The “dark proteome”:
Discovering new protein-coding genes in non-canonical open-reading frames

Technical Journal Club
Lukas Frick
9 June 2020
How many protein-coding genes exist?

- There are ~\textbf{20’000} protein-coding genes in the human genome...
  ... but the exact number is unknown!

- The catalog of protein-coding genes is derived mainly from the analysis of
  \textbf{coding sequences}
  - bioinformatics pipelines
    ± manual review
  - The algorithms have \textbf{blind spots}!

\textit{GENCODE reference annotation for the human and mouse genomes,}
\textit{Nucleic Acids Research, 2018}
DNA

Non-transcribed

“Junk DNA”
- Retrotransposons, e.g. LINE-1
- ...

Regulatory DNA
- Promoter
- Enhancer
- Silencer
- Centromere
- Telomere
- ...

Transcribed

Non-coding genes
- pri-miRNA
- lncRNA
- rRNA
- tRNA
- snoRNA
- ...

mRNA

~1% of DNA in humans is translated into proteins

(~80% of DNA in prokaryotes)
Only the **CDS** (coding DNA sequence – a subset of the exonic sequence) is translated into protein.
The simplest way to find potential protein-coding sequences is to look for (long) open reading frames (ORFs)

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

An open reading frame is a continuous stretch of codons that begins with a start codon and ends with a stop codon.

AUG

UAA / UAG / UGA

CUG / UUG / GUG are non-canonical
In addition to the start codon, the surrounding nucleotides, i.e. the **Kozak sequence**, determine the initiation of translation.

![Diagram of the Kozak sequence](image)

Consensus recognition site: 

5′-ACCAUGG-3′

If a start codon deviates from this sequence, it may sometimes be skipped → **leaky scanning** by the ribosome!
In ~90% of cases, translation begins at the first AUG start codon.

In eukaryotes, the ribosome binds to the 5’ cap of mRNA and starts scanning.

Molecular Biology of the Cell
6th edition
In a random DNA sequence, the median ORF is 23 codons long. 3/64 triplets are stop codons, which equals a 5% probability. Very few ORFs ≥ 100 amino acids in size are expected by chance! Among the millions of small ORFs in our genome, only a tiny fraction code for proteins. (Couso and Patraquim, 2017)
Criteria for identifying canonical protein-coding ORFs

1. Length > 100 amino acids
2. Canonical start codon & Kozak motif
3. **Homologies** to known proteins
4. ± Mass spectrometry confirmation

MS is poorly suited for...
- Newly identifying proteins
- Detecting low-abundance proteins
Bioinformaticians rely heavily on evolutionary relationships to known protein-coding genes.

Sequence preservation among organisms

PhyloCSF score, Lin et al, 2011

Homologs

Orthologs

Paralogs

Similarity to known protein domains
When comparing human and mouse sequences, a large fraction of **synonymous substitutions** indicates a protein-coding gene!

Non-coding sequences / RNA genes will accumulate mutations that do **not** conserve the amino acid sequence!
The existing pipelines have a bias against small, new and non-canonical ORFs!

“Functional small ORFs are often not annotated because they have not been experimentally corroborated, and they have not been corroborated because they are not annotated...”

Couso and Patraquim, 2017
Because of their small size, microproteins usually have regulatory functions.

Example:

**Phospholamban** (52 aa) and **myoregulin** (46 aa) inhibit SERCA, which pumps Ca^{2+} back to the sarcoplasmic reticulum to terminate muscle contraction.

Phospholamban and myoregulin are paralogs.
Localization of newly discovered non-canonical ORFs

1. On **IncRNAs** (long non-coding RNAs)

2. On transcribed **pseