth, metabolism, and sexual development and function (Neave, 2007). They act as extracellular signaling molecules and bind to a receptor protein that is embedded in the plasma membrane of the target cell. The inside portion of the receptor undergoes a conformational change which activates intracellular signaling reactions, involving signaling proteins. One or more signaling proteins alter the activities of effector proteins and thereby the cell behaviour. Immune cells synthesize, store and secrete hormones.

Hormones, including adrenocorticotropic hormone (ACTH), endorphin and triiodothyronine (T3) were found in Natural killer cells and activated T cells, and suggested a need for a decrease in the levels of these hormones for the killing of target cells (Pállinger, and Csaba, 2008).

Small peptides and proteins that are secreted and play an important role in cell signaling are loosely categorized as Cytokines. Cytokine is a general name, it includes chemokines (cytokines with chemotactic activities), interferons (cytokines in response to virus infection), interleukins (cytokines made by one leukocyte and acting on other leukocytes), lymphokines (cytokines made by lymphocytes), monokine (cytokines made by monocytes) and tumour necrosis factors, but generally not hormones. Immune cells including macrophages, B lymphocytes, T lymphocytes and mast cells produce cytokines (Abbas et al., 2014). Cytokines have been shown to act on the cells that secrete them (autocrine action), on nearby cells (paracrine action), or in some instances on distant cells (endocrine action) as immunomodulating agents. Cytokines are involved in health and disease, specifically in development (Saito, 2001), host responses to infection, inflammation, trauma, and sepsis (Dinarello, 2000), also they are linked to schizophrenia, depression (Dowlati et al., 2010), Alzheimer's disease (Swardfager et al., 2010) and cancer (Locksley et al., 2001). Chemokines are a family of cytokines with length ranging from 80-100 AA in different species (Abbas et al., 2014). They are essential signaling molecules in both innate and adaptive immune responses serving as messengers for intracellular communication and recruiting leukocytes to move towards sites of infection or inflammation. Interleukins (ILs) are larger cytokines than chemokines, with the majority of them between 100 and 200 AA in length. They promote the development and differentiation of T and B lymphocytes, and hematopoietic cells. For example, interleukin 4 (IL4) produced principally by CD4+ T cells, is important in promoting B cell responses including B cell proliferation, class switch recombination and somatic hypermutation, as well as the differentiation of B cells into plasma cells (Yokota et al., 1986).

1.2 Small open reading frames (smORFs) and micropeptides

The concept of a gene has been continuously refined, evolved and has become more complex. Initially genes were identified as DNA sequences which contain open reading frames (ORFs).

Those ORFs are potentially translatable DNA region that begins with a start codon (e.g. AUG) and ends with one of the three stop codons (UAA, UGA, and UAG), with no stop codons in between, they are termed CDS (from coding DNA sequence) and direct the sequence of amino acids in a protein. As the understanding of the molecular mechanisms of genes deepens, the updated view has included regulatory regions and elements (promotor, enhancer, chromatin structure) as well as transcripts (Gerstein et al, 2007). Furthermore, non-coding transcripts have been discovered from distinct genomic loci, they bear signatures of mRNAs, including 5' capping, spliced via canonical splice motifs, and polyadenylation, but one important difference between ncRNAs and protein-coding mRNAs is their low level of nucleotide sequence conservation (Carninci et al., 2005). These non-coding genes and their transcript products (known as non-coding RNAs or ncRNAs) have revolutionized our understanding of gene regulation (Derrien et al., 2012; Guttman and Rinn, 2012).

The advent of next-generation sequencing technologies and proteomic approaches has led to a more comprehensive annotation of genes, transcripts and their translated protein products. Several large-scale genomic studies have revealed that a much larger fraction of the genome is transcribed and translated than was anticipated (Carninci et al., 2005; Kim et al., 2014; Ingolia et al., 2014). In recent years, a class of genetic elements has emerged to challenge the understanding of the coding potential of the genome: translated functional small ORFs (smORFs or sORFs) of 100 codons and fewer (Basrai et al., 1997). Apart from annotated CDSs, the genomes of many metazoans, including mouse and human, contain millions of putative smORF sequences (Kastenmayer et al., 2006; Frith et al., 2006; Ladoukakis et al., 2011). The discovery of smORFs and their protein products points to a fundamental gap in our knowledge of protein-coding genes.

The sizes of smORFs can range from 2 (a theoretical lower boundary) to 100 codons in length (Andrews and Rothnagel, 2014). The shortest coding smORF reported to date has 6 codons, it is an upstream open reading frame (uORF) on S-Adenosylmethionine decarboxylase (AdoMetDC) mRNA. AdoMetDC is a key enzyme in the pathway of polyamine biosynthesis. The cellular levels of the polyamines regulate AdoMetDC translation. The AdoMetDC uORF, which encodes a peptide of sequence MAGDIS, is specifically required for translational control of AdoMetDC by polyamines (Ruan et al., 1996; Law et al., 2001; Raney et al., 2002). There is no consensus for the upper limit of a smORF, some studies have described smORFs of 150-200 codons (Hayden and Bosco, 2008; Yang et al., 2011). The protein products of smORFs are referred to as SEPs (from smORF-encoded polypeptides) or micropeptides (Saghatelian and Couso, 2015; Anderson et al., 2015; Mackowiak et al., 2015). Micropeptides differ from classical bioactive peptides in how they are biochemically synthesized (**Figure 1.1**) (Saghatelian and Couso, 2015; Makarewich and Olson, 2017). Classical bioactive peptides such as neuropeptides, peptide hormones and growth factors are often enzymatically cleaved from longer precursor proteins by proteolysis to form their final active structures. Take insulin as an example, it is firstly synthesized as a single polypeptide called preproinsulin in pancreatic β-cells, subsequently, the signal peptide is cleaved to form proinsulin. To form the mature insulin, the proinsulin is then cleaved at two positions to yield two polypeptide chains linked by 2 disulphide bonds. The resulting mature insulin is secreted to the outside of the cell (Steiner and Oyer, 1967). Bioactive micropeptides in principle are directly translated and released in the cytoplasm and mitochondria (Aspden et al., 2014) as well as nucleus (Slavoff et al., 2014) without being processed (**Figure 1.1**) (Hashimoto et al., 2008).

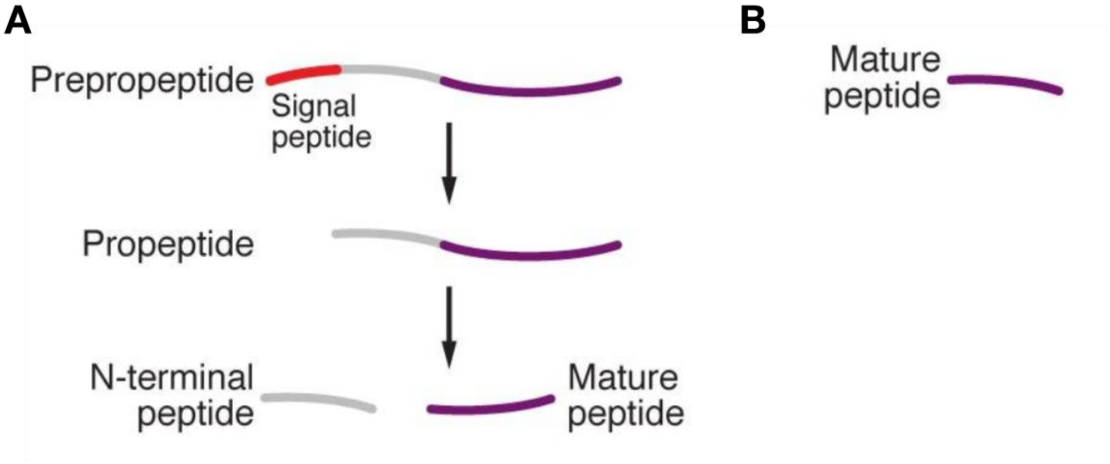


Figure 1.1 | **Micropeptides and classical bioactive peptide biosynthesis.** (A) Peptide hormones are cleaved from longer precursor prepropeptides to form their final active structures. (B) Micropeptides are directly translated and released without being processed.

Recently, a large number of micropeptides with diverse functions have been discovered in different organisms (**Table 1.1**) (Duncan and Mata, 2014; Hsu et al., 2018; Finkel et al., 2018; Delcourt et al., 2018; Erpf and Fraser, 2018; van Heesch et al., 2019). The smORFs that encode them were identified from sequences originally annotated as non-coding regions of the genome including untranslated regions of an mRNA or a ncRNA, with ncRNA being a major source (Anderson et al., 2015; Nelson et al., 2016; D'lima et al., 2017; Matsumoto et al., 2017; van Heesch et al., 2019). Newly discovered micropeptides have been shown to influence development (Kondo et al., 2007,2010; Chng et al., 2013; Pauli et al., 2014; Chaunt-Delalande et al., 2014), DNA repair (Slavoff et al., 2014), mRNA decapping (D'Lima et al., 2017), muscle calcium homeostasis (Magny et al., 2013; Anderson et al., 2015,2016; Nelson et al., 2015,2016), metabolism (Lee et al, 2015), stress signalling (Matsumoto et al., 2017), cancer (Huang et al., 2017) and inflammatory diseases (Jackson et al., 2018). These new discoveries emphasize the functional potential of this unexplored class of biomolecules and complement already characterised peptides and small proteins known to be important biological regulators.

Micropeptide gene name	**Conservation**	**Method of identification/ characterization**	**Function**	**Size (number of amino acids)**	**Reference**
HAMP (Hepcidin)	Vertebrates	MS assay	Regulates iron metabolism and mediator of anemia of inflammation	25	Krause et al., 2000; Park et al., 2001
ENOD40-1	Plants	*In vitro* translation	Associates with a subunit of sucrose synthase in root nodule	12 and 24	Röhrig et al., 2002
PLS (POLARIS)	Plants	Gene expression analysis	Leaf morphogenesis	36	Casson et al., 2002
Brick1 (Brk)	Plants and animals	Mutation analysis	Leaf morphogenesis	76	Frank and Smith, 2002
MT-RNR2 (Humanin)	Mammals (only 6: Bonobo, Cat, Chimpanzee, Gelada, Tiger, Green monkey)	Functional expression screening of cDNA library	Neuroprotective factor and involve in programmed cell death	24	Tajima et al., 2002; Guo et al., 2003
ROT4	Plants	Screening of a mutant in *Arabidopsis thaliana*	Leaf morphogenesis	53	Narita et al., 2004

tal (ftarsal-less/Polished rice/Pri)	Insects	Mutation analysis	Activates an essential transcription factor, driving formation of cuticle structures during embryo development	11-32 (Three of 11; one of 32 which is rarely translated)	Galindo et al., 2007; Kondo et al., 2007, 2010
Wfdc21	Mammals (non-human)	Bioinformatics (microarray analysis)	Promotes activation of the metalloproteinase MMP2	63	Wu et al., 2008
C12orf75 (AGD3)	Mammals	Bioinformatics (microarray and RNA-Seq analysis)	Involves in stem cell differentiation	63	Kikuchi et al., 2009
CYREN (MRI-2)	Mammals	MS screening and RNA-Seq	DNA repairing process (non-homologous end joining pathway)	69	Slavoff et al., 2014
Toddler	Vertebrates	Bioinformatics (mined zebrafish genomic data sets for previously non-annotated translated open reading frames)	Activates a G protein–coupled receptor to promote migration of mesendodermal cells in the developing embryo	58	Pauli et al., 2014
MLN (Myoregulin)	Mammals	Bioinformatics (screen for uncharacterized skeletal muscle-specific genes)	Calcium homeostasis	46	Anderson et al., 2015
MT-RNR1 (MOTS-c)	Vertebrates	Bioinformatics (in silico search for potential smORFs in 12S rRNA)	Regulates insulin sensitivity and metabolic homeostasis	16	Lee et al., 2015
STRIT1	Mammals	Bioinformatics	Interacts with	35	Nelson et al.,

(DWORF)		(PhyloCSF search)	and enhances calcium pump activity in muscle cells		2016; Makarewich et al., 2018
NBDY (NoBody)	Mammals	MS screening and RNA-Seq	mRNA decapping process	68	D'Lima et al., 2017
SPAAR	Mammals	Proteomics	Regulates mTORC1 and muscle regeneration	90	Matsumoto et al., 2017
HOXB-AS3	Primates	Ribo-Seq	Inhibits colon cancer growth	53	Huang et al., 2017
MYMX (Myomixer/Minion)	Mammals	CRISPR mediated loss of function screening of genes required for myoblast fusion	Mediates cell fusion and muscle formation	84	Bi et al., 2017; Zhang et al., 2017
Aw112010	Mammals (non-human)	Ribo-Seq	Controls mucosal inflammatory response	82	Jackson et al., 2018
MTLN (Mitoregulin /MPM)	Vertebrates	Bioinformatics (in silico search for potential smORFs in transcripts detected in mouse skeletal muscle)	Enhances mitochondrial respiratory activity and promotes myogenic differentiation	56	Stein et al., 2018; Lin et al., 2019
PIGBOS1	Mammals	Proteomics	Regulates unfolded protein response	54	Chu et al., 2019
NCBP2AS2/KRASIM	Vertebrates and *Drosophila*	Proteomics	promotes tumor angiogenesis in cancer-associated fibroblasts	99	Kugeratski et al., 2019; Prensner et al., 2020
POLGARF	Mammals	Ribo-Seq and	Unknown	64	Loughran et

		MS	function but potentially a regulatory protein		al., 2020
BRAWNIN (BR)	Vertebrates	Bioinformatics and overexpression of the smORF in cell line	Essential for respiratory chain complex III (CIII) assembly	71	Zhang et al., 2020

Table 1.1 | **Examples of characterized micropeptides and their biological functions.**

1.3 Identification of smORFs and micropeptides

High quality gene annotation requires the power to correctly identify open reading frames that encode genuine protein products, and to discriminate between them and the vast number of ORFs that are untranslated in a particular context. This challenge becomes particularly acute when applied to smORFs. The high numbers of smORFs and lack of experimental validation present a challenge for annotation and curation. The difficulty for smORF study is that functional smORFs are often discarded by genome annotations because they have not been experimentally validated or no homology with other protein-coding genes.

Putative ORFs can exist in any DNA sequence by chance. Stop codons occur at a frequency of roughly 1 in 20 in random sequences, ORFs of up to 60 codons will occur frequently by chance (5%) and even ORFs of 150 codons will appear by chance in a large genome (0.05%) (Kamvysselis, 2003). Many putative ORFs do not encode proteins. Traditional computational prediction of protein-coding ORFs relies on a number of stringent criteria to remove meaningless ORFs, such as size cutoff of 300 nucleotides, AUG start codon usage, and sequence similarity (Gish and States, 1993; Kochetov, 2004), rendering them inappropriate for smORF detection.

The ORF length is a fundamental criterion used to distinguish *bona fide* protein-coding ORFs from short putative ORFs that occur by chance (Dinger et al., 2008). The likelihood that an ORF encodes an authentic protein increases with its length (Lipman et al., 2002). The choice of a 100-codon cutoff as the minimum size for detection was made historically by ORF-discovery pipelines to search for long ORFs that are unlikely to have occurred by chance. This cutoff has also been based on the assumption that peptides of shorter than 100 AA have exceedingly low probability to fold into stable structures to perform robust biological functions (Ingolia et al., 2014). This arbitrary threshold is consistent with the observation that > 95% of proteins in public databases such as UniProtKB/Swiss-Prot (UniProt Consortium, 2018) are > 100 AA in length, and has subsequently been shown to display a high level of concordance with more sophisticated discrimination methods (Frith et al., 2006). ORFs of less than 100 codons have been disregarded by such filtering and it will potentially result in the misclassification of some protein coding transcripts as ncRNAs.

Translation is a key feature for of smORFs in its expression. Selection of the translation initiation site (TIS) is a crucial step during translation. In the classic view of eukaryotic translation, ribosomes almost always initiate at the first AUG codon on an mRNA and translate a single, long open reading frame (Hinnebusch, 2014). However, exceptions have been known since the 1980s that translation can initiate at a non-AUG codon, even though at a much lower efficiency (Zitomer et al., 1984; Peabody, 1987, 1989; Clements et al., 1988; Hann et al., 1988). In most of these cases, near-cognate codons CUG, GUG, and UUG are used. Interestingly, not all near-cognate start codons are equally efficient, CUG is generally most efficient (Kearse and Wilusz., 2017). Recent advancements in ribosome footprint mapping have revealed that non-AUG start codons are used at an astonishing frequency (~60%) (Ingolia et al., 2009, 2011). Methods solely based on AUG start codon have limited the smORF discovery.

Searching for sequence conservation is a commonly used way to verify whether an ORF is actually a protein-coding sequence, because conservation across species is a strong indication of function. This can be done by detecting either its similarity to annotated protein sequences in a pairwise alignment manner or its conservation across species by multiple sequence alignment. However, these methods do not work well to differentiate coding smORFs from non-coding smORFs. For example, BLAST tool is size dependent, it measures the absolute amount of conservation, i.e., the number of conserved amino acid positions (Wheeler et al., 2006), so short sequences are physically unable to obtain high conservation scores as an indication of functionality. BLAST penalizes the identification of protein sequences of fewer than 80 AA and fails below 20 AA (Ladoukakis et al., 2011). K_a/K_s ratio is another example. K_a is the number of non-synonymous (a nucleotide mutation that alters the amino acid sequence of a protein) substitutions per non-synonymous site per time period, K_s is the number of synonymous (a nucleotide mutation that do not alter the amino acid sequence of a protein) substitutions per synonymous site in the same time period. Amino acid sequences of canonical protein-coding ORFs are conserved across different species, and the K_a/K_s ratio measures this purifying selection at the nucleotide level. It is expected to see a prevalence of synonymous versus non-synonymous codon substitutions ($K_a/K_s<1$). However, it is difficult to score statistically significant values for very short sequences because the number of possible changes is low, such that K_a/K_s loses predictive power below 100 AA (Couso, 2015).

Translation of smORFs is not direct evidence for function of the peptide produced. Experimental evidence for smORF function was and is difficult to obtain. Because of their small size, smORFs in model organisms such as mice, flies, and zebrafish are less likely to be hit in random mutagenesis screens than larger ORFs, and the large number of smORFs in the genome makes it impractical to carry out systematic mutagenesis, meaning their functions are less likely to be revealed. As for the micropeptides themselves, the standard practice for isolation is to use electrophoresis to separate peptides by size, this method fails to detect peptides below 10kDa (Couso and Patraquim, 2017). Small peptides would often run off the gel or masked by degraded peptides from large proteins.

Despite the challenges, smORFs and micropeptides have been uncovered in an *ad hoc* manner over the years. Searching for regulators responsible for certain phenotypes resulted in the unexpected discovery of a few micropeptides (**Table 1.1**) (Galindo et al., 2007; Kondo et al., 2007,2010; Wadler and Vanderpool, 2007; Maki et al., 2010; Rice and Vanderpool, 2011). Studies in *Drosophila melanogaster* revealed micropeptides have a crucial regulatory role in larval epidermal differentiation (Galindo et al., 2007; Kondo et al., 2007, 2010) and Cardiac physiology (Schiemann et al., 2019). The Tal gene was previously annotated as a non-coding RNA. Four tandem smORFs in the Tal transcript are independently translated to micropeptides of 11 (translated from the first three smORFs) and 32 AA (the fourth smORF) in epidermal cells, the fourth smORF is rarely translated. The transcription factor Ovo binds to the promoter of a target gene to repress gene expression and prevent differentiation of larval epidermal cells, and no trichomes are formed. The Tal micropeptides promote the post-translational modification of Ovo; they promote cleavage of Ovo repressor domain, which turns Ovo into an activator to switch on target gene expression that induces trichome formation (Kondo et al., 2010). Phylogenetic analysis as the evidence for conservation showed that the Tal gene belongs to a gene family that is at least 440 million years old (Galindo et al., 2007). Recently another micropeptide named Myoregulin was identified when a bioinformatics screening for uncharacterized skeletal muscle-specific genes was carried out (Anderson et al., 2015). Similar to Tal, this gene was annotated as a putative long non-coding RNA (lncRNA). Myoregulin is a member of SERCA-inhibitory micropeptide family and is conserved at the structural and functional level. It has a role in regulating muscle performance by inhibiting the activity of SERCA which is the membrane pump that controls muscle relaxation by regulating Ca^{2+} uptake. Following Myoregulin gene knockout, mice show improved exercise performance and Ca^{2+} handling in muscle (Anderson et al., 2015).

Approaches to identify protein-coding smORFs

The identification of smORFs that are translatable and that are likely to encode micropeptides remains a major challenge. With the advancement of technology, the challenge has begun to be addressed. Recent computational and experimental approaches have been developed to increase our ability to infer the translational state and coding potential of smORFs and detect the micropeptides generated from translation. Three complementary approaches that are typically used to discover functional smORFs are bioinformatics, transcriptomics and proteomics (**Table 1.2**). However, these techniques are useful for identification of smORF and micropeptides and not for direct functional characterization.

Approaches to identify smORFs	**Methods & Metrics**	**Description**	**Reference**
Bioinformatics	sORF finder, CPC	Tools to locate smORFs having coding potential	Hanada et al., 2009; Kong et al., 2007
	PhyloCSF	A computational method examining evolutionary conservation of a smORF across species	Lin et al., 2011
Transcriptomics/ translatomics	Ribo-Seq	A deep sequencing-based method of ribosome protected fragments to obtain global snapshot of translation	Ingolia et al., 2009
	Poly-Ribo-Seq	A combination of Ribo-Seq and polysome to	Aspden et al., 2014

		enrich more potent protein-coding ORFs	
	Ribosome Release Score (RSS)	A metric to detect the termination of translation at the stop codon of an ORF using Ribo-Seq	Guttman et al., 2014
	FLOSS score	A method designed to distinguish true coding from non-coding sequences based on the RPF-length distribution	Ingolia et al., 2014
	ORFScore, ORF-RATER, RiboORF, RiboTaper, RP-BP, RiboCode, Ribotricer	These are methods to identify true protein coding ORFs based on triplet periodicity pattern in Ribo-Seq data	Bazzini et al., 2014; Fields et al., 2015; Ji et al., 2015; Calviello et al., 2016; Malone et al., 2017; Xiao et al., 2018; Choudhary et al., 2020
	GWIPS-viz, TISdb, uORFdb, RPFdb, sORFs.org, SmProt, HRPDviewer	Databases to collect Ribo-Seq data and genome annotations derived from the data	Michel et al., 2013; Wan and Qian, 2013; Wethmar et al., 2013; Xie et al., 2015; Olexiouk et al., 2015; Hao et al., 2017; Wu et al., 2018
Proteomics/ peptidomics	Proteogenomics	A method that combines proteomics, genomics, and transcriptomics	Slavoff et al., 2013

Table 1.2 | **Computational and experimental approaches to identify smORFs.**

Bioinformatics

A large collection of putative translatable smORFs have been identified by bioinformatics methods based on the level of DNA and protein sequence conservation across species and synonymous (nucleotide substitutions that do not change the coded amino acid) versus nonsynonymous substitution (Kimura, 1980; Ina, 1995; Makalowski and Boguski, 1998), coding potential (Karlin et al., 1998; Bateman et al., 2004; Skarshewski et al., 2014), sites of transcripts and context of the initiation codon (Kozak, 1987; Brent and Guigó, 2004).

As mentioned earlier, tools such as BLAST and K_a/K_s ratio do not work well to differentiate coding smORFs from non-coding smORFs. More recently, bioinformatics tools were designed to overcome the limitations. sORF finder is a package for identifying smORFs with coding potential based on their similarity in nucleotide composition to known coding sequences (Hanada et al., 2009). It calculates the likelihood of a smORF appearing in the coding regions of a genome using Bayesian estimation. This method was initially applied to two small protein gene datasets in S. cerevisiae and A. thaliana and showed low false negative rate (~9%). CPC (Coding Potential Calculator) is a support vector machine classifier that incorporates six sequence features to discriminate coding versus non-coding ORFs (Kong et al., 2007). Three of the features score the quality of the ORF (size, coverage, integrity) and the remaining three, assessed by BLASTX, are based on the putative homologous protein sequences which are conserved in other species (number of hits, quality of hits, frame distribution of hits). PhyloCSF (phylogenetic codon substitution frequency) is a vigorous conservation-based method. It evaluates the likelihood of an ORF to be a conserved protein-coding sequence by analysing multiple alignment of nucleotide sequences incorporate phylogenetic distance and a model of codon substitution frequencies (Lin et al., 2011). PhyloCSF is a method to determine whether a multi-species nucleotide sequence alignment is likely to represent a protein-coding region, it provides a conservation score for all six reading frames (three on the forward strand and three on the reverse strand) of a given genomic sequence. PhyloCSF has been used to identify several novel micropeptides, including non-annotated P-body dissociating polypeptide (NBDY) (D'Lima et al., 2017) (**Figure 1.2**). PhyloCSF has been integrated to UCSC genome browser, which makes easy access for the community (Cabili et al., 2011; Pauli et al., 2012).

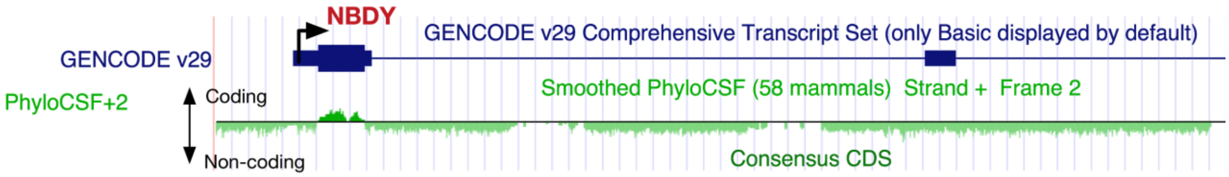


Figure 1.2 | **Identification of micropeptide by PhyloCSF.** PhyloCSF is a computational tool to identify potential coding genes based on the evolutionary conservation of their nucleotide sequence. PhyloCSF has been used to identify several novel micropeptides, including non-

annotated P-body dissociating polypeptide (NBDY). NBDY scores positively on PhyloCSF (upward deflection) in exon 1 that encodes the functional NBDY protein. NBDY transcript was annotated as noncoding before it was discovered, however it shows strong conservation in PhyloCSF. By combining RNA-Seq and MS, NBDY was identified in a screening, then it was characterized by performing immunoprecipitation and MS analysis (IP-MS) on the co-precipitated proteins. NBDY is a component of the mRNA decapping protein complex cross-linking to EDC4 (enhancer of mRNA decapping 4) and regulates the P-body number in cells by interacting with decapping proteins.

In silico micropeptide feature searches and modelling have been applied to amino acid sequences in recent studies (Jackson et al., 2018; van Heesch et al., 2019). TargetP (Emanuelsson et al., 2000) and DeepLoc (Almagro Armenteros et al., 2017) are online tools for prediction of protein subcellular localizations. Prediction of signal peptides can be performed by SignalP and transmembrane helices by TMHMM (Petersen et al., 2011). InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites (Mitchell et al., 2018).

Studies that used these tools have predicted hundreds, even thousands of putative novel smORFs in difference genomes (Frith et al., 2006; Ulitsky et al., 2011; Pauli et al., 2012). However, this will not work for all smORFs. For example, those that are encoded by lncRNAs are less likely to be conserved given lncRNAs themselves do not have high sequence conservation (Makarewich and Olson, 2017). These methods are also used together with experimental methods to validate the identified smORFs.

Transcriptomics

RNA sequencing (RNA-Seq) and ribosome profiling (Ribo-seq) are transcriptomic-based experimental methods for finding potential translated smORFs and micropeptides. RNA-Seq uses next-generation sequencing (NGS) to determine which RNAs are expressed in a given cell, tissue, or organism at a specific time point. This collection of data, known as a transcriptome, can then be used as transcriptional evidence to combine *in silico* prediction for finding potential smORFs (Ladoukakis et al., 2011). Ribosome profiling (Ribo-Seq), an approach based on massively parallel deep sequencing of isolated ribosome-protected fragments (RPFs, or referred to as “ribosome footprints”), provides a “snapshot” of genome-wide protein synthesis *in* *vivo* with single-nucleotide resolution thus allows detailed and accurate analysis of protein production (Ingolia et al., 2009,2012,2014). In this method, firstly samples are either treated with a protein synthesis inhibitor to stall translating ribosomes or with no-drug treatment like thermal freezing (Michel et al., 2013), ribosome-bound RNAs then are extracted from cell lysates and undergo nuclease digestion to generate RPFs. These RPFs are isolated and purified for sequencing. The sequencing reads will be mapped to a reference genome or transcriptome to identify the precise position of the ribosome at the time the translation was halted (**Figure 1.3**).

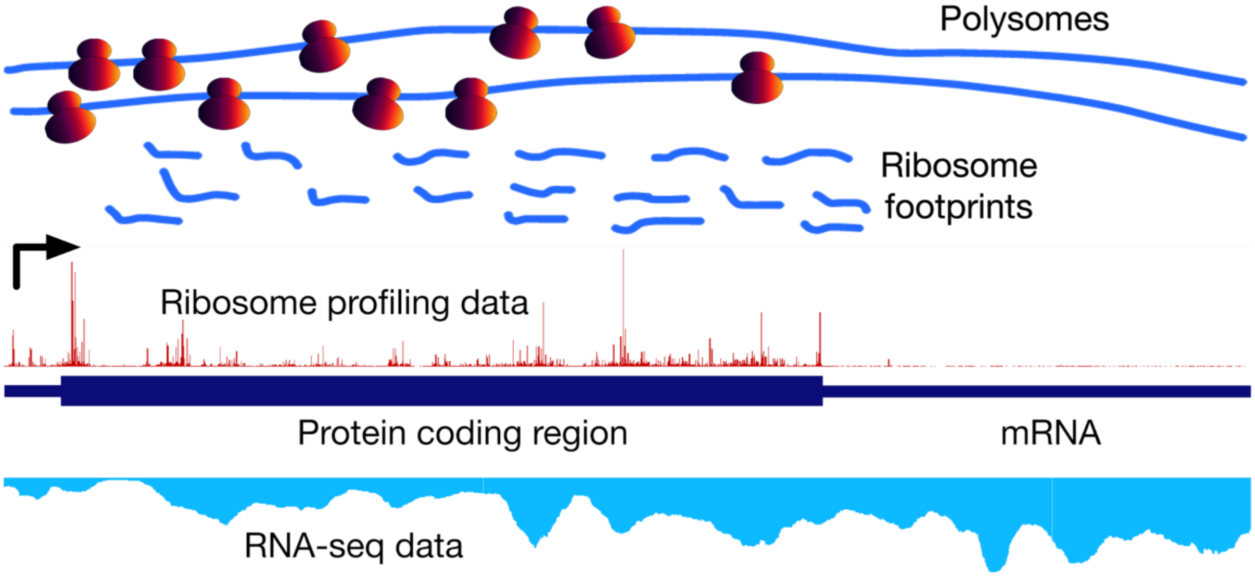


Figure 1.3 | **An overview of Ribo-Seq and RNA-seq.** Ribo-Seq enables the identification of actively translated smORFs. During translation, ribosomes are moving on the transcript codon by codon, the sample is treated with protein synthesis inhibitor to block elongation. After nuclease digestion, ribosome protected fragments or footprints were purified and sequenced. Footprints were then mapped back to the reference genome in the protein coding region as shown in the red track. The signal indicates the ribosome density at each nucleotide on the transcript. RNA-Seq data as shown in the blue track has covered the whole mRNA.

Ribosome profiling studies in a wide variety of species including flies, zebrafish, mice and humans and cell cultures have revealed that translation occurs in a pervasive manner (Ingolia et al., 2011; Lee et al., 2012; Stern-Ginossar et al., 2012; Dunn et al., 2013; Aspden et al., 2014; Bazzini et al., 2014; Juntawong et al., 2014; Smith et al., 2014; Vasquez et al., 2014; Ji et al., 2015; Fields et al., 2015; Johnstone et al., 2016). Ribosome footprints were detected in lncRNAs, in the upstream and downstream regions, and even overlapping the CDS of annotated coding transcripts. Moreover, studies revealed that translation from non-canonical start sites, including internal sites of CDSs and non-AUG start codons, is widespread.

Ribo-Seq can identify either initiating ribosomes (initiation Ribo-Seq) or elongating ribosomes (elongation Ribo-Seq) by using different inhibitors (Michel and Baranov, 2013). Initiation Ribo-Seq has been used to map translation initiation sites (TISs). This method uses compounds such as harringtonine (Ingolia et al., 2011), lactimidomycin (Lee et al., 2012) or puromycin (Clamer et al., 2018) to stop ribosomes at translation initiation sites, which indicate where active translation is taking place. Surprisingly complex organization of translation initiation sites in eukaryotes has been revealed by using this method including non-AUG sequences that initiate translation, the generation of N-terminally extended and truncated isoforms of annotated proteins as well as the translation of numerous open reading frames from host transcripts. Elongation Ribo-Seq uses translation elongation inhibitors, such as cycloheximide (Ingolia et al., 2009) or emetine (Ingolia et al., 2011) as well as no-drug by thermal freezing (Oh et al., 2011), to obtain ribosome footprints which are more likely to be the result in a translated ORF. In addition, this method also provides quantitative information of translation including translation efficiency (TE) and differential gene expression at the level of translation. Modification has been applied to the original Ribo-Seq, a method called Poly-Ribo-Seq enriches polysomes that are more likely to be actively translating mRNA into proteins. Poly-Ribo-Seq was successfully used to identify several smORFs in the *Drosophila* genome (Aspden et al., 2014).

While Ribosome profiling provides data on many putatively functional translated ORFs, including smORFs, ribosome occupancy does not automatically imply true coding potential or biological function at the peptide level. A sequencing read is not necessary an actively translated RNA fragment. A read could be obtained by a scanning ribosome which is a genuine RPF, or other RNA-binding proteins (Ingolia et al., 2014; Ji et al., 2016), or could represent technical or biological noise. Consequently, several ribosome profiling guided computational approaches and metrics based on the triplet periodicity pattern, sequence conservation, RPF-length distribution and other features - were devised and utilised to assess the coding potential of smORFs (**Table 1.2**). ORFScore is a metric to quantify the bias of the trinucleotide periodicity pattern of RPFs towards the first reading frame in a smORF (Bazzini et al., 2014). ORFScore determines whether RPFs are uniformly distributed in all three reading frames or preferentially accumulate in one frame. Using the periodicity pattern, several algorithms and pipelines have been developed including ORF-RATER (Fields et al., 2015), RiboORF (Ji et al., 2015), RiboTaper (Calviello et al., 2016), RP-BP (Malone et al., 2017), RiboCode (Xiao et al., 2018) and Ribotricer (Choudhary et al., 2020). The Ribosome Release Score (RRS) is a metric to detect the termination of translation at the stop codon of an ORF and has shown to robustly distinguish protein-coding transcripts from ncRNAs (Guttman et al., 2014). In addition to the approaches mentioned above, several databases have been developed to collect Ribo-Seq data and genome annotations derived from the data, including GWIPS-viz (Michel et al., 2013), TISdb (Wan and Qian, 2013), uORFdb (Wethmar et al., 2013), RPFdb (Xie et al., 2015), sORFs.org (Olexiouk et al., 2015), SmProt (Hao et al., 2017) and HRPDviewer (Wu et al., 2018). Those databases provide rich resources for the community, in the meantime, continued optimization of these methods and combination with other emerging technologies will enhance the power to identify functional smORFs.

Proteomics

Mass spectrometry (MS) is the gold standard for proteomics research, it is a powerful technique for direct detection and quantification of peptides and proteins. Mass spectrometry based peptidomics and proteomics have been implemented for micropeptides discovery in recent years, several micropeptides encoded by smORFs have been directly validated in *Drosophila melanogaster* (Aspden et al., 2014; Pueyo et al., 2016), zebrafish (Bazzini et al., 2014), human tissues and cell lines (Slavoff et al., 2013,2014; Na et al., 2018). MS is able to determine if polypeptides are, in fact, translated from a smORF comparing to Ribo-Seq, thus show direct evidence of protein-coding potential of the transcript. In MS experiments, peptide mapping is usually performed to identify proteins. Proteins will be digested and fragmented into peptides by using enzymes (e.g. trypsin), the molecular weights of the peptides will be accurately measured. These experimental masses of the peptides are compared to masses generated from an *in silico* digest of annotated proteins contained within a database. A protein will be confirmed if several of the masses match those for a specific protein in the database. Interestingly, in proteomics studies, there are currently many peptides are not matched to any protein, and one possible reason is that some of them may belong to micropeptides which have not yet been annotated. Proteogenomics is a research method that combines proteomics, genomics, and transcriptomics to improve the identification and validation of micropeptides (Castellana and Bafna, 2010; Woo et al., 2013; Branca et al., 2014). In a study carried out by Slavoff et al., peptidomics and RNA-Seq were combined to identify smORFs in human K562 cells (Slavoff et al., 2013). The authors first built *de novo* assembly of the K562 transcriptome using RNA-Seq, and then added this transcriptome on top of the annotated human transcriptome from public database RefSeq (Pruitt et al., 2013) to build a custom database for all possible peptides and proteins. They then performed liquid chromatography followed by tandem MS (LC-MS/MS) in a modified protocol to enrich small polypeptides and matched the results against the custom protein database. Through this strategy, 86 unannotated micropeptides were identified in human K562 cells.

Although MS-based proteomics has made great progress in micropeptide identification, there are still some difficulties to consider. In general, proteomics is limited in sensitivity and some micropeptides do not have suitable tryptic peptides (Martinez et al., 2019). At the same time, micropeptides may be short lived, their average abundance is low in samples, they are often lost in sample preparation, therefore absent from detection, they may also have tissue- and time-specific expression patterns, which further impedes their identification. Improved purification methods may be more efficient at extracting micropeptides, for example, Schwaid et al. described an affinity-based approach that is able to enrich cysteine-containing human micropeptides, they identified 16 novel micropeptides in the study (Schwaid et al., 2013).

The best strategy to date for detecting micropeptides is likely to combine computational and experimental approaches, and the methods described above have been successfully used to identify putative micropeptide that could have diverse biological functions.

1.4 Functional characterization of micropeptides

Advancement of technologies have led to discovery of hundreds or even thousands of potential novel micropeptides, however, the existence of a peptide does not imply it has a function. Each of these micropeptides needs to be studied independently. Experimental demonstration is essential to reveal their biological effects. So far, only a small number of micropeptides have been fully characterized and found to play important roles in fundamental biological processes (**Table 1.1**).

Ideally, an antibody against a micropeptide can be generated and validated to demonstrate its specificity, however there might be lack of available antibodies as well as means to generate custom antibodies. Firstly, the small size of micropeptide provides limited choices for designing antibodies, secondly, the 3D structure of the micropeptides is unknown, it might limit the regions for epitope design. An additional concern is that techniques that make use of antibodies, e.g. Western blot, if a micropeptide is expressed at a low level, the antibody may not be sufficient to generate strong enough signals for detection.

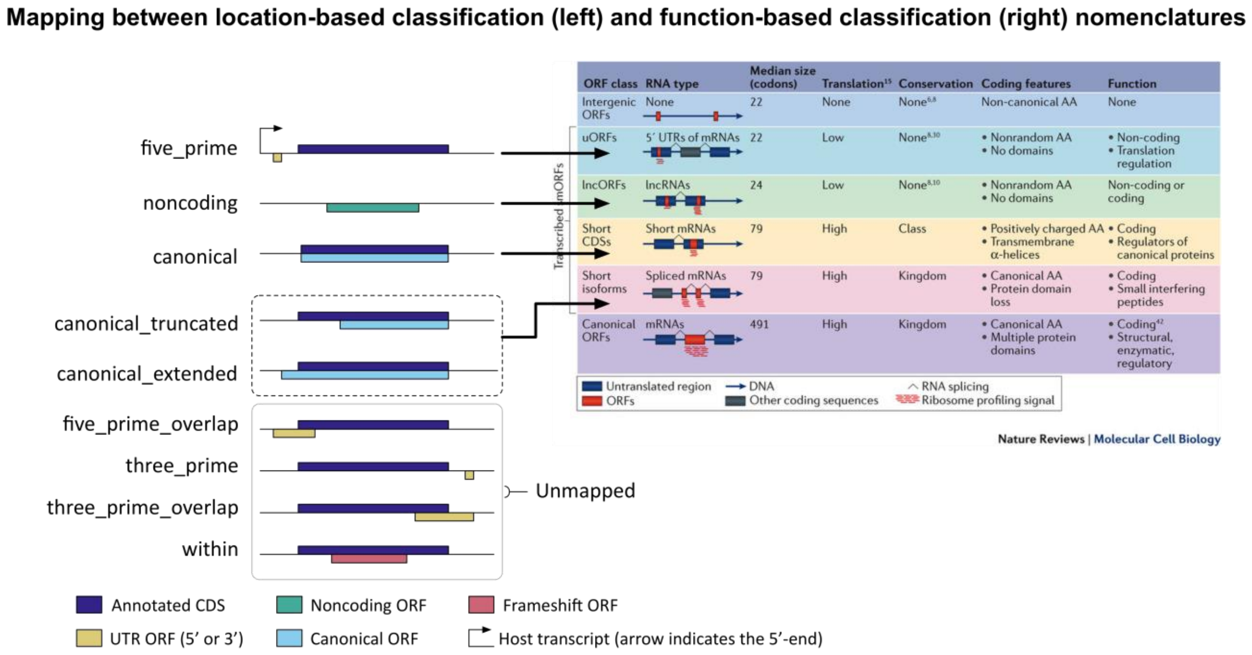
In addition to antibody-based validation of micropeptides, the coding potential of smORFs can be assessed by *in vitro* translation assays (https://www.thermofisher.com/uk/en/home/references/ambion-tech-support/large-scale-transcription/general-articles/the-basics-in-vitro-translation.html). The full-length cDNA of a smORF is cloned into a vector containing a phage polymerase promoter, and then expression of the construct is evaluated using a cell-free protein-synthesizing system in the presence of ^{35}S-methionine. The protein products are analysed by gel electrophoresis and autoradiography is performed to visualize the synthesis of a ^{35}S-labeled micropeptide. Introducing a frame-shift mutation in the smORF and subsequently the predicted peptide is not produced will strengthen the results. This is a valuable method to screen potential candidates, however the results should be interpreted with caution, as it is possible that the smORF can be translated *in vitro* but not *in vivo* (Anderson et al., 2015,2016; van Heesch et al., 2019).

CRISPR-Cas9 mediated gene-editing strategies can be designed to insert an epitope tag into the endogenous locus of the micropeptide in-frame with the encoding smORF using homology-directed repair *in vitro* or *in vivo* (Ran et al., 2013). The method has been used to generate fusion proteins that can be detected by Western blot and provides evidence that the micropeptide host transcript is actively transcribed from its native chromosome and translated into stable peptides (Galindo et al., 2007; Anderson et al., 2015; Matsumoto et al., 2017). The position of the knock-in tag (N-, C-terminal or internal) as well as the size and biochemical properties of the micropeptide are critical factors to consider when designing the experiment and these modifications have potential to change the biochemical properties of the micropeptide.

The coding potential of micropeptides can be demonstrated using the methods mentioned above, the next step is to find their biological relevance. Several functionally characterized micropeptides have been shown to engage with, or modulate, larger proteins or protein complexes; therefore, the key to elucidating their function often lies in identifying their interacting proteins. Functional proteomics has been successfully employed to identify binding partners of candidate micropeptides (Matsumoto et al., 2017; D'Lima et al., 2017; van Heesch et al., 2019) For example, the biological significance of a novel micropeptide named NBDY or NoBody (non-annotated P-body dissociating polypeptide) was characterized by performing immunoprecipitation and MS analysis (IP-MS) on the co-precipitated proteins (D'Lima et al., 2017), the researchers found NoBody is a component of the mRNA decapping protein complex cross-linking to EDC4 (enhancer of mRNA decapping 4). The mRNA decapping complex removes the 5′ cap from mRNAs to promote 5′-3′ decay. Molecular components of this pathway localize to p-bodies. Manipulation of NoBody expression is anti-corelated with the P-body number. NoBody regulates the P-body number in cells by interacting with decapping proteins. Even though finding interacting proteins of micropeptides has been shown a useful method, it is not always the case that binding means function; robust functional validation will be needed after the binding partners of micropeptides are identified.

1.5 smORF categories

Small ORFs can be present at different positions on their host transcripts, relative to other longer and usually annotated ORFs, including the annotated 5’-untranslated region (5’UTR), 3’-untranslated region (3’UTR), or the CDS of an mRNA. Additionally, they may be found in ncRNAs or transcripts annotated as pseudogenes. Based on this smORFs have been classified by their location (**Figure 1.4** and **Table 1.3**) (Ji et al., 2015; Fields et al., 2015; Delcourt et al., 2017).



Mapping between location-based classification (left) and function-based classification (right) nomenclatures

ORF class	RNA type	Median size (codons)	Translation[15]	Conservation	Coding features	Function
Intergenic ORFs	None	22	None	None[6,8]	Non-canonical AA	None
uORFs	5′ UTRs of mRNAs	22	Low	None[8,10]	• Nonrandom AA • No domains	• Non-coding • Translation regulation
lncORFs	lncRNAs	24	Low	None[8,10]	• Nonrandom AA • No domains	Non-coding or coding
Short CDSs	Short mRNAs	79	High	Class	• Positively charged AA • Transmembrane α-helices	• Coding • Regulators of canonical proteins
Short isoforms	Spliced mRNAs	79	High	Kingdom	• Canonical AA • Protein domain loss	• Coding • Small interfering peptides
Canonical ORFs	mRNAs	491	High	Kingdom	• Canonical AA • Multiple protein domains	• Coding[42] • Structural, enzymatic, regulatory

Nature Reviews | Molecular Cell Biology

Figure 1.4 | **smORF classification.** smORFs can be present at different positions on their host transcripts. They will be classified by their location relative to the transcript (left). Another way is a functionally relevant classification based on transcript type, size, conservation, rate of translation (number of amino acid residues per second), peptide structure properties and function (right, adapted from Couso and Patraquim, 2017). There is direct mapping between these two types of classification for canonical, extended, truncated and noncoding smORFs, but not for overlapping uORF, dORF, overlapping dORF and within (internal) smORFs.

Class (location-based)	**Description**
canonical	an ORF which exactly coincides with an annotated CDS
canonical_extended or extended	an ORF starts upstream of an annotated CDS and has the same stop codon as the CDS
canonical_truncated or truncated	an ORF starts downstream of an annotated CDS, have the same stop codon as CDS
five_prime or uORF	an ORF which is completely in the annotated 5’UTR of a protein-coding transcript and does not overlap the annotated CDS
five_prime_overlap or overlapping uORF (ouORF)	an ORF in the annotated 5’UTR of a protein-coding transcript but which overlaps the

	annotated CDS
three_prime or dORF	an ORF in the annotated 3’UTR of a protein-coding transcript and does not overlap the annotated CDS
three_prime_overlap or overlapping dORF (odORF)	an ORF in the annotated 3’UTR of a protein-coding transcript but which overlaps the annotated CDS
within or internal	an ORF in the interior of an annotated CDS, but in a different frame relative the CDS
noncoding or ncORF	an ORF from a transcript annotated as noncoding, such as a lncRNA or pseudogene

Table 1.3 | **smORF classification.** Table shows the smORF classes which is based on the relative location to their host transcripts, and the description of each class.

A functionally relevant classification of smORFs was introduced by linking their sequence to biochemical properties and molecular functions (Couso and Patraquim, 2017). It was proposed to classify smORFs based on distinctive transcript organization, size, conservation, mode of translation, amino acid usage and peptide structure properties (**Figure 1.4**). The translation products of smORFs are micropeptides. Micropeptides will be categorised to the annotated and the non-annotated. Among the annotated, some are having known biological function, some are having unknown function. For the non-annotated ones, we do not know their function yet and we can not rule out the possibility that some of them do not have function (**Figure 1.5**).

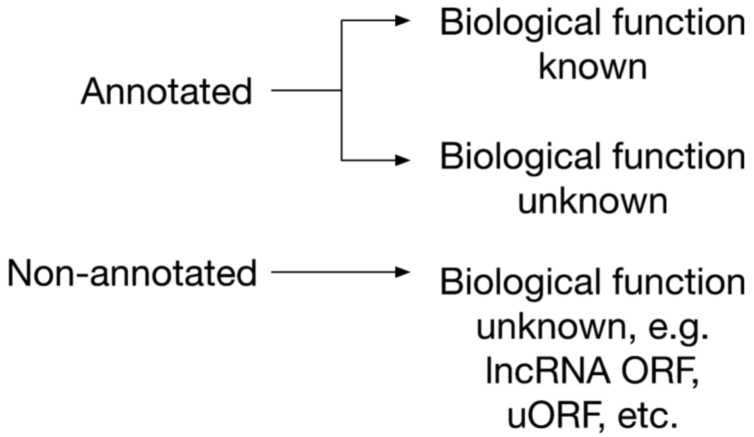

Annotated
Biological function known
Biological function unknown
Non-annotated
Biological function unknown, e.g. lncRNA ORF, uORF, etc.

Figure 1.5 | **Micropeptide categories.** Micropeptides can be grouped as annotated and non-annotated. Annotated micropeptides have known or unknown biological functions, and non-annotated have not yet been identified.

Canonical smORFs

Canonical smORFs are annotated ORFs of 100 codons and fewer. We divided canonical smORFs into “short CDS” and “short isoforms”. Short CDSs are located on monocistronic transcripts with higher probability; and their host transcripts are structurally shorter and simpler compared with canonical mRNAs (Couso and Patraquim, 2017). They appear translated as frequently and as strongly as canonical longer proteins sand appear to be conserved. There are hundreds of putative short CDSs in flies, mice and humans, but only a small fraction has been functionally characterized, the examples suggest that they have membrane-related functions as regulators of canonical proteins. short isoforms are generated by alternative splicing of canonical protein-coding gene. Annotated short isoform amino acid sequences are closer to canonical proteins as expected, but alternative splicing can result in loss of protein domains. Short isoforms have the potential for functions related to their canonical protein paralogues if the functional domains retain.

Upstream ORFs

The presence of smORFs within the 5’ untranslated regions (UTRs) of mRNAs is common. They were referred to as upstream ORFs or uORFs. They were noted in the first systematic survey of mRNA sequences (Kozak, 1987). uORFs act to attenuate the main ORF (usually a CDS in the downstream of a protein-coding transcript) translation in an inhibitory manner. In the standard scanning model (Kozak, 1989), translation usually initiates via 5’-cap of the mRNA and scan from 5’ to 3’ until the first initiation codon AUG is recognized. Ribosomal reinitiation efficiency of downstream CDS after translation of an uORF is constrained partly by the length, sequence arrangement, structural features of uORF, and intercistronic distance (Luukkonen et al., 1995; Kozak et al., 2001; Jackson et al., 2010). Re-initiation efficiency decreases quite abruptly with increasing length of the uORF (Luukkonen et al., 1995), or if the uORF includes stable RNA secondary structures that cause pausing of elongation (Kozak et al., 2001). It suggests it is the time taken to translate the uORF rather than the length that is crucial; translation initiation factors (e.g. eIF4F and eIF4B) will not be in place if the uORF translation takes longer to complete (Jackson et al., 2010). In the simplest case, the uORF peptides do not appear to participate, but the reduction in efficiency of downstream CDS translation could be an important regulatory feature (Morris and Geballe, 2000; Iacono et al., 2005). Conservation of position and length, but not sequence, of an uORF could be taken as an indication that its translation is important (Crowe et al., 2006). Translation of uORFs has been reported in several studies (Wang and Rothnagel, 2004; Calvo et al., 2009; Fritsch et al., 2012; Ji et al., 2015). In general, the process of uORF translation might be important, but the encoded peptide itself might not be functional (Zhang et al., 2019). However, a small fraction of uORF-encoded micropeptides has been confirmed by mass spectrometry (Oyama et al., 2004, 2007; Slavoff et al., 2013; Vanderperre et al., 2013; Andrews and Rothnagel, 2014; Johnstone et al., 2016), and for an even smaller fraction their subcellular localization, protein interactions and cellular function have been revealed (Jousse et al., 2001; Diba et al., 2001, Akimoto et al., 2013, Chen et al., 2020). One study shows that some uORF-encoded peptides formed complexes with the proteins encoded by the corresponding main ORFs (Chen et al., 2020), however the functional importance of such interactions remains to be tested. Another example shows that there is evidence to support the hypothesis that uORF peptides have physical interactions with ribosome complex to cause it to pause or disassociate from the transcript (Jousse et al., 2001). The sequence conservation of uORFs is an import feature to take into account during the identification of uORF-encoded peptides.

Downstream smORFs

In contrast to the study of 5’-UTR, the study of 3’-UTR has attracted little attention with respect to identifying and characterizing downstream smORFs or dORFs because they were considered not to be translated nor indeed translatable (Ingolia et al., 2011). However, as most 3’-UTR sequences are generally much longer than 5’-UTR (Crowe et al., 2006; Mercer et al., 2010), they could be expected to contain more smORFs. There is not yet any characterized dORF.

smORFs in non-coding RNAs

Non-coding ORFs (ncORFs) are smORFs that are found in annotated lncRNAs and pseudogenes. Tens of thousands of lncRNAs and transcripts of unknown function (TUFs) (Carninci et al., 2005; Willingham et al., 2006; ENCODE Project Consortium, 2007; Kapranov et al., 2007) are identified by genome-wide analysis. By definition, non-coding RNAs are not translated into protein. However, annotated lncRNAs have been predicted from their sequences to contain six smORFs on average (Couso and Patraquim, 2017). Recent studies have suggested that ncRNAs represent the greatest source for smORFs, which were previously overlooked because of their small size and the lack of evidence for "codingness" (Frith et al., 2006; Cohen, 2014; Pauli et al., 2015). Several cases of RNAs initially classified as long non-coding have been shown to actually encode and translate peptides with biomedically important functions in development and physiology, and to be conserved (**Table 1.1**).

1.6 Importance of smORFs to health and disease

In recent years, studies have indicated a diverse range of functions for smORF-encoded micropeptides. These include muscle regeneration, DNA replication, phagocytosis, metabolism and cancer. For example, micropeptide Myoregulin specifically expressed in skeletal muscle. Loss of function studies revealed the importance of Myoregulin in vivo, as mice lacking this micropeptide had increased endurance when compared to their WT counterparts. These studies reveal new research directions to the specific regulation of contraction in different muscles and muscle types and may prove important in the future development of therapeutic approaches to muscle diseases or aging. Another example is the identification of the smORF-encoding gene Boymaw, which is linked to an inherited form of schizophrenia (Ji et al., 2015). Boymaw activity affects rRNA expression and protein translation and is found at high levels in the post-mortem brains of people with neuropsychiatric diseases. Interestingly, the Boymaw micropeptide also localizes to mitochondria, and in flies, both mitochondrial localisation and putative electron transport functions appeared as favoured amongst translated micropeptides (Aspden et al., 2014). These highly interesting findings reveal the potential for micropeptides to regulate mitochondrial-based physiology. In addition, biochemical studies demonstrate that micropeptides utilize short sequences (usually 2–4 amino acids) (Arnoult et al., 2017) to bind to more massive protein complexes to regulate biology. Interactions that utilize short peptide interactions are amenable for small molecule inhibition, and, therefore, microprotein-protein interactions will reveal new druggable targets for medicine. These examples indicate that micropeptides are essential for cell functions and could be used to develop new therapeutics (Rathore et al., 2018).

1.7 Research objectives

As of August 2019, UniProtKB (UniProt Consortium, 2018) listed 1,987,752 entries (56,792 manually annotated and reviewed) for possible peptides and small proteins of less than 100 AA in all organisms, of which 37,841 are in human (748 reviewed) and 13,585 are in mouse (484 reviewed). Currently experimental evidence shows that, only approximately 1.2% of smORFs are expressed. However, even this small percentage of functional smORFs could theoretically produce tens of thousands of as yet uncharacterized peptides. Even if only a small proportion of these peptides have biological activity, we could be missing hundreds of peptides that could shed light on many aspects of biology and medicine. Thus, micropeptides offer an area of significant interest that currently is largely unexplored.

In the immune system, we propose important micropeptides are yet to be discovered. We are interested to find out how widespread micropeptides are in the immune system and what their functional roles might be. To answer these questions, a search for this class of micropeptides will be undertaken. Furthermore, identifying their functions will be essential and potentially lead to useful applications.

The identification of novel biologically active micropeptides in the immune system will be an important discovery, we are highly motivated to carry out this project. We aim to address the questions proposed above by firstly taking an *in-silico* approach to predict novel functional smORFs from lymphocytes and then categorise smORFs and learn their properties bioinformatically, secondly validate the existence of the prediction and thirdly perform functional characterization.

Chapter Two

Materials and Methods

2.1 Datasets

In the Turner lab, we have successfully generated ribosome profiling libraries (see **Appendix B** for additional information about Ribo-Seq and RNA-Seq experiments). We were among the first to publish datasets of this kind using immune cells (Diaz-Muñoz et al., 2015; Tiedje et al., 2016). Ribosome profiling and RNA-Seq (total RNA sequencing) or mRNA-Seq (polyA-selected RNA sequencing) experiments have been carried out on mouse lymphocytes including (*ex vivo*) resting B cells; two independently generated datasets of LPS-activated B cells; stimulated naïve CD4+ T cells (see **Table 2.1** and Materials and Methods). In addition, a published time-course dataset from Elke Glasmacher's group of Th1 T cells re-stimulated with anti-CD3+anti-CD28 was used in our study (Davari et al., 2017).

Cell type	**Source**	**Treatment**	**Experiment type**	**Number of biological replicates (N)**	**Illumina sequencing run type**	**Sequencing read count**
B cell setup 1	Spleen	Resting	Ribo-Seq	4	50bp Single-End	29,776,606
						19,730,128
						28,869,948
						29,399,992
			mRNA-Seq	4	100bp Single-End	39,791,941
						30,347,645
						38,143,451
						41,826,586
		LPS+IL-4-activated (48H)	Ribo-seq	5	50bp Single-End	27,802,753
						29,904,842
						24,847,188
						24,521,162
						27,010,038
			mRNA-Seq	4	100bp Single-End	38,437,221
						46,565,378
						35,764,459
						36,383,343
B cell setup 2	Lymph nodes	LPS+IL-4+IL-5-activated (48H)	Ribo-seq	5	50bp Single-End	43,527,344
						34,062,948
						38,992,152
						44,782,997
						37,566,709
			RNA-Seq	5	100bp	44,896,153

					Single-End	45,275,501
						69,059,677
						54,317,942
						25,724,660
CD4$^+$ T cell	Spleen, peripheral and mesenteric lymph nodes	anti-CD3/CD28 stimulated (24H)	Ribo-Seq	5	50bp Paired-End	46,252,019
						101,717,087
						28,346,186
						35,123,808
						32,225,196
			mRNA-Seq	3	100bp Single-End	37,223,659
						47,677,540
						44,258,974

Table 2.1 | **Sample summary.**

There are two independent experimental setups for B cells. For the first setup, purified B cells from spleen are a mixed collection of follicular and marginal zone B-2 cells. At resting status, cells from 4 biological replicates (mice) were processed individually for ribosome profiling libraries, and 4 different biological replicates were processed for mRNA-seq. To induce strong proliferation, B cells were treated with mitogen LPS (Lipopolysaccharide) to be stimulated for 48 hours. At 40 hours, stimulated cells went into their first cycle of cell division. At 48 hours, most cells had divided once, and some started second division. As in this setup, experiment was not designed for paired Ribo-Seq and RNA-Seq, 5 biological replicates were processed for ribosome profiling libraries, and 4 biological replicates from a different group of mice were processed for mRNA-Seq. For the second setup, B cells are from lymph nodes, they are mainly follicular B-2 cells. Cells were treated with LPS plus interleukin IL-4 and IL-5 to be stimulated for strong proliferation for 48 hours. There is no major difference between LPS treated only cells and LPS+IL-4+IL-5 treated cells at 48 hours. The difference will appear at a later stage in terms of class switch recombination. Five biological replicates were processed. Ribosome profiling and RNA-Seq libraries were prepared from the same sample. For the stimulated naïve CD4$^+$ T cell experiment, the purified naïve CD4$^+$ T cells are from multiple sources, including spleen, peripheral and mesenteric lymph nodes, because they are scarce. In addition, we have downloaded a time-course dataset of Th1 T cells re-stimulated with anti-CD3+anti-CD28 (Davari et al., 2017). We extended our prediction from mouse to human to address questions such as whether predicted smORFs are also conserved in human and are there human specific smORFs. We collected Ribo-Seq and RNA-Seq data of human lymphoma cell lines (Activated

B-cell (ABC)-Diffuse large B-cell lymphoma (DLBCL), germinal center B-cell (GCB)-DLBCL and Burkitt's lymphoma, 29 cell lines in total) and human primary B cells from collaborators (unpublished).

Ribo-Seq and RNA-Seq assays were prepared with ARTseq^{TM} Ribosome Profiling Kit-Mammalian (Epicentre, Illumina) (see Appendix B). Ribosome profiling is commonly performed using cycloheximide (CHX) to stall elongating ribosomes on mRNA. A different protocol enables the identification of alternative start codons through drug treatment using harringtonine (HARR) or lactimidomycin (LTM) that immobilises initiating ribosomes at the translation initiation site (TIS). In our experimental setups, elongation was targeted using protein synthesis inhibitor CHX.

Datasets have been submitted or are in the process of submission to Gene Expression Omnibus (GEO) database **(Table 2.2).**

Cell type	**RNA-Seq**	**Ribo-Seq**
B cell setup 1	GSE62129	GSE62134
B cell setup 2	Prepare for submission	Prepare for submission
CD4^{+} T cell	Prepare for submission	Prepare for submission
Th1 reactivation	GSE83351	GSE83351

Table 2.2 | **Sequencing data GEO primary accession codes.**

2.2 Reference genome, transcriptome and annotation

GENCODE (Harrow et al., 2012) reference genome sequences (mouse GRCm38/mm10 and human GRCh38/hg38) are downloaded from the GENCODE website (**Table 2.3**). Transcriptome sequences and gene annotation are used to search for putative ORFs, they are also downloaded from the same GENCODE source, note that the transcriptome sequences are cDNA sequences. Ribosomal RNA (rRNA) and Transfer RNA (tRNA) content in Ribo-Seq library need to be removed by mapping the sequencing reads to rRNA and tRNA sequences. tRNA sequences are downloaded from UCSC Table Browser (Karolchik et al., 2004). rRNA sequences are downloaded from GENCODE (version M20, we have also tested M13 and M15) as well as published studies (Bazzini et al., 2014; Fields et al., 2015). The transcriptome is defined as the collection of all transcripts on the reference chromosomes. GENCODE Transcript biotypes are defined here - https://www.gencodegenes.org/pages/biotypes.html. In our pipeline, we remove the following biotypes:

- IG_* and TR_* (Immunoglobulin variable chain and T-cell receptor genes)
- miRNA
- misc_RNA
- Mt_rRNA and Mt_tRNA
- rRNA and ribozyme
- scaRNA, scRNA, snoRNA, snRNA and sRNA
- nonsense_mediated_decay
- Non_stop_decay

File	**Type**	**Region**	**Source**
Genome sequence, primary assembly	Nucleotide sequence of the GRCm38 primary genome assembly (Fasta format)	PRI (reference chromosomes and scaffolds)	GENCODE
Transcript sequences	Nucleotide sequences of all transcripts (Fasta format)	CHR (reference chromosomes only)	GENCODE
Protein-coding transcript sequences	Nucleotide sequences of coding transcripts (Fasta format)	CHR	GENCODE
LncRNA transcript sequences	Nucleotide sequences of lncRNA transcripts (Fasta format)	CHR	GENCODE
Comprehensive gene	The main annotation	CHR	GENCODE

annotation	file (GTF and GFF format)		
LncRNA gene annotation	comprehensive gene annotation of lncRNA genes (GTF and GFF format)	CHR	GENCODE
tRNA sequences	Nucleotide sequences of tRNA genes predicted by UCSC using tRNAscan-SE	CHR	UCSC Table Browser
rRNA sequences	Nucleotide sequences of rRNA. In Ensembl Biomart, restrict search Gene biotype as rRNA will give the organism specific list of rRNA regions in the genome.	CHR	Ensembl Biomart
Small RNA sequences	Nucleotide sequences of snRNA, snoRNA, misc_RNA and miRNA	CHR	GENCODE

Table 2.3 | **List of public sequences and annotation files used in the pipeline.** Reference genome and transcriptome sequences, gene annotation (mouse and human) are from GENCODE. Ribosomal RNA (rRNA) and Transfer RNA (tRNA) sequences are from UCSC Table Browser. rRNA sequences are from Ensembl.

2.3 Identifying putative smORFs

Using the nucleotide sequences of all transcripts downloaded from GENCODE (release M13) (Frankish et al., 2018) as a reference, we searched for ORFs that begin with a start codon (XUG) and end with a stop codon, with no intervening stop codon, in all three reading frames for each transcript. All ORFs that have 100 codons or fewer were designated putative smORFs.

2.4 Sequencing data processing

Quality control: Raw sequencing data from RNA-Seq and Ribo-Seq was demultiplexed, adaptor trimmed with Trim Galore v0.4.5 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore), quality checked with FastQC v 0.11.8 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc).

Contaminant removal: Ribo-Seq reads were aligned to mouse rRNA or tRNA sequences using Bowtie v1.2.2 (Langmead et al., 2009), the reads that were not aligned were kept for alignment to the reference genome.

Alignment to reference genome: The reads were mapped to the GRCm38/mm10 reference genome using the STAR aligner v2.5.2a (Dobin et al., 2013). The aligner only reports uniquely mapped reads. The following shows an example command, and parameters are in bold:

```
STAR --runThreadN $THREAD \
     --genomeDir $REFGENOMESTAR \
     --readFilesIn $OUTPATH/bowtie-contanminant-
removal/${NAME}_trimmed_unfiltered.fq.gz --readFilesCommand zcat \
     --outReadsUnmapped Fastx \
     --outFileNamePrefix $OUTPATH/star-genome/$NAME/ \
     --alignIntronMin $ALIGNINTRON_MIN \
     --alignIntronMax $ALIGNINTRON_MAX \
     --alignEndsType EndToEnd \
     --outFilterMismatchNmax $MISMATCH_MAX \
     --outFilterMismatchNoverLmax $MISMATCH_NOVERL_MAX \
     --outFilterType $FILTER_TYPE \
     --outFilterIntronMotifs RemoveNoncanonicalUnannotated \
     --outSAMattributes $SAM_ATTR \
     --outSAMtype BAM SortedByCoordinate \
     --outBAMsortingThreadN $THREAD
```

Transcript expression quantification: In each experiment, sequence alignments (in BAM format) of all biological replicates were combined for both RNA-Seq and Ribo-Seq. After that transcript expression was quantified using StringTie v1.3.6 (Pertea et al., 2015) in RPKM (Reads Per Kilobase per Million for single-end read) or FPKM (for paired-end read). From a given dataset, a minimal expression level was set to RPKM > 0.5 (Hart et al., 2013) to exclude non-expressed transcripts.

P-site offset determination: Majority of RPFs are having a length between 28-31 nucleotides (nt). P-site offsite defined as a distance from the 5’ end of an RPF to the P-site of the ribosome was estimated for each read length using plastid python library v0.4.8 (Dunn et al., 2016). We observed P-site offsets are 12 nt long for RPF in 28-31 nt in our experiments.

2.5 ORF discovery

RPF coverage: To filter ORFs which are insufficiently covered by reads, we calculated the proportion of codons being covered by RPFs. We consider a codon covered if there is a mapped RPF with the P-site aligned to nucleotide 1 of that codon. An ORF is discarded if the ratio of covered codons to the total number of codons in the ORF < 0.1 (Bazzini et al., 2014).

ORFScore: ORFScore was proposed by Bazzini and colleagues (Bazzini et al., 2014), I re-implemented the ORFScore algorithm in R. The ORFScore was then calculated as:

$$\text{ORFscore} = \log_2\left(\left(\sum_{i=1}^{3}\frac{\left(F_i - \overline{F}\right)^2}{\overline{F}}\right) + 1\right) \times \begin{cases} -1, \text{if } (F_1 < F_2) \cup (F_1 < F_3) \\ 1, \text{otherwise} \end{cases}$$

where F_n is the number of reads in reading frame n, $\overline{F}$ is the total number of reads across all three frames divided by 3. RPFs were counted at each position within an ORF, excluding the first and last coding codons. To filter out putative artefactual peaks, the most abundant read position was masked if reads aligning to that position comprised more than 70% of the total reads in the ORF. The ORFScore is a log-scaled chi-squared goodness of fit test statistic, the p-values associated with the test were adjusted using Benjamini-Hochberg FDR-controlling method and smORFs with ORFScore > 0 and adjusted p-value < 0.01 were retained.

Ribosome Release Score (RRS): Firstly, I defined the 3’UTRs of smORFs. For canonical smORFs, we used annotated 3’UTRs. For other classes of smORFs, their 3’UTRs were defined as the region between the stop codon and the next possible start codon in any frame. The RRS score is defined as the ratio of the two normalized ratios and calculated with the following equation: RRS = (RPKM_RF ORF/RPKM_RF 3’-UTR)/(RPKM_RNA ORF/RPKM_RNA 3’-UTR). Based on the original study, smORF with RRS > 5 is considered to be translated (Guttman et al., 2013).

Inside/outside read ratio: The ribosome footprints typically show precise positioning between the start and the stop codon of translated ORFs. Low density of footprints before start codons and after stop codons and high inside/outside ratio is expected. By considering read distribution of the nearest 3 upstream codons outside and the first 3 codons inside an ORF, we devised a feature called inside/outside read ratio (total RPF of inside codons/total RPF of outside codons) to assess whether genuine translation takes place. ORFs will retain if the ratio ≤ 1 (more reads mapping outside than inside).

2.6 Analysis of predicted smORFs

Translation efficiency (TE): A measure of the rate of translation for a given feature (e.g. the CDS of a mRNA or a smORF), obtained in ribosome profiling experiments. It was calculated as the base 2 logarithmic ratio of RPF expression (RPKM) over mRNA expression (RPKM).

Conservation of the amino acid sequences: To examine the conservation of smORF-encoded micropeptide sequences between species, we performed PhyloCSF (Lin et al., 2011), a likelihood-based method to analyse signatures of evolutionary conservation in multiple species sequence alignments. PhyloCSF assigned a score to each smORF based on conservation within 100 vertebrate species (https://github.com/mlin/PhyloCSF/wiki#available-phylogenies). For each smORF, the corresponding multiple species alignments were obtained from a publicly available whole genome multiple alignment using Galaxy “stitch gene blocks” tool (Blankenberg et al., 2011). smORFs were considered conserved if their PhyloCSF score was > 50 (Guttman et al., 2010), and weakly conserved if they had a PhyloCSF score > 0. PhyloCSF score = 0 indicates that there is no DNA sequence alignment cross species and PhyloCSF score < 0 is considered not conserved.

Gene ontology (GO) enrichment analysis: We used the g:Profiler server (Raudvere et al., 2019) to perform GO analysis in two unranked lists of genes mode. The background list comprised the combined expressed transcripts (RPKM > 0.5) of B and T cells. The target list contains the host transcripts of the smORFs. For the significance threshold, we chose the default option g:SCS threshold and the default value 0.05.

Secreted micropeptide prediction: I used the SignalP 4.1/5.0 server (Petersen et al., 2011; Armenteros et al., 2019) to predict signal peptides present at the N-terminus of the micropeptide amino acid sequences. I used default parameters. For selected candidates, we ran prediction of transmembrane helices using the TMHMM 2.0 Server (Krogh et al., 2001) (default parameters) to rule out transmembrane peptides.

2.7 Plasmid design and smORF cloning

We have designed a customized plasmid based on a commercial gene expression vector from Cyagen/VectorBuilder (**Figure 2.1A** and **Table 2.4**). We inserted a 3xFLAG-tag and a following stop codon to the vector. Unique BamHI and BglII sites were added around the tag, this adds the option to remove the tag if required (**Figure 2.1B**). A unique EcoRI site was added, so the smORF sequence without a stop codon can be cloned between EcoRI and BamHI. During protein expression, 3xFLAG-tag will be added to the C-terminus of the micropeptide. We validated the construct by Sanger sequencing to be certain the new inserts were correctly placed.

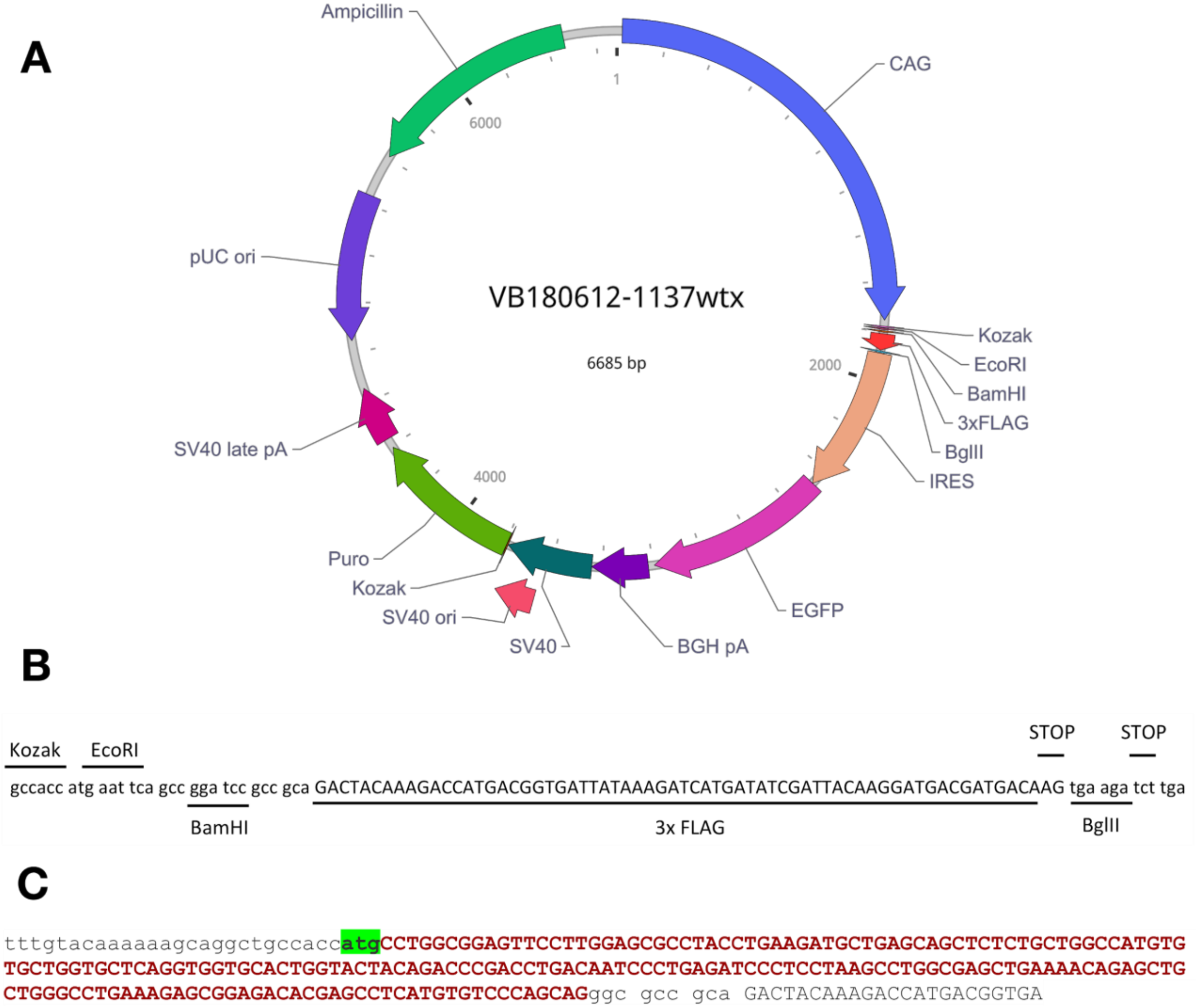
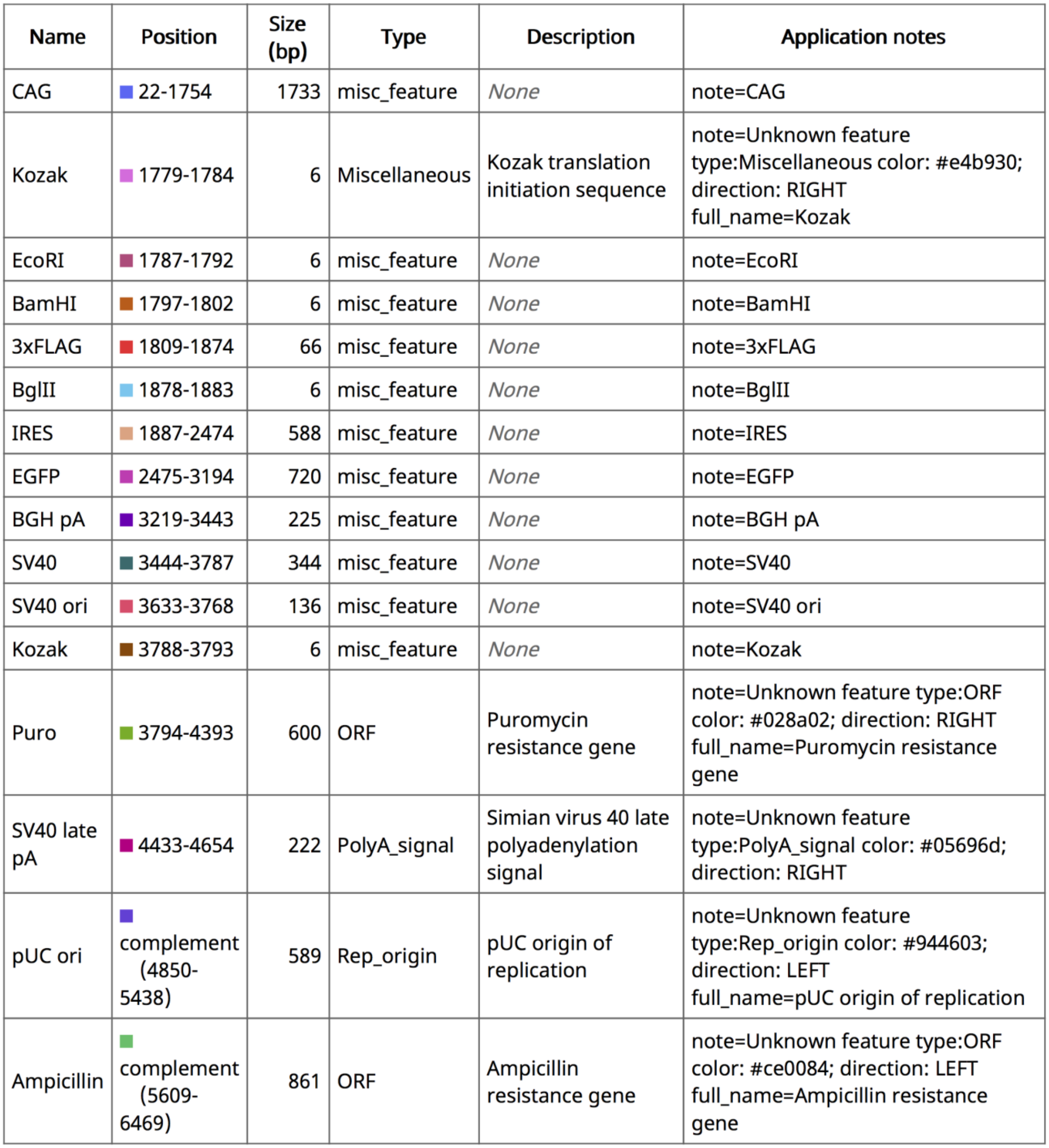

A
Ampicillin
CAG
1
6000
pUC ori
VB180612-1137wtx
6685 bp
2000
Kozak
EcoRI
BamHI
3xFLAG
BglII
IRES
SV40 late pA
4000
Puro
Kozak
SV40 ori
SV40
BGH pA
EGFP
B
STOP
STOP
Kozak
EcoRI
gccacc atg aat tca gcc gga tcc gcc gca GACTACAAAGACCATGACGGTGATTATAAAGATCATGATATCGATTACAAGGATGACGATGACAAG tga aga tct tga
BamHI
3x FLAG
BglII
C
tttgtacaaaaaagcaggctgccaccatgCCTGGCGGAGTTCCTTGGAGCGCCTACCTGAAGATGCTGAGCAGCTCTCTGCTGGCCATGTG
TGCTGGTGCTCAGGTGGTGCACTGGTACTACAGACCCGACCTGACAATCCCTGAGATCCCTCCTAAGCCTGGCGAGCTGAAAACAGAGCTG
CTGGGCCTGAAAGAGCGGAGACACGAGCCTCATGTGTCCCAGCAGggc gcc gca GACTACAAAGACCATGACGGTGA

Figure 2.1 | **Expression vector design.** (A) Vector map. (B) Sequence to show the three restriction sites and 3xFLAG tag. (C) Example of a smORF cDNA was optimized for codon usage (red, start codon highlighted in green) and overlapping ends were added for Gibson Assembly. Vector was designed together by Fengyuan Hu and Alexander Saveliev.

In order to have a stronger production of the micropeptides, smORFs sequences were optimized to improve codon usage in mouse and human cell lines using IDT Codon Optimization Tool. This tool was written using a codon sampling strategy in which the reading frame is recoded based on the frequencies of each codon's usage in the new organism (Robison, 2009). As an example, codon optimizations of sequences that will be expressed in human cell lines assign the phenylalanine codon UUU 46% and UUC 54% of the time. The optimized sequence then was added overlapping ends (**Figure 2.1C**) and synthesized by a commercial service (Integrated DNA Technologies). Then the synthetic DNAs were cloned into the vector by the Gibson Assembly method (Gibson et al., 2009).



Name	Position	Size (bp)	Type	Description	Application notes
CAG	■ 22-1754	1733	misc_feature	*None*	note=CAG

Kozak	■ 1779-1784	6	Miscellaneous	Kozak translation initiation sequence	note=Unknown feature type:Miscellaneous color: #e4b930; direction: RIGHT
					full_name=Kozak

EcoRI	■ 1787-1792	6	misc_feature	*None*	note=EcoRI
BamHI	■ 1797-1802	6	misc_feature	*None*	note=BamHI
3xFLAG	■ 1809-1874	66	misc_feature	*None*	note=3xFLAG
BglII	■ 1878-1883	6	misc_feature	*None*	note=BglII
IRES	■ 1887-2474	588	misc_feature	*None*	note=IRES
EGFP	■ 2475-3194	720	misc_feature	*None*	note=EGFP
BGH pA	■ 3219-3443	225	misc_feature	*None*	note=BGH pA
SV40	■ 3444-3787	344	misc_feature	*None*	note=SV40
SV40 ori	■ 3633-3768	136	misc_feature	*None*	note=SV40 ori
Kozak	■ 3788-3793	6	misc_feature	*None*	note=Kozak
Puro	■ 3794-4393	600	ORF	Puromycin resistance gene	note=Unknown feature type:ORF color: #028a02; direction: RIGHT full_name=Puromycin resistance gene
SV40 late pA	■ 4433-4654	222	PolyA_signal	Simian virus 40 late polyadenylation signal	note=Unknown feature type:PolyA_signal color: #05696d; direction: RIGHT
pUC ori	■ complement (4850-5438)	589	Rep_origin	pUC origin of replication	note=Unknown feature type:Rep_origin color: #944603; direction: LEFT full_name=pUC origin of replication

Ampicillin	■ complement (5609-6469)	861	ORF	Ampicillin resistance gene	note=Unknown feature type:ORF color: #ce0084; direction: LEFT
					full_name=Ampicillin resistance gene

Table 2.4 | **Vector components.**

Chapter Three

Computational Pipeline to Predict Actively Translated smORFs

3.1 Summary

In order to address the question “how widespread smORFs are in the immune system”, I planned to take an *in-silico* approach to discover novel functional smORFs from lymphocytes using next generation sequencing data. Here I describe a new analytical pipeline “ORFLine” that performs a comprehensive and systematic analyses of RNA-Seq and Ribo-Seq to identify actively translated smORFs **(Figure 3.1)**. In comparison to previously published pipelines, this pipeline is more stringent at smORF prediction. We have applied ORFLine to data from mouse B and T cells and discovered 5744 actively translated smORFs and their predicted translation products. smORFs were classified, and for each class, we performed analyses to look at their conservation, translation efficiency, and the biological processes linked to them. It has been shown in the UniProt database that a small subset of chemokines and majority of cytokines are between 101 and 200 AA long. With this in mind, we extended our analysis to candidate proteins of up to 200 AA in length and found evidence for translation of 945 such polypeptides. We further investigate whether the predicted micropeptides possess features of signal peptides which have a potential to be secreted and could act as immune regulators.

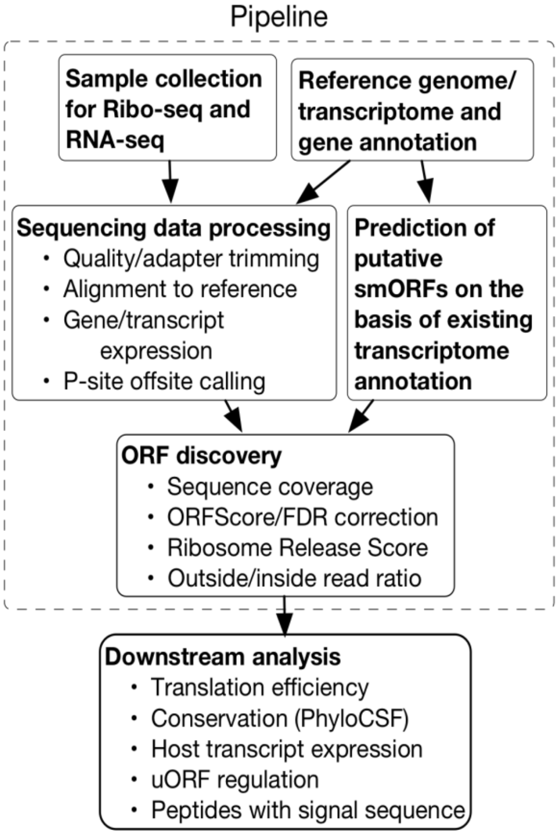

Pipeline
Sample collection for Ribo-seq and RNA-seq
Reference genome/ transcriptome and gene annotation
Sequencing data processing
• Quality/adapter trimming
• Alignment to reference
• Gene/transcript expression
• P-site offsite calling
Prediction of putative smORFs on the basis of existing transcriptome annotation
ORF discovery
• Sequence coverage
• ORFScore/FDR correction
• Ribosome Release Score
• Outside/inside read ratio
Downstream analysis
• Translation efficiency
• Conservation (PhyloCSF)
• Host transcript expression
• uORF regulation
• Peptides with signal sequence

Figure 3.1 | **Computational pipeline to identify translated smORFs.** Sequencing data for RNA-Seq and Ribosome profiling is processed and the reads mapped to the reference genome. In parallel, putative smORFs were predicted by scanning the annotated transcriptome. Several experimental metrics for each putative smORF were quantified and the smORFs exceeding a high confidence level for each metric were kept for downstream analysis.

3.2 Overview of the pipeline

The pipeline takes Ribo-Seq data and the paired RNA-Seq data, reference genome, transcriptome and gene annotation as input. The output of the pipeline is a list of predicted smORFs with genomic coordinates and classification. There are three main pipeline components to process the raw Illumina sequences and perform smORF prediction:

- Prediction of putative smORFs
- Sequencing data QC and processing
- ORF discovery

Prediction of putative smORFs and sequencing data processing are independent components and can be executed in parallel. ORF discovery component utilizes the output of previous two components as input (**Table 3.1**). The pipeline is built for general purpose and potentially applicable to data from any species. I have applied it to mouse and human data.

Component	**Step**	**Description**	**Cut-offs and rationales**	**Input**	**Output**
Putative smORF prediction	Computational prediction	Search for all theoretical smORFs *in silico*	Scan all annotated transcripts	Transcriptome sequences; user defined start codons (default: “AUG”, “TUG”, “CUG”, “GUG”)	List of genomic regions for predicted smORFs
Ribo-Seq	Quality and	Trim	Trim “N” bases	Ribo-Seq and	Trimmed

and RNA-Seq QC and processing	adapter trimming	sequencing adapter and low-quality bases	at the end of reads	RNA-Seq FASTQ files	FASTQ files
	Ribo-Seq contamination removal	Remove rRNA/tRNA reads	Align all reads to rRNA/tRNA sequences	Trimmed Ribo-Seq FASTQ files; rRNA/tRNA sequences	Contamination removed FASTQ files; rRNA/tRNA alignment BAM files
	Alignment to reference genome	Align reads to the genome	Unique mapped reads are kept; read length < 36 nt	Contamination removed FASTQ files	Genome alignment BAM files; unmapped FASTQ files
	Ribo-Seq P-site offset calling	Infer P-site offset for each read length	5’ end mapping rule	Genome alignment BAM files	List of P-site offset values for each read length
	Ribo-Seq read phasing estimation	Ribosome protected fragments (RPFs) show triplet periodicity	First reading frame shows greater read proportion (> 50%)	Genome alignment BAM files	Summary of read proportion for each reading frame
	Transcript expression estimation	Estimate FPKM value for each transcript	FPKM > 0.5 is considered expressed (Hart et al., 2013)	Genome alignment BAM files	List of FPKM values for each

					transcript
ORF discovery	Read length filter	Filter out read lengths that do not show triplet periodicity	Only keep read lengths that show triplet periodicity	Merged genome alignment BAM file; read phasing summary	Read length filtered BAM files
	Read count filter	Filter out smORF regions where there is no read aligned to	Read count > 0	Read length filtered BAM files; full list of predicted smORFs	Read count filtered smORFs
	RPF count filter	Filter out smORF regions where there is no RPF aligned to	RPF count > 0. Not all reads are RPFs.	Filtered BAM files; read count filtered smORFs	RPF count filtered smORFs
	Transcript expression filter	Filter out smORFs whose host transcripts are not expressed	Only keep smORFs whose host transcript FPKM > 0.5	Host transcript FPKM values; RPF count filtered smORFs	Transcript expression filtered smORFs
	Class assignment	Assign class to a smORF based on its relative location on its host transcript	Class is added as annotation, the number of smORFs are the same as last step	Transcript expression filtered smORFs	Class annotated smORFs
	ORFScore	Calculate	RPF coverage >	Class	ORFScore

	filter	ORFScore for each smORFs	0.1, ORFScore > 0 and adjusted p-value < 0.01	annotated smORFs	filtered smORFs
	Region filter	Test if a smORF is overlapped with a CDS	The estimated ratio (RPF $\text{count}_{\text{CDS}}$/RPF $\text{count}_{\text{smORF}}$) > 1	Read length filtered BAM files; ORFScore filtered smORFs	Region filtered smORFs
	Ribosome release score (RRS) filter	Calculated RRS	RRS > 5 (Guttman et al., 2013)	Read length filtered BAM files; RNA-Seq alignment BAMs; Region filtered smORFs	RRS filtered smORFs
	Nested filter	smORFs can possibly be nested in other smORFs (same stop code but different start codons)	the one with the maximum ORFScore is retained, otherwise a smORF with AUG start codon is retained	RRS filtered smORFs	Nested filtered smORFs
	Inside/outside (I/O) ratio filter	Translated regions in a Ribo-Seq data typically show higher read density outside start codon	inside/outside read ratio ≤ 1 (more reads mapping outside than inside)	Read length filtered BAM files; nested filtered smORFs	I/O ratio filtered smORFs

Table 3.1 | **Pipeline step summary.**

3.3 Prediction of putative smORFs

Given transcriptome sequences (see Materials and Methods 2.3), I exhaustively search for theoretical putative ORFs beginning with a start codon (“AUG”, “TUG”, “CUG”, “GUG”) and ending with a stop codon ("UAG", "UAA", "UGA") without an intervening stop codon in between in each of the three reading frames. We then remove ORFs that are not n*3 (n > 1) nucleotides long and keep the ones that are 100 codons or shorter in length as putative smORFs. The ORF coordinates are initially transcript coordinates and are converted to genomic coordinates given exon location information in the gene annotation (in GTF/GFF format), the output of the smORFs are their genomic coordinates and strands in BED format (https://genome.ucsc.edu/FAQ/FAQformat.html#format1). Each ORF will be assigned two different identifiers, one is called **RegionId**, the second is called **ORFId**. RegionId is created based on genomic coordinates, ORFId is created based on the transcript coordinates. An ORF has a unique genomic location, so RegionId is unique, but it may be from multiple transcripts (overlapping transcripts), so it may have multiple ORFIds **(Figure 3.2)**. This step is carried out only once and needs to be updated when transcriptome annotation is changed.

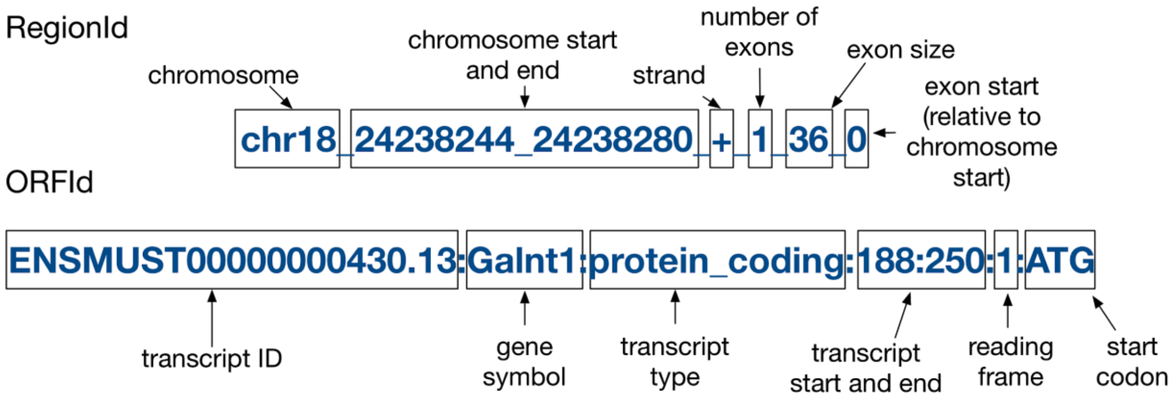


Figure 3.2 | **RegionId and ORFId explained.** RegionId is genomic-based, it indicts the unique location of a smORF on the genome. ORFId is transcript-based, it contains information regarding the smORF’s relative position on its host transcript.

3.4 Sequencing data QC and processing

Raw Illumina sequencing data is in FASTQ format (see Materials and Methods 2.2). I assessed the similarity between biological replicates by calculating their correlation (estimated transcript FPKM values from Ribo-Seq and RNA-Seq alignment BAM files) and showed that they are closely related (**Figure 3.3**). Then Illumina adapter sequences are trimmed off from the raw reads. The resultant reads are then mapped to a collection of rRNA/tRNA (see Materials and Methods 2.3) and small RNA sequences to filter out contaminants. The remaining reads are aligned to the reference genome (GRCm38). The alignment files (BAM format) will be the input for ORF calling steps.

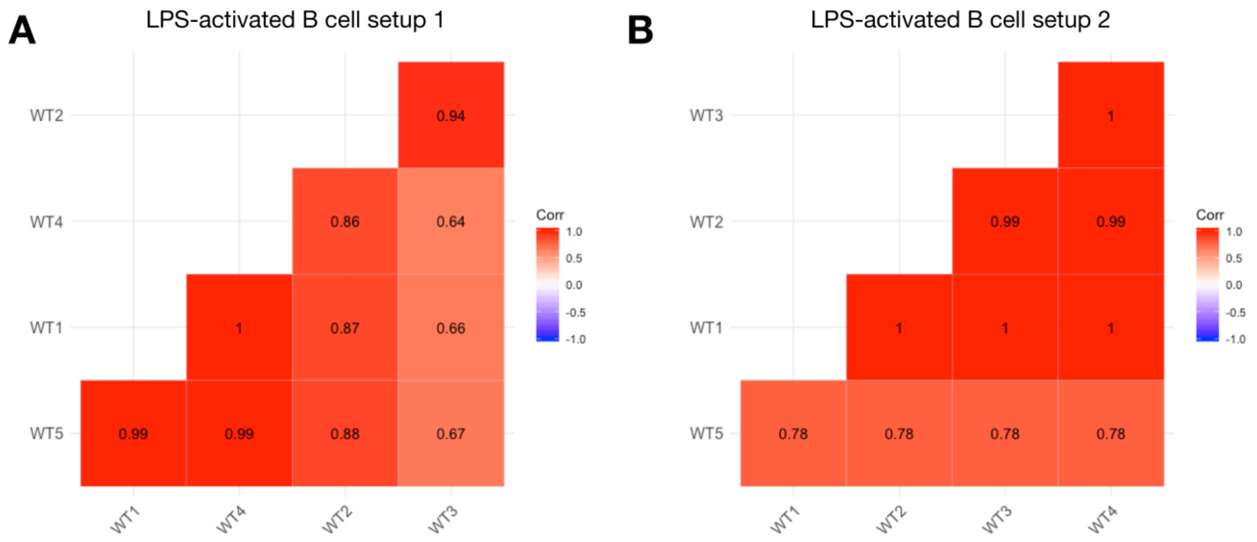


Figure 3.3 | **Ribo-Seq biological replicates correlation.** Example of LPS-activated B cell experiments (two setups, setup 1: N=5, setup 2: N=5). Replicate name is prefixed with WT. Correlation coefficients are added.

Adapter trimming

We trimmed sequencing adapters using TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and Cutadapt (Martin, 2011) and QC reporting programme FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality trimming programmes assume a loss of quality from the 3’ end. As long as there are good quality bases further into the read it will trim back the 3’ end until it gets to better quality sequence. I observed a lone poor quality “N” base (unknown quality) at the 5’ end of some of the reads (<0.2%). “N” bases were trimmed by Cutadapt which provides the flexibility of trimming low quality bases at 5’ end of the read.

According to the manufacturer’s specification (ARTseq^{TM} Ribosome Profiling Kit-Mammalian protocol), RPFs will be ~28-30 nt in length. We have kept trimmed reads that have a length between 25 and 35 nt, as they account for ~75% of the total reads on average.

Contaminant removal

rRNA is the most abundant unwanted data in a sequencing library. Depletion of rRNAs biochemically can remove 99.5% of them from the library. However, rRNAs are still highly dominant in a typical ribosome profiling sample. In addition to rRNA, a sample can have other contaminating sequences, such as tRNA. Due to the compact nature, size (~75 nt), and stable structure of tRNAs, RNase I digestion can cleave the individual tRNA molecules in half. This results in two fragments that are roughly similar in size to RPFs and can thus become a major contaminant in the samples. The use of sufficiently high levels of ribonuclease can overcome this problem, although a significant fraction of tRNA contamination can remain.

In order to remove and quantify rRNA/tRNA content or other contaminants in the sample, I use Bowtie 1 (Langmead et al., 2009) to align the trimmed reads against specific contaminant sequences assembled from a collection of rRNA, tRNA, Mt_rRNA and Mt_tRNA snRNA, snoRNA, misc_RNA and miRNA sequences.

FastQ screen

After the contaminant removal step, it is useful to confirm the composition of our sequencing libraries matches with mouse rather than other species. To do this, I sample a subset of our libraries (10,000 reads) and search them against a set of standard reference set including mouse, rat, human, rRNA, Phix, vectors and other model organisms using FastQ Screen (FastQ Screen allows the user to screen a library of sequences in FastQ format against a set of sequence databases so the user can see if the composition of the library matches with what he/she expects. https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/). I expect to see the reads would be mainly mapped to the mouse genome.

Sequence alignment to the reference genome

We use STAR aligner (Dobin et al., 2013) to align filtered reads (non-contaminant) to the reference genome. We only keep the uniquely mapping reads (mapping quality MAPQ = 255). For Ribo-Seq, parameters are tuned for reads that are shorter than 35 nt (see Materials and Methods).

Metagene Analysis

A metagene analysis is an average (typically median) of quantitative data over one or more genomic features/regions (e.g. genes or transcripts) aligned at some internal features (e.g. start codon). Metagene analysis reveals patterns across features that may not be obvious when looking at any individual feature. We firstly fetch vectors of quantitative data – the raw read counts surrounding start codon of each annotated CDS. Secondly, we normalize each vector to the same scale by dividing by the total number of aligned reads in a window (200 nt downstream of start codon). Then we align each vector at the start codon. Finally, we take the median of all the normalized vectors at each aligned nucleotide position (**Figure 3.4**), this is useful to estimate P-site offsite.

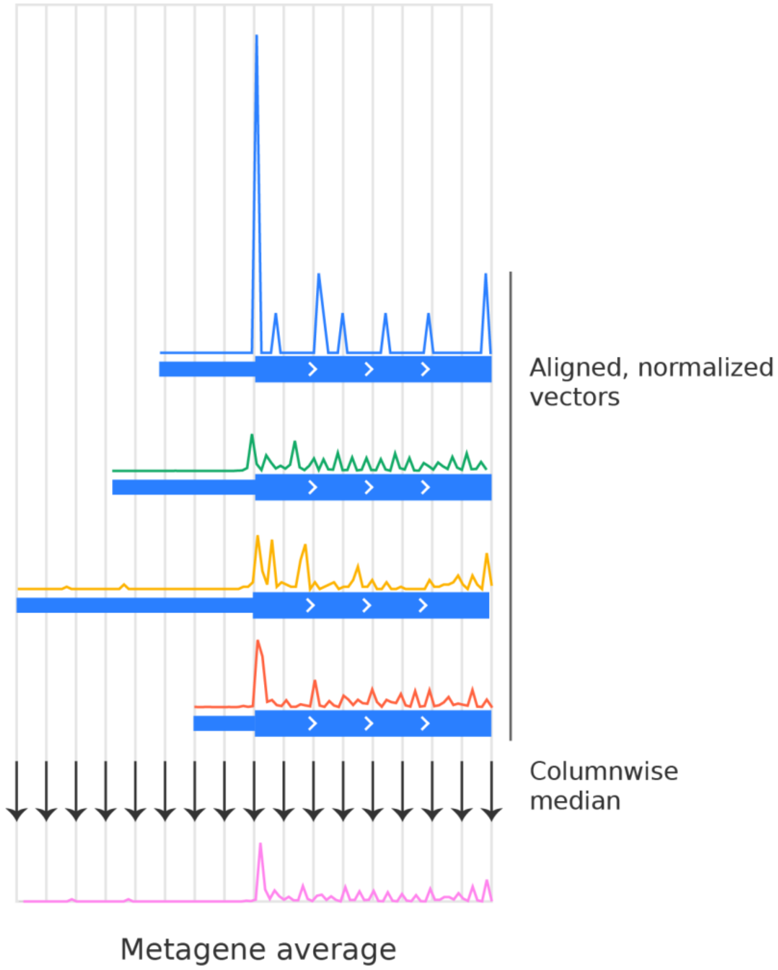


Figure 3.4 | **Metagene analysis of ribosome protected fragments.** A metagene plot is used as a summary statistic to visualize the distribution of ribosome protected fragments along the positions of a gene often starting/ending at the start/stop codon. This is useful for estimating P-site offsets. Blue boxes: protein-coding transcript models. Thick boxes: coding regions. Thin boxes: 5' UTRs. Above transcript models, normalized vectors of RPFs were shown. Final metagene average over the four transcripts shown at the bottom.

RPF mapping rules and P-site offsets calculation

When mapping RPFs to reference sequences, reads typically are mapped to their 5' or 3' ends, there are other mapping rules including variable 5' end mapping, Stratified variable 5' end mapping, entire or centre-weighted mapping (https://plastid.readthedocs.io/en/latest/concepts/mapping_rules.html). 5' end mapping has been applied in previous studies (Ingolia, et al., 2011; Bazzini et al., 2014), each read alignment is mapped at a fixed distance from its 5' end, where the distance is determined by the length of the read, we use this mapping in our study as well. The distance is called P-site offset. The P-site (P for peptidyl) is the ribosomal site most frequently occupied by peptidyl-tRNA, i.e. the tRNA carrying the growing peptide chain, it is also where peptide elongation starts. P-site covers start codon of a translatable ORF. Ribosome profiling reads are frequently mapped to their P-sites. The 5' P-site offset is the distance from the 5' end of the read to the ribosomal P-site (**Figure 3.5**).

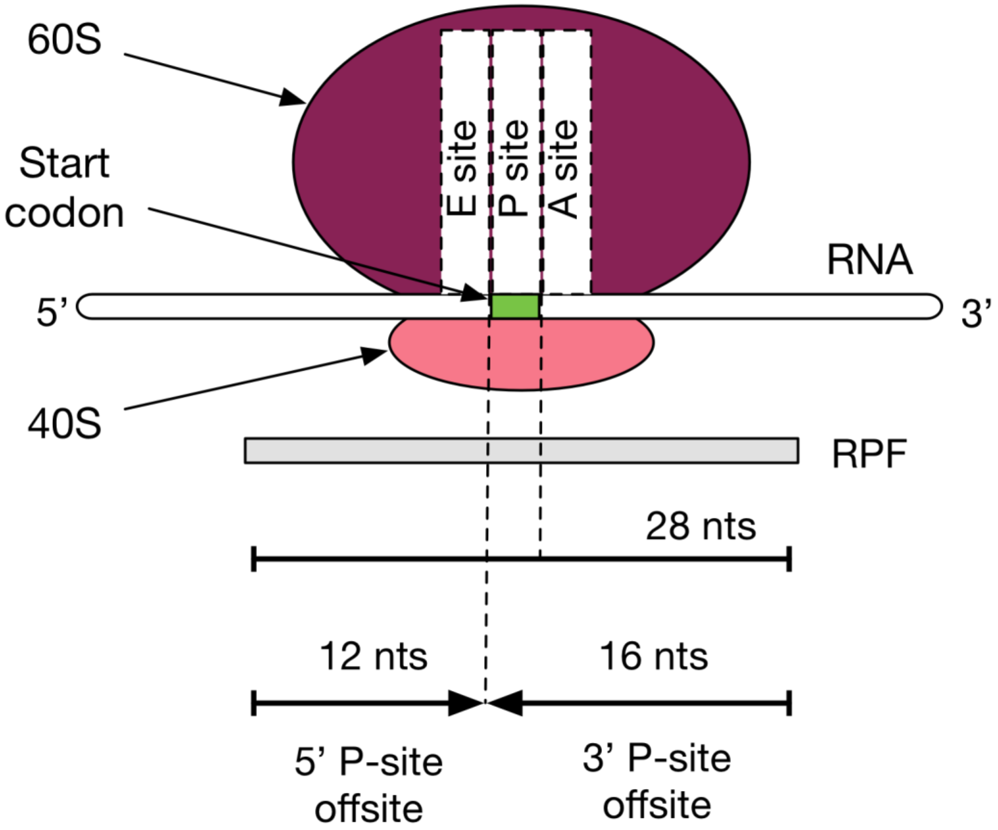


Figure 3.5 | **A ribosome containing a footprint after digestion.** This shows an example of a 28 nt long read with the 5' P-site offset 12 nt.

The strategy we used to determine P-site offset is as follows:

1. Separate footprints into groups based upon their lengths
2. For each length:
    a. Perform a metagene analysis at the start codon, in which the footprints are mapped to their 5' ends.

b. Measure the distance between the highest peak 5’ of the start codon and the start codon. This distance is the offset to use for reads of this length.

Read phasing in Ribo-Seq

Triplet periodicity or sub-codon phasing is a feature of ribosome profiling data. Ribosomes step three nucleotides along the mRNA in each cycle of peptide elongation and this physical process creates triplet periodicity which becomes observable when the reads are aligned to their P-site offsets (Ingolia et al., 2019).

This feature allows inference of the reading frame(s) in which a coding region is translated. This is also the principle of several ORF calling methods (**Table 1.2**). To calculate phasing proportion, we use read alignments and annotated canonical transcripts (with start codon, CDS and stop codon locations) as input. For each transcript and each read length, we counted the number of reads that are aligned at each position of each codon in the CDS region, then summed over three codon positions (frame 1, 2 and 3 or phase 0, 1 and 2).

Transcript expression estimation

We estimate transcript expression (FPKM value) using StringTie (Pertea et al., 2015). The cut-off for expressed transcripts is FPKM > 0.5 (Hart et al., 2013). We will only consider smORF(s) that locate on an expressed transcript.

3.5 ORF discovery

This component takes the gene annotation, putative smORFs, Ribo-Seq and RNA-Seq alignment as input to predict actively translated smORFs (ORF calling). We combine alignment files of all biological replicates to increase the signal intensity in case the smORFs are lowly expressed. This component consists of several metrics and filters, putative smORFs that have exceeded a confidence threshold for each metric (see Materials and Methods) were kept.

ORF discovery steps

Read length filter: Sequencing reads of different read lengths show different phasing patterns (see above). We determine the read length populations that show a strong periodicity pattern towards reading frame 1. Reads that are not in these lengths are filtered out.

Read count filter: This step filters out smORFs that have no read aligned to their regions. The filtered smORFs are considered not expressed. Rather than give an arbitrary cut off (e.g. 10 reads), I set read count equals zero.

RPF count filter: A sequencing read could be obtained by a scanning ribosome, or other RNA-binding proteins (Ji et al., 2016), the ones that are protected by ribosomes are truly RPFs. A RPF will be mapped to P-site at each codon in an ORF. A smORF with no reads mapping (0 RPF) is considered not translated and will be filtered out (same reasoning as read count filter above).

Transcript expression filter: Expressed host transcripts are retained after expression values are estimated. smORFs that are not from expressed transcripts are filtered out.

Class assignment: smORFs are classified to nine categories according to their relative location to an annotated CDS (**Table 1.3**). Each smORF is assigned a class label.

ORFScore filter: ORFscore is a model to compare the triplet periodicity distribution in an ORF to a uniform distribution (**Figure 3.6**). ORFScore for each smORF is calculated and smORFs with RPF coverage > 0.1, ORFScore > 0 and adjusted p-value < 0.01 were retained (see Materials and Methods).

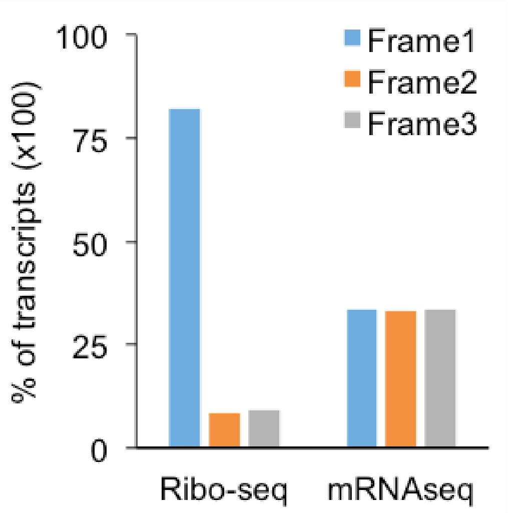


Figure 3.6 | **Triplet periodicity pattern appears in Ribo-Seq data.** Because ribosomes step three nucleotides in each cycle of translation elongation, in our ribosome profiling datasets a triplet periodicity is observable in the distribution of ribosome-protected footprints, in which 70-90% of the reads on a codon fall within the first of the three codon positions. This allows deduction of translation reading frames, if the reading frame is not known a priori. Comparing to Ribo-Seq, reads of RNA-Seq are uniformly distributed. Figure shows the whole transcriptome triplet periodicity distribution in Ribo-Seq compared to a uniform distribution in mRNA-Seq in LPS-activated B cell setup 1 sample.

Region filter: If a smORF region is overlapping with a CDS, the proportion of signal it absorbs from the CDS is estimated as (RPF count$_{CDS}$/RPF count$_{smORF}$), if the ratio is greater than 1, it means that the smORF is inside the CDS and absorbs all its signal from the CDS, this smORF will be filtered out.

Ribosome release score (RRS) filter: Ribosomes release at *bona fide* stop codons. Unlike RNA-Seq, we expect a very low level of reads (none or background noise) at an ORF’s 3’UTR. The Ribosome Release Score measures ribosome release to detect translation through ORFs. A pseudo 3’UTR was defined for noncoding ORFs (see Materials and Methods).

Nested filter: ORFs may have the same stop codon but different start codon, the short ones are nested in the longest one, among the nested ORFs, we examine the ORFScores and start codons of those nested ORFs, the one with the maximum ORFScore is retained, if more than one ORFs have the same ORFScore, an ORF with AUG start codon is retained.

Inside/outside ratio filter: Ribosome footprints typically show precise positioning between start codon and stop codon of a coding region (Brar and Weissman, 2015). Protein synthesis inhibitors stall ribosome elongation, the elongating ribosomes pause and accumulate at the start codon, a peak of reads can be seen in the sequence alignment. By considering read distribution or density outside and inside an ORF, we devised a feature called inside/outside read ratio to assess whether a genuine translation takes.

3.6 Pipeline Output

The output of the pipeline is a list of smORFs that have passed the filters in ORF discovery component. They are identified as actively translated smORFs supported by strong experimental evidence. In the output file, the genomic location and splicing information (including number of exons and exon lengths) of a smORF is clearly annotated and can be loaded and visualized in a genome browser. The quantitative information about a smORF is also calculated including translation efficiency, RNA expression and Ribosome expression (FPKM value). The nucleotide sequences are retrieved and translated into amino acid sequences (**Table 3.2**). The information will be used for downstream analysis including translation efficiency, cross-species conservation, uORF regulation, host transcript expression and signal sequence prediction.

Column	**Description**
1 - 12	The first 12 columns are in BED12 format, the fields are described here - https://genome.ucsc.edu/FAQ/FAQformat.html#format1. The 4th column is ORFId (transcript-based).
13	smORF class
14	Peptide length
15	RegionId (genomic-based)
16	Ensembl transcript Id
17	Gene symbol

18	Gene description
19	ORF score
20	Ribosome release score
21	Ribo FPKM
22	RNA FPKM
23	Translation efficiency (TE)
24	CDS TE (NA if host transcript is noncoding)
25	AA sequence

Table 3.2 | **Pipeline final output format.**

3.7 Results

Ribosome profiling data quality control

I carried out data quality control (QC) for all RNA-Seq and Ribo-Seq data. The reads are 50 bp single end. Firstly, I examined the sequencing base quality using FastQC. In the raw data, base calls that fall in the green area are of very good quality (**Figure 3.7A**). The blue line represents the mean quality that drops towards the sequencing adapter at the 3’ end of the reads. I trimmed the sequencing adapter and kept reads that are likely protected by ribosomes (read length between 25 nt and 35 nt). I also removed reads that are from rRNA/tRNA and other classes of small RNAs as described above. The remaining reads are of good quality for downstream analysis (**Figure 3.7B** and **Table 3.3**).

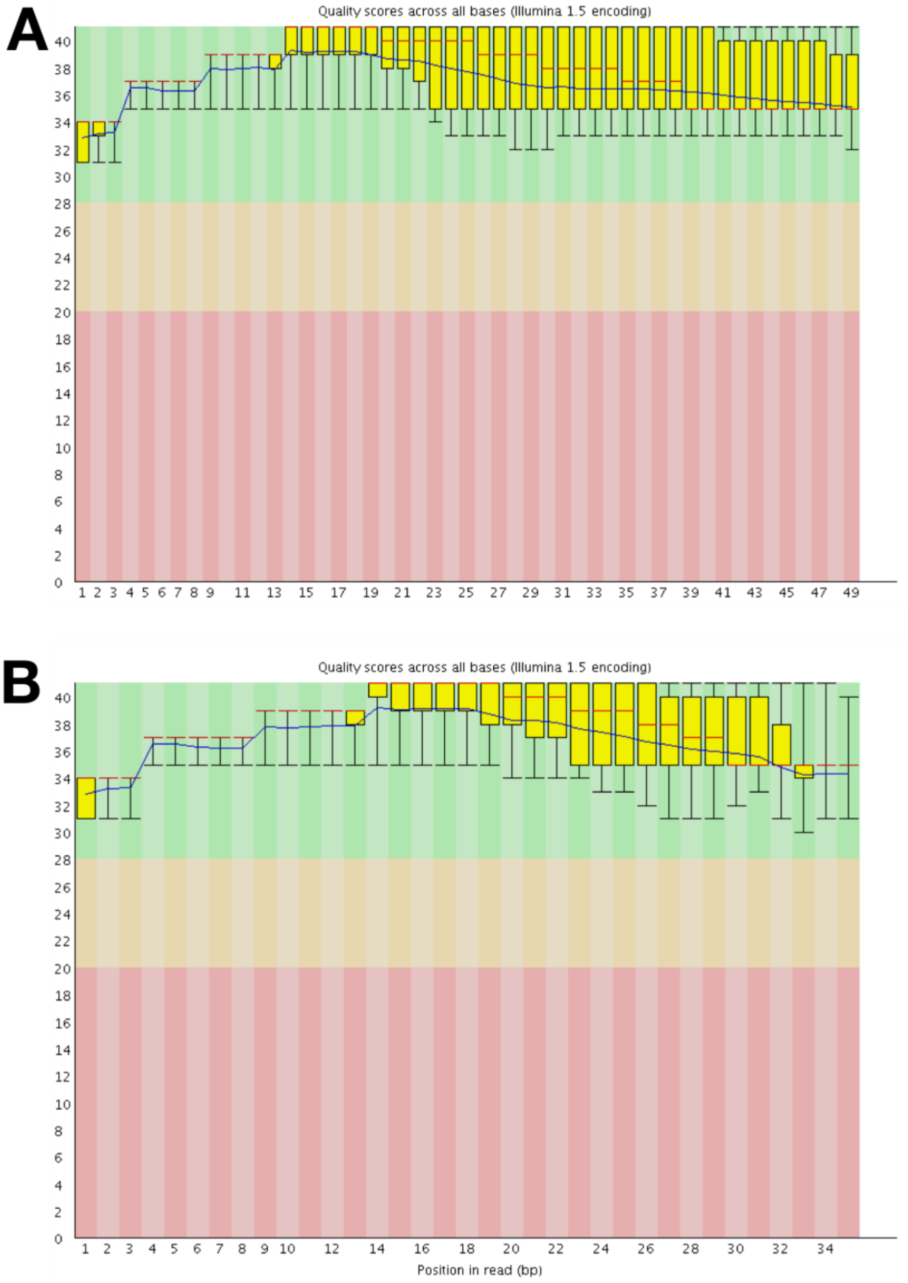


Figure 3.7 | **Per base quality of Ribosome Profiling sequencing data.** (A) Base quality of the raw Ribo-Seq data of LPS-activated B cell setup 1 sample (WT1). (B) Base quality after adapter trimming and contaminant removal.

Experiment	**Raw read count**	**Read count and proportion after adapter trimming**	**Read count and proportion after contaminant removal**
B cell setup 1 Resting (N=4)	29,776,606	29,776,606 (100%)	13,380,423 (44.9%)
	19,730,128	19,730,128 (100%)	8,326,016 (42.2%)
	28,869,948	28,869,948 (100%)	12,833,039 (44.5%)
	29,399,992	29,399,992 (100%)	11,605,385 (39.5%)
B cell setup 1 LPS-activated (N=5)	27,802,753	17,041,834 (61.3%)	12,187,070 (43.8%)
	29,904,842	22,943,205 (76.7%)	15,547,757 (52%)
	24,847,188	17,676,038 (71.1%)	12,374,227 (49.8%)
	24,521,162	19,784,374 (80.7%)	14,200,808 (57.9%)
	27,010,038	21,866,707 (81%)	16,338,358 (60.5%)
B cell setup 2 LPS-activated (N=5)	43,527,344	35,769,394 (82.2%)	23,375,471 (53.7%)
	34,062,948	28,962,088 (85%)	20,807,177 (61.1%)
	38,992,152	33,953,299 (87.1%)	24,536,458 (62.9%)
	44,782,997	38,213,769 (85.3%)	26,139,506 (58.4%)
	37,566,709	31,760,261 (84.5%)	22,708,490 (60.4%)
CD4$^+$ T cell (N=5)	46,252,019	34,073,573 (73.7%)	32,653,025 (70.6%)
	101,717,087	87,210,021 (85.7%)	84,155,705 (82.7%)
	28,346,186	23,265,159 (82.1%)	22,109,646 (78%)
	35,123,808	31,637,542 (90.1%)	30,249,887 (86.1%)
	32,225,196	27,771,252 (86.2%)	26,270,808 (81.5%)

Table 3.3 | **Ribo-Seq QC summary.** Read count of the raw sequencing data and read count and proportion to the raw read count after each step of QC. The remaining reads will be used for genome alignment. For B cell setup 1, adapters have already been removed when I obtained the Fastq files.

Missing rRNA reference sequences

I screened mouse libraries against a collection of reference genomes using FastQ Screen. I found that around 25% of reads mapped to the rat genome due to the overall similarity between mouse and rat. I also noticed that in a LPS-activated B cell sample, 5% of the reads were mapped to the human genome only at multiple locations, while 40% of the reads didn’t find a place in the reference genomes (no hits) which means they are from unknown sources other than the reference set (**Figure 3.8A**). In a resting B cell sample, 25% of the reads were mapped to the human genome only at multiple locations, 20% of the reads were no hits (**Figure 3.8B**).

I investigated the nature of the suspicious alignment to human and no hits. Regarding the reads that aligned to human genome only, we noticed that the reported mapping qualities were low. Mapping quality estimates the probability that the alignment does not correspond to the read's true origin. Low mapping quality means a read can be mapped to multiple locations, thus indicates the read is a low-complexity sequence. I did not see reads accumulated at any specific regions with the exception of the mitochondrial genome (**Figure 3.8C**). The human mitochondrial genome sequence is GC-rich and GC-rich reads are a class of low-complexity sequence. This issue was noted in the case that the library is composed of low complexity or short sequences which are very easily mapped (Andrews, 2016). In a LPS-activated B cell sample, where more transcripts were expressed, more reads were mapped to mouse specifically.

I searched the most overrepresented no hits reads using NCBI BLAST. The top hits were:

- TPA_exp: Mus musculus ribosomal DNA, complete repeating unit (BK000964.3)
- Mus musculus 45S pre-ribosomal RNA (Rn45s), ribosomal RNA (NR_046233.2)
- Mus musculus 28S ribosomal RNA (Rn28s1), ribosomal RNA (NR_003279.1)
- Mus musculus strain BALB/c 45S ribosomal RNA region genomic sequence (GU372691.1)

Those sequences were not included in our initial rRNA sequences. I added them to the incomplete annotation and now have a better filter for rRNA contaminants.

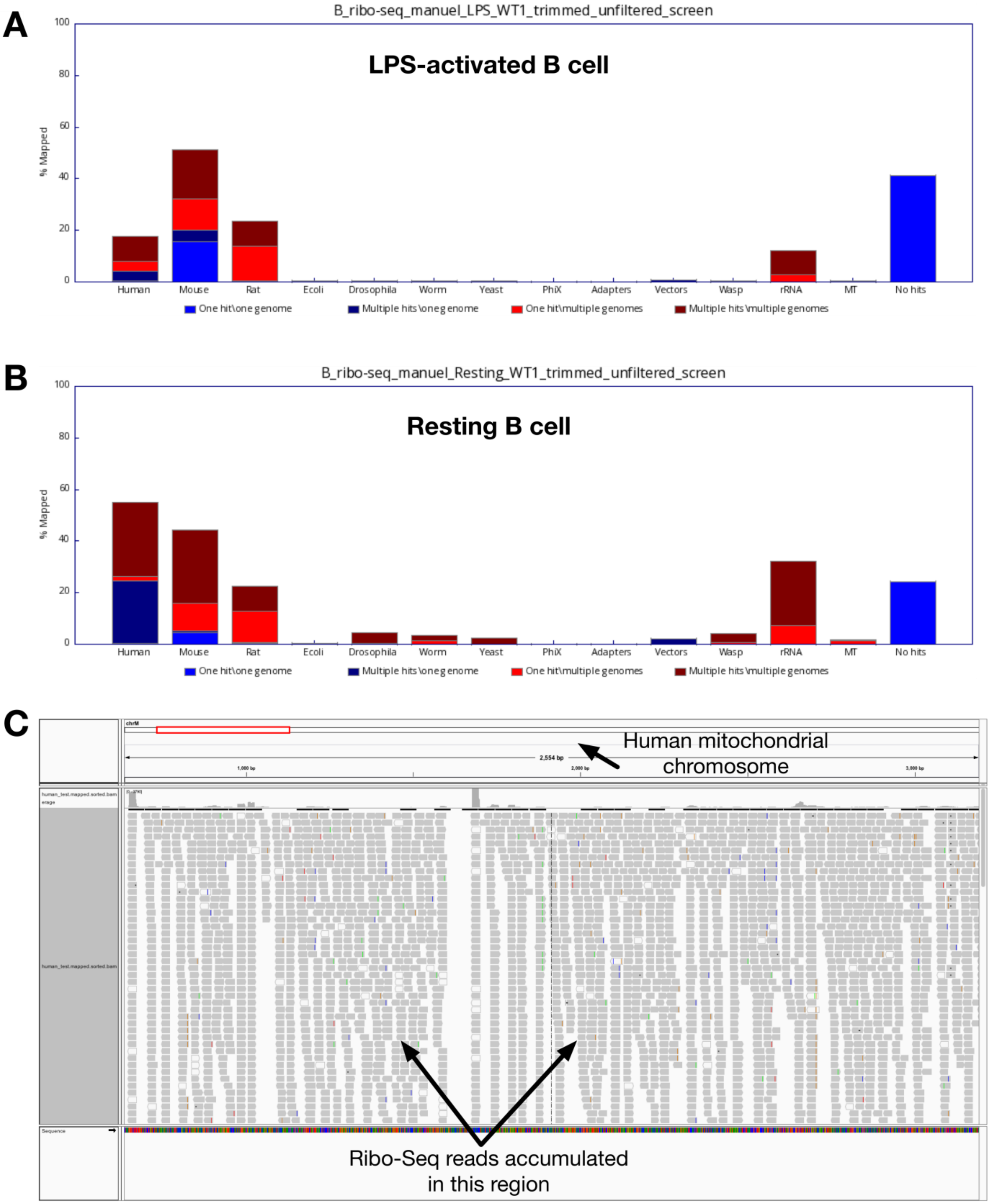

A
B_ribo-seq_manuel_LPS_WT1_trimmed_unfiltered_screen
LPS-activated B cell
% Mapped
100
80
60
40
20
0
Human
Mouse
Rat
Ecoli
Drosophila
Worm
Yeast
PhiX
Adapters
Vectors
Wasp
rRNA
MT
No hits
One hit\one genome
Multiple hits\one genome
One hit\multiple genomes
Multiple hits\multiple genomes
B
B_ribo-seq_manuel_Resting_WT1_trimmed_unfiltered_screen
Resting B cell
% Mapped
100
80
60
40
20
0
Human
Mouse
Rat
Ecoli
Drosophila
Worm
Yeast
PhiX
Adapters
Vectors
Wasp
rRNA
MT
No hits
One hit\one genome
Multiple hits\one genome
One hit\multiple genomes
Multiple hits\multiple genomes
C
chrM
2,554 bp
1,000 bp
2,000 bp
3,000 bp
Human mitochondrial chromosome
human_test.mapped.sorted.bam Coverage
human_test.mapped.sorted.bam
Sequence
Ribo-Seq reads accumulated in this region

Figure 3.8 | **Sequence contaminant estimation using FastQ Screen against a collection of reference genomes.** (A) A read was mapped to multiple genomes due to sequencing similarity between species. In LPS-activated B cell setup 1 sample (WT1), 5% of the reads were mapped to the human genome only at multiple locations. (B) Followed the same alignment procedure, in a resting B cell sample, 25% of the reads were mapped to human only multiple times. (C) Ribo-Seq reads accumulated at human mitochondrial genome. Legend for (A) and (B): blue – one hit\one genome; navy – multiple hits\one genome; red – one hit\multiple genomes; crimson – multiple hits\multiple genomes.

Genome alignment

Ribo-Seq raw sequencing data was processed through adapter trimming, size selection and contaminant removal. After contaminant removal, reads were mapped to the mouse genome. Ribo-Seq alignment looks different to RNA-Seq, with reads positioned precisely from the start codon (**Figure 3.9**). Taking the data from LPS-activated- and resting-B cells as an example, on average, 25% of the reads were discarded after trimming adapters and size selection to retain reads between 25-35 nt. On average 52% of reads from LPS-activated B cell samples and 44% from resting B cell samples remained after rRNAs and tRNAs been removed. On average 26.9% (varying from 22.5% to 31.3%) of reads from LPS-activated B cell samples and 19.7% (varying from 16.9% to 21.5%) from resting B cell samples were uniquely mapped to genome respectively. As expected, more reads were mapped for LPS-activated B cell samples than resting B cells as LPS-activated B cell expressed more transcripts. I also tried to map the reads to reference transcriptome using Bowtie, but the alignment showed a higher count than genome alignment because a read would be double-counted if it was mapped to multiple transcript isoforms of the same gene **(Figure 3.10)**.

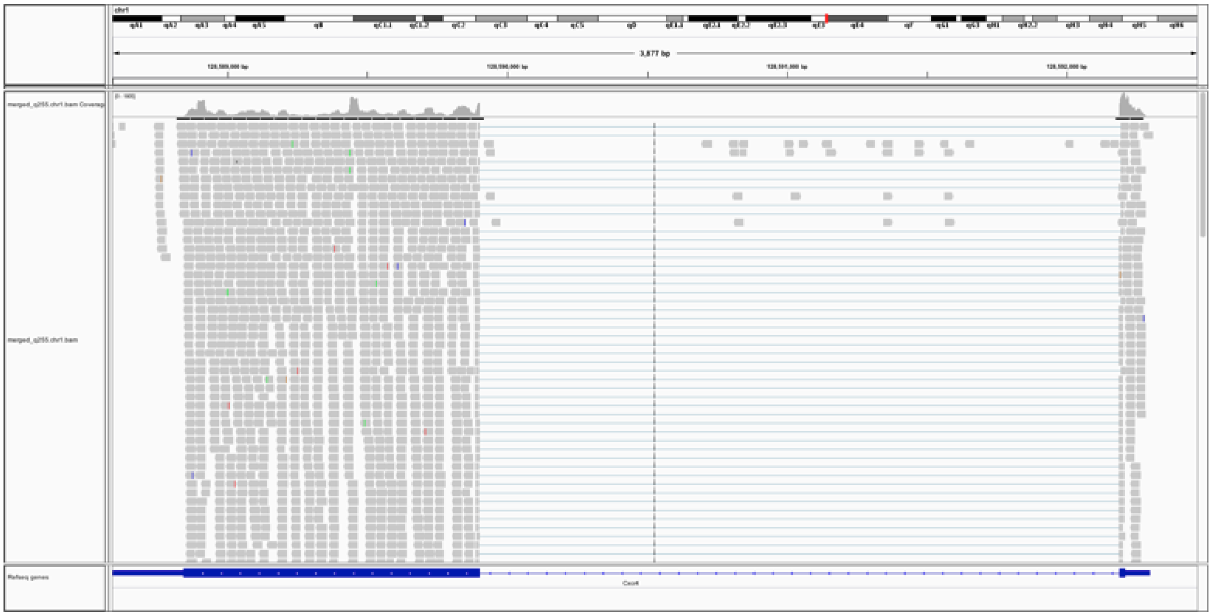


Figure 3.9 | **Ribos-Seq reads mapped to Cxcr5 transcript.** Reads position precisely from the start codon of Cxcr5 CDS (the thick blue block on the bottom track represents CDS, thin block represents UTRs)

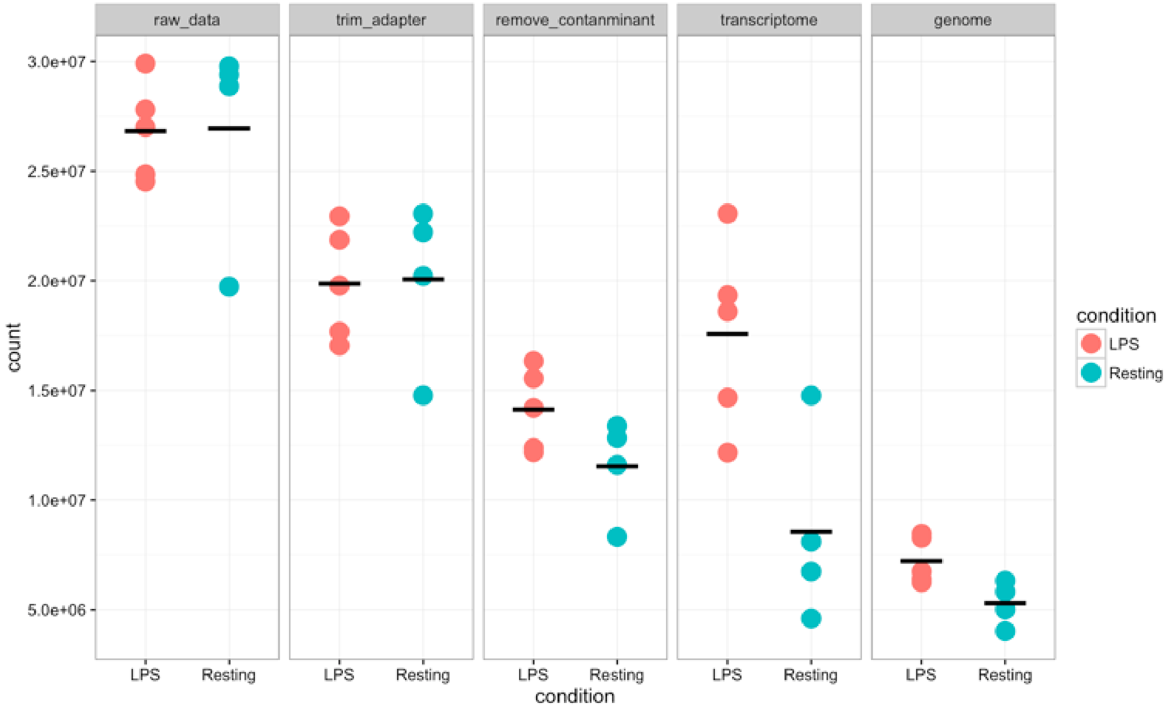


Figure 3.10 | **Read counts of LPS-activated (N=5) and resting (N=4) B cell samples at each step of QC.** Firstly, reads were trimmed and those were not of 25-35 nt were removed. Secondly, contaminant reads were removed. The remaining reads that were not uniquely mapped to the reference genome were removed. Black horizontal bar represents the mean.

There are different options of sequence alignment tools for genome alignment. I tested three widely used tools - TopHat (Trapnell et al., 2009), HISAT2 (Kim et al., 2015) and STAR aligner. It was noted that TopHat has incorrectly reported reads as unique mapping (Andrews, 2016). TopHat initially maps reads to the transcriptome, and only if it does not find a hit does it then map to the genome. Reads can be reported as uniquely mapped to the transcriptome when they are actually mapped to many locations within the genome. On the other hand, HISAT2 and STAR do not have this limitation as they map directly to the genome. By comparing HISAT2 and STAR alignment, I noticed that STAR aligned additional reads at splice junctions, for example, gene Mrpl15 (**Figure 3.11**). We searched some of those reads in the transcriptome alignment and found them to be aligned uniquely to the transcript of Mrpl15. The reason for differential alignment between HISAT2 and STAR is unknown and we have raised this question with the authors of HISAT2 and STAR. If STAR alignment is true, it will help to increase the coverage thus to improve ORF detection. If the alignments of STAR are genuine, it might have implications for other work such as studies of splicing.

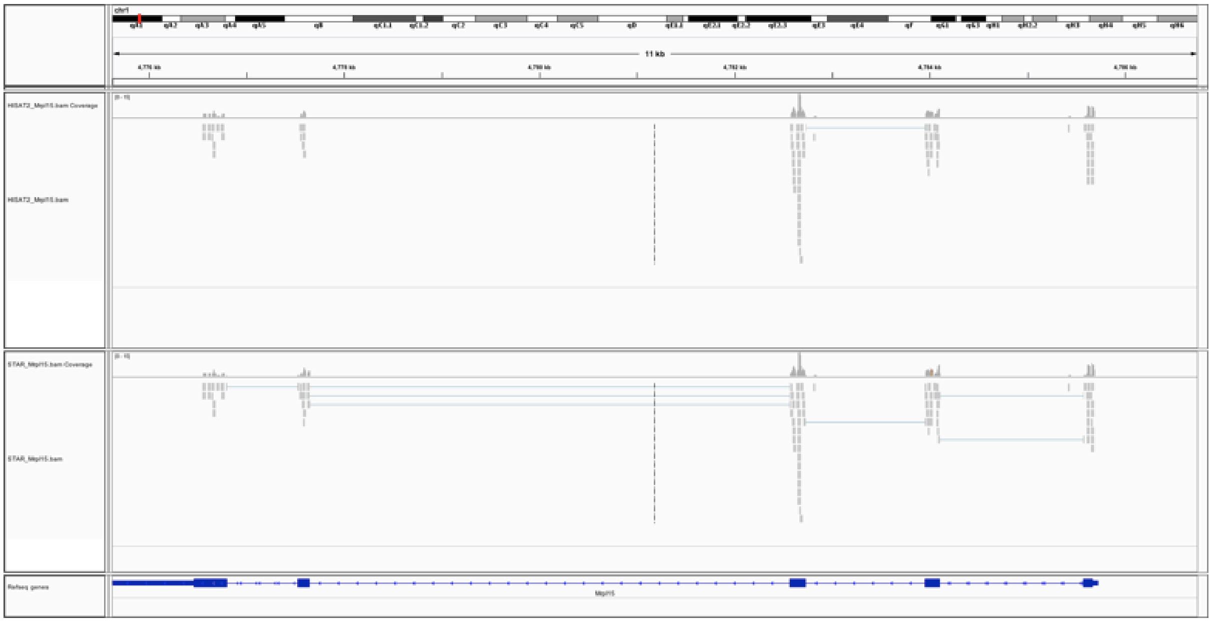


Figure 3.11 | **Ribo-Seq reads mapped to Mrpl15 transcript.** STAR aligned additional reads at splice junctions and HISAT2 missed the splice junction alignment. The upper track shows HISAT2 alignment. The lower track shows STAR alignment.

Determine P-site offset and sub-codon phasing

Ribo-Seq reads are frequently mapped to their P-sites. In the initial study of ribosome profiling (Ingolia et al., 2009), it was found that the positions of the 5′ ends of the footprints started abruptly 12 to 13 nt upstream of the start codon. Therefore, I would expect to see a peak at a fixed distance upstream of the start codon. In our data, reads with length from 29 to 32 nt showed P-site offset of 12 or 13 nt, matching our expectation and indicating good data quality (**Figure 3.12**). Reads that had low or no peaks were likely not RPFs. Using P-site offset results, I calculated the sub-codon phasing proportion of each read length. It showed that 29-mers to 33-mers were abundant and highly phased in frame 1 (or phase 0) (**Figure 3.13** and **Table 3.4**). So I used read length from 29 to 33 nt for the downstream ORF calling.

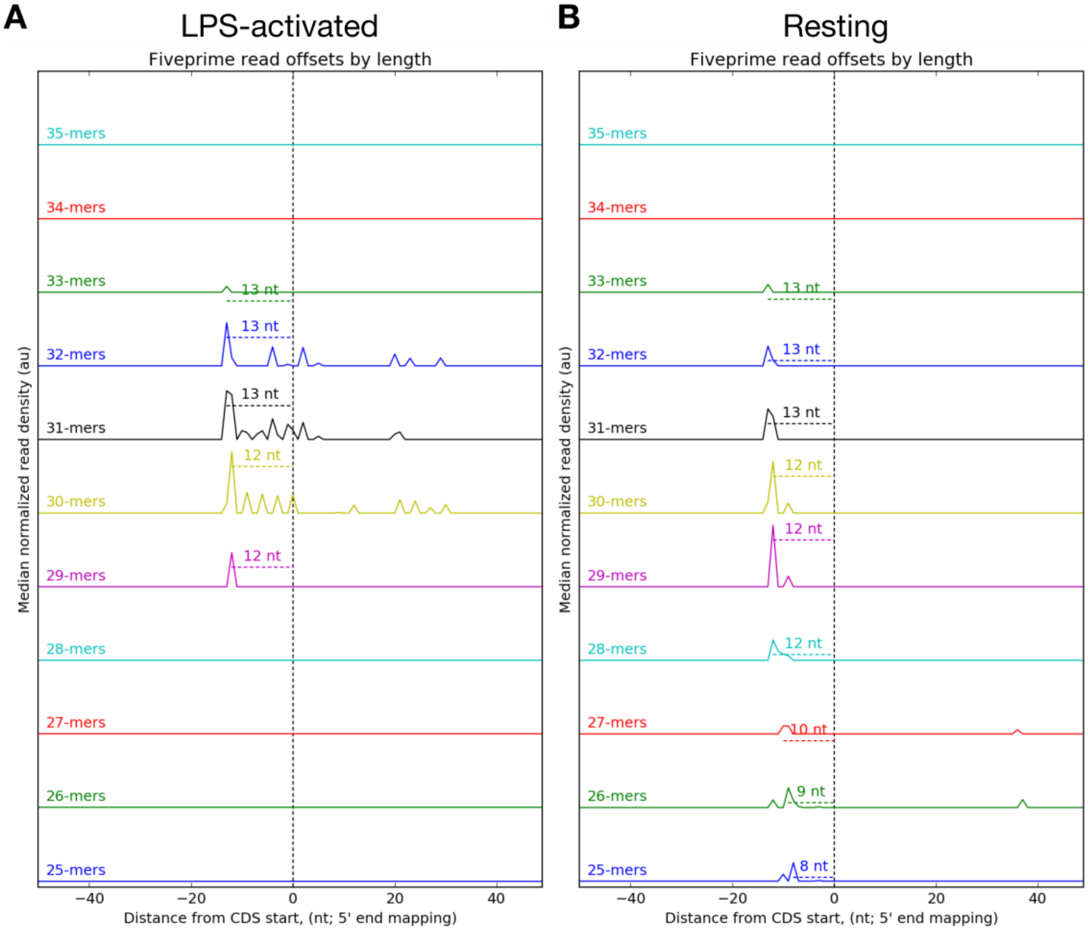


Figure 3.12 | **P-site offsets of reads in different length.** (A) 29-mers to 32-mers showed strong peaks from LPS-activated B cell setup 1 sample (WT1). (B) 28-mers to 32-mers showed strong peaks from a resting B cell setup 1 sample (WT1). Vertical dotted line cross first base of start codon.

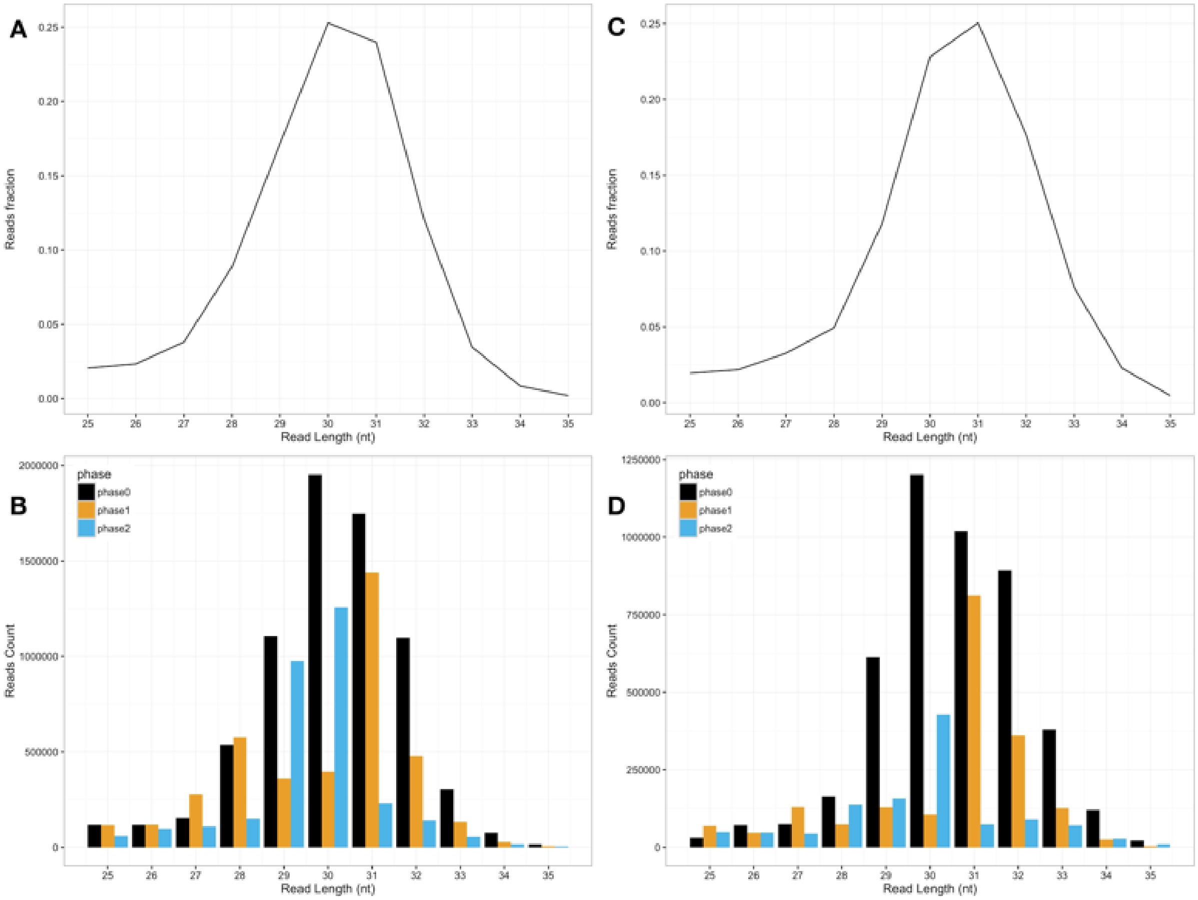


Figure 3.13 | **Sub-codon phasing.** (A) and (B) show the read length distribution and triplet periodicity of LPS-activated B cell setup 1 sample (WT1) respectively. (C) and (D) show the same information for resting B cell setup 1 sample (WT1).

Read length	**P-site offsite**	**Reads counted**	**Fraction reads counted**	**Phase 0**	**Phase 1**	**Phase 2**
25	7	19781	0.015942	0.332946	0.443304	0.223750
26	8	26675	0.021498	0.331621	0.264105	0.404274

27	9	47856	0.038568	0.588829	0.177094	0.234077
28	12	97003	0.078176	0.461099	0.166923	0.371978
29	12	216829	0.174745	0.607806	0.175069	0.217125
30	12	374353	0.301695	0.699292	0.086186	0.214522
31	12	298826	0.240827	0.534723	0.052951	0.412327
32	13	121240	0.097709	0.592610	0.329627	0.077763
33	13	30141	0.024291	0.567732	0.313991	0.118277
34	13	6519	0.005254	0.557908	0.312931	0.129161
35	13	1609	0.001297	0.517091	0.324425	0.158484

Table 3.4 | **P-site offset and sub-codon phasing of LPS-activated B cell setup 1 sample (WT1).**

Sufficient read coverage to predict smORFs

I would like to find out if our datasets have enough reads to call smORFs. One way is to compare our data to published datasets of similar cell types or data that has previously been used for ORF prediction. Firstly, we compared our activated T cell dataset to a published reactivated Th1 cell dataset (Davari et al., 2017), it showed our data had 30% more reads on average. We also looked at another study by Crappé and colleagues (Crappé et al., 2013), they have identified smORFs using a mouse embryonic stem cells (mESCs) dataset (Ingolia et al., 2011). In order to have a fair comparison, we looked at genes that expressed in both their dataset and our dataset. For example, we selected one predicted smORF on a non-coding RNA Snhg12 and loaded their data and one of our LPS-activated B cell data into a genome browser. It showed our data has better coverage in this case **(Figure 3.14)**, it indicated that we have stronger signal to call this smORF. The comparisons increased our confidence that our datasets have sufficient reads to make predictions.

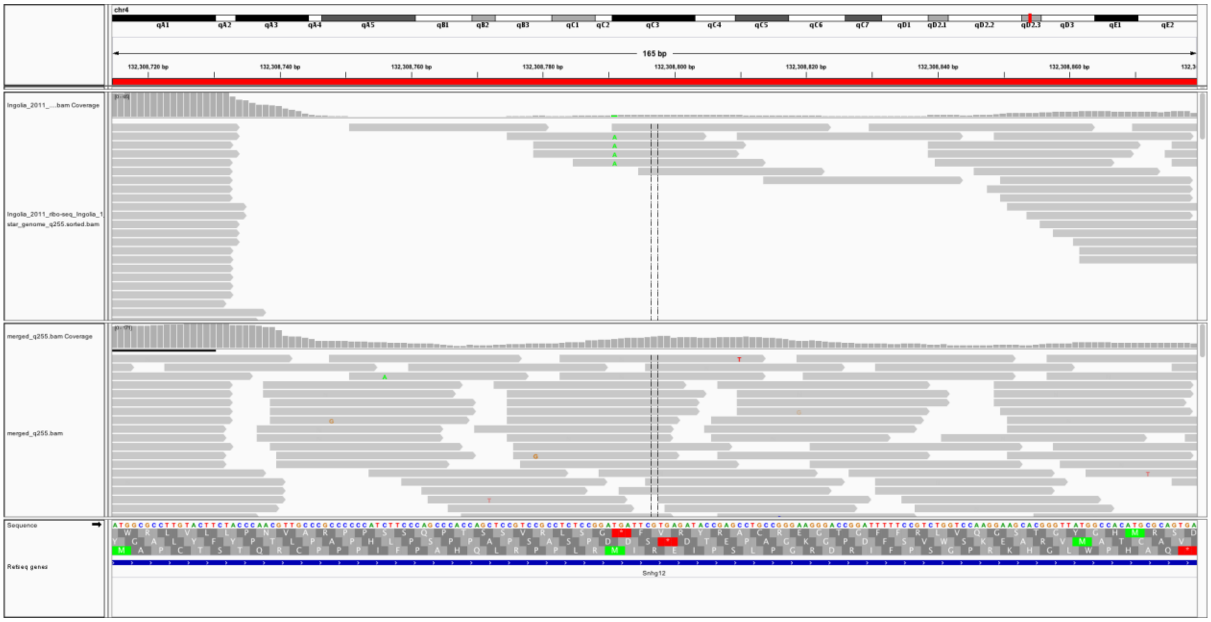

chr4
165 bp
Ingolia_2011_....bam Coverage
Ingolia_2011_ribo-seq_Ingolia_1 star_genome_q255.sorted.bam
merged_q255.bam Coverage
merged_q255.bam
Sequence
Refseq genes
Snhg12

Figure 3.14 | **Sequencing read coverage of two datasets on a smORF on Snhg12 transcript.** The read block shows the smORF region. The upper track shows a public dataset (Ingolia et al., 2011). The lower track shows our LPS-activated B cell data.

Reproduce the published smORFs

We tested our pipeline to see if we can reproduce the smORFs predicted from a published study by Fields and colleagues (Fields et al., 2015). The cell type is mouse bone marrow derived dendritic cells. There are Ribo-Seq data of nine time points (unactivated and LPS-activated for 0.5h, 1h, 2h, 4h, 6h, 8h, 9h or 12h). We processed the raw sequencing data of all nine time points and aligned them to the reference genome allowing multiple mapping (N=64, same as the original study setting), then we combined at all time points for smORF prediction in ORFLine. Out of 46 smORFs reported in their study, we reproduced 23 (50%). We manually checked the other 23 that we had not predicted, they were all on long terminal repeat (LTR) sequences (ERV1 or ERVL-MaLR), and they all have low expression levels. The 23 reproducible smORFs are mainly from annotated lincRNAs, processed transcripts and protein-coding transcripts. Among the reproducible smORFs, there is only one LTR smORF which has a very high expression. We were uncertain if those LTR-derived smORFs truly encode peptides or they were false positives by absorbing a huge amount of signals (reads which are supported to be aligned to their origins). We searched the literature for peptides encoded in LTR regions in mammalian genomes, but could not find any supportive information. In practice, LTRs regions always result in wrong sequence alignment and they are commonly ignored.

Comparison between ORFLine and RiboCode

We applied a recently published ORF-detection pipeline called RiboCode (Xiao et al., 2018) its default settings on the B and T cell datasets. RiboCode assesses the triplet periodicity of RPFs in an ORF with modified Wilcoxon signed-rank test and is claimed to outperform other existing pipelines including RiboTaper, Rp-Bp and ORF-RATER. There are differences between ORFLine and RiboCode to predict smORFs. ORFLine scans a transcript sequence from the 5' end, when it detects a start codon it will continue and stop when meeting a stop codon. RiboCode does it in a reverse way, it finds a stop codon first, then goes from 3' to 5', so RiboCode will predict nested ORFs with the same start codon. For example, if an ORF starting with AUG exists on a transcript, at the same time there is another AUG inside this ORF in the same reading frame, then it is a shorter ORF nests in the first ORF, both ORFs will be predicted by RiboCode, but ORFLine will only predict the longer ORF. RiboCode maps Ribo-Seq reads to the transcriptome, but ORFLine maps reads uniquely to the genome. As we mentioned earlier, for transcriptome alignment, reads can be mapped to multiple transcripts and potentially this may increase false signals. As RiboCode does not take RNA-Seq into account, it will not be able to estimate host transcript expression and potentially maps reads to non-expressed transcripts.

In total, using the B and T cell datasets RiboCode predicted 15920 unique smORFs, in which 3667 are smORFs nested in longer smORFs in the same reading frame and 48 smORFs are from non-expressed transcripts. We removed those 3715. In the remaining 12205 smORFs, 3337 were predicted as internal or frameshift smORFs. These are found nested in the CDS, but in a different reading frame. Considering that frameshift translation is a rare event (Michel et al., 2012), they are not included in our results. We removed all 3337 frameshift smORFs predicted by RiboCode and compared the remaining 8868 non-internal smORFs predicted by RiboCode with 5744 predicted from our pipeline (**Figure 3.15**). Of these, 1957 (22% in RiboCode and 34% in ORFLine) are found as exact genomic coordinate matches by both pipelines. For the un-annotated smORFs, we are not certain they are translated, and we lack a reference set of true-positives, therefore we sampled the smORFs which are different between the two pipelines and noticed that smORFs predicted by RiboCode typically have low RPF coverage or are assigned a low or negative ORFScore, or low RRS, and they are filtered out by our pipeline. Our criteria for metrics have shown to be robust in smORF prediction in previous studies (Bazzini et al., 2014, Guttman et al., 2014). We also predict smORFs encoded by low abundance transcripts that are not predicted by RiboCode (**Figure 3.16).** Therefore, it appears our pipeline is more stringent at predicting smORFs.

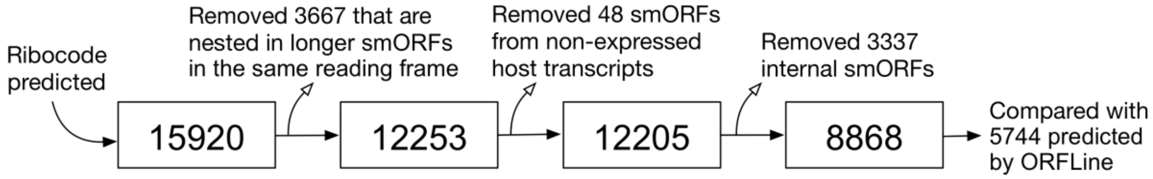

Ribocode predicted
Removed 3667 that are nested in longer smORFs in the same reading frame
Removed 48 smORFs from non-expressed host transcripts
Removed 3337 internal smORFs
15920
12253
12205
8868
Compared with 5744 predicted by ORFLine

Figure 3.15 | **Number of smORFs used for a comparison between RiboCode and ORFLine.** Initially, 15920 smORFs were predicted by RiboCode, 3367 were removed as they were nested in longer smORFs in the same frame frame. 48 were removed as they were from non-expressed host transcripts, and 3337 were removed as they were internal smORFs. The remaining 8868 were used to compare with ORFLine result.

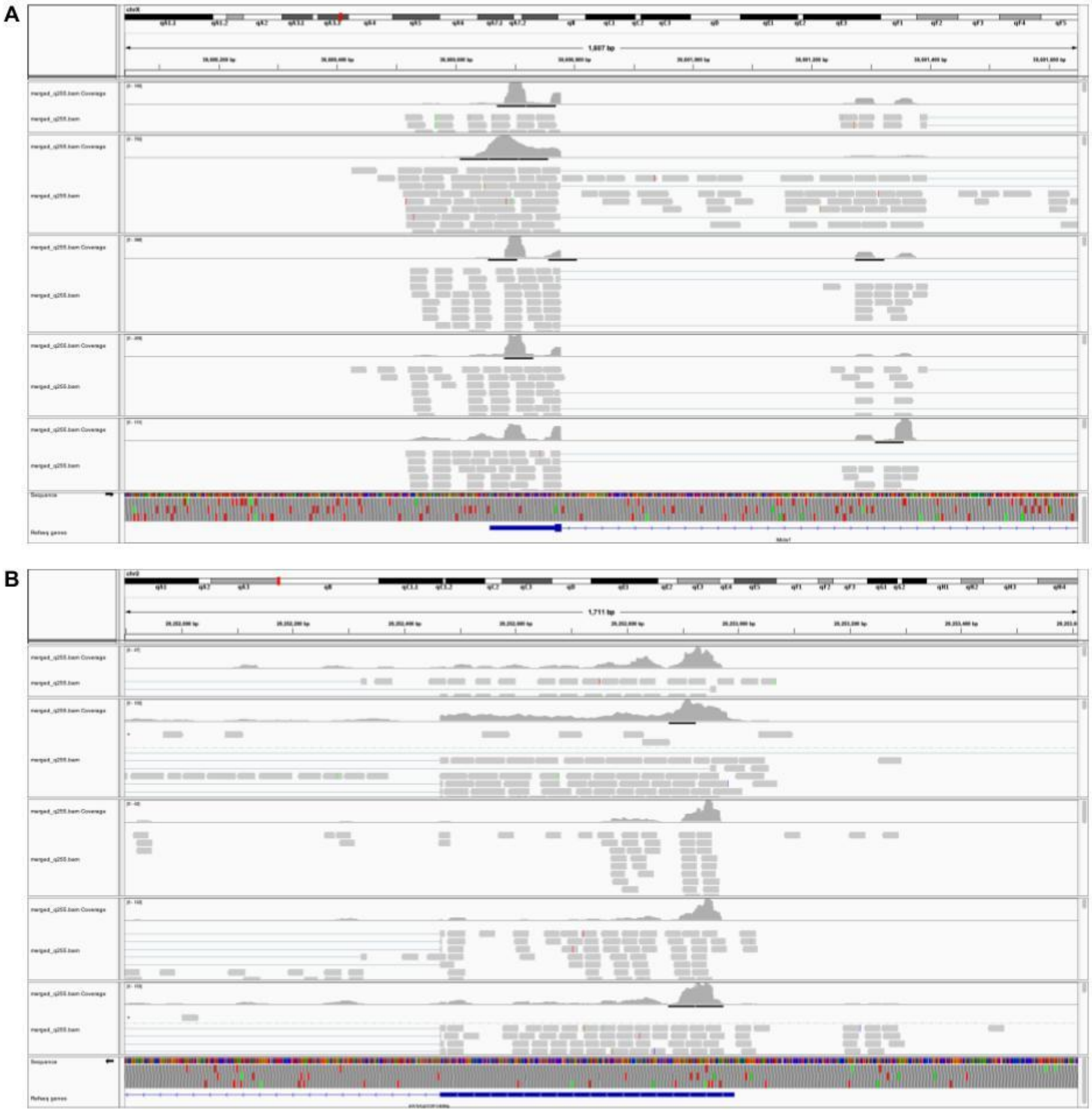


Figure 3.16 | **smORFs predicted by ORFLine but not by RiboCode.** (A) uORF from Mnt transcript. (B) smORF from noncoding transcript 6530402F18Rik. Tracks show alignment of the datasets used in the pipeline comparison.

3.8 Pipeline availability

Pipeline code is publicly available on the source code hosting platform GitHub. The URL is https://github.com/boboppie/ORFLine. We also create a Singularity image (https://singularity.lbl.gov/) which enables the users to execute and test the pipeline easily in a virtual environment. All dependencies including bioinformatics tools are pre-installed in the image, the URL is https://github.com/boboppie/ORFLine-singularity.

3.9 Discussion

In a typical ribosome profiling library, it was noted that many sequencing reads do not correspond to translated regions (Ji et al., 2016). Ribosomes are not specifically selected during the biochemical isolation procedure for ribosome profiling, and therefore non-ribosomal RNA-protein complexes (e.g. RNA binding proteins) may also be present. There are different ways to purify ribosomes, for example, we can generate a transgenic mouse model to add a tag (e.g. GFP protein) to the ribosome and fish them out using antibodies, however it is expensive and we risk to changing the ribosome’s properties. The sequencing reads can be a mix of RPFs and non-ribosomal RNA-protein complex protected fragments. RPFs span the entire translated region and show triplet periodicity. In contrast, non-ribosomal RNA-protein complex protected fragments should be highly localized (Ji et al., 2016). It is possible to carry out further QC steps to distinguish different RNA species, so far, we retain the reads that show strong triplet periodicity.

Regarding gene annotation, another commonly used annotation is NCBI RefSeq (Pruitt et al., 2013), in our study, we chose GENCODE instead of RefSeq because a comparison study between those two has been recently carried out and shown that the GENCODE Comprehensive set is richer in alternative splicing, novel coding sequences (CDSs), novel exons and has higher genomic coverage than RefSeq (Frankish et al., 2015). It is also possible to use a combined set, but the issue will be that the more complete we try to make our set, the more we will include transcripts which are not functionally relevant and which might add substantial numbers of false positives to our study. We think making use of higher quality annotation is more advantageous in this study.

Regarding novel transcripts, we would expect to discover novel transcripts or alternative isoforms expressed in our data which potentially have ORFs embedded. We have tried to use RNA-Seq to assemble *de novo* the transcriptome using Cufflinks, a few hundreds of novel transcripts were predicted, but we decided not to include them to the reference transcriptome. Studies have shown that when using computational approaches (both genome-guided such as Cufflinks and de novo assembly) to infer the set of transcripts expressed in RNA samples using RNA-seq, those approaches produce a large number of artefacts (false positives), which absorbed a substantial proportion of the reads from truly expressed transcripts and were assigned large expression estimates (Jänes et al., 2015, Steijger et al., 2013). It is known that the validation of those novel transcripts is not trivial. Computational predictions could be made more conservative by using reconstructed transcripts detected by several methods (Steijger et al., 2013). Alternatively, the use of long read technologies (Oxford Nanopore sequencing or PacBio SMRT sequencing) to identify the distinct set of transcripts in a sample combined with RNA-Seq to estimate expression levels may be the optimal approach for the time being for accurate characterization of RNA samples. One way to validate novel ORFs and transcripts is to detect triplet periodicity pattern in them.

In this chapter we have developed and validated a new small ORF-calling pipeline. In the following chapter we will use this to characterise smORF in the lymphocyte datasets.

Chapter Four

Properties of smORFs and Functional Validation of Micropeptides

4.1 Summary

I systematically characterized smORFs in mouse lymphocytes using data from (*ex vivo*) resting B cells; two independently generated datasets of LPS-activated B cells; stimulated naïve CD4+ T cells; and a time-course of Th1 T cells re-stimulated with anti-CD3+anti-CD28 (Davari et al., 2017). I processed the data using the pipeline described in the previous chapter and I identified a total of 5744 unique smORFs in all samples analysed (union of 2607 smORFs predicted in B cells and 4935 smORFs predicted in T cells). A lower number (568) of smORFs were predicted for the resting B cells than for LPS-activated B cells (2444), most likely reflecting the elevated rates of transcription and translation in activated B cells. I aim to categorise smORFs as it is an important step allows us to group and properly learn their properties in each group bioinformatically, next, with the help with other members in the Turner lab, we can validate the existence of the predicted smORFs and further perform functional characterization.

4.2 Predicted smORFs

In total, 5744 unique smORFs were identified in B and T cells (**Table 4.1**). 1291 smORFs were identified in LPS-activated B cell setup 1 and 1859 in setup 2. 706 smORFs appear in both datasets and 2444 unique smORFs were identified in the union of the two LPS-activated B cell datasets (**Figure 4.1A**). The different sample treatments between these two setups (setup 1 was LPS+IL-4, setup 2 was LPS+IL-4+IL-5) might have resulted in different expression profiles and influenced the smORF prediction. 568 smORFs were identified in resting B cells and 415 of these were also found in LPS-activated B cells (**Figure 4.1B**). It is shown that the number of smORFs identified in Th1 reactivation 2h dataset (1084) is ~40% less than other two time points (2580 for 0h, 2697 for 4h) (**Figure 4.1C**), this was caused by the low read depth in this dataset (~40% less reads compared to 0h and 4h).

Experiment	**Canonical**	**Extended**	**uORF**	**ouORF**	**dORF**	**odORF**	**ncORF**	**Total**
Resting B cell	175	6	330	40	4	0	13	568
LPS-activated B cell (setup 1)	180	14	907	119	9	1	65	1295

LPS-activated B cell (setup 2)	207	15	1367	101	38	6	132	1866
Stimulated CD4$^+$ T cell	220	9	1675	120	77	2	187	2290
Th1 cell reactivation 0h	257	14	1812	171	102	7	217	2580
Th1 cell reactivation 2h	135	9	774	98	19	0	49	1084
Th1 cell reactivation 4h	256	17	1917	188	114	5	200	2697
Merged (unique)	**338**	**30**	**4174**	**441**	**243**	**17**	**501**	**5744**

Table 4.1 | **Predicted smORFs in each experiment.**

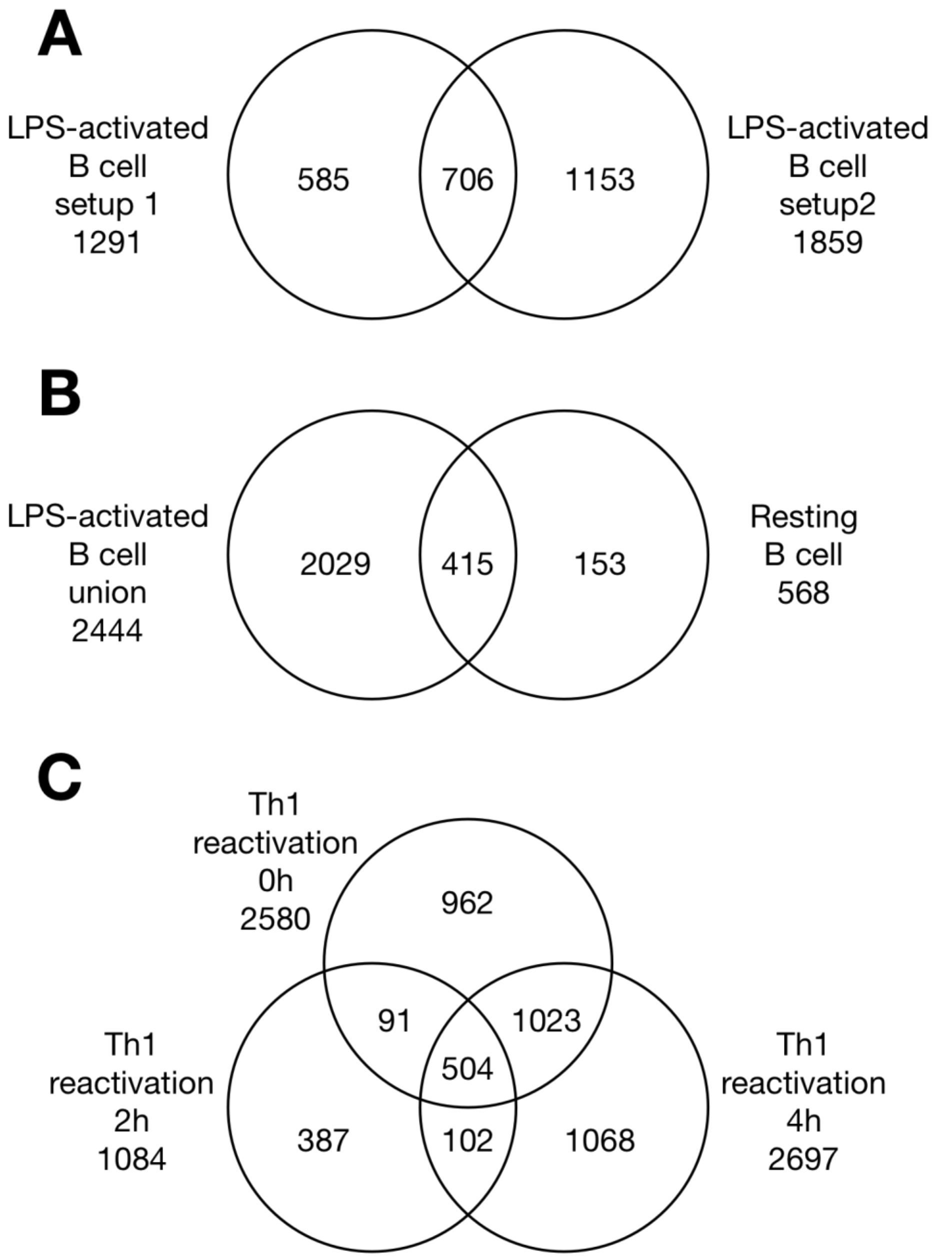


Figure 4.1 | **Number of actively translated smORFs in this study.** (A) Predicted smORFs were classified into seven groups according to their relative location in the host transcript. The number of smORFs in each class is shown in parenthesis. (B) Pie chart showing the proportion of smORFs of different classes. (A) 706 smORFs are present in both LPS-activated B cell datasets. (B) 415 smORFs are predicted in both resting B cell and LPS-activated B cell (union). (C) 504 smORFs are present in all three time points of Th1 reactivation datasets.

4.3 smORF classification

We classified smORFs according to their relative position within and the nature of their host transcript (**Figure 4.2A, B**). We predicted canonical smORFs and extended variants of annotated coding DNA sequences (CDS) of 100 codons or less in protein-coding mRNAs. We also find upstream ORFs (uORFs) and uORFs overlapping with coding regions (ouORFs) located in the 5' untranslated region (5' UTR) of annotated protein-coding mRNAs. uORFs are known to be prevalent in the genome and, in our data, they represent 80% of all smORFs found (**Figure 4.2B**). In addition, we predicted downstream ORFs (dORFs) and overlapping dORFs (odORFs) that are located in 3'UTRs of known protein-coding mRNAs as the rarest class of smORFs in this study. Lastly, 501 smORFs in putative non-coding RNAs (long non-coding RNAs and pseudogenes) were predicted, which are termed ncORF. Direct biochemical and functional evidence is available for only a fraction (~7%) of canonical smORFs in protein databases such as UniProt (UniProt Consortium, 2018) for their protein products, it includes diverse entities such as chemokines and subunits of mitochondrial complexes. The remainder (~5340) have either not been functionally characterised or have not been annotated at all.

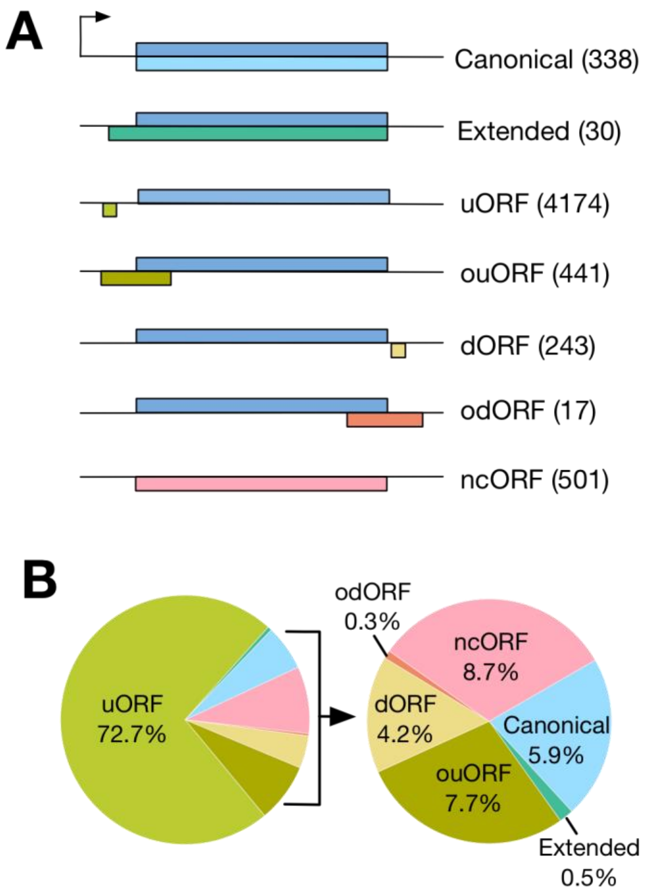


Figure 4.2 | **Identification of different classes of actively translated smORFs in this study.** (A) Predicted smORFs were classified into 7 groups according to their relative location in the host transcript. The number of translated smORFs in each class is shown in parentheses. (B) Pie chart showing the proportion of smORFs of different classes.

4.4 Start codon usage in smORFs

Alternate start codons (non-AUG) are very rare in eukaryotic genomes, naturally occurring non-AUG start codons have been reported for some cellular mRNAs (Ivanov et al., 2011). However, recent advancements in Ribo-Seq have revealed a strong enrichment (~60%) for non-AUG start codons at initiation sites (Ingolia et al., 2009, 2011). In a Ribo-Seq dataset of mouse embryonic stem cells, it has been shown that AUG is the most efficient start codon, followed by CUG, GUG, UUG, ACG, AGG, AUC, AUU, AAG, AUA (Ingolia et al., 2011). Another study also has shown near-cognate start codons CUG, ACG and AUU were frequently used in yeast, *Neurospora crassa* and mammalian cell line HEK293T (Kearse and Wilusz., 2017).

In our study, AUG is the most dominant start codon used by translated smORFs overall (**Figure 4.3A**), as well as in uORFs (**Figure 4.3B**), dORFs (**Figure 4.3C**) and ncORFs (**Figure 4.3D**). Not all near-cognate start codons are equally enriched, CUG are most enriched followed by GUG and UUG, also different smORFs have different distribution. Our pipeline has the flexibility to search for putative smORF with alternative start codons, it is possible to predict smORFs with a start codon beyond XUG.

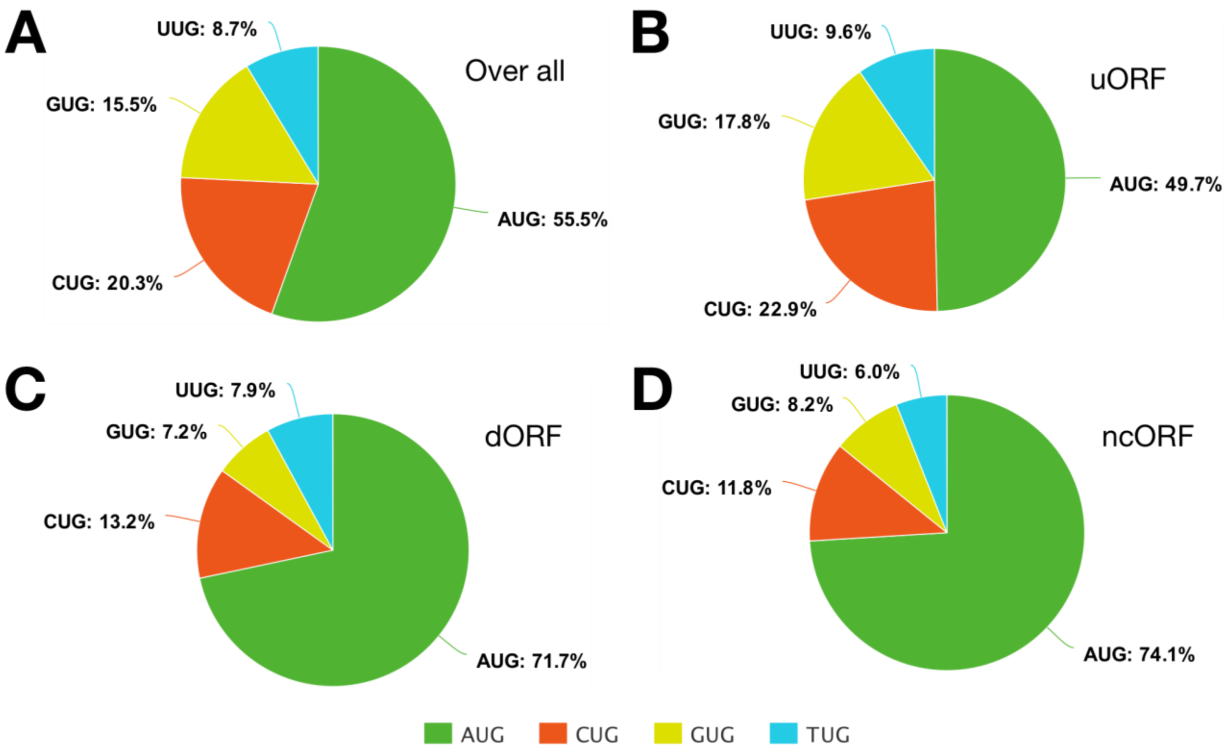


Figure 4.3 | **Distribution of AUG start codons and near-cognate start codons in smORFs.** AUG is the most dominant start codon used by smORFs overall (A), as well as in uORFs (B), dORFs (C) and ncORFs (D) near-cognate start codons are not enriched equally, with CUG being most enriched, followed by GUG and UUG.

4.5 smORF conservation

To examine the conservation of smORF-encoded micropeptides between species, we employed PhyloCSF to analyse signatures of evolutionary conservation. We prepared a cross-species nucleotide sequence alignment (or whole-genome multiple alignments) of smORFs as input to PhyloCSF. We used the Galaxy “Stitch Gene blocks” tool (Blankenberg et al., 2011) to extract alignments from pre-cached whole-genome multiple alignments cross 100 species. To match the 100-species alignments, we selected PhyloCSF “100 vertebrate-phylogenies” (https://github.com/mlin/PhyloCSF/wiki#available-phylogenies). PhyloCSF outputs a score that is positive if the alignment is likely to represent a conserved coding region and negative otherwise.

11.4% of smORFs showed strong evidence of conservation (PhyloCSF score > 50 which has been shown to accurately separate known protein-coding genes from known noncoding sequences) (**Figure 4.4A**), with canonical smORFs being enriched among them (**Figure 4.4B**). A small subset (~6.5%) of uORFs, ncORFs and dORFs shows high PhyloCSF scores, pointing to the smORFs that may produce functional micropeptides. There are over 60% of smORFs lacking signs of selective pressure to maintain their amino acid sequences (no cross-species sequence alignment and not conserved, **Figure 4.4A**), in which uORFs, ncORFs and dORFs are enriched (**Figure 4.4B**). The majority of smORFs are shorter than 100 nt (**Figure 4.5A**). The median length of canonical smORFs is 79 codons, however, the median length of uORF, dORF and ncORF are 24, 34 and 33 codons respectively. By comparison with other classes, canonical smORFs are, on average, longer and more highly conserved (**Figure 4.4C, Figure 4.5B**). Having distinct transcript organization, size, conservation and peptide structure, canonical smORFs, uORFs, dORFs and ncORFs are likely to have different cellular and molecular functions. Below we speculate the potential roles of individual classes.

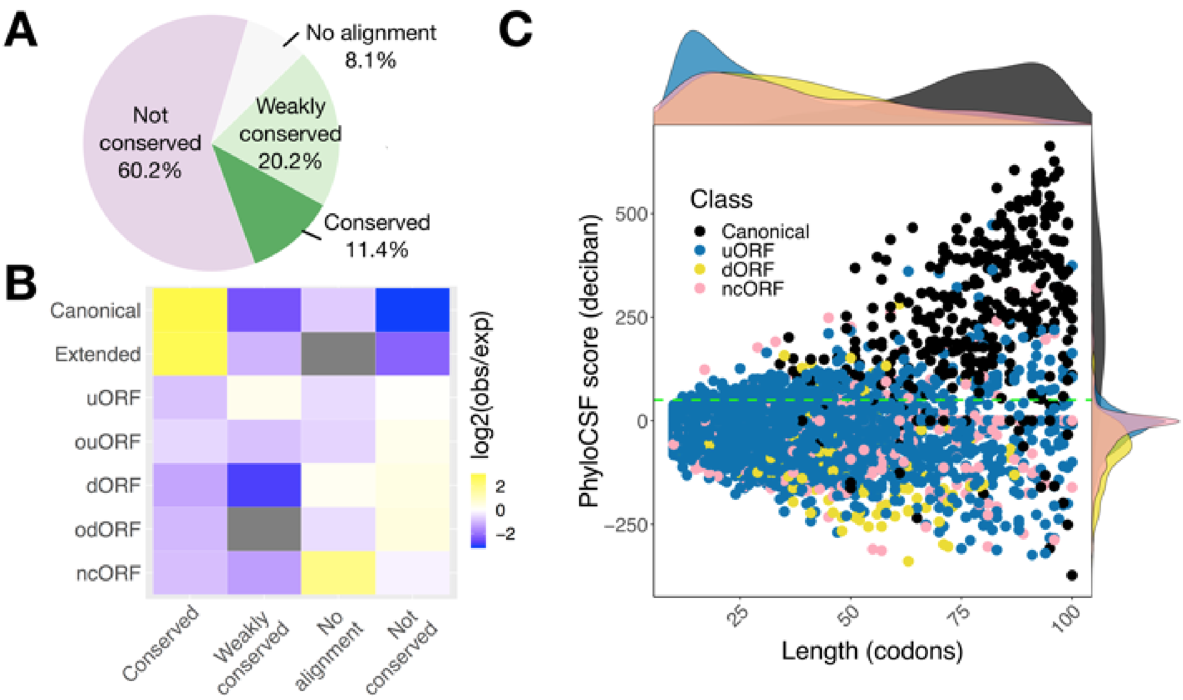


Figure 4.4 | **smORFs showing different conservation and length distributions according to their classes.** (A) Most smORFs are not conserved at the peptide level. Pie chart represents the coding potential (PhyloCSF score). smORFs with PhyloCSF score ≥ 50 are considered conserved. smORFs are considered weakly conserved if their PhyloCSF scores are positive but smaller than the threshold 50. (B) Canonical and extended smORFs are enriched in conserved peptides. Enrichment heatmap depicts log 2 ratio of the number of smORF observed (obs) to the number of smORF that would be expected (exp) by chance given overall distributions of smORF classes and conservation levels. (C) Scatter plot shows the distributions of codon length and PhyloCSF score for each smORF type. Marginal densities of length and PhyloCSF score are also shown on the top and the right-hand side of the scatter plot. Green dashed line indicates PhyloCSF score of 50. Here the original classification in Fig 3.1A was simplified by combining the canonical and canonical extended ORFs as canonical; uORF and ouORF as uORF; and dORF and odORF as dORF. Canonical smORFs are on average longer and more conserved than other type.

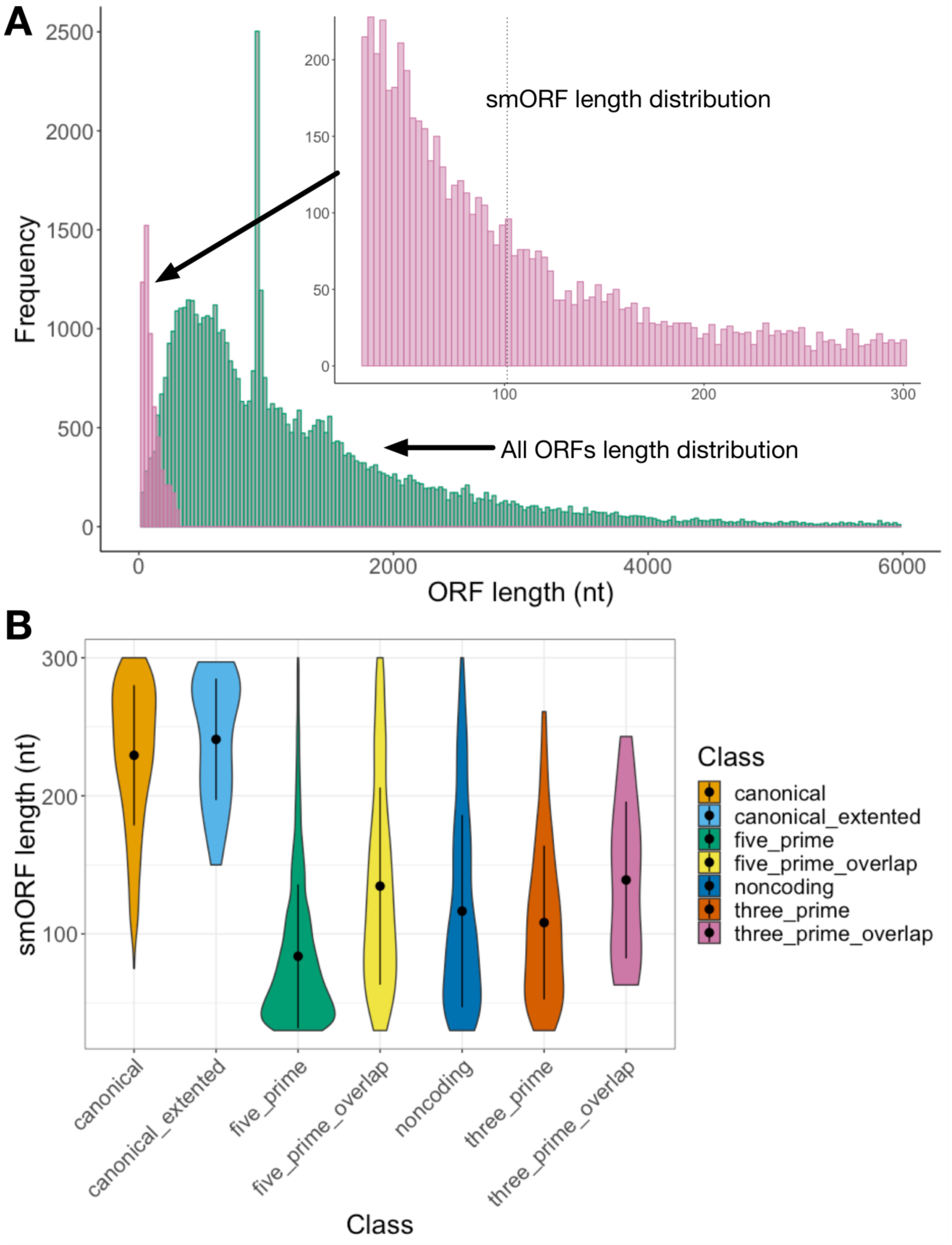

A
Frequency
2500
2000
1500
1000
500
0
0
2000
4000
6000
ORF length (nt)
smORF length distribution
200
150
100
50
0
100
200
300
All ORFs length distribution
B
smORF length (nt)
300
200
100
canonical
canonical_extented
five_prime
five_prime_overlap
noncoding
three_prime
three_prime_overlap
Class
Class
canonical
canonical_extented
five_prime
five_prime_overlap
noncoding
three_prime
three_prime_overlap

Figure 4.5 | **smORF length distribution.** (A) Overall smORF length distribution and all annotated ORF (GENCODE) length distribution. Majority of smORFs are shorter than 100 nt. (B) Violin plots of smORF length distribution for each class. Canonical and canonical extended are longer on average. Non-canonical smORFs tend to be shorter.

4.6 Canonical smORFs

A total of 338 canonical smORFs were predicted in B and T cells. The majority (88%) of these are conserved or weakly conserved between species (**Figure 4.6A**). As suggested by Couso and Patraquim (Couso and Patraquim, 2017), I divided canonical smORFs into “short CDS” and “short isoforms”. There are hundreds of putative short CDSs in mouse and human; they are located on monocistronic transcripts with higher probability; and their host transcripts are structurally shorter and simpler compared with canonical mRNAs (Couso and Patraquim, 2017). We have predicted 184 short CDSs and they have a median size of 79 codons. By contrast, short isoforms are the products of alternative splicing of transcripts from genes annotated as encoding proteins > 100 amino acids. We find these have a median size of 80 codons and resemble short CDSs in size and conservation (**Figure 4.6B**). However, short isoforms are distinct from short CDSs in that they share conserved amino acid sequences with their long canonical protein isoforms, thus they have the potential for functions that are directly related to their longer protein isoforms. Among the predicted canonical smORFs, 54.4% are short CDSs and 45.6% are short isoforms.

I calculated the translation efficiency of short CDSs and short isoforms. When compared to long CDSs of expressed protein-coding transcripts, we found their median translation efficiency to be greater (**Figure 4.6C**) for LPS-activated B cells. This is also the case for other conditions in our datasets). We also conducted GO term enrichment analysis separately for 159 short CDS and 136 short isoforms against 3481 background genes. The top hits of short CDS are related to chemokine activity and mitochondrial biology (**Figure 4.6D**). Seven chemokines are predicted (Ccl1, Ccl22, Ccl3, Ccl4, Ccl5, Cxcl10, Cxcl11). We also see gene products enriched in mitochondrial complexes, for example, Uqcr10 is a subunit of Coenzyme Q:cytochrome c oxidoreductase (Complex III); this complex has a critical role in oxidative phosphorylation pathway for the generation of ATP. Another mitochondrial protein is Romo1, which is located in the mitochondrial membrane and is responsible for increasing the level of reactive oxygen species (ROS) in cells (Na et al., 2008). Romo1 also has antimicrobial activity against a variety of bacteria by penetrating the bacterial membrane (Sha et al., 2012). Short isoform encoding genes are associated with a broad range of GO terms, however, there are no GO biological processes terms enriched in for them.

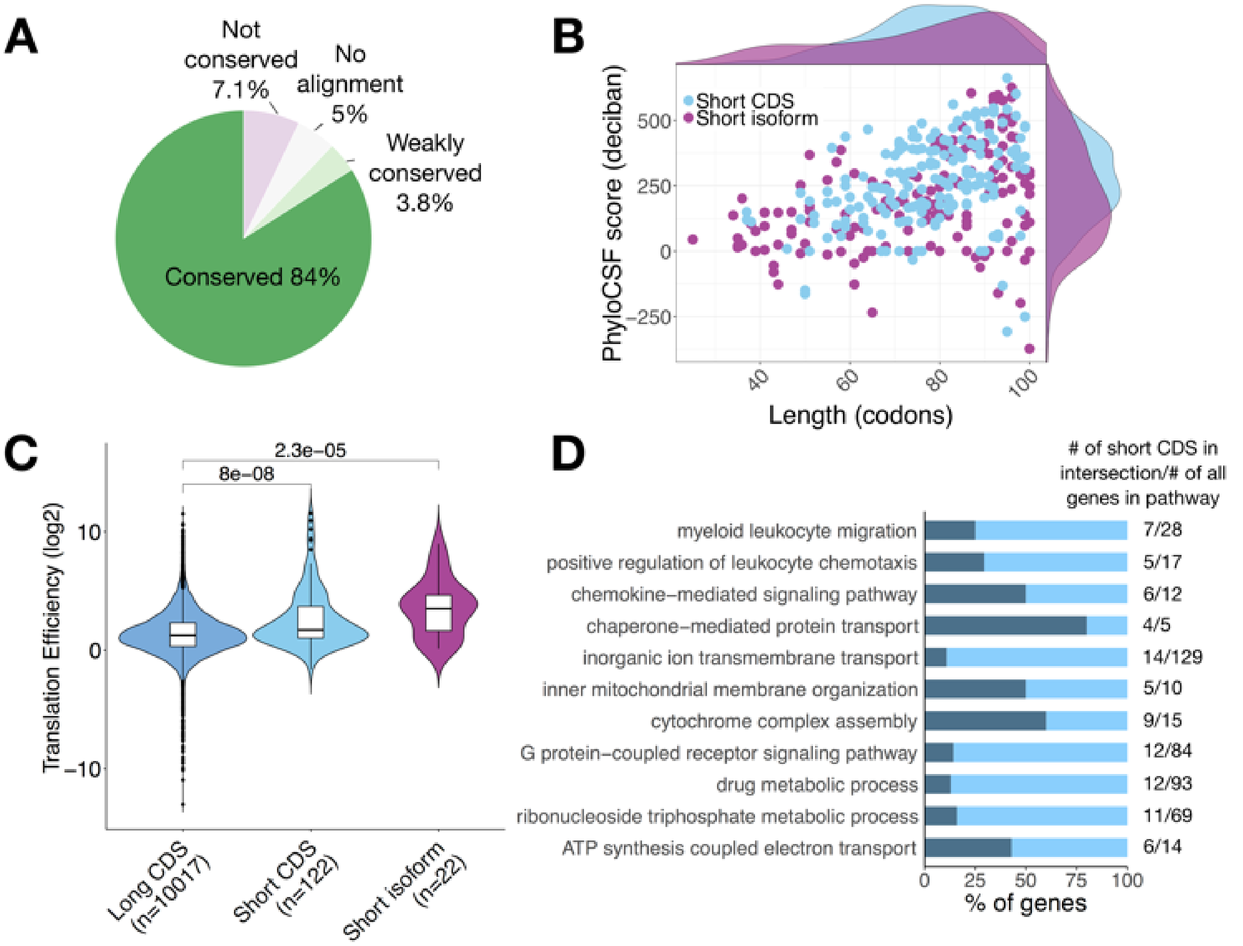


Figure 4.6 | **Canonical smORFs consist of short CDSs and small isoforms.** (A) Pie chart shows 87.8% of canonical smORFs are conserved or weakly conserved. (B) Canonical smORFs were further divided to short CDSs (54.4%) and small isoforms (45.6%). Short CDSs are annotated ORFs of 100 codons or fewer. Small isoforms are ORFs of 100 codons and fewer, which are products of alternative splicing of canonical mRNAs. (C) Translation efficiency (log2) distributions of long CDS (CDSs greater than 100 codons), short CDSs and small isoforms. Mean and standard deviation are shown. Significance was computed using two-sided Wilcoxon test. (D) Biological process gene ontology terms found to be significantly enriched in the short CDS gene list.

4.7 uORFs

Approximately 50% of annotated animal mRNAs contain uORFs (Andrews and Rothnagel, 2014; Johnstone et al., 2016; Couso and Patraquim, 2017) and translation of uORFs has been widely reported in different organisms (Wang et al., 2004; Calvo et al., 2009; Johnstone et al., 2016). We have predicted 4615 translated uORFs (including ouORFs) and about 30.4% of these are considered conserved or weakly conserved (**Figure 4.7A**). We observed that the median translation efficiency of uORFs is greater than that of long CDS (**Figure 4.7B**). About 4% of the uORFs have a high PhyloCSF score and TE (above the median TE of long CDS) and potentially encode conserved functional micropeptides (**Figure 4.7C**, for LPS-activated B cells). However, the sequences of the majority of uORFs are not conserved, suggesting that any potential function is largely independent of the encoded peptide. It has been demonstrated uORFs may regulate the translation of the downstream CDS. Several studies have shown a repressive effect of uORFs on the translation of CDS (Johnstone et al., 2016; Chew et al., 2016; Zhang et al., 2019). The proportion of expressed uORF-containing transcript in B cells and T cells is between 6.2% and 12.4%, except resting B cells (2.7%), we analysed the effect of uORFs on mRNA translation by comparing the translation efficiency of the CDS in all uORF-containing transcripts versus those lacking uORFs. As expected, the presence of uORFs and overlapping uORFs was associated with a translation repression (**Figure 4.7D**). We performed GO enrichment analysis for all uORF-containing genes to discover their associated biological processes (2881 target genes against 3481 background genes) and these genes are mostly enriched in protein modification process, regulation of gene expression and cellular response to stimulus (**Figure 4.7E**). This indicates that uORF-containing genes are broadly involved in complex biological pathways such as protein or RNA production and cell signalling. Regulatory uORFs may be suited to allow the rapid expression of genes in response to stress and environmental stimuli.

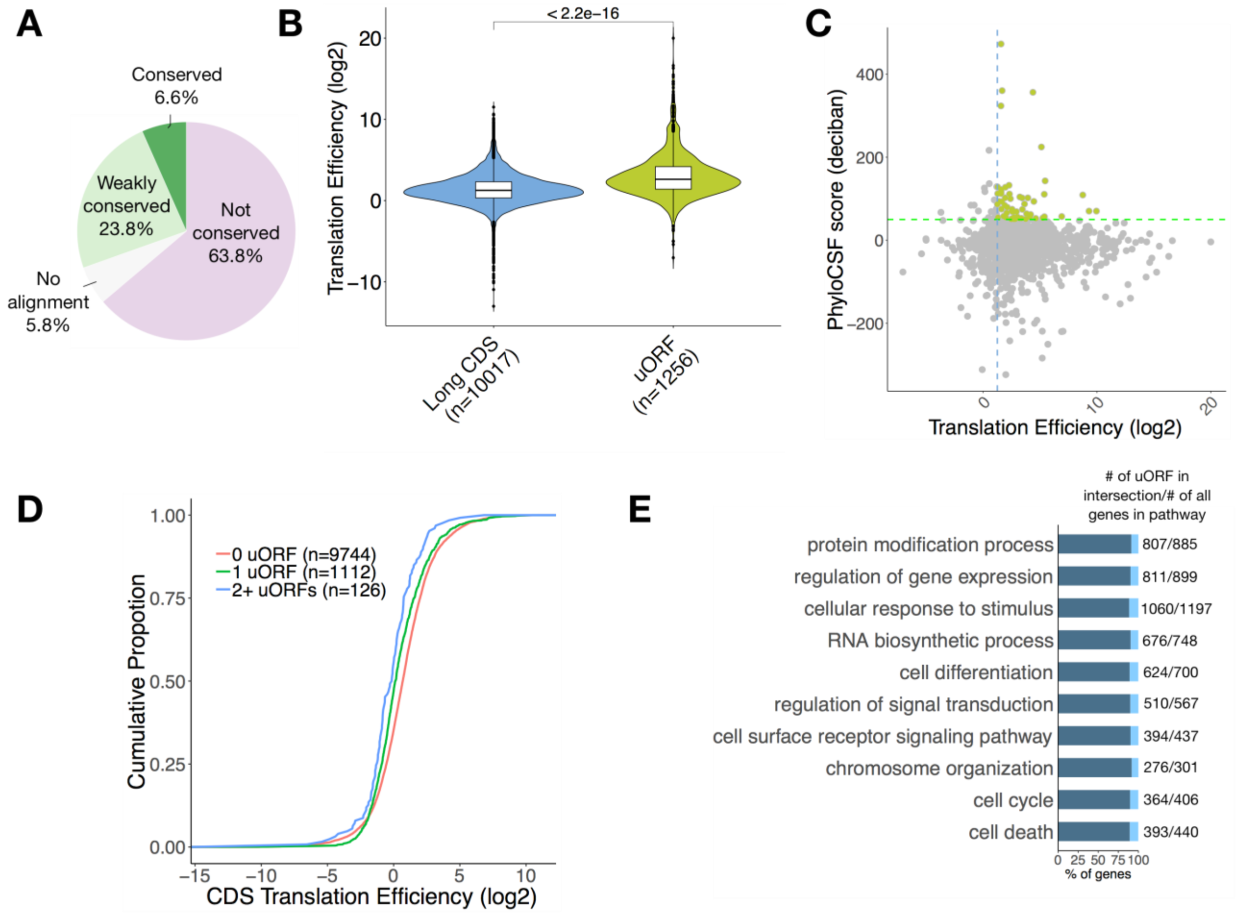


Figure 4.7 | **uORFs regulate the translation of their downstream CDS.** (A) Pie chart shows 14.4% of noncoding smORFs are conserved or weakly conserved. (B) Translation efficiency distributions of long CDS and uORF. Significance was computed using two-sided Mann-Whitney test. (C) Scatter plot of uORF translation efficiency and PhyloCSF score. Green broken line represents a PhyloCSF score value of 50 used as a threshold for conservation, blue broken line represents the median TE of long CDS. uORFs that are conserved and having high TE are highlighted. (D) Cumulative distribution of translation efficiency in expressed uORF-containing transcripts versus transcripts lacking uORFs as control. Significance was computed using two-sample Kolmogorov–Smirnov test for each uORF set compared to the control. (1 uORF P = 1.321e-14, 2+ uORFs P = 1.828e-6). (E) Biological process gene ontology terms found to be significantly enriched in the uORF-containing gene list.

Dynamic regulation of CDS by uORFs

We also investigated the influence of uORFs on the downstream CDS during the first few hours of T cell activation (**Figure 4.8**) and between LPS-activated and resting B cells. Upon T cell activation, RNA abundance is increased (comparison between 2h vs 0h and 4h vs 2h) for both non-uORF-containing (P < 2.2E-16, Mann-Whitney test) transcripts and transcripts containing 1 uORF (P = 1.349E-11, it is not statistically significant for transcripts containing 2+ uORFs, P = 0.5179). However translational efficiency is decreased for all transcripts (P < 2.2E-16 for non-uORF-containing transcripts; P < 2.2E-16 for transcripts containing 1 uORF; P = 0.001625 for transcripts containing 2+ uORFs). We then noticed that the RPF abundance of uORF-containing transcripts did not change (P = 0.1573 for transcripts containing 1 uORF; P = 0.3934 for transcripts containing 2+ uORFs; but statistically significant for non-uORF-containing transcripts, P = 2.034E-12). As translation efficiency is the level of mRNA translated into protein, it suggests that by 4 hours, transcription is increased, but translation machinery is limited and not caught up. Further experimental evidence will be required to validate the hypothesis.

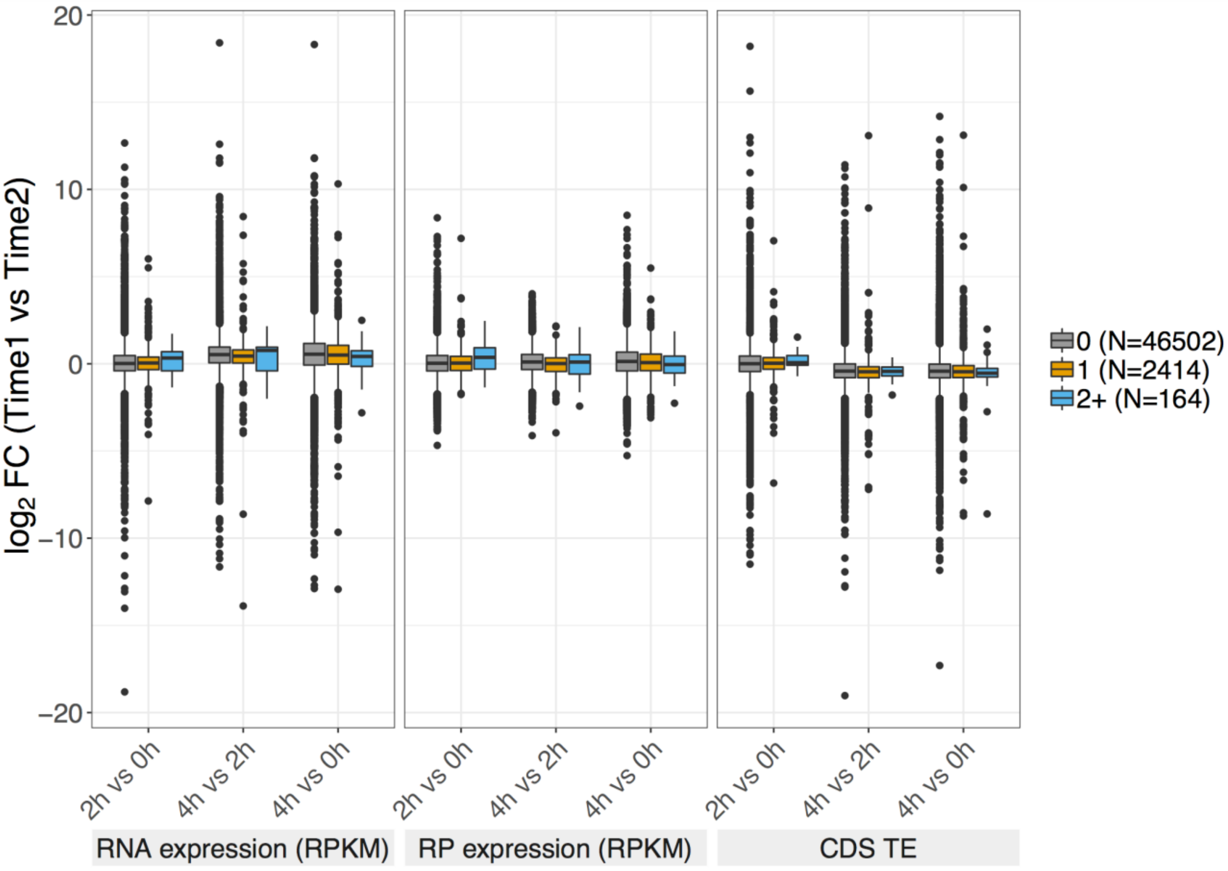


Figure 4.8 | **Dynamic regulation of downstream CDSs by uORFs during T cell activation.** Logarithmic transformed fold change of CDS RNA abundance (RPKM), RPF abundance (RPKM) and TE between two time points (2h vs 0h, 4h vs 2h and 4h vs 0h) in 0,1, 2+ uORF-containing transcripts.

4.8 smORFs in non-coding RNAs

Non-coding ORFs (ncORFs) are smORFs that are found in annotated long non-coding RNAs (lncRNAs) and pseudogenes. They are typically short with an median length of 33 codons. By definition, non-coding RNAs are not translated into protein. However, annotated lncRNAs have been predicted from their sequences to contain six smORFs on average (Couso and Patraquim, 2017). We have predicted 501 translated ncORFs and about 14.4% of these are considered conserved or weakly conserved (**Figure 4.9A**). We noticed very different distributions of size and PhyloCSF score between ncORFs and canonical smORFs (**Figure 4.9B**). The translation efficiency distribution for ncORFs is also different from long CDS, the median TE of ncORFs is greater than long CDS (**Figure 4.9C**). Three ncORFs Cct6a, Gm16675 and 6330418K02Rik were found to have a high PhyloCSF score (> 100) and TE ($\log_2$TE > 2), so we infer them to be functional micropeptides (**Figure 4.9D**). We searched the micropeptides they encode in NCBI BLASTp database (Altschul et al., 1990), but did not find any match for Gm16675. The 6330418K02Rik gene is annotated as an antisense lncRNA gene in GENCODE, only one match was found for its predicted micropeptide (124 AA). The micropeptide was fully aligned to part of an uncharacterized protein of 201 AA with 88.5% identity in *Mus caroli*. The third smORF's host transcript Cct6a (chaperonin containing Tcp1-subunit 6a), which is annotated as a processed transcript (defined as a noncoding transcript that does not contain an ORF) and has a human ortholog which is annotated as protein-coding and encodes two isoforms of 486 AA and 532 AA in length respectively. The Cct6a micropeptide (45 AA) was 100% identical to a small part of human CCT6A protein, which may suggest this micropeptide has a function in a protein complex.

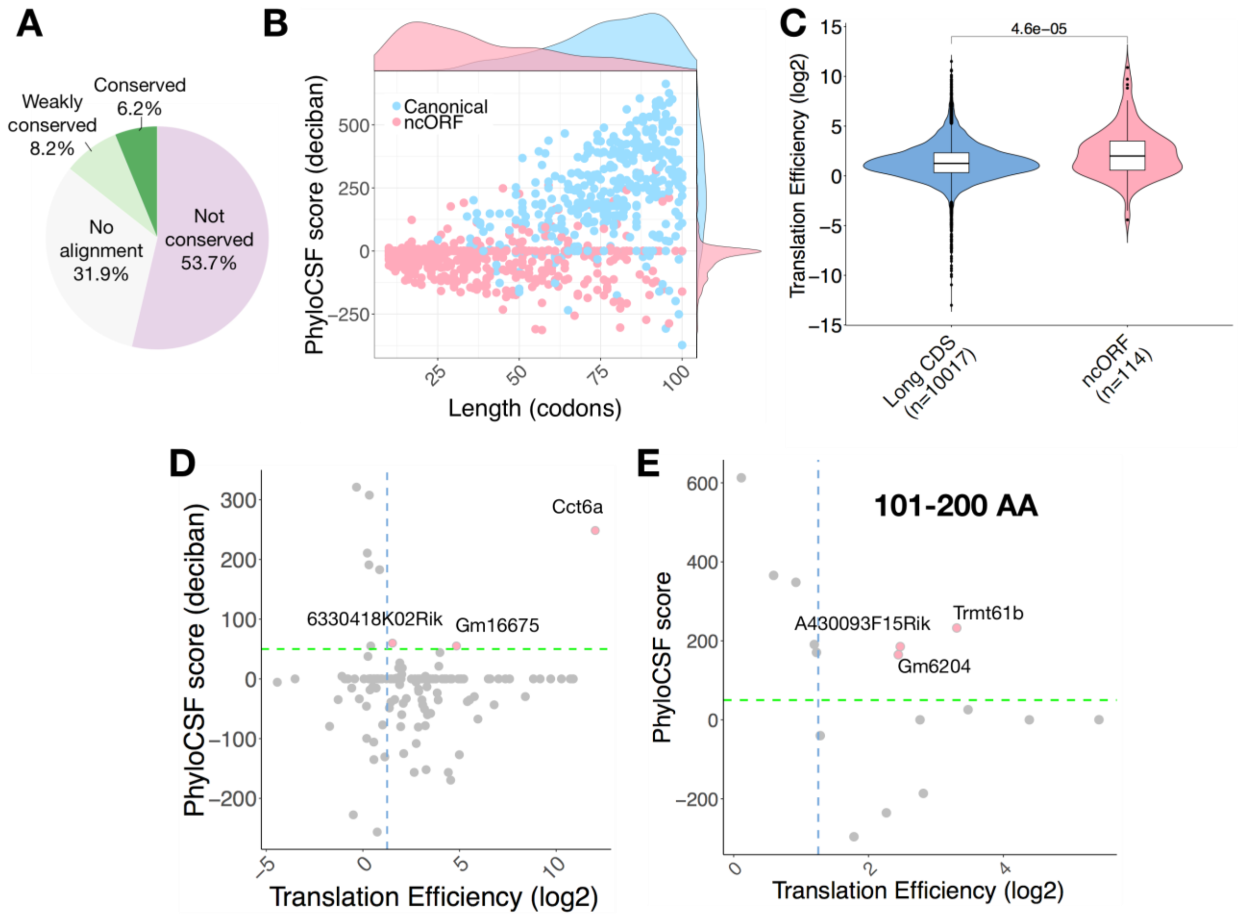


Figure 4.9 | **Translated smORFs predicted in noncoding RNAs.** (A) Pie chart shows 14.4% of noncoding smORFs are conserved or weakly conserved. (B) Canonical smORFs and ncORFs showing very different distributions in length and PhyloCSF score. (C) Translation efficiency distributions of long CDS and ncORF. Significance was computed using two-sided Mann-Whitney test. (D) Translation efficiency and PhyloCSF score are shown for ncORFs (LPS-activated B cells). Scatter plot of ncORF translation efficiency and PhyloCSF score. Green broken line represents a PhyloCSF score value of 50 used as a threshold for conservation, blue broken line represents the median TE of long CDS. ncORFs that are conserved and having high TE are highlighted. Three genes (Cct6, 6330418K02Rik, Gm16675) potentially encode micropeptides. (E) Translation efficiency and PhyloCSF score are shown for ncORFs of 101-200 codon in length (T cell activation). Three genes (Trmt61b, A430093F15Rik, Gm6204) potentially encode proteins between 101 and 200 AA.

4.9 dORFs

243 downstream ORFs and 17 overlapping downstream ORFs were predicted. The median length is 34 AA. Only 20 (~7.7%) are conserved or weakly conserved. The translation efficiency of dORFs is lower than the long CDSs in general. In transcripts that contain multiple ORFs, a translation re-initiation mechanism is able to prevent recycling of some or all ribosome subunits upon termination of the first translated ORF and thereby enable the translation of the dORF (Gunišová et al., 2018). The low TE indicates a very low level of translational re-initiation after the stop codon of the upstream CDS. We are also interested to know whether or not dORFs play a regulatory role for their upstream CDSs' translation similar to uORFs for their downstream CDSs. We compared the TE of the CDS in all dORF-containing transcripts versus those lacking dORFs, however we did not find evidence of dORFs having a repressive effect for CDSs (**Figure 4.10**).

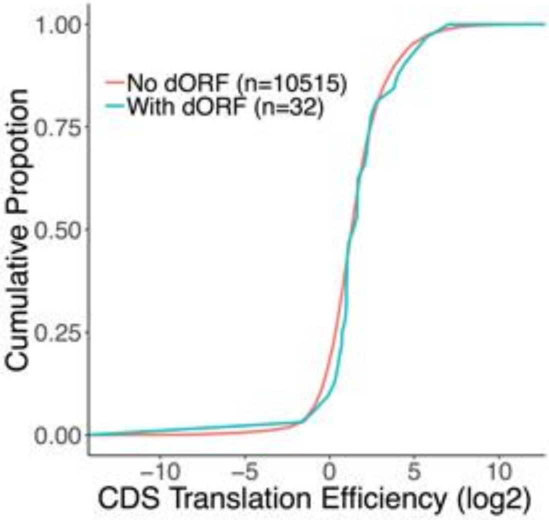


Figure 4.10 | **dORF-containing transcripts are not translationally repressed.** Cumulative distribution of translation efficiency in expressed dORF-containing transcripts versus transcripts lacking dORFs as control. Significance was computed using two-sample Kolmogorov–Smirnov test, P = 0.437.

4.10 Signal sequence containing micropeptides

An N-terminal signal peptide sequence of 16-30 amino acids is characteristic of proteins destined to be secreted or resident within cellular membranes. We predicted the presence of signal peptides in amino acid sequences of micropeptides using SignalP server (Petersen et al., 2011; Armenteros et al., 2019). This predicted 80 candidates including known chemokines (CCL-1, -2, -4, -5 and -22) and the cell surface protein CD52, as well as a recently identified lncRNA encoded Aw112010 micropeptide (Jackson et al., 2018) (**Table 4.1**). In total, we predicted 28 canonical micropeptides which typically have high levels of conservation. By contrast, the majority (77%) of non-canonical micropeptides have poor conservation (**Figure 4.11A**).

Source	**Annotation status**	**Total number of proteins**	**Signal peptide containing proteins**	**Secreted proteins**
Human proteome – all	UniProt reviewed (manually annotated)	20,365	3,596 (17.7%)	1,864 (9.2%)
Human proteome - micropeptides		745	157 (21.1%)	137 (18.4%)
Mouse proteome - all		17,038	3,153 (18.5%)	1,438 (8.4%)
Mouse proteome - micropeptides		485	146 (31.9%)	133 (29%)
Mouse lymphocyte micropeptides	Predicted	5,744	80 (13.9%)	15 (0.3%)

Table 4.1 | **Proportion of signal peptide containing protein/micropeptide in different datasets.** Human and mouse proteome information was queried from UniProt (date to 03 May 2020), and only manually annotated data was considered. Signal peptide containing proteins include secreted proteins and other types (e.g. transmembrane proteins). Our mouse lymphocyte micropeptides are predicted using our in-house pipeline and SignalP server.

RNA expression in a particular cell is a proxy for protein expression when direct quantitative information for the proteins of interest is not available. In order to examine the secreted micropeptide-host transcript expression in lymphocytes in various conditions, we used public mouse RNA-Seq datasets of B cell terminal differentiation including Follicular B cell (FoB), marginal zone B cell (MZB), B-1 cell (B1), germinal center B cell (GCB), spleen plasma cell (SplPC) and bone marrow plasma cell (BMPC) (Shi et al., 2015), Th1 cell activation at three time points (0h, 2h and 4h) (Davari et al., 2017), resting and activated regulatory T cells (Luo et al., 2016) as well as an epidermis cell dataset (Sendoel et al., 2017). These data revealed dynamic expression patterns for several of the host transcripts (**Figure 4.11B**). For example, BC031181 was downregulated during B cell development but upregulated during Th1 cell activation, it was also highly expressed in epidermis cells (**Figure 4.11B**). Host transcript expression pattern provides a lead to where and what stage of cell differentiation micropeptides may be produced and can help with experimental validation of micropeptide prediction.

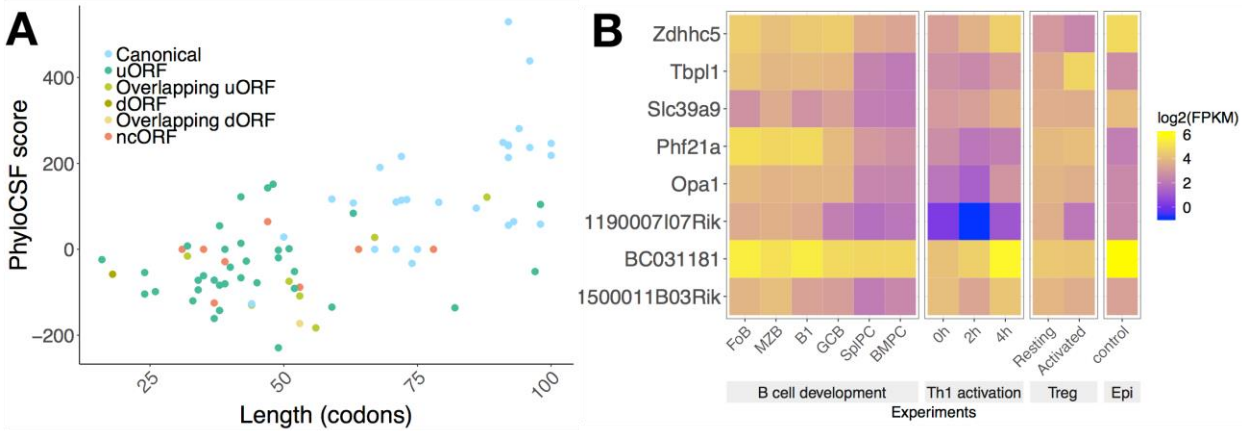
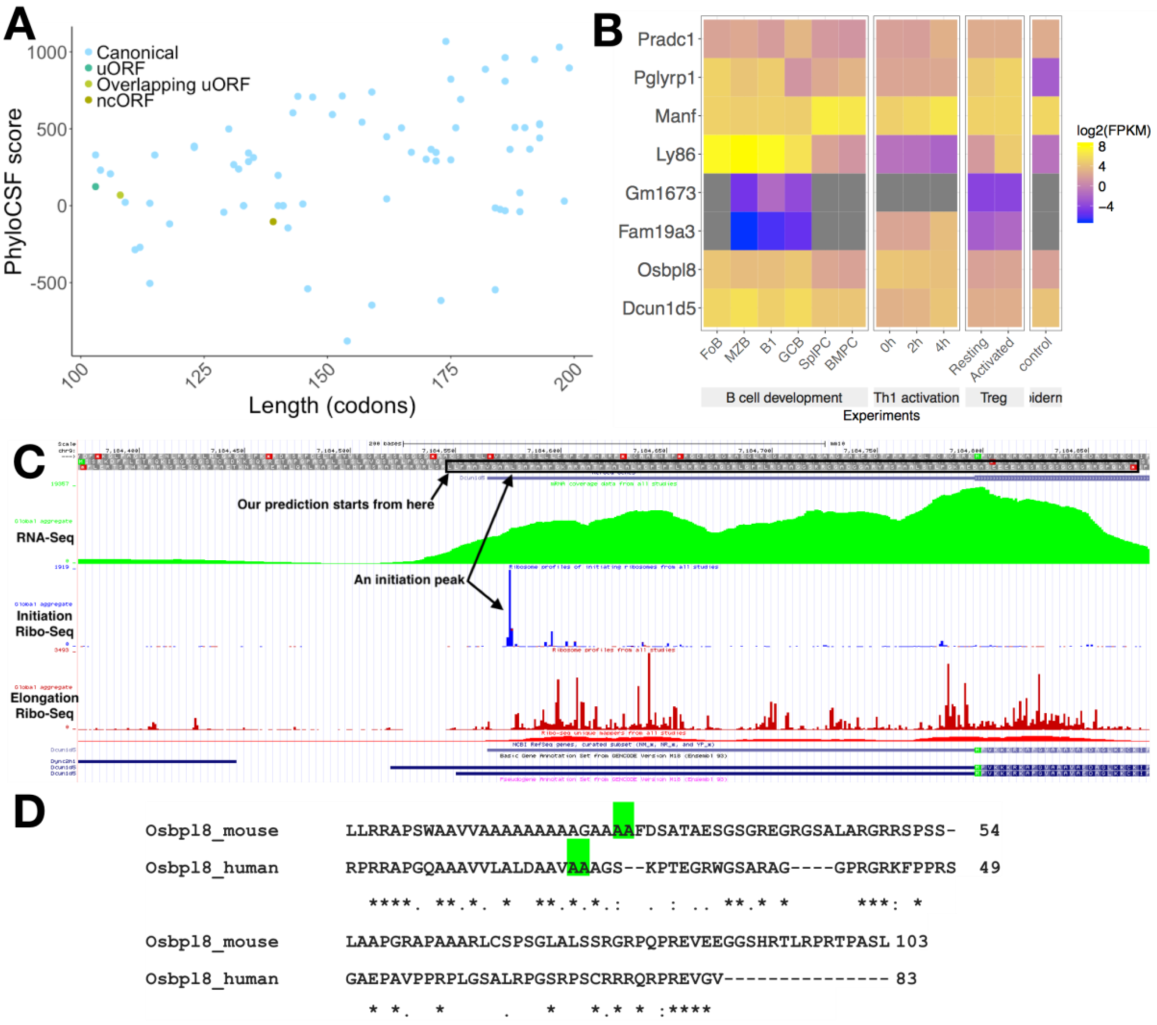


Figure 4.11 | **Predicted signal sequence containing micropeptides and their host transcripts expression under different conditions.** (A) Scatter plots show the distributions of length (codon) and PhyloCSF score for each predicted signal peptide containing micropeptides. (B) Heatmap analysis of host transcript expression during B cell terminal differentiation, Th1 cell activation, resting/activated regulatory T cells and epidermal cells (Epi). Selected micropeptides are shown in the heatmap, they are conserved in humans and there is limited or no information regarding their function. They are ordered by the length.

4.11 Signal sequence containing proteins of 101-200 AA

The power of ribosome profiling is not limited to detecting smORFs but can detect novel ORFs encoding larger proteins. We therefore used our pipeline to predict candidate signal sequence containing proteins of between 101 and 200 AA in length. 74 candidates were predicted, among which 71 are annotated and 3 are unannotated ORFs (**Figure 4.12A**). The 71 annotated peptides included the chemokine Ccl9, interferon gamma and four interleukins (Il3, Il13, Il17, Il22) which confirmed the predictive power of our pipeline. We also predicted mesencephalic astrocyte-derived neurotrophic factor (Manf) which was originally identified as a secreted trophic factor for dopamine neurons (Petrova et al., 2003). In the expression heatmap, we noticed that Phf21a and Ly86 are downregulated during B cell differentiation, and Manf is upregulated during B cell differentiation as well as Th1 cell activation at RNA level (**Figure 4.12B**), it indicates that Manf might play a role in plasma cells. The three unannotated ORFs are an uORF (103 AA) in Osbpl8; an overlapping uORF (108 AA) in Dcun1d5 and a ncORF (139 AA) in 4930481A15Rik. All three unannotated ORFs start from CUG. PhyloCSF analysis shows that both uORFs are conserved (PhyloCSF score for Osbpl8 uORF is 147.4 and for Dcun1d5 ouORF is 61.5) but the ncORF is not. Further manual curation has shown that, for Dcun1d5 ouORF, the true translation might start from the downstream of the start codon we predicted. We looked at the published initiation Ribo-Seq data in GWIPZ (Michel et al., 2013) and noticed a peak at a downstream CUG of the predicted ORF (**Figure 4.12C**). However, signal sequence was not predicted in the resulted truncated protein. We also found that the Osbpl8 uORF human sequence was predicted to contain a signal peptide, but the alignment between mouse and human became poor towards the C-terminus (**Figure 4.12D**), indicating that the functional domain is not conserved.



```
Osbpl8_mouse        LLRRAPSWAAVVAAAAAAAAAGAAAAFDSATAESGSGREGRGSALARGRRSPSS-   54
Osbpl8_human        RPRRAPGQAAAVVLALDAAVAAAGS--KPTEGRWGSARAG----GPRGRKFPPRS   49
                      ****. **.*. *  **.*.*.:  . : .. **.* *      ***: *

Osbpl8_mouse        LAAPGRAPAAARLCSPSGLALSSRGRPQPREVEEGGSHRTLRPRTPASL 103
Osbpl8_human        GAEPAVPPRPLGSALRPGSRPSCRRRQRPREVGV--------------- 83
                     * *.  *     .   *   *.* * :****
```

Figure 4.12 | **Predicted signal sequence containing proteins of 101-200 AA.** (A) Scatter plots show the distributions of length (codon) and PhyloCSF score for each predicted signal peptide containing proteins. (B) Heatmap analysis of host transcript expression during B cell terminal differentiation, Th1 cell activation, resting/activated regulatory T cells and epidermal cells. Selected proteins are shown in the heatmap, they are conserved in humans and there is limited or no information regarding their function. They are ordered by the length. (C) Initiation Ribo-Seq peak shows there exists a truncated protein of the Dcun1d5 ouORF predicted by ORFLine. (D) Pair-wise amino acid alignment between mouse and human of Osbpl8 uORF. The alignment becomes poor towards C-terminus. Predicted cleavage site is highlighted in green.

4.12 Functional validation of candidate secreted micropeptides

Being an immunology lab, we are particularly interested in knowing whether the micropeptides are secreted, as they might be candidate immunoregulators. We have further established the prediction of signal sequence containing micropeptides and proteins (101-200 AA). Among the candidate secreted micropeptides, we selected eight to test their secretion (**Table 4.2**). They were prioritized as they are conserved in humans as well as having limited or no information regarding their function. Among the longer proteins (101-200 AA), Manf has shown a striking expression profile in plasma cells (**Figure 4.13A**).

Gene name	**Class**	**Length**	**AA sequence**
Zdhhc5	uORF	37	**MSYTLICLTLHGFHLQLFACIQPTVC**LHVLNCTSCVS
Tbpl1	uORF	42	**METGERTRFIFILVLQLLLRVRR**NQQQRCRRVLYDRPVFPRM
Slc39a9	uORF	43	**MKRCHLAAMAAVVLATQGQGLA**EGSTMGSTGCRAETASCRLCC
Phf21a	uORF	51	**MKKSSLLLLLLLLLLRVPASS**CQGGQPASSRRGTGELKERQLLQNWTSQNL
Opa1	uORF	67	**MRHWEGLGGCSMPLLLRA**SSWVIVGAGIGLGPTRGSPRGRLSACVWSALAGCGEQVGRPWPVKSANP
1190007I07Rik	Canonical	68	**MPGGVPWSAYLKMLSSSLLAMCAGA**QVVHWYYRPDLTIPEIPPKPGELKTELLGLKERRHEPHVSQQ
BC031181	Canonical	72	**MVCIPCIVIPVLLWIFKKFLEP**YIYPVVSRIWPKKAVQQSGDKNMSKVDCKGAGTNGLPTKGPTEVSDKKKD
1500011B03Rik	Canonical	72	**MLRSGWMRLLPMLCSLLLGRA**EAPSPGVPPEQSQPYAVLRRQSLVLMGTIFSILLVTVLLMAFCVYKPIRRR

Table 4.2 | **Eight candidate secreted micropeptides.** Signal peptide is in bold.

4.13 CRISPR/Cas9-mediated knockout of Manf in plasmablasts

From the prediction of secreted proteins within the 101-200 range, we found the 182 AA mesencephalic astrocyte-derived neurotrophic factor (MANF). MANF was firstly described as a survival-promoting factor for embryonic midbrain dopaminergic neurons (Petrova et al., 2003). It is later reported to have high expression in *Drosophila* hemocytes and CD11b$^+$ innate immune cells in mouse (Neves et al., 2016), as well as human peripheral white blood cells (Chen et al., 2015), human plasma cells and macrophages in the spleen (Liu et al., 2015). Recent studies suggest that MANF plays a role in endoplasmic reticulum (ER) stress response (Tadimalla et al., 2008; Glembotski et al., 2012; Cheng et al., 2013; Zhao et al., 2013; Yan et al., 2019). Under most conditions, MANF was not secreted but was retained in cells, however, its expression and secretion were upregulated by ER calcium depletion (Glembotski et al., 2012), secreted MANF could function in an autocrine and/or paracrine manner to protect cells from death in response to ER calcium depletion (Apostolou et al., 2008). To achieve the secretion of extensive levels of immunoglobulins, the ER of plasma cells undergoes expansion in a process that requires continuous ER stress and activation of the unfolded protein response. The expression of MANF in plasma cells may be required to maintain ER homeostasis (Cheng et al., 2013). Recently MANF was reported to act directly on immune cells and modulate their inflammatory phenotype by reducing pro-inflammatory signalling and promote pro-reparative activation of macrophages (Sousa-Victor et al., 2018).

It has shown that *Manf* expression is upregulated during B cell differentiation (**Figure 4.12B**). In addition, the Immunological Genome Project (Heng et al., 2008; Yoshida et al., 2019) RNA-Seq data also reveals Manf mRNA is most abundant in spleen plasma cells (**Figure 4.13A**). *Manf* is found to be expressed in human lymphoma cell lines and primary B cells (unpublished RNA-Seq) (**Figure 4.13B**) as well as in human tissues (Genotype-Tissue Expression (GTEx) project RNA-Seq) (Yizhak et al., 2019) (**Figure 4.13C**). MANF was also identified by mass spectrometry in a proteomics study of plasmablasts (unpublished, in this M/S datasets, in total 2853 annotated proteins are identified, among them, 27 are micropeptides. 13 out 27 are predicted by my pipeline as canonical smORFs) carried out in the Turner lab. Based on this information, we brought forth a hypothesis that *Manf* plays a role in plasmablasts.

To test the hypothesis, my colleague David Turner performed CRISPR/Cas9-mediated knockout of *Manf* during mouse B cell differentiation to plasmablasts (CD138+) (Nojima et al., 2011). We observed a reduction in the proportion of cells that were plasmablasts compared to the non-targeting guide RNA on day 8 (**Figure A.1A** in Appendix B). As a positive control we observed a reduced proportion of cells were plasmablasts upon CRISPR/Cas9 knockout of the transcription factor IRF4 which initiates plasmablast differentiation (Sciammas et al., 2006) and is required for plasmablast survival (Tellier et al., 2016). Our assay did not distinguish between a role for *Manf* in differentiation or reduced survival of cells that have differentiated. To test whether *Manf* has a cell intrinsic effect we co-cultured Cas9+ and Cas9- B cells and observed that CRISPR/Cas9 knockout of *Manf* specifically lead to a reduction in the proportion of plasmablasts in the Cas9+ population (**Figure A.1B**). Previous studies have described a cell intrinsic role for *Manf* in response to ER stress through interaction with the major ER chaperone GRP78 (Cheng et al., 2013; Lindström et al., 2016). The loss of MANF protein may lead to an aberrantly regulated unfolded protein response and consequently reduced plasmablast viability.

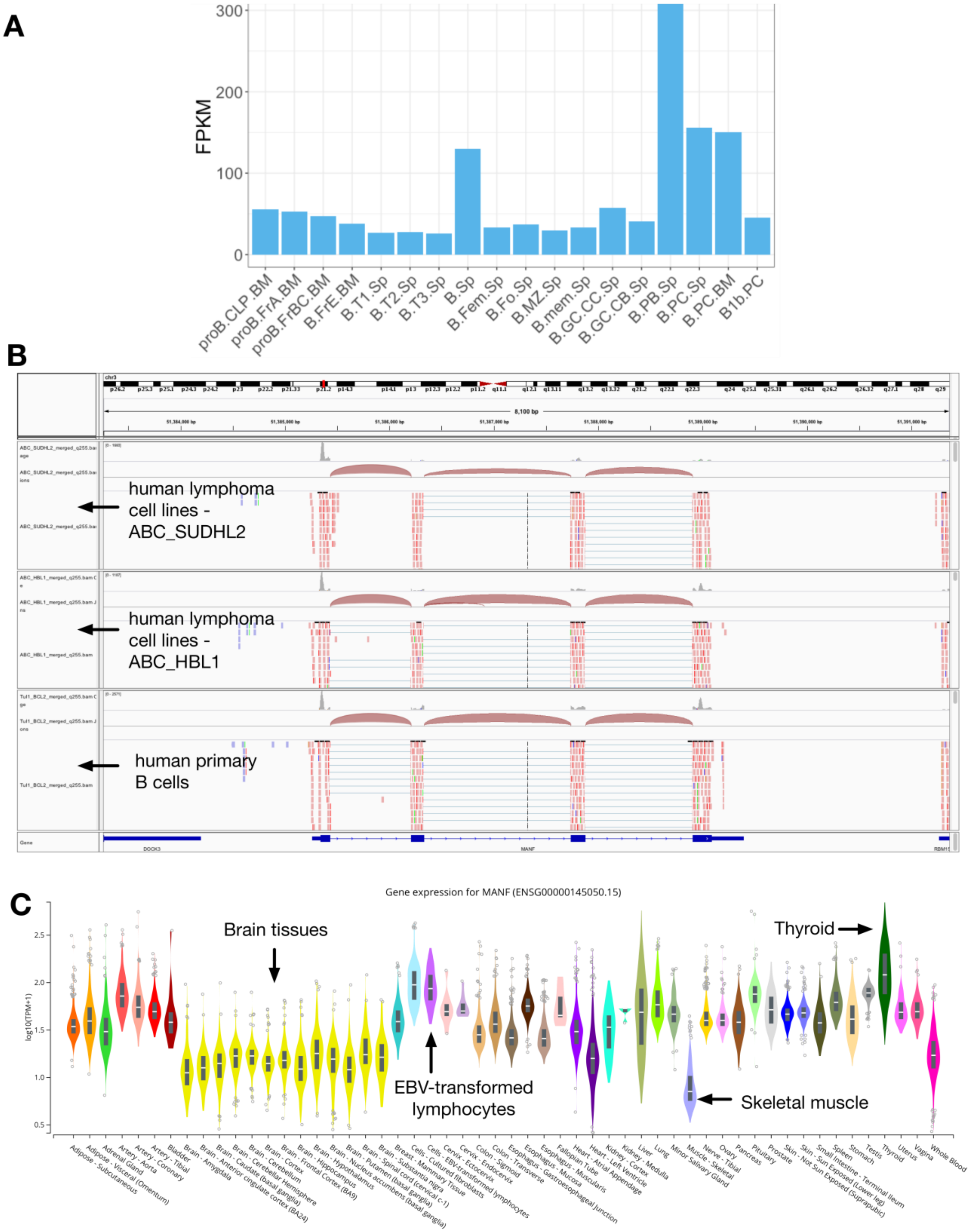


Figure 4.13 | **Manf RNA expression patterns.** RNA-Seq data has shown Manf were widely expressed in (A) mouse immune cells (Immunological Genome Project), and much higher in

spleen plasma cells (B.PB.Sp), (B) human lymphoma cell lines (ABC SUDHL2 and HBL1) and primary B cells (unpublished data from collaborator) and (C) human tissues (Genotype-Tissue Expression project or GTEx). It shows MANF expression (log transformed TPM value) in various tissues and cell types (colours based on tissue types), for example, MANF is highly expressed in EBV-transformed lymphocytes (purple) and thyroid (dark green), but lowly expressed in brain tissues (yellow) and skeletal muscle cells (light blue).

4.14 Expression of secreted micropeptides

We designed an expression vector which the synthetic DNA of smORFs can be cloned into (see Materials and Methods). smORFs will be translated and secreted peptides will be produced in mammalian cell lines and harvested from tissue culture supernatants. The design of the ORFs will allow a polypeptide protein tag (e.g. epitope tags) to be added, the resulting epitope tag allows the antibody to find the micropeptide for localization (e.g. cell surface binding), and further molecular characterization, e.g. *in vitro* assays of proliferation, survival, differentiation and chemotaxis (**Figure 4.14**).

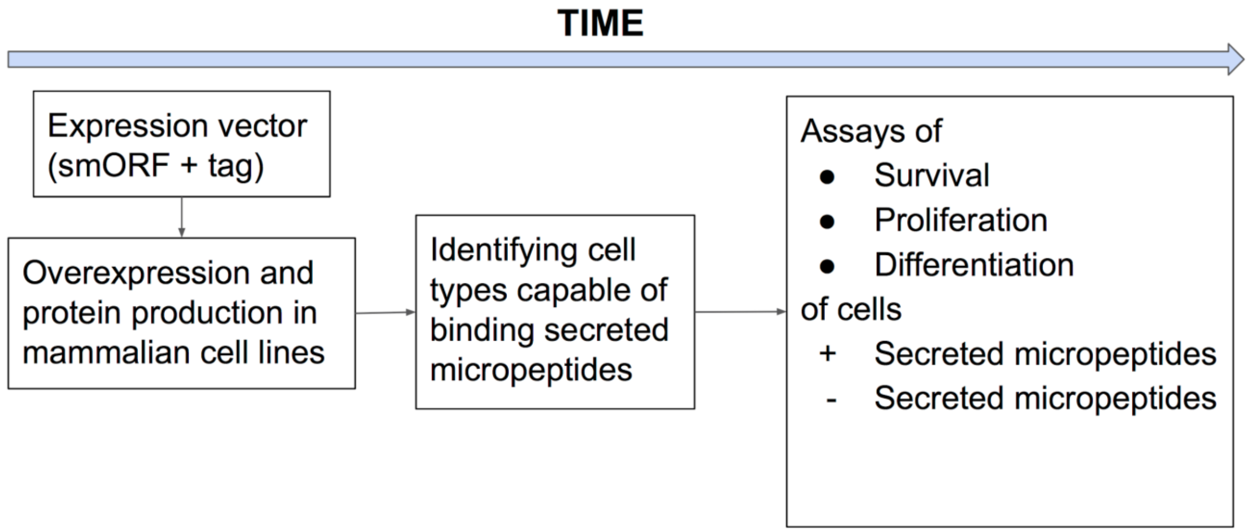

TIME
Expression vector (smORF + tag)
Overexpression and protein production in mammalian cell lines
Identifying cell types capable of binding secreted micropeptides
Assays of
● Survival
● Proliferation
● Differentiation
of cells
+ Secreted micropeptides
- Secreted micropeptides

Figure 4.14 | **Proposed approaches to study the function of secreted micropeptides.**

Expression plasmid validation

My colleague Jia Lu performed experiments to test the expression plasmid (**Figure A.2A** in Appendix B) and developed western-blot-based detection of cellular and secreted micropeptides. 293T cells were mock transfected and transfected with empty vector, and two candidate secreted micropeptides (1500011B03Rik and Phf21a), both micropeptides have shown high probabilities to be secreted in a *in silico* prediction (**Figure 4.15**). 44 hours post transfection, supernatant was harvested, the protein products were then analysed by gel electrophoresis to visualize the secreted micropeptides. Firstly, FLAG and GFP signals were detected which demonstrated that the plasmid works. Both Phf21a and 1500011B03Rik were detected by anti-FLAG antibodies in the supernatant (**Figure A.2B**). We noticed a band between 17 and 26.6 kDa for 1500011B03Rik from supernatant, it is possibly a dimer. We also observed GFP signals in supernatant in both constructs (**Figure A.2C**), it might be caused when the transfection condition was not optimized.

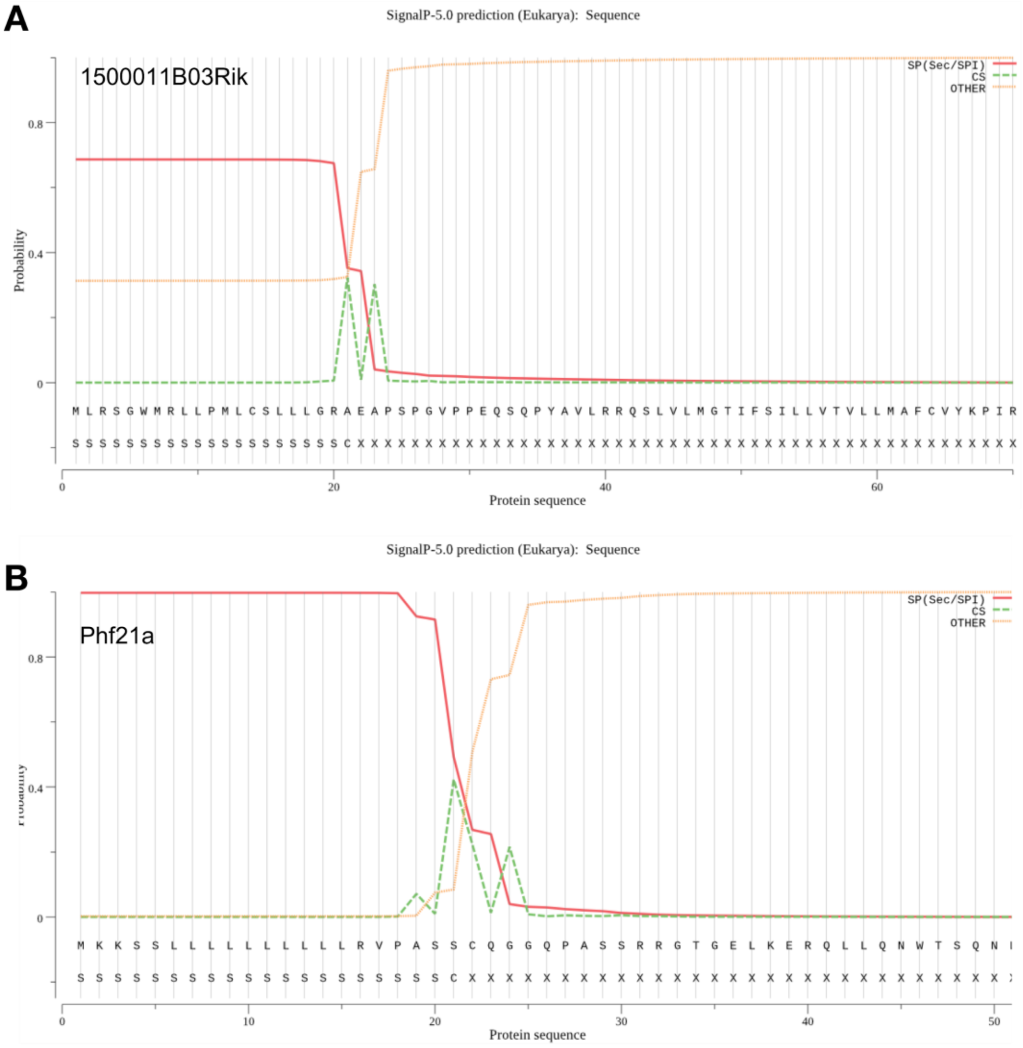


Figure 4.15 | **Signal peptide prediction for 1500011B03Rik and Phf21a by SignalP server.** Red line shows the probability of a sequence being signal peptide, green dash line indicates the cleavage site. SP – signal peptide, CS – cleavage site.

4.15 Discussion

In this study I have predicted 5744 unique smORFs that show evidence of transcription and translation in B and T lymphocytes. Apart from 368 being annotated as short CDSs or isoforms, the others are novel and located in long non-coding RNAs, pseudogenes and the 5'UTR and 3'UTRs of canonical protein coding transcripts. By assessing the conservation of the amino acid sequences compared with long proteins I can infer whether the translation products of these smORFs have any potential functions.

Among the predicted smORFs, 80% were located within 5'UTRs. The biological functions of the majority of uORFs are unknown, but specific examples are known which play important roles in gene expression; that is regulating the translation of the downstream ORF. For example, MDM2, which is an important negative regulator of the p53 tumor suppressor, is mainly expressed in normal cells from a transcript isoform that contains two uORFs. However, following a switch in promoter usage, a transcript isoform without uORFs produces more MDM2 protein in human soft tissue tumours (Brown et al., 1999). CD36 encodes a cell-surface receptor expressed by B cells that regulates uptake of lipids and modulates antibody responses during bacterial infection (Won et al., 2008). A study shows this uORF is involved in atherosclerosis development in diabetics (Griffin et al., 2001). Under high glucose conditions, due to ribosomal re-initiation following translation of this uORF, CD36 main CDS translation efficiency is increased thus resulting in increased expression of CD36, and providing a mechanism for accelerated atherosclerosis in diabetics. In addition, a recent study has shown that non-canonical Hoogsteen-paired G-quadruplex (rG4) structures are present upstream of uORFs and promote 80S ribosome formation on upstream start codons, causing inhibition of translation of the downstream main CDSs (Murat et al., 2018). Searching for rG4 motifs in an uORF upstream context will help to distinguish regulatory uORFs.

About four percent of our predicted smORFs are dORFs. A number of ribosome profiling and mass spectrometry studies reported translation events in the 3′ UTR and dORF-encoded micropeptide detection (Slavoff et al., 2013; Ingolia, 2016). Amongst these, human protein

HTD2 (hydroxyacyl-thioester dehydratase type 2, 168 AA), is the most completely functionally characterized one to date (Autio et al. 2008). HTD2 was identified as a 3' open reading frame on the RPP14 transcript which is known to encode the RPP14 (ribonuclease P protein subunit p14), a subunit of human ribonuclease P (RNase P) complex (**Figure 4.16**). HTD2 has been shown to be involved in mitochondrial fatty acid biosynthesis (Autio et al., 2008). In our prediction for 101-200 AA small proteins, mouse Rpp14 canonical protein (122 AA) and a downstream protein (158 AA, annotated as a Rpp14 isoform in Ensembl) are predicted. RPP14 downstream protein is 79.6% identical to human HTD2 protein. Phylogenetic analysis of RPP14 and HTD2 sequences highlight a conserved bicistronic relationship over 400 million years and therefore suggest a functional link between RNA processing and vertebrate mitochondrial biology.

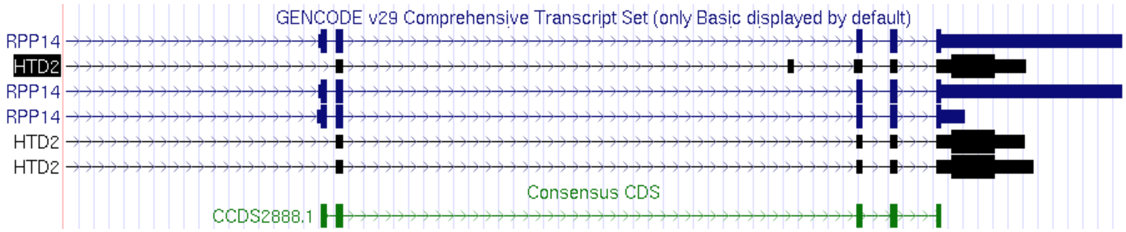

GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
RPP14
HTD2
RPP14
RPP14
HTD2
HTD2
Consensus CDS
CCDS2888.1

Figure 4.16 | **Human HTD2 and RPP14 genes.** HTD2 was identified as a dORF on the RPP14 transcript which is known to encode the RPP14 (ribonuclease P protein subunit p14). Thick blocks represent the coding sequences.

We predicted that ~9% of smORFs reside within annotated noncoding RNAs. Several micropeptides have been identified from transcripts previously annotated as noncoding (Anderson et al., 2015; Nelson et al., 2016; D'lima et al., 2017; Matsumoto et al., 2017). We predicted Nbdy (68 AA), a recently discovered micropeptide that binds to the mRNA decapping complex and promotes dispersal of P-body components (D'lima et al., 2017). A recent study characterised the Aw112010 peptide from a lncRNA as being secreted in macrophages and playing a role in host defence and inflammatory disease models (Jackson et al., 2018). We also discovered Aw112010*,* but we noted that this gene does not have a human homolog.

A classical model for the structure of a eukaryotic gene is that it codes for a single polypeptide (Beadle and Tatum, 1941). An evolutionary conserved micropeptide called tarsal-less (tal) has been identified in *Drosophila*, in which four tandem smORFs located on the tal transcript are independently translated to micropeptides of 11 and 32 AA (Savard et al., 2006; Galindo et al., 2007; Kondo et al., 2007). Additionally, several mammalian polycistronic mRNAs have been characterized in recent years (Karginov et al., 2017). Studies have provided evidence that uORF- and dORF-encoded micropeptides can be expressed and function in trans (Andreev et al., 2015; Ma et al., 2016; Autio et al., 2008). Traditionally polycistronic mRNAs were thought to occur mainly in prokaryotes, a classical example is the lac operon of *E. coli.* However, emerging evidence is revealing that there are many more polycistronic mRNAs in eukaryotes than was originally thought. We have predicted hundreds of uORFs in mouse, however the function of their translation products is still unknown and needs more effort to investigate.

Validating the coding potential of smORFs after they were identified by computational approaches is an essential step towards the characterization of their function. *In vitro* translation assays have been reported to assess smORF coding potential (Anderson et al., 2015,2016; van Heesch et al., 2019). The full-length cDNA of a smORF is cloned into a vector and then expression of the construct is evaluated using a cell-free protein-synthesizing system in the presence of ^{35}S-methionine. The protein products are analysed by gel electrophoresis and autoradiography is performed to visualize the synthesis of a ^{35}S-labeled micropeptide. This is a valuable method to screen potential candidates, however there is a possibility that the smORF can be translated *in vitro* but not *in vivo*. Our approach takes one step forwards, the vector is expressed in mammalian cells which maintain the translation apparatus and mechanism, it will show stronger evidence that the micropeptides are expressed *in vivo*.

We used antibodies to detect FLAG-tagged micropeptides using gel electrophoresis. One concern is that if a micropeptide is expressed at a low level, the antibody may not be sufficient to generate strong enough signals for detection. During the experiment, we experienced issues in gel resolution. 15% Tricine seemed to give desired molecular weight range for separation, but the resolution is poor for proteins of 1-6 kDa. We could try to increase cross-linker percentage to increase resolution at low molecular weight, this will also allow analysis of signal peptide cleavage.

In this project, we will focus on developing DNA based approaches including the expression of smORFs and micropeptide production in mammalian cell lines. Once the coding potential of smORFs is tested. Those that are secreted will be tested to identify their localization, for example, whether the peptide binds to the cell surface and what cell types they bind to. This is critical as the information will guide us to design experiments to identify their biological functions.

In the meanwhile, we are interested in protein-based approaches using synthetic peptides and single-domain antibodies, and this will be a route to develop the project further.

Chapter Five

General Discussion and Future Directions

Recent advances in computational and experimental techniques have revealed that a much larger portion of the genome is translated than was previously recognized. Functional smORFs have emerged as a class of genetic elements to deepen the understanding of the coding potential of the genome. They are now representing a frontier in biochemistry, molecular biology, and physiology that is at its inception. It is likely that many more smORFs and micropeptides await discovery and characterization. smORFs have been relatively neglected as genome annotations arbitrarily excluded ORFs of less than 100 codons. However, this is changing rapidly as more investigation conducted in recent years have indicated a diverse range of functions for smORF-encoded micropeptides. These include muscle regeneration, DNA replication, phagocytosis, metabolism and cancer. These examples indicate that micropeptides are essential for cell functions and could be used to develop new therapeutics. Thus, micropeptides offer an area of significant interest that currently is largely unexplored.

The immune system as a host defence system protects organisms against disease. Several peptides and small proteins including host defense antimicrobial peptides, hormones and cytokines that are known to have important functions in normal and pathological immune reactions. However, little was known how widespread micropeptides are in the immune system and what their functional roles might be. In this study, we have tried to address those questions.

We have taken an *in-silico* approach to discover novel functional smORFs from lymphocytes. We have built a computational pipeline “ORFLine” to systematically analyse RNA-Seq and Ribosome profiling to identify actively translated smORFs. ORFLine was applied to mouse B (resting and LPS-activated) and T (activated CD4+ T cells and reactivated Th1 cells at different time points). We have considered two classes based on their size: micropeptides encoded by smORFs of 100 codons or fewer and small proteins in the size range of 101-200 amino acids. In total, 5744 actively translated smORFs and their predicted translation products as well as 945 of 101-200 AA were identified and described. Among the 5744 smORFs, 338 are canonical and annotated (5.9%), 5404 are unannotated and novel (94.1%). Specifically, micropeptides possess signal peptides which are potentially to be secreted were further investigated. Our identification of thousands of translated smORFs in the mouse lymphocytes provides an entry point to investigate their functions *in vivo*.

Although there is now robust evidence for the translation of smORFs, there is still a large amount of work that needs to be carried out to experimentally characterize each of the smORFs to understand their biological function. Based on the current state of the field, several questions and challenges remain to be addressed, and several future directions seem likely.

Firstly, methods for the elucidation of smORFs only reveal their existence, but not provide insight into the functions of these genes. With so many smORFs already discovered, higher throughput methods in the form of gain- or loss-of-function screens with smORFs are needed to find the most interesting smORFs and micropeptides for further investigation. What is the fraction of smORFs that are translated to stable micropeptides and function in their own right, versus those that are incidental unstable by-products of random translational events that merely transcriptional/translational noise? If they have a function, complete understanding of their action may play an important role in therapeutic purposes, where a drug may be designed by modulating or mimicking their functions to regulate any biological pathway they may be involved in or inhibit their activities.

Secondly, how will researchers overcome the many unique technical obstacles that come with working with micropeptides and small proteins? They might be in low abundance or short lived as their stability is unknown. It has been suggested that many peptide products are selectively and rapidly degraded within cells, and hence are difficult to detect biochemically (Oyama et al., 2007; Slavoff et al., 2013). These factors impede their identification by mass spectrometry as they are often lost in the sample preparation thus not available for detection. Ideally, an antibody against a micropeptide can be generated and validated to demonstrate its specificity, however there might be lack of available antibodies and means to generate custom antibodies for reasons including that the small size of micropeptide provides limited choices for designing antibodies and the 3D structure of the micropeptides is unknown, it might limit the regions for epitope design.

Thirdly, the integration of smORFs into big data will provide additional methods to identify and prioritise interesting smORFs. For example, combining smORF discovery with GWAS data can identify disease-associated smORFs, or mining expression profiling data can identify smORFs that are up or down regulated in different diseases (Jackson et al., 2017; Jagannathan et al., 2019; Whiffin et al., 2019). A related question is that whether we are missing smORFs that express under specific conditions and timing? Comparison between smORFs identified in resting B cells and LPS-activated B cells shows more smORFs expressed in an activated condition. The transcription profile of cells will certainly be different during perturbation (e.g.

stimulation and stress) compared to a normal state. Ideally, Ribo-Seq will be generated for lymphocytes responding to different stimulations, and potential smORFs in specific condition can be identified, however, the public available Ribo-Seq datasets are limited for lymphocytes. It is also useful to consider datasets of cell types that are directly interacting with lymphocyte or in the same immune response context but secret novel cytokines to bind to lymphocytes.

Secreted micropeptides and small proteins have particularly drawn our attention. Living cells communicate with their surroundings by the secretion of biomolecules including proteins and peptides. Chemokines and cytokines are paradigms of this class and have proven to be a rich source of therapeutic targets. For example, selective inhibition TNF (tumor necrosis factor) by monoclonal antibodies or soluble TNF receptor analogues has been clinically and commercially highly successful in a number of diseases. The monoclonal antibody Mogamulizumab is approved for the treatment of cutaneous T-cell lymphoma and has the potential to treat allergic disease, as it targets CCR4 chemokine receptors necessary for T-helper type 2 cell entry into the lung. Thus, blocking cytokines/chemokines can have utility in multiple diseases. In addition to antibodies small molecule approaches to cytokines/chemokines or their receptors have proven efficacy. Plerixafor is a small molecule CXCR4 antagonist that has approval to mobilize hematopoietic stem cells. Maraviroc is an FDA-approved chemokine receptor-targeting drug clinically used for the treatment of HIV-1 infection. The discovery of new cytokines or chemokines offers a starting point for the development of further therapeutic modalities or biomarkers.

A recent study characterised the Aw112010 peptide as being important for mucosal immunology (Jackson et al., 2018). We also discovered Aw112010, but we noted that this gene does not have a human homolog. This raises the question of why lymphocytes produce this. In addition to known micropeptides, we identified further candidate-secreted micropeptides that are conserved between human and mouse; some of these originate from transcripts annotated as non-coding and upstream ORFs of protein-coding transcripts. Amongst the group of proteins sized between 101 and 200 amino acids we identified known cytokines and chemokines (e.g. IFNγ), but also candidate or known secreted proteins not previously associated with lymphocytes including the 182 AA Mesencephalic Astrocyte-derived Neurotrophic Factor as well as a number of genes of unknown function.

Our informatic approach has suggested the existence of numerous secreted proteins which have either remained undiscovered, and are thus totally novel, as well as secreted proteins that have been previously identified but not yet assigned a role in lymphocyte biology. Here we propose to pursue the biology of the novel class of micropeptides by hypothesising that they will have immunoregulatory cytokine-like function. As such, they will turn out to be entities for which it will be desirable to inhibit or augment their function in human disease.

We have described the effort to apply DNA based approach including the expression of smORFs and micropeptide production in mammalian cell lines to validate coding potential of the candidate secreted micropeptide. We will further develop this approach. Following the validation, we will test micropeptides binding to cells. Supernatants will be screened for binding to a variety of human and mouse cell lines and cells from peripheral blood by microscopy in the case of adherent cells or by flow cytometry for non-adherent cells. Positive binding will be subject to competitive inhibition with increasing amounts of supernatant derived from cells expressing a version of the protein with a different tag. Although there is precedent that C-terminal tagging of chemokines does not inhibit binding (Kawamura et al., 2014), it is possible that a tag may interfere with the binding and as an alternative approach direct labelling of lysine or cysteine residues or a different tagging strategy could be used. Given that we can analyse numerous proteins with this approach, it seems probable that we will identify candidates to take to the next stage of the analysis. The information gained in this part of the project will guide the design of functional experiments where we will test the capacity of micropeptides to mediate an effect on cell function.

Cell types bound by micropeptides will be exposed to different dilutions of the supernatant from micropeptide expressing transfectants. The exact nature of the assay will depend on the cell types under study, but we envisage using assays of cell proliferation and apoptosis as an initial screen. Additional assays of lymphocyte properties can include the expression of cell surface markers indicative of activation and differentiation state. For example, culturing naïve mouse T cells in the presence of blocking antibodies to IFNγ and then re-stimulating them and staining for intracellular IL-4 as an assay of Th2 differentiation.

smORF and micropeptide research is basic research, even though the identification of a novel biologically active micropeptide must still be an important discovery. It would raise the question of whether the micropeptide was involved in human disease and whether augmenting or inhibiting its function was desirable, or whether the micropeptide might be a biomarker of disease state. We strongly believe that continued investigations will begin to find more smORFs and micropeptides linked to human disease, and in the future new medicines may emerge from these studies, all the fundamental research will be translated to benefit humanity. The identification of a cell type to which the micropeptide bound would open the door to the isolation of molecular clones of the receptor. We will pursue this biology further, in particular using animal models of loss of function of the micropeptide, or its receptor if we find one. The generation of monoclonal antibodies will also be pursued and careful consideration will be given how exactly to do this as antibodies may turn out the be key reagents for commercial development. If promising results turn up, we will approach potential collaborators in medicine to understand the biology of the micropeptide in human disease and seek means for commercialization.

References

Abbas, Abul K., Andrew H. Lichtman, and Shiv Pillai. *Cellular and molecular immunology E-book*. Elsevier Health Sciences, 2014.

Akimoto, Chizuru, Eiji Sakashita, Katsumi Kasashima, Kenji Kuroiwa, Kaoru Tominaga, Toshiro Hamamoto, and Hitoshi Endo. "Translational repression of the McKusick–Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites." *Biochimica et Biophysica Acta (BBA)-General Subjects* 1830, no. 3 (2013): 2728-2738.

Almagro Armenteros, Jose Juan, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. "DeepLoc: prediction of protein subcellular localization using deep learning." *Bioinformatics* 33, no. 21 (2017): 3387-3395.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. "Basic local alignment search tool." *Journal of molecular biology* 215, no. 3 (1990): 403-410.

Andreev, Dmitry E., Patrick BF O'Connor, Ciara Fahey, Elaine M. Kenny, Ilya M. Terenin, Sergey E. Dmitriev, Paul Cormican, Derek W. Morris, Ivan N. Shatsky, and Pavel V. Baranov. "Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression." *Elife* 4 (2015): e03971.

Anderson, Douglas M., Kelly M. Anderson, Chi-Lun Chang, Catherine A. Makarewich, Benjamin R. Nelson, John R. McAnally, Prasad Kasaragod et al. "A micropeptide encoded by a putative long noncoding RNA regulates muscle performance." *Cell* 160, no. 4 (2015): 595-606.

Anderson, Douglas M., Catherine A. Makarewich, Kelly M. Anderson, John M. Shelton, Svetlana Bezprozvannaya, Rhonda Bassel-Duby, and Eric N. Olson. "Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides." *Sci. Signal.* 9, no. 457 (2016): ra119-ra119.

Andrews, Shea J., and Joseph A. Rothnagel. "Emerging evidence for functional peptides encoded by short open reading frames." *Nature Reviews Genetics* 15, no. 3 (2014): 193-204.

Andrews, Simon (2016). Contamination with a different species you can guess. https://sequencing.qcfail.com/articles/contamination-with-a-different-species-you-can-guess

Andrews, Simon (2016). Mapping to a transcriptome can incorrectly report reads as mapping uniquely. https://sequencing.qcfail.com/articles/mapping-to-a-transcriptome-can-incorrectly-report-reads-as-mapping-uniquely

Apostolou, Andria, Yuxian Shen, Yan Liang, Jun Luo, and Shengyun Fang. "Armet, a UPR-upregulated protein, inhibits cell proliferation and ER stress-induced cell death." *Experimental cell research* 314, no. 13 (2008): 2454-2467.

Armenteros, José Juan Almagro, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. "SignalP 5.0 improves signal peptide predictions using deep neural networks." *Nature biotechnology* 37, no. 4 (2019): 420.

Arnoult, Nausica, Adriana Correia, Jiao Ma, Anna Merlo, Sara Garcia-Gomez, Marija Maric, Marco Tognetti et al. "Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN." Nature 549, no. 7673 (2017): 548-552.

Aspden, Julie L., Ying Chen Eyre-Walker, Rose J. Phillips, Unum Amin, Muhammad Ali S. Mumtaz, Michele Brocard, and Juan-Pablo Couso. "Extensive translation of small open reading frames revealed by Poly-Ribo-Seq." *Elife* 3 (2014): e03528.

Autio, Kaija J., Alexander J. Kastaniotis, Helmut Pospiech, Ilkka J. Miinalainen, Melissa S. Schonauer, Carol L. Dieckmann, and J. Kalervo Hiltunen. "An ancient genetic link between vertebrate mitochondrial fatty acid synthesis and RNA processing." *The FASEB Journal* 22, no. 2 (2008): 569-578.

Basrai, Munira A., Philip Hieter, and Jef D. Boeke. "Small open reading frames: beautiful needles in the haystack." *Genome research* 7, no. 8 (1997): 768-771.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna et al. "The Pfam protein families database." *Nucleic acids research* 32, no. suppl_1 (2004): D138-D141.

Bazzini, Ariel A., Timothy G. Johnstone, Romain Christiano, Sebastian D. Mackowiak, Benedikt Obermayer, Elizabeth S. Fleming, Charles E. Vejnar et al. "Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation." *The EMBO journal* 33, no. 9 (2014): 981-993.

Beadle, George W., and Edward L. Tatum. "Genetic control of biochemical reactions in Neurospora." *Proceedings of the National Academy of Sciences of the United States of America* 27, no. 11 (1941): 499.

Berg, Jordan A., Jonathan R. Belyeu, Jeffrey T. Morgan, Yeyun Ouyang, Alex J. Bott, Aaron R. Quinlan, Jason Gertz, and Jared Rutter. "XPRESSyourself: Enhancing, Standardizing, and Automating Ribosome Profiling Computational Analyses Yields Improved Insight into Data." *BioRxiv* (2019): 704320.

Bi, Pengpeng, Andres Ramirez-Martinez, Hui Li, Jessica Cannavino, John R. McAnally, John M. Shelton, Efrain Sánchez-Ortiz, Rhonda Bassel-Duby, and Eric N. Olson. "Control of muscle formation by the fusogenic micropeptide myomixer." *Science* 356, no. 6335 (2017): 323-327.

Blankenberg, Daniel, James Taylor, Anton Nekrutenko, and Galaxy Team. "Making whole genome multiple alignments usable for biologists." *Bioinformatics* 27, no. 17 (2011): 2426-2428.

Boonen, Kurt, John W. Creemers, and Liliane Schoofs. "Bioactive peptides, networks and systems biology." *Bioessays* 31, no. 3 (2009): 300-314.

Branca, Rui MM, Lukas M. Orre, Henrik J. Johansson, Viktor Granholm, Mikael Huss, Åsa Pérez-Bercoff, Jenny Forshed, Lukas Käll, and Janne Lehtiö. "HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics." *Nature methods* 11, no. 1 (2014): 59.

Brar, Gloria A., and Jonathan S. Weissman. "Ribosome profiling reveals the what, when, where and how of protein synthesis." *Nature reviews Molecular cell biology* 16, no. 11 (2015): 651-664.

Brent, Michael R., and Roderic Guigo. "Recent advances in gene structure prediction." *Current opinion in structural biology* 14, no. 3 (2004): 264-272.

Brown, Cheryl Y., Gregory J. Mize, Mario Pineda, Donna L. George, and David R. Morris. "Role of two upstream open reading frames in the translational control of oncogene mdm2." *Oncogene* 18, no. 41 (1999): 5631.

Bürger, Marco, and Joanne Chory. "Stressed out about hormones: how plants orchestrate immunity." *Cell host & microbe* 26, no. 2 (2019): 163-172.

Cabili, Moran N., Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L. Rinn. "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." *Genes & development* 25, no. 18 (2011): 1915-1927.

Cabrera-Quio, Luis Enrique, Sarah Herberg, and Andrea Pauli. "Decoding sORF translation–from small proteins to gene regulation." *RNA biology* 13, no. 11 (2016): 1051-1059.

Calviello, Lorenzo, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler. "Detecting actively translated open reading frames in ribosome profiling data." *Nature methods* 13, no. 2 (2016): 165.

Calvo, Sarah E., David J. Pagliarini, and Vamsi K. Mootha. "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans." *Proceedings of the National Academy of Sciences* 106, no. 18 (2009): 7507-7512.

Carninci, Pea, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, Norihiro Maeda, Rieko Oyama et al. "The transcriptional landscape of the mammalian genome." *Science* 309, no. 5740 (2005): 1559-1563.

Casson, Stuart A., Paul M. Chilley, Jennifer F. Topping, I. Marta Evans, Martin A. Souter, and Keith Lindsey. "The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning." *The Plant Cell* 14, no. 8 (2002): 1705-1721.

Castellana, Natalie, and Vineet Bafna. "Proteogenomics to discover the full coding content of genomes: a computational perspective." *Journal of proteomics* 73, no. 11 (2010): 2124-2135.

Chanut-Delalande, Hélène, Yoshiko Hashimoto, Anne Pelissier-Monier, Rebecca Spokony, Azza Dib, Takefumi Kondo, Jérôme Bohère et al. "Pri peptides are mediators of ecdysone for the temporal control of development." *Nature cell biology* 16, no. 11 (2014): 1035.

Chen, Jin, Andreas-David Brunner, J. Zachery Cogan, James K. Nuñez, Alexander P. Fields, Britt Adamson, Daniel N. Itzhak et al. "Pervasive functional translation of noncanonical human open reading frames." *Science* 367, no. 6482 (2020): 1140-1146.

Chen, Lijian, Lijie Feng, Xia Wang, Jian Du, Ying Chen, Wen Yang, Chengyue Zhou et al. "Mesencephalic astrocyte-derived neurotrophic factor is involved in inflammation by negatively regulating the NF-κB pathway." *Scientific reports* 5 (2015): 8133.

Cheng, Lei, Hua Zhao, Wen Zhang, Ben Liu, Yi Liu, Yingjun Guo, and Lin Nie. "Overexpression of conserved dopamine neurotrophic factor (CDNF) in astrocytes alleviates endoplasmic reticulum stress-induced cell damage and inflammatory cytokine secretion." Biochemical and biophysical research communications 435, no. 1 (2013): 34-39.

Chew, Guo-Liang, Andrea Pauli, and Alexander F. Schier. "Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish." *Nature communications* 7 (2016): 11663.

Choudhary, Saket, Wenzheng Li, and Andrew D. Smith. "Accurate detection of short and long active ORFs using Ribo-seq data." *Bioinformatics* 36, no. 7 (2020): 2053-2059.

Chu, Qian, Thomas F. Martinez, Sammy Weiser Novak, Cynthia J. Donaldson, Dan Tan, Joan M. Vaughan, Tina Chang et al. "Regulation of the ER stress response by a mitochondrial microprotein." *Nature communications* 10, no. 1 (2019): 1-13.

Chng, Serene C., Lena Ho, Jing Tian, and Bruno Reversade. "ELABELA: a hormone essential for heart development signals via the apelin receptor." *Developmental cell* 27, no. 6 (2013): 672-680.

Clamer, Massimiliano, Toma Tebaldi, Fabio Lauria, Paola Bernabo, Rodolfo F. Gómez-Biagi, Marta Marchioretto, Divya T. Kandala et al. "Active ribosome profiling with RiboLace." *Cell Reports* 25, no. 4 (2018): 1097-1108.

Clements, J. M., T. M. Laz, and F. Sherman. "Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae." *Molecular and cellular biology* 8, no. 10 (1988): 4533-4536.

Cohen, Stephen M. "Everything old is new again:(linc) RNAs make proteins!." *The EMBO journal* 33, no. 9 (2014): 937-938.

Couso, Juan Pablo. "Finding smORFs: getting closer." *Genome biology* 16, no. 1 (2015): 189.

Couso, Juan-Pablo, and Pedro Patraquim. "Classification and function of small open reading frames." *Nature reviews Molecular cell biology* 18, no. 9 (2017): 575.

Crappé, Jeroen, Wim Van Criekinge, Geert Trooskens, Eisuke Hayakawa, Walter Luyten, Geert Baggerman, and Gerben Menschaert. "Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs." *BMC genomics* 14, no. 1 (2013): 648.

Crappé, Jeroen, Wim Van Criekinge, and Gerben Menschaert. "Little things make big things happen: a summary of micropeptide encoding genes." *EuPA Open Proteomics* 3 (2014): 128-137.

Crowe, Mark L., Xue-Qing Wang, and Joseph A. Rothnagel. "Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides." *BMC genomics* 7, no. 1 (2006): 16.

Cunha, Fernanda M., Denise A. Berti, Zulma S. Ferreira, Clécio F. Klitzke, Regina P. Markus, and Emer S. Ferro. "Intracellular peptides as natural regulators of cell signaling." *Journal of biological Chemistry* 283, no. 36 (2008): 24448-24459.

Davari, Kathrin, Johannes Lichti, Christian Gallus, Franziska Greulich, N. Henriette Uhlenhaut, Matthias Heinig, Caroline C. Friedel, and Elke Glasmacher. "Rapid genome-wide recruitment of RNA polymerase II drives transcription, splicing, and translation events during T cell responses." *Cell reports* 19, no. 3 (2017): 643-654.

Delcourt, Vivian, Antanas Staskevicius, Michel Salzet, Isabelle Fournier, and Xavier Roucou. "Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA." *Proteomics* 18, no. 10 (2018): 1700058.

Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec et al. "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." *Genome research* 22, no. 9 (2012): 1775-1789.

Diaz-Muñoz, Manuel D., Sarah E. Bell, Kirsten Fairfax, Elisa Monzon-Casanova, Adam F. Cunningham, Mar Gonzalez-Porta, Simon R. Andrews et al. "The RNA-binding protein HuR is essential for the B cell antibody response." *Nature Immunology* 16, no. 4 (2015): 415.

Diba, Fantahun, Cheryl S. Watson, and Bahiru Gametchu. "5′ UTR sequences of the glucocorticoid receptor 1A transcript encode a peptide associated with translational regulation of the glucocorticoid receptor." *Journal of cellular biochemistry* 81, no. 1 (2001): 149-161.

Dijkstra, Johannes M., and Keith T. Ballingall. "Non-human lnc-DC orthologs encode Wdnm1-like protein." *F1000Research* 3 (2014).

Dinarello, Charles A. "Proinflammatory cytokines." *Chest* 118, no. 2 (2000): 503-508.

Dinger, Marcel E., Ken C. Pang, Tim R. Mercer, and John S. Mattick. "Differentiating protein-coding and noncoding RNA: challenges and ambiguities." *PLoS computational biology* 4, no. 11 (2008): e1000176.

D'Lima, Nadia G., Jiao Ma, Lauren Winkler, Qian Chu, Ken H. Loh, Elizabeth O. Corpuz, Bogdan A. Budnik, Jens Lykke-Andersen, Alan Saghatelian, and Sarah A. Slavoff. "A human microprotein that interacts with the mRNA decapping complex." *Nature chemical biology* 13, no. 2 (2017): 174.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29, no. 1 (2013): 15-21.

Dowlati, Yekta, Nathan Herrmann, Walter Swardfager, Helena Liu, Lauren Sham, Elyse K. Reim, and Krista L. Lanctôt. "A meta-analysis of cytokines in major depression." *Biological psychiatry* 67, no. 5 (2010): 446-457.

Duncan, Caia DS, and Juan Mata. "The translational landscape of fission-yeast meiosis and sporulation." *Nature structural & molecular biology* 21, no. 7 (2014): 641.

Dunn, Joshua G., Catherine K. Foo, Nicolette G. Belletier, Elizabeth R. Gavis, and Jonathan S. Weissman. "Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*." *Elife* 2 (2013): e01179.

Emanuelsson, Olof, Henrik Nielsen, Søren Brunak, and Gunnar Von Heijne. "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence." *Journal of molecular biology* 300, no. 4 (2000): 1005-1016.

ENCODE Project Consortium. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature* 447, no. 7146 (2007): 799.

Erpf, Paige E., and James A. Fraser. "The long history of the diverse roles of short ORFs: sPEPs in Fungi." *Proteomics* 18, no. 10 (2018): 1700219.

Fields, Alexander P., Edwin H. Rodriguez, Marko Jovanovic, Noam Stern-Ginossar, Brian J. Haas, Philipp Mertins, Raktima Raychowdhury et al. "A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation." *Molecular cell* 60, no. 5 (2015): 816-827.

Finkel, Yaara, Noam Stern-Ginossar, and Michal Schwartz. "Viral short ORFs and their possible functions." *Proteomics* 18, no. 10 (2018): 1700255.

Frith, Martin C., Alistair R. Forrest, Ehsan Nourbakhsh, Ken C. Pang, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, Timothy L. Bailey, and Sean M. Grimmond.

"The abundance of short proteins in the mammalian proteome." *PLoS genetics* 2, no. 4 (2006): e52.

Fritsch, Claudia, Alexander Herrmann, Michael Nothnagel, Karol Szafranski, Klaus Huse, Frank Schumann, Stefan Schreiber et al. "Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting." *Genome research* 22, no. 11 (2012): 2208-2218.

Frank, Mary J., and Laurie G. Smith. "A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells." *Current Biology* 12, no. 10 (2002): 849-853.

Frankish, Adam, Barbara Uszczynska, Graham RS Ritchie, Jose M. Gonzalez, Dmitri Pervouchine, Robert Petryszak, Jonathan M. Mudge et al. "Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction." *BMC genomics* 16, no. S8 (2015): S2.

Fricker, Lloyd D. "Neuropeptide-processing enzymes: applications for drug discovery." In *Drug Addiction*, pp. 497-509. Springer, New York, NY, 2008.

Galindo, Máximo Ibo, José Ignacio Pueyo, Sylvaine Fouix, Sarah Anne Bishop, and Juan Pablo Couso. "Peptides encoded by short ORFs control development and define a new eukaryotic gene family." *PLoS biology* 5, no. 5 (2007).

Ganz, Tomas, Michael E. Selsted, Dorothy Szklarek, S. S. Harwig, Kathleen Daher, Dorothy F. Bainton, and Robert I. Lehrer. "Defensins. Natural peptide antibiotics of human neutrophils." *The Journal of clinical investigation* 76, no. 4 (1985): 1427-1435.

Gerstein, Mark B., Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder. "What is a gene, post-ENCODE? History and updated definition." *Genome research* 17, no. 6 (2007): 669-681.

Gibson, Daniel G., Lei Young, Ray-Yuan Chuang, J. Craig Venter, Clyde A. Hutchison, and Hamilton O. Smith. "Enzymatic assembly of DNA molecules up to several hundred kilobases." *Nature methods* 6, no. 5 (2009): 343-345.

Gish, Warren, and David J. States. "Identification of protein coding regions by database similarity search." *Nature genetics* 3, no. 3 (1993): 266-272.

Glembotski, Christopher C., Donna J. Thuerauf, Chengqun Huang, John A. Vekich, Roberta A. Gottlieb, and Shirin Doroudgar. "Mesencephalic astrocyte-derived neurotrophic factor protects the heart from ischemic damage and is selectively secreted upon sarco/endoplasmic reticulum calcium depletion." *Journal of Biological Chemistry* 287, no. 31 (2012): 25893-25904.

Griffin, Erik, Alessandro Re, Nance Hamel, Chenzong Fu, Harry Bush, Timothy McCaffrey, and Adam S. Asch. "A link between diabetes and atherosclerosis: glucose regulates expression of CD36 at the level of translation." *Nature medicine* 7, no. 7 (2001): 840-846.

Gunišová, Stanislava, Vladislava Hronová, Mahabub Pasha Mohammad, Alan G. Hinnebusch, and Leoš Shivaya Valášek. "Please do not recycle! Translation reinitiation in microbes and higher eukaryotes." *FEMS microbiology reviews* 42, no. 2 (2018): 165-192.

Guo, Bin, Dayong Zhai, Edelmira Cabezas, Kate Welsh, Shahrzad Nouraini, Arnold C. Satterthwait, and John C. Reed. "Humanin peptide suppresses apoptosis by interfering with Bax activation." *Nature* 423, no. 6938 (2003): 456-461.

Guttman, Mitchell, and John L. Rinn. "Modular regulatory principles of large non-coding RNAs." *Nature* 482, no. 7385 (2012): 339-346..

Guttman, Mitchell, Pamela Russell, Nicholas T. Ingolia, Jonathan S. Weissman, and Eric S. Lander. "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins." *Cell* 154, no. 1 (2013): 240-251.

Hanada, Kousuke, Kenji Akiyama, Tetsuya Sakurai, Tetsuro Toyoda, Kazuo Shinozaki, and Shin-Han Shiu. "sORF finder: a program package to identify small open reading frames with high coding potential." *Bioinformatics* 26, no. 3 (2010): 399-400.

Hancock, Robert EW, Evan F. Haney, and Erin E. Gill. "The immunology of host defence peptides: beyond antimicrobial activity." *Nature Reviews Immunology* 16, no. 5 (2016): 321.

Hann, Stephen R., Michael W. King, David L. Bentley, Carl W. Anderson, and Robert N. Eisenman. "A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas." *Cell* 52, no. 2 (1988): 185-195.

Hao, Yajing, Lili Zhang, Yiwei Niu, Tanxi Cai, Jianjun Luo, Shunmin He, Bao Zhang et al. "SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci." *Briefings in bioinformatics* 19, no. 4 (2018): 636-643.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken et al. "GENCODE: the reference human genome annotation for The ENCODE Project." *Genome research* 22, no. 9 (2012): 1760-1774.

Hart, Traver, H. Kiyomi Komori, Sarah LaMere, Katie Podshivalova, and Daniel R. Salomon. "Finding the active genes in deep RNA-seq gene expression studies." BMC genomics 14, no. 1 (2013): 778.

Hashimoto, Yoshiko, Takefumi Kondo, and Yuji Kageyama. "Lilliputians get into the limelight: novel class of small peptide genes in morphogenesis." *Development, growth & differentiation* 50 (2008): S269-S276.

Hayden, Celine A., and Giovanni Bosco. "Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila melanogaster* and other dipteran species." *BMC genomics* 9, no. 1 (2008): 61.

Heng, Tracy SP, Michio W. Painter, Kutlu Elpek, Veronika Lukacs-Kornek, Nora Mauermann, Shannon J. Turley, Daphne Koller et al. "The Immunological Genome Project: networks of gene expression in immune cells." *Nature Immunology* 9, no. 10 (2008): 1091-1094.

Hernández, Greco, Vincent G. Osnaya, and Xochitl Pérez-Martínez. "Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes." *Trends in biochemical sciences* (2019).

Hinnebusch, Alan G. "The scanning mechanism of eukaryotic translation initiation." *Annual review of biochemistry* 83 (2014): 779-812.

Hobeika, E., S. Thiemann, B. Storch, H. Jumaa, P. J. Nielsen, R. Pelanda, and M. Reth. "Testing gene function early in the B cell lineage in mb1-cre mice." *Proceedings of the National Academy of Sciences* 103, no. 37 (2006): 13789-13794.

Hsu, Polly Yingshan, and Philip N. Benfey. "Small but mighty: functional peptides encoded by small ORFs in plants." *Proteomics* 18, no. 10 (2018): 1700038.

Huang, Jin-Zhou, Min Chen, De Chen, Xing-Cheng Gao, Song Zhu, Hongyang Huang, Min Hu, Huifang Zhu, and Guang-Rong Yan. "A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth." *Molecular cell* 68, no. 1 (2017): 171-184.

Iacono, Michele, Flavio Mignone, and Graziano Pesole. "uAUG and uORFs in human and rodent 5′ untranslated mRNAs." *Gene* 349 (2005): 97-105.

Ina, Yasuo. "New methods for estimating the numbers of synonymous and nonsynonymous substitutions." *Journal of molecular evolution* 40, no. 2 (1995): 190-226.

Ingolia, Nicholas T., Sina Ghaemmaghami, John RS Newman, and Jonathan S. Weissman. "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling." *Science* 324, no. 5924 (2009): 218-223.

Ingolia, Nicholas T., Liana F. Lareau, and Jonathan S. Weissman. "Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes." *Cell* 147, no. 4 (2011): 789-802.

Ingolia, Nicholas T., Gloria A. Brar, Noam Stern-Ginossar, Michael S. Harris, Gaëlle JS Talhouarne, Sarah E. Jackson, Mark R. Wills, and Jonathan S. Weissman. "Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes." *Cell reports* 8, no. 5 (2014): 1365-1379.

Ivanov, Ivaylo P., Andrew E. Firth, Audrey M. Michel, John F. Atkins, and Pavel V. Baranov. "Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences." *Nucleic acids research* 39, no. 10 (2011): 4220-4234.

Jackson, Felix, Matthew T. Wayland, and Sudhakaran Prabakaran. "Identification And Prioritisation Of Variants In The Short Open-Reading Frame Regions Of The Human Genome." *bioRxiv* (2017): 133645.

Jackson, Richard J., Christopher UT Hellen, and Tatyana V. Pestova. "The mechanism of eukaryotic translation initiation and principles of its regulation." *Nature reviews Molecular cell biology* 11, no. 2 (2010): 113-127.

Jackson, Ruaidhri, Lina Kroehling, Alexandra Khitun, Will Bailis, Abigail Jarret, Autumn G. York, Omair M. Khan et al. "The translation of non-canonical open reading frames controls mucosal immunity." *Nature* 564, no. 7736 (2018): 434-438.

Jagannathan, N. Suhas, Narendra Meena, Kethaki Prathivadi Bhayankaram, and Sudhakaran Prabakaran. "Proteins encoded by Novel ORFs have increased disorder but can be biochemically regulated and harbour pathogenic mutations." *bioRxiv* (2019): 562835.

Jänes, Jürgen, Fengyuan Hu, Alexandra Lewin, and Ernest Turro. "A comparative study of RNA-seq analysis strategies." *Briefings in bioinformatics* 16, no. 6 (2015): 932-940.

Ji, Zhe, Ruisheng Song, Aviv Regev, and Kevin Struhl. "Many lncRNAs, 5’UTRs, and pseudogenes are translated and some are likely to express functional proteins." *elife* 4 (2015): e08890.

Ji, Zhe, Ruisheng Song, Hailiang Huang, Aviv Regev, and Kevin Struhl. "Transcriptome-scale RNase-footprinting of RNA-protein complexes." *Nature biotechnology* 34, no. 4 (2016): 410.

Johnstone, Timothy G., Ariel A. Bazzini, and Antonio J. Giraldez. "Upstream ORFs are prevalent translational repressors in vertebrates." *The EMBO journal* 35, no. 7 (2016): 706-723.

Jousse, Céline, Alain Bruhat, Valérie Carraro, Fumihiko Urano, Marc Ferrara, David Ron, and Pierre Fafournoux. "Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5′ UTR." *Nucleic acids research* 29, no. 21 (2001): 4341-4351.

Juntawong, Piyada, Thomas Girke, Jérémie Bazin, and Julia Bailey-Serres. "Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis." *Proceedings of the National Academy of Sciences* 111, no. 1 (2014): E203-E212.

Karlin, Samuel, Allan M. Campbell, and Jan Mrazek. "Comparative DNA analysis across diverse genomes." *Annual review of genetics* 32, no. 1 (1998): 185-225.

Kamvysselis, Manolis. "Computational comparative genomics: genes, regulation, evolution." PhD diss., Massachusetts Institute of Technology, 2003.

Kapranov, Philipp, Aarron T. Willingham, and Thomas R. Gingeras. "Genome-wide transcription and the implications for genomic organization." *Nature Reviews Genetics* 8, no. 6 (2007): 413-423.

Karginov, Timofey A., Daniel Parviz Hejazi Pastor, Bert L. Semler, and Christopher M. Gomez. "Mammalian polycistronic mRNAs and disease." *Trends in Genetics* 33, no. 2 (2017): 129-142.

Karolchik, Donna, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. "The UCSC Table Browser data retrieval tool." *Nucleic acids research* 32, no. suppl_1 (2004): D493-D496.

Kastenmayer, James P., Li Ni, Angela Chu, Lauren E. Kitchen, Wei-Chun Au, Hui Yang, Carole D. Carter et al. "Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae." *Genome research* 16, no. 3 (2006): 365-373.

Katsanou, Vicky, Stavros Milatos, Anthie Yiakouvaki, Nikos Sgantzis, Anastasia Kotsoni, Maria Alexiou, Vaggelis Harokopos, Vassilis Aidinis, Myriam Hemberger, and Dimitris L. Kontoyiannis. "The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development." *Molecular and cellular biology* 29, no. 10 (2009): 2762-2776.

Kawamura, Tetsuya, Bryan Stephens, Ling Qin, Xin Yin, Michael R. Dores, Thomas H. Smith, Neil Grimsey et al. "A general method for site specific fluorescent labeling of recombinant chemokines." *PLoS One* 9, no. 1 (2014).

Kearse, Michael G., and Jeremy E. Wilusz. "Non-AUG translation: a new start for protein synthesis in eukaryotes." *Genes & development* 31, no. 17 (2017): 1717-1731.

Kikuchi, Kunio, Makiha Fukuda, Tomoya Ito, Mitsuko Inoue, Takahide Yokoi, Suenori Chiku, Toutai Mitsuyama, Kiyoshi Asai, Tetsuro Hirose, and Yasunori Aizawa. "Transcripts of unknown function in multiple-signaling pathways involved in human stem cell differentiation." *Nucleic acids research* 37, no. 15 (2009): 4987-5000.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. "HISAT: a fast spliced aligner with low memory requirements." *Nature methods* 12, no. 4 (2015): 357-360.

Kimura, Motoo. "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." *Journal of molecular evolution* 16, no. 2 (1980): 111-120.

Kochetov, Alex V. "AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context." *Bioinformatics* 21, no. 7 (2005): 837-840.

Kondo, Takefumi, Yoshiko Hashimoto, Kagayaki Kato, Sachi Inagaki, Shigeo Hayashi, and Yuji Kageyama. "Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA." *Nature cell biology* 9, no. 6 (2007): 660-665.

Kondo, T., S. Plaza, J. Zanet, E. Benrabah, P. Valenti, Y. Hashimoto, S. Kobayashi, F. Payre, and Y. Kageyama. "Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis." *Science* 329, no. 5989 (2010): 336-339.

Kong, Lei, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." *Nucleic acids research* 35, no. suppl_2 (2007): W345-W349.

Kozak, Marilyn. "At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells." *Journal of molecular biology* 196, no. 4 (1987): 947-950.

Kozak, Marilyn. "An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs." *Nucleic acids research* 15, no. 20 (1987): 8125-8148.

Kozak, Marilyn. "The scanning model for translation: an update." *The Journal of cell biology* 108, no. 2 (1989): 229-241.

Kozak, Marilyn. "Constraints on reinitiation of translation in mammals." *Nucleic acids research* 29, no. 24 (2001): 5226-5232.

Krause, Alexander, Susanne Neitz, Hans-Jürgen Mägert, Axel Schulz, Wolf-Georg Forssmann, Peter Schulz-Knappe, and Knut Adermann. "LEAP-1, a novel highly disulfide-bonded human peptide, exhibits antimicrobial activity." *FEBS letters* 480, no. 2-3 (2000): 147-150.

Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." *Journal of molecular biology* 305, no. 3 (2001): 567-580.

Kugeratski, Fernanda G., Samuel J. Atkinson, Lisa J. Neilson, Sergio Lilla, John RP Knight, Jens Serneels, Amelie Juin et al. "Hypoxic cancer–associated fibroblasts increase NCBP2-AS2/HIAR to promote endothelial sprouting through enhanced VEGF signaling." *Sci. Signal.* 12, no. 567 (2019): eaan8247.

Lee, Changhan, Jennifer Zeng, Brian G. Drew, Tamer Sallam, Alejandro Martin-Montalvo, Junxiang Wan, Su-Jeong Kim et al. "The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance." *Cell metabolism* 21, no. 3 (2015): 443-454.

Lee, Sooncheol, Botao Liu, Soohyun Lee, Sheng-Xiong Huang, Ben Shen, and Shu-Bing Qian. "Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution." *Proceedings of the National Academy of Sciences* 109, no. 37 (2012): E2424-E2432.

Ladoukakis, Emmanuel, Vini Pereira, Emile G. Magny, Adam Eyre-Walker, and Juan Pablo Couso. "Hundreds of putatively functional small open reading frames in *Drosophila*." *Genome biology* 12, no. 11 (2011): R118.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome biology* 10, no. 3 (2009): R25.

Law, G. Lynn, Alexa Raney, Carrie Heusner, and David R. Morris. "Polyamine regulation of ribosome pausing at the upstream open reading frame of S-adenosylmethionine decarboxylase." *Journal of Biological Chemistry* 276, no. 41 (2001): 38036-38043.

Legrand, Carine, and Francesca Tuorto. "RiboVIEW: a computational framework for visualization, quality control and statistical analysis of ribosome profiling data." *Nucleic acids research* 48, no. 2 (2020): e7-e7.

Lin, Michael F., Irwin Jungreis, and Manolis Kellis. "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." *Bioinformatics* 27, no. 13 (2011): i275-i282.

Lin, Yi-Fang, Man-Huan Xiao, Hua-Xing Chen, Yu Meng, Na Zhao, Liang Yang, Haite Tang et al. "A novel mitochondrial micropeptide MPM enhances mitochondrial respiratory activity and promotes myogenic differentiation." *Cell death & disease* 10, no. 7 (2019): 1-11.

Lipman, David J., Alexander Souvorov, Eugene V. Koonin, Anna R. Panchenko, and Tatiana A. Tatusova. "The relationship of protein conservation and sequence length." *BMC Evolutionary Biology* 2, no. 1 (2002): 20.

Liu, Hao, Xiaolei Tang, and Lei Gong. "Mesencephalic astrocyte-derived neurotrophic factor and cerebral dopamine neurotrophic factor: New endoplasmic reticulum stress response proteins." *European journal of pharmacology* 750 (2015): 118-122.

Locksley, Richard M., Nigel Killeen, and Michael J. Lenardo. "The TNF and TNF receptor superfamilies: integrating mammalian biology." *Cell* 104, no. 4 (2001): 487-501.

Loughran, Gary, Alexander V. Zhdanov, Maria S. Mikhaylova, Fedor N. Rozov, Petr N. Datskevich, Sergey I. Kovalchuk, Marina V. Serebryakova et al. "Unprecedentedly efficient CUG initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF." *BioRxiv* (2020).

Luo, Chong T., Will Liao, Saida Dadi, Ahmed Toure, and Ming O. Li. "Graded Foxo1 activity in T reg cells differentiates tumour immunity from spontaneous autoimmunity." *Nature* 529, no. 7587 (2016): 532-536.

Luukkonen, B. G., Wei Tan, and Stefan Schwartz. "Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance." *Journal of virology* 69, no. 7 (1995): 4086-4094.

Ma, Jiao, Jolene K. Diedrich, Irwin Jungreis, Cynthia Donaldson, Joan Vaughan, Manolis Kellis, John R. Yates III, and Alan Saghatelian. "Improved identification and analysis of small open reading frame encoded polypeptides." *Analytical chemistry* 88, no. 7 (2016): 3967-3975.

Mackowiak, Sebastian D., Henrik Zauber, Chris Bielow, Denise Thiel, Kamila Kutz, Lorenzo Calviello, Guido Mastrobuoni et al. "Extensive identification and analysis of conserved small ORFs in animals." *Genome biology* 16, no. 1 (2015): 179.

Magny, Emile G., Jose Ignacio Pueyo, Frances MG Pearl, Miguel Angel Cespedes, Jeremy E. Niven, Sarah A. Bishop, and Juan Pablo Couso. "Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames." *Science* 341, no. 6150 (2013): 1116-1120.

Makałowski, Wojciech, and Mark S. Boguski. "Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences." *Proceedings of the National Academy of Sciences* 95, no. 16 (1998): 9407-9412.

Makarewich, Catherine A., and Eric N. Olson. "Mining for micropeptides." *Trends in cell biology* 27, no. 9 (2017): 685-696.

Makarewich, Catherine A., Amir Z. Munir, Gabriele G. Schiattarella, Svetlana Bezprozvannaya, Olga N. Raguimova, Ellen E. Cho, Alexander H. Vidal, Seth L. Robia, Rhonda Bassel-Duby, and Eric N. Olson. "The DWORF micropeptide enhances contractility and prevents heart failure in a mouse model of dilated cardiomyopathy." *Elife* 7 (2018): e38319.

Maki, Kimika, Teppei Morita, Hironori Otaka, and Hiroji Aiba. "A minimal base-pairing region of a bacterial small RNA SgrS required for translational repression of ptsG mRNA." *Molecular microbiology* 76, no. 3 (2010): 782-792.

Malone, Brandon, Ilian Atanassov, Florian Aeschimann, Xinping Li, Helge Großhans, and Christoph Dieterich. "Bayesian prediction of RNA translation from ribosome profiling." *Nucleic acids research* 45, no. 6 (2017): 2960-2972.

Martin, Marcel. "Cutadapt removes adapter sequences from high-throughput sequencing reads." *EMBnet. journal* 17, no. 1 (2011): 10-12.

Martinez, Thomas F., Qian Chu, Cynthia Donaldson, Dan Tan, Maxim N. Shokhirev, and Alan Saghatelian. "Accurate annotation of human protein-coding small open reading frames." *Nature Chemical Biology* (2019): 1-11.

Matsumoto, Akinobu, Alessandra Pasut, Masaki Matsumoto, Riu Yamashita, Jacqueline Fung, Emanuele Monteleone, Alan Saghatelian, Keiichi I. Nakayama, John G. Clohessy, and Pier Paolo Pandolfi. "mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide." *Nature* 541, no. 7636 (2017): 228-232.

Mercer, Tim R., Dagmar Wilhelm, Marcel E. Dinger, Giulia Solda, Darren J. Korbie, Evgeny A. Glazov, Vy Truong et al. "Expression of distinct RNAs from 3′ untranslated regions." *Nucleic acids research* 39, no. 6 (2011): 2393-2403.

Michel, Audrey M., Gearoid Fox, Anmol M. Kiran, Christof De Bo, Patrick BF O’Connor, Stephen M. Heaphy, James PA Mullan, Claire A. Donohue, Desmond G. Higgins, and Pavel V. Baranov. "GWIPS-viz: development of a ribo-seq genome browser." *Nucleic acids research* 42, no. D1 (2014): D859-D864.

Michel, Audrey M., Kingshuk Roy Choudhury, Andrew E. Firth, Nicholas T. Ingolia, John F. Atkins, and Pavel V. Baranov. "Observation of dually decoded regions of the human genome using ribosome profiling data." *Genome research* 22, no. 11 (2012): 2219-2229.

Michel, Audrey M., and Pavel V. Baranov. "Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale." *Wiley Interdisciplinary Reviews: RNA* 4, no. 5 (2013): 473-490.

Mitchell, Alex L., Teresa K. Attwood, Patricia C. Babbitt, Matthias Blum, Peer Bork, Alan Bridge, Shoshana D. Brown et al. "InterPro in 2019: improving coverage, classification and access to protein sequence annotations." *Nucleic acids research* 47, no. D1 (2019): D351-D360.

Morris, David R., and Adam P. Geballe. "Upstream open reading frames as regulators of mRNA translation." *Molecular and cellular biology* 20, no. 23 (2000): 8635-8642.

Murat, Pierre, Giovanni Marsico, Barbara Herdy, Avazeh Ghanbarian, Guillem Portella, and Shankar Balasubramanian. "RNA G-quadruplexes at upstream open reading frames cause DHX36-and DHX9-dependent translation of human mRNAs." *Genome biology* 19, no. 1 (2018): 229.

Na, Ah Ram, Young Min Chung, Seung Baek Lee, Seon Ho Park, Myeong-Sok Lee, and Young Do Yoo. "A critical role for Romo1-derived ROS in cell proliferation." *Biochemical and biophysical research communications* 369, no. 2 (2008): 672-678.

Na, Chan Hyun, Mustafa A. Barbhuiya, Min-Sik Kim, Steven Verbruggen, Stephen M. Eacker, Olga Pletnikova, Juan C. Troncoso et al. "Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini." *Genome research* 28, no. 1 (2018): 25-36.

Narita, Noriyuki N., Sally Moore, Gorou Horiguchi, Minoru Kubo, Taku Demura, Hiroo Fukuda, Justin Goodrich, and Hirokazu Tsukaya. "Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in Arabidopsis thaliana." *The Plant Journal* 38, no. 4 (2004): 699-713.

Neave, Nick. *Hormones and behaviour: a psychological approach*. Cambridge University Press, 2007.

Nelson, Benjamin R., Catherine A. Makarewich, Austin L. Reese, Benjamin R. Winders, Douglas M. Anderson, John R. McAnally, Ege T. Kavalali, Rhonda Bassel-Duby, and Eric N. Olson. "DWORF: Discovery and Characterization of a Cardiac Micropeptide Encoded in a Putative Long Noncoding RNA." *Circulation Research* 117, no. suppl_1 (2015): A189-A189.

Nelson, Benjamin R., Catherine A. Makarewich, Douglas M. Anderson, Benjamin R. Winders, Constantine D. Troupes, Fenfen Wu, Austin L. Reese et al. "A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle." *Science* 351, no. 6270 (2016): 271-275.

Neves, Joana, Jie Zhu, Pedro Sousa-Victor, Mia Konjikusic, Rebeccah Riley, Shereen Chew, Yanyan Qi, Heinrich Jasper, and Deepak A. Lamba. "Immune modulation by MANF promotes tissue repair and regenerative success in the retina." *Science* 353, no. 6294 (2016): aaf3646.

Nojima, Takuya, Kei Haniuda, Tatsuya Moutai, Moeko Matsudaira, Sho Mizokawa, Ikuo Shiratori, Takachika Azuma, and Daisuke Kitamura. "In-vitro derived germinal centre B cells differentially generate memory B or plasma cells in vivo." *Nature communications* 2, no. 1 (2011): 1-11.

Oh, Eugene, Annemarie H. Becker, Arzu Sandikci, Damon Huber, Rachna Chaba, Felix Gloge, Robert J. Nichols et al. "Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo." *Cell* 147, no. 6 (2011): 1295-1308.

Olexiouk, Volodimir, Jeroen Crappé, Steven Verbruggen, Kenneth Verhegen, Lennart Martens, and Gerben Menschaert. "sORFs. org: a repository of small ORFs identified by ribosome profiling." *Nucleic acids research* 44, no. D1 (2016): D324-D329.

Oyama, Masaaki, Chiharu Itagaki, Hiroko Hata, Yutaka Suzuki, Tomonori Izumi, Tohru Natsume, Toshiaki Isobe, and Sumio Sugano. "Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs." *Genome research* 14, no. 10b (2004): 2048-2052.

Oyama, Masaaki, Hiroko Kozuka-Hata, Yutaka Suzuki, Kentaro Semba, Tadashi Yamamoto, and Sumio Sugano. "Diversity of translation start sites may define increased complexity of the human short ORFeome." *Molecular & Cellular Proteomics* 6, no. 6 (2007): 1000-1006.

Pállinger, Éva, and György Csaba. "A hormone map of human immune cells showing the presence of adrenocorticotropic hormone, triiodothyronine and endorphin in immunophenotyped white blood cells." *Immunology* 123, no. 4 (2008): 584-589.

Park, Christina H., Erika V. Valore, Alan J. Waring, and Tomas Ganz. "Hepcidin, a urinary antimicrobial peptide synthesized in the liver." *Journal of biological chemistry* 276, no. 11 (2001): 7806-7810.

Patraquim, Pedro, Muhammad Ali S. Mumtaz, Jose I. Pueyo, Julie L. Aspden, and Juan Pablo Couso. "Developmental regulation of Canonical and small ORF translation from mRNAs." *bioRxiv* (2019): 727339.

Pauli, Andrea, Eivind Valen, Michael F. Lin, Manuel Garber, Nadine L. Vastenhouw, Joshua Z. Levin, Lin Fan et al. "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis." *Genome research* 22, no. 3 (2012): 577-591.

Pauli, Andrea, Megan L. Norris, Eivind Valen, Guo-Liang Chew, James A. Gagnon, Steven Zimmerman, Andrew Mitchell et al. "Toddler: an embryonic signal that promotes cell movement via Apelin receptors." *Science* 343, no. 6172 (2014): 1248636.

Pauli, Andrea, Eivind Valen, and Alexander F. Schier. "Identifying (non-) coding RNAs and small peptides: Challenges and opportunities." *Bioessays* 37, no. 1 (2015): 103-112.

Peabody, David S. "Translation initiation at an ACG triplet in mammalian cells." *Journal of Biological Chemistry* 262, no. 24 (1987): 11847-11851.

Peabody, David S. "Translation initiation at non-AUG triplets in mammalian cells." *Journal of Biological Chemistry* 264, no. 9 (1989): 5031-5035.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar Von Heijne, and Henrik Nielsen. "SignalP 4.0: discriminating signal peptides from transmembrane regions." *Nature methods* 8, no. 10 (2011): 785.

Petrova, Penka S., Andrei Raibekas, Jonathan Pevsner, Noel Vigo, Mordechai Anafi, Mary K. Moore, Amy E. Peaire et al. "MANF." *Journal of Molecular Neuroscience* 20, no. 2 (2003): 173-187.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." *Nature biotechnology* 33, no. 3 (2015): 290.

Pierre-Jerome, Edith, Colleen Drapek, and Philip N. Benfey. "Regulation of division and differentiation of plant stem cells." *Annual review of cell and developmental biology* 34 (2018): 289-310.

Platt, Randall J., Sidi Chen, Yang Zhou, Michael J. Yim, Lukasz Swiech, Hannah R. Kempton, James E. Dahlman et al. "CRISPR-Cas9 knockin mice for genome editing and cancer modeling." *Cell* 159, no. 2 (2014): 440-455.

Prensner, John R., Oana M. Enache, Victor Luria, Karsten Krug, Karl R. Clauser, Joshua M. Dempster, Amir Karger et al. "Non-canonical open reading frames encode functional proteins essential for cancer cell survival." *bioRxiv* (2020).

Pruitt, Kim D., Garth R. Brown, Susan M. Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M. Farrell et al. "RefSeq: an update on mammalian reference sequences." *Nucleic acids research* 42, no. D1 (2014): D756-D763.

Pueyo, Jose I., Emile G. Magny, and Juan P. Couso. "New peptides under the s (ORF) ace of the genome." *Trends in biochemical sciences* 41, no. 8 (2016): 665-678.

Ran, F. Ann, Patrick D. Hsu, Jason Wright, Vineeta Agarwala, David A. Scott, and Feng Zhang. "Genome engineering using the CRISPR-Cas9 system." *Nature protocols* 8, no. 11 (2013): 2281.

Raney, Alexa, G. Lynn Law, Gregory J. Mize, and David R. Morris. "Regulated translation termination at the upstream open reading frame in S-adenosylmethionine decarboxylase mRNA." *Journal of Biological Chemistry* 277, no. 8 (2002): 5988-5994.

Rathore, Annie, Thomas F. Martinez, Qian Chu, and Alan Saghatelian. "Small, but mighty? Searching for human microproteins and their potential for understanding health and disease." (2018): 963-965.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. "g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)." *Nucleic acids research* 47, no. W1 (2019): W191-W198.

Rice, Jennifer B., and Carin K. Vanderpool. "The small RNA SgrS controls sugar–phosphate accumulation by regulating multiple PTS genes." *Nucleic acids research* 39, no. 9 (2011): 3806-3819.

Robison K. (2009) Omics! Omics! http://omicsomics.blogspot.com/2009/04/is-codon-optimization-bunk.html.

Röhrig, Horst, Jürgen Schmidt, Edvins Miklashevichs, Jeff Schell, and Michael John. "Soybean ENOD40 encodes two peptides that bind to sucrose synthase." *Proceedings of the National Academy of Sciences* 99, no. 4 (2002): 1915-1920.

Roosild, Tarmo P., Samantha Castronovo, and Senyon Choe. "Structure of anti-FLAG M2 Fab domain and its use in the stabilization of engineered membrane proteins." *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 62, no. 9 (2006): 835-839.

Ruan, Hangjun, Lisa M. Shantz, Anthony E. Pegg, and David R. Morris. "The upstream open reading frame of the mRNA encoding S-adenosylmethionine decarboxylase is a polyamine-responsive translational control element." *Journal of Biological Chemistry* 271, no. 47 (1996): 29576-29582.

Saghatelian, Alan, and Juan Pablo Couso. "Discovery and characterization of smORF-encoded bioactive polypeptides." *Nature chemical biology* 11, no. 12 (2015): 909.

Saito, Shigeru. "Cytokine cross-talk between mother and the embryo/placenta." *Journal of reproductive immunology* 52, no. 1-2 (2001): 15-33.

Savard, Joël, Henrique Marques-Souza, Manuel Aranda, and Diethard Tautz. "A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides." *Cell* 126, no. 3 (2006): 559-569.

Sendoel, Ataman, Joshua G. Dunn, Edwin H. Rodriguez, Shruti Naik, Nicholas C. Gomez, Brian Hurwitz, John Levorse et al. "Translation from unconventional 5′ start sites drives tumour initiation." *Nature* 541, no. 7638 (2017): 494-499.

Schiemann, Ronja, Kay Lammers, Maren Janz, Jana Lohmann, Achim Paululat, and Heiko Meyer. "Identification and in vivo characterisation of cardioactive peptides in *Drosophila melanogaster."* International journal of molecular sciences 20, no. 1 (2019): 2.

Schittek, Birgit, Rainer Hipfel, Birgit Sauer, Jürgen Bauer, Hubert Kalbacher, Stefan Stevanovic, Markus Schirle et al. "Dermcidin: a novel human antibiotic peptide secreted by sweat glands." *Nature Immunology* 2, no. 12 (2001): 1133-1137.

Schwaid, Adam G., D. Alexander Shannon, Jiao Ma, Sarah A. Slavoff, Joshua Z. Levin, Eranthie Weerapana, and Alan Saghatelian. "Chemoproteomic discovery of cysteine-containing human short open reading frames." *Journal of the American Chemical Society* 135, no. 45 (2013): 16750-16753.

Sciammas, Roger, A. L. Shaffer, Jonathan H. Schatz, Hong Zhao, Louis M. Staudt, and Harinder Singh. "Graded expression of interferon regulatory factor-4 coordinates isotype switching with plasma cell differentiation." *Immunity* 25, no. 2 (2006): 225-236.

Sha, Jibin, Guang Zhao, Xiaojuan Chen, Weiping Guan, Yanling He, and Zhaoqing Wang. "Antibacterial potential of hGlyrichin encoded by a human gene." *Journal of Peptide Science* 18, no. 2 (2012): 97-104.

Shi, Wei, Yang Liao, Simon N. Willis, Nadine Taubenheim, Michael Inouye, David M. Tarlinton, Gordon K. Smyth, Philip D. Hodgkin, Stephen L. Nutt, and Lynn M. Corcoran. "Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells." *Nature Immunology* 16, no. 6 (2015): 663.

Shigenaga, Alexandra M., and Cristiana T. Argueso. "No hormone to rule them all: Interactions of plant hormones during the responses of plants to pathogens." In *Seminars in Cell & Developmental Biology*, vol. 56, pp. 174-189. Academic Press, 2016.

Skarshewski, Adam, Mitchell Stanton-Cook, Thomas Huber, Sumaya Al Mansoori, Ross Smith, Scott A. Beatson, and Joseph A. Rothnagel. "uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation." *BMC bioinformatics* 15, no. 1 (2014): 36.

Slavoff, Sarah A., Andrew J. Mitchell, Adam G. Schwaid, Moran N. Cabili, Jiao Ma, Joshua Z. Levin, Amir D. Karger, Bogdan A. Budnik, John L. Rinn, and Alan Saghatelian. "Peptidomic discovery of short open reading frame–encoded peptides in human cells." *Nature chemical biology* 9, no. 1 (2013): 59.

Slavoff, Sarah A., Jinho Heo, Bogdan A. Budnik, Leslyn A. Hanakahi, and Alan Saghatelian. "A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining." *Journal of Biological Chemistry* 289, no. 16 (2014): 10950-10957.

Smedley, Damian, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz et al. "The BioMart community portal: an innovative alternative to large, centralized data repositories." *Nucleic acids research* 43, no. W1 (2015): W589-W598.

Smith, Jenna E., Juan R. Alvarez-Dominguez, Nicholas Kline, Nathan J. Huynh, Sarah Geisler, Wenqian Hu, Jeff Coller, and Kristian E. Baker. "Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae." *Cell reports* 7, no. 6 (2014): 1858-1866.

Sousa-Victor, Pedro, Heinrich Jasper, and Joana Neves. "Trophic factors in inflammation and regeneration: the role of MANF and CDNF." *Frontiers in physiology* 9 (2018): 1629.

Steijger, Tamara, Josep F. Abril, Pär G. Engström, Felix Kokocinski, Martin Akerman, Tyler Alioto, Giovanna Ambrosini et al. "Assessment of transcript reconstruction methods for RNA-seq." *Nature methods* 10, no. 12 (2013): 1177-1184.

Stein, Colleen S., Pooja Jadiya, Xiaoming Zhang, Jared M. McLendon, Gabrielle M. Abouassaly, Nathan H. Witmer, Ethan J. Anderson, John W. Elrod, and Ryan L. Boudreau. "Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency." *Cell reports* 23, no. 13 (2018): 3710-3720.

Steiner, Donald F., and Philip E. Oyer. "The biosynthesis of insulin and a probable precursor of insulin by a human islet cell adenoma." *Proceedings of the National Academy of Sciences of the United States of America* 57, no. 2 (1967): 473.

Stern-Ginossar, Noam, Ben Weisburd, Annette Michalski, Vu Thuy Khanh Le, Marco Y. Hein, Sheng-Xiong Huang, Ming Ma et al. "Decoding human cytomegalovirus." *Science* 338, no. 6110 (2012): 1088-1093.

Swardfager, Walter, Krista Lanctôt, Lana Rothenburg, Amy Wong, Jaclyn Cappell, and Nathan Herrmann. "A meta-analysis of cytokines in Alzheimer's disease." *Biological psychiatry* 68, no. 10 (2010): 930-941.

Tadimalla, Archana, Peter J. Belmont, Donna J. Thuerauf, Matthew S. Glassy, Joshua J. Martindale, Natalie Gude, Mark A. Sussman, and Christopher C. Glembotski. "Mesencephalic astrocyte-derived neurotrophic factor is an ischemia-inducible secreted endoplasmic reticulum stress response protein in the heart." *Circulation research* 103, no. 11 (2008): 1249-1258.

Tajima, Hirohisa, Takako Niikura, Yuichi Hashimoto, Yuko Ito, Yoshiko Kita, Kenzo Terashita, Kazuto Yamazaki, Atsuo Koto, Sadakazu Aiso, and Ikuo Nishimoto. "Evidence for in vivo production of Humanin peptide, a neuroprotective factor against Alzheimer's disease-related insults." *Neuroscience letters* 324, no. 3 (2002): 227-231.

Tellier, Julie, Wei Shi, Martina Minnich, Yang Liao, Simon Crawford, Gordon K. Smyth, Axel Kallies, Meinrad Busslinger, and Stephen L. Nutt. "Blimp-1 controls plasma cell function through the regulation of immunoglobulin secretion and the unfolded protein response." *Nature Immunology* 17, no. 3 (2016): 323.

Tiedje, Christopher, Manuel D. Diaz-Muñoz, Philipp Trulley, Helena Ahlfors, Kathrin Laaß, Perry J. Blackshear, Martin Turner, and Matthias Gaestel. "The RNA-binding protein TTP is a global post-transcriptional regulator of feedback control in inflammation." *Nucleic acids research* 44, no. 15 (2016): 7418-7440.

Uhlén, Mathias, Max J. Karlsson, Andreas Hober, Anne-Sophie Svensson, Julia Scheffel, David Kotol, Wen Zhong et al. "The human secretome." *Science signaling* 12, no. 609 (2019).

Ulitsky, Igor, Alena Shkumatava, Calvin H. Jan, Hazel Sive, and David P. Bartel. "Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution." *Cell* 147, no. 7 (2011): 1537-1550.

UniProt Consortium. "UniProt: a worldwide hub of protein knowledge." *Nucleic acids research* 47, no. D1 (2019): D506-D515.

van Heesch, Sebastiaan, Franziska Witte, Valentin Schneider-Lunitz, Jana F. Schulz, Eleonora Adami, Allison B. Faber, Marieluise Kirchner et al. "The translational landscape of the human heart." *Cell* 178, no. 1 (2019): 242-260.

Vanderperre, Benoît, Jean-François Lucier, Cyntia Bissonnette, Julie Motard, Guillaume Tremblay, Solène Vanderperre, Maxence Wisztorski, Michel Salzet, François-Michel Boisvert, and Xavier Roucou. "Direct detection of alternative open reading frames translation products in human significantly expands the proteome." *PloS one* 8, no. 8 (2013).

Vasquez, Juan-Jose, Chung-Chau Hon, Jens T. Vanselow, Andreas Schlosser, and T. Nicolai Siegel. "Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages." *Nucleic acids research* 42, no. 6 (2014): 3623-3637.

Wadler, Caryn S., and Carin K. Vanderpool. "A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide." *Proceedings of the National Academy of Sciences* 104, no. 51 (2007): 20454-20459.

Wan, Ji, and Shu-Bing Qian. "TISdb: a database for alternative translation initiation in mammalian cells." *Nucleic acids research* 42, no. D1 (2014): D845-D850.

Wang, Xue-Qing, and Joseph A. Rothnagel. "5′-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation." *Nucleic acids research* 32, no. 4 (2004): 1382-1391.

Wethmar, Klaus, Adriano Barbosa-Silva, Miguel A. Andrade-Navarro, and Achim Leutz. "uORFdb—a comprehensive literature database on eukaryotic uORF biology." *Nucleic acids research* 42, no. D1 (2014): D60-D67.

Wheeler, David L., Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church et al. "Database resources of the national center for biotechnology information." *Nucleic acids research* 36, no. suppl_1 (2007): D13-D21.

Whiffin, Nicola, Konrad J. Karczewski, Xiaolei Zhang, Sonia Chothani, Miriam J. Smith, D. Gareth Evans, Angharad M. Roberts et al. "Characterising the loss-of-function impact of 5'untranslated region variants in whole genome sequence data from 15,708 individuals." *BioRxiv* (2019): 543504.

Willingham, A. T., S. Dike, J. Cheng, J. R. Manak, I. Bell, E. Cheung, J. Drenkow et al. "Transcriptional landscape of the human and fly genomes: nonlinear and multifunctional modular model of transcriptomes." In *Cold Spring Harbor symposia on quantitative biology*, vol. 71, pp. 101-110. Cold Spring Harbor Laboratory Press, 2006.

Won, Woong-Jai, Martin F. Bachmann, and John F. Kearney. "CD36 is differentially expressed on B cell subsets during development and in responses to antigen." *The Journal of Immunology* 180, no. 1 (2008): 230-237.

Woo, Sunghee, Seong Won Cha, Gennifer Merrihew, Yupeng He, Natalie Castellana, Clark Guest, Michael MacCoss, and Vineet Bafna. "Proteogenomic database construction driven from large scale RNA-seq data." *Journal of proteome research* 13, no. 1 (2014): 21-28.

Wu, Wei-Sheng, Yu-Xuan Jiang, Jer-Wei Chang, Yu-Han Chu, Yi-Hao Chiu, Yi-Hong Tsao, Torbjörn EM Nordling, Yan-Yuan Tseng, and Joseph T. Tseng. "HRPDviewer: human ribosome profiling data viewer." *Database* 2018 (2018).

Wu, Yu, and Cynthia M. Smas. "Wdnm1-like, a new adipokine with a role in MMP-2 activation." *American Journal of Physiology-Endocrinology and Metabolism* 295, no. 1 (2008): E205-E215.

Xiao, Zhengtao, Rongyao Huang, Xudong Xing, Yuling Chen, Haiteng Deng, and Xuerui Yang. "De novo annotation and characterization of the translatome with ribosome profiling data." *Nucleic acids research* 46, no. 10 (2018): e61-e61.

Xie, Shang-Qian, Peng Nie, Yan Wang, Hongwei Wang, Hongyu Li, Zhilong Yang, Yizhi Liu, Jian Ren, and Zhi Xie. "RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling." *Nucleic acids research* 44, no. D1 (2015): D254-D258.

Xu, Wenli, Bing Deng, Penghui Lin, Chang Liu, Bin Li, Qiaojuan Huang, Hui Zhou, Jianhua Yang, and Lianghu Qu. "Ribosome profiling analysis identified a KRAS-interacting microprotein that represses oncogenic signaling in hepatocellular carcinoma cells." *Science China Life Sciences* (2019): 1-14.

Yan, Yahui, Claudia Rato, Lukas Rohland, Steffen Preissler, and David Ron. "MANF antagonizes nucleotide exchange by the endoplasmic reticulum chaperone BiP." *Nature communications* 10, no. 1 (2019): 541.

Yang, Xiaohan, Timothy J. Tschaplinski, Gregory B. Hurst, Sara Jawdy, Paul E. Abraham, Patricia K. Lankford, Rachel M. Adams et al. "Discovery and annotation of small proteins using genomics, proteomics, and computational approaches." *Genome research* 21, no. 4 (2011): 634-641.

Yizhak, Keren, François Aguet, Jaegil Kim, Julian M. Hess, Kirsten Kübler, Jonna Grimsby, Ruslana Frazer et al. "RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues." *Science* 364, no. 6444 (2019): eaaw0726.

Yokota, Takashi, Takeshi Otsuka, Timothy Mosmann, Jacques Banchereau, Thierry DeFrance, Dominique Blanchard, Jan E. De Vries, Frank Lee, and K. I. Arai. "Isolation and characterization of a human interleukin cDNA clone, homologous to mouse B-cell stimulatory factor 1, that expresses B-cell-and T-cell-stimulating activities." *Proceedings of the National Academy of Sciences* 83, no. 16 (1986): 5894-5898.

Yoshida, Hideyuki, Caleb A. Lareau, Ricardo N. Ramirez, Samuel A. Rose, Barbara Maier, Aleksandra Wroblewska, Fiona Desland et al. "The cis-regulatory atlas of the mouse immune system." *Cell* 176, no. 4 (2019): 897-912.

Zhang, Hong, Yirong Wang, and Jian Lu. "Function and evolution of upstream ORFs in eukaryotes." *Trends in biochemical sciences* (2019).

Zhang, Qiao, Ajay A. Vashisht, Jason O'Rourke, Stéphane Y. Corbel, Rita Moran, Angelica Romero, Loren Miraglia et al. "The microprotein Minion controls cell fusion and muscle formation." *Nature communications* 8, no. 1 (2017): 1-15.

Zhang, Shan, Boris Reljić, Chao Liang, Baptiste Kerouanton, Joel Celio Francisco, Jih Hou Peh, Camille Mary et al. "Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly." *Nature Communications* 11, no. 1 (2020): 1-16.

Zhao, Hua, Yi Liu, Lei Cheng, Ben Liu, Wen Zhang, Ying-Jun Guo, and Lin Nie. "Mesencephalic astrocyte-derived neurotrophic factor inhibits oxygen–glucose deprivation-induced cell damage and inflammation by suppressing endoplasmic reticulum stress in rat primary astrocytes." *Journal of molecular neuroscience* 51, no. 3 (2013): 671-678.

Zitomer, R. S., D. A. Walthall, B. C. Rymond, and C. P. Hollenberg. "Saccharomyces cerevisiae ribosomes recognize non-AUG initiation codons." *Molecular and cellular biology* 4, no. 7 (1984): 1191-1197.

Appendix A. Pipeline Instructions

The most updated pipeline source code and instructions are available on GitHub at – https://github.com/boboppie/ORFLine.

We also create a Singularity image which enables the users to execute and test the pipeline easily in a virtual environment, All dependencies including bioinformatics tools are pre-installed in the image, the URL is https://github.com/boboppie/ORFLine-singularity.

In case the above links are broken, the source code and instructions can be downloaded from a shared folder - https://bit.ly/2mSh6Fm.

ORFLine

This repository holds the pipeline for prediction of actively translated small open reading frames (smORFs) in the immune system.

Obtaining

To download the source code, please use git to download the most recent development tree. Currently, the tree is hosted on github, and can be obtained via:

```
git clone git://github.com/boboppie/ORFLine.git
```

Dependencies

- Samtools and HTSlib
- bedtools
- BEDOPS
- Bowtie
- STAR
- FastQC
- Trim Galore
- plastid

- StringTie
- EMBOSS
- GNU Parallel
- R
- Bioconductor

R/Bioconductor packages:

- riboSeqR
- GenomicFeatures
- rtracklayer

Dataset

We will use *Diaz-Muñoz et al, 2015* LPS activated B cell dataset as an example to demonstrate typical workflow.

Download raw sequencing data from EBI:

```
RNA-Seq  -
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR160/001/SRR1605271/SRR1605271.fastq.gz

Ribo-Seq -
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR160/004/SRR1605304/SRR1605304.fastq.gz
```

Workflow

1. Check if all the dependencies are installed

```
        bash ./module-check.sh
```

2. Download and generate files that are used in the pipeline

```
        bash ./ref-download.sh -o mouse -r M22 -t 4
```

3. Generate putative ORFs

```
        bash ./orf-prediction.sh -o \"Mus musculus\" -t 8
```

4. Ribosome profiling (Ribo-Seq) data processing

```
        bash ./riboseq-process.sh -f ./out/data/ribo-seq/ribo.fastq.gz -a
AAAAAAAAAAA -t 4
```

5. RNA-Seq data processing

```
bash ./rnaseq-process.sh -f ./out/data/rna-seq/rna.fastq.gz -t 4
```

6. ORF calling

```
bash ./orf-calling.sh -o mouse -x 10090 -m 32 -n 28 -t 8
```

Output

The final output file in *info_table* directory is in BED12 format with extension.

Column	**Description**
1 - 12	The first 12 columns are in BED12 format, the fields are described here - https://genome.ucsc.edu/FAQ/FAQformat.html#format1. The 4th column is ORFId (transcript-based).
13	smORF class, including canonical, five_prime...
14	Peptide length
15	RegionId (genomic-based)
16	Ensembl transcript Id
17	Gene symbol
18	Gene description
19	ORF score
20	Ribosome release score
21	Ribo FPFM
22	RNA FPKM
23	Translation efficiency (TE)
24	CDS TE (NA if host transcript is noncoding)
25	AA sequence

Run the pipeline in a virtual machine

We recommend running a test on a virtual machine, e.g. VirtualBox with a minimal ISO (e.g. CentOS 7 minimal). Users can install all dependencies via Miniconda, for example:

```
# Tools to install on CentOS before miniconda
yum -y install gcc tar bzip2 git which

curl -fsSL https://repo.anaconda.com/miniconda/Miniconda2-latest-Linux-x86_64.sh -o
miniconda2.sh

# assume miniconda is installed in the home directory
bash miniconda2.sh -b -p ~/miniconda2

export PATH=~/miniconda2/bin:$PATH
export PYTHONPATH=~/miniconda2/lib/python2.7/site-packages

conda install -y -c conda-forge wget
conda install -y -c conda-forge parallel
conda install -y -c bioconda samtools
conda install -y -c bioconda htslib
conda install -y -c bioconda bedtools
conda install -y -c bioconda bedops
conda install -y -c bioconda bowtie
conda install -y -c bioconda fastqc
conda install -y -c bioconda cutadapt
conda install -y -c bioconda trim-galore
conda install -y -c bioconda star
conda install -y -c bioconda stringtie
conda install -y -c bioconda sra-tools
conda install -y -c bioconda emboss
conda install -y -c bioconda plastid
conda install -y -c bioconda bioconductor-rhtslib
Rscript -e 'install.packages("BiocManager", repos="http://cran.us.r-project.org");
BiocManager::install(c("riboSeqR", "GenomicFeatures", "rtracklayer"))'
```

We have a main.sh script to run all the steps mentioned above, you can simply pull the source code and run it as:

```
git clone https://github.com/boboppie/ORFLine.git
cd orf-discovery
chmod +x *.sh

bash ./main.sh
```

Run the pipeline in a Singularity image

An easier way to test the pipeline is to run a Singularity image we created (see the following section “ORFLine Singularity Image” for more information). This will avoid installing all the dependencies.

ORFLine Singularity Image

We have created a Singularity image, the image was automatically build by Singularity Hub (https://singularity-hub.org/). During image build, all ORFLine dependencies were installed, specifically, bioinformatics tools were installed via Miniconda to */opt/miniconda/bin* and pipeline source code was pulled to *~/project/*.

Usage

Install singularity via Conda:

```
# Assuming Linux and root privilige
curl -fsSL https://repo.anaconda.com/miniconda/Miniconda2-latest-Linux-x86_64.sh -o
miniconda2.sh

# Install miniconda to user home directory
bash miniconda2.sh -b -p ~/miniconda2

# Add conda bin to $PATH
export PATH=~/miniconda2/bin:$PATH

# Install singularity
conda install -y -c bioconda singularity
```

Pull the container to your machine:

```
singularity pull shub://boboppie/ORFLine-singularity
```

Shell into the container:

```
singularity shell ORFLine-singularity_latest.sif
```

Run the container:

```
singularity run ORFLine-singularity_latest.sif
```

Appendix B. Additional information of the materials and experiments

RNA-Seq and Ribo-Seq experiments

Mice

Mice used in the experimental setups were on the C57BL/6 background. B cell setup 1: These B cells were from Elavl1$^{fl/fl}$ mice (Diaz-Muñoz et al., 2015). B cell setup 2: Zfp36l1$^{fl/fl}$ mice were used. CD4^{+} T cell setup: Zfp36$^{fl/fl}$Zfp36l1$^{fl/fl}$ mice. These mice were littermate controls used for comparison with B or T cell specific Cre mice that are not part of this study. For CRISPR/Cas9-mediated knockout of Manf the mice used were derived by crossing strains: Cd79a^{cre} (Cd79a$^{tm1(cre)Reth}$) (Hobeika et al., 2006) and Cas9-GFP (Gt(ROSA)26Sor$^{tm1(CAG\text{-}cas9*,\text{-}EGFP)Fezh}$) (Platt et al., 2014).

Tissue culture

B cells from spleen or peripheral lymph nodes (LNs) were isolated using the B Cell Isolation Kit (Miltenyi Biotec). For activation, B cells were cultured for 48 hours in RPMI 1640 Medium (Dutch Modification) supplemented with 10% FCS, 100 IU/ml penicillin, 100 µg/ml streptomycin, 2 mM L-GlutaMAX (Gibco), 1 mM Sodium Pyruvate and 50 µM β-mercaptoethanol in the presence of 10 mg/ml of LPS (Sigma, E. Coli 0127: B8), 10 ng/ml of IL4 and 10 ng/ml of IL5. T cells from spleen, peripheral and mesenteric LNs were isolated with CD4+CD62L+ T Cell Isolation Kit (Miltenyi Biotec) and stimulated in the same medium as for B cells using plate bound anti-CD3 (2C11) and 1 µg/ml of anti-CD28 (37.51) for 24 hours.

Sequencing library preparation

RNA-Seq libraries were obtained using TruSeq Stranded mRNA Sample Prep Kit (Illumina Inc). After isolation B cells were either processed directly for RNA extraction (*ex vivo* samples, n=4) or were stimulated with LPS (two experiments one with n=3 and one n=5) prior to RNA preparation. Ribo-Seq libraries were prepared with ARTseq™ Ribosome Profiling Kit (Epicentre, Illumina). *Ex vivo* or LPS-activated B cells (n=4-5) were treated with cycloheximide (CHX, 100 µg/ml) three minutes before a rapid cooling of the culturing plate; then cells were collected, and RNA extracts were prepared according to the kit manual. cDNA libraries were sequenced using Illumina HiSeq2000 system in a 100-bp single-end (RNA-Seq) or 50-bp single-end (Ribo-Seq) mode.

CRISPR/Cas9-mediated knockout of Manf in B cells

iGB cell culture

B cells from Cd79^{cre}+ and Cd79^{cre}- Cas9-GFP mice were purified via negative isolation (Miltenyi 130-090-862) and seeded at a ratio of 4:1 (3x10^{4} cells per well in a 12 well plate) in the presence of pre-seeded irradiated 40LB cells (120Gy). From day 0 to day 4 B cells were cultured with 10 ng/ml rIL-4 (Peprotech 214-14). On day 3, B cells underwent retrovirus transduction of sgRNA constructs by spinfection (1000 g, 32 ^{o}C, 45 minutes) in the presence of 4ng/ml polybrene. On day 4 cells were re-plated on pre-seeded irradiated 40LB cells. From day 4 to day 8 B cells were cultured with 10 ng/ml rIL-21 (Peprotech 210-21). Throughout the culture B cells were maintained in a humidified atmosphere at 37 °C with 5% CO_2 in RMPI-1640 medium (Gibco 21870076) supplemented with 10% FBS (Gibco 12657011), 50µM 2-ME (Gibco 31350010), 100 units ml^{-1} penicillin (Gibco 15070063) and 100 µg ml^{-1} streptomycin (Gibco 15070063).

Figure A.1 | **CRISPR/Cas9-mediated knockout of Manf in B cells limit plasmablast numbers in culture.**

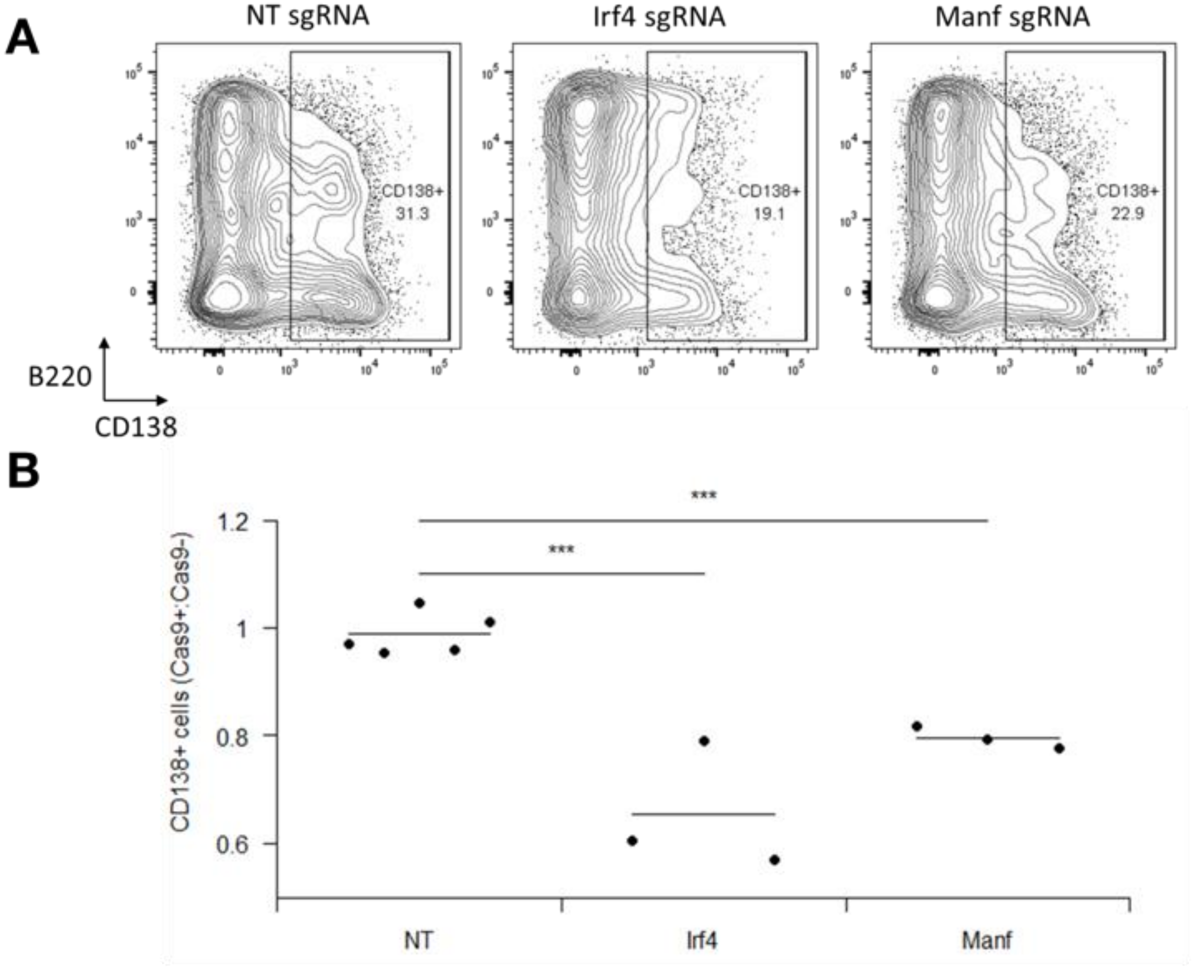


Figure A.1 | **CRISPR/Cas9-mediated knockout of Manf in B cells limit plasmablast numbers in culture.** (A) Flow-cytometry analysis, pre-gated on CD19+Cas9+Thy1.1+ cells, of the expression of the plasmablast marker CD138 in iGB cells transduced with non-targeting sgRNA, a sgRNA against transcription factor Irf4; and a sgRNA against Manf. Numbers in outlined areas indicate percentage of cells in each gate. (B) A ratio between the CD138+ proportions of Cas9+ and Cas9- populations in co-culture. Analysis limited to CD19+Thy1.1+ cells. Each symbol represents an individual sgRNA. *** P < 0.001. Significance was computed using two-tailed unpaired t-test. Data shown are three independent sgRNAs and representative of two independent experiments. Cell culture and flow cytometry was performed by David Turner.

Expression of micropeptides in 293T cells

293T cells were seeded at 7.5 × 10^5 cells/well in 6-well plates overnight. Cells were transfected with 1 µg of empty vector (EV), 1500011B03Rik (clone 6) or Phf21a (clone 12), with 3 µl TransIT 293 reagent, mixed in 250 µl OptiMEM for 30 min at room temperature before added to 2.5 ml antibiotics-free complete media drop-wise. 4 hours post transfection, one well of 1500011B03Rik and Phf21a each were replaced with 2.5 ml OptiMEM containing 2% FBS, other wells replaced with 2.5 ml antibiotics-free complete media containing 10% FBS. 44 hours post transfection, supernatant was harvested by spinning at 300 x g for 5 minutes at 4 and collected supernatant. Total cell lysates were washed twice with PBS, spun at 300 × g for 5 minutes at 4 ^{o}C, resuspended with 50 µl PBS and lysed with 50 µl 2x RIPA buffer containing 1:100 protease inhibitors cocktails. Cells were lysed on ice for 10 minutes and centrifuged at 21,000 × g for 5 minutes at 4 ^{o}C and protein concentrations determined by BCA assay. 80 µg total cell lysates and 30 µl supernatant were analysed. Samples were separated by Tris/Tricine SDS-PAGE (15% T, 2.6% C) at 120V, transferred with 30 minutes 0.2 A constant for membrane 1, followed by 70 minutes 0.2 A constant for membrane 2. Membranes were stained with Ponceau S, blocked with Odyssey® blocking buffer, then stained with 1:1000 mouse anti-FLAG antibody (Sigma) (Roosild et al., 2006) at 4 ^{o}C overnight. After 3 × washes with 1× TBST, Goat anti-mouse 800CW was added at 1:10000 in blocking buffer and incubated at room temperature for 1 hour. The membranes were washed twice with 1 × TBST and once with 1 × TBS and milliQ water before drying and scanned. Rabbit anti-GFP (Clontech 8367) were added to membrane at 1:1000 followed by first scan and developed with Goat anti-rabbit 680IR at 1:10000.

Figure A.2 | **In vitro expression of epitope-tagged micropeptides**

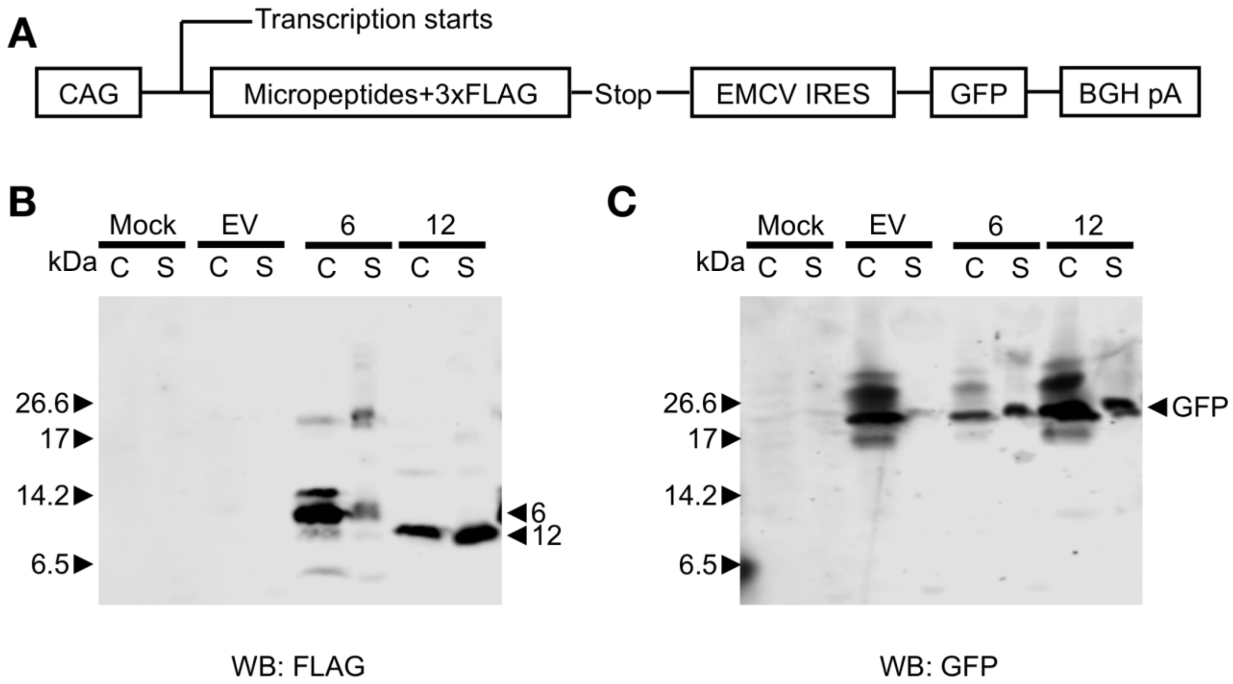


Figure A.2 | **In vitro expression of epitope-tagged micropeptides.** (A) Dicistronic mammalian expression constructs of micropeptides containing C' 3xFLAG tags upstream of EMCV IRES and GFP. 293T were mock transfected or transfected with plasmids encoding: empty vector (EV), 1500011B03Rik (clone 6) or Phf21a (clone 12). 44 hours post transfection, total cell lysates (C) and supernatant (S) were collected and analysed by western blot against: B) FLAG, and C) GFP.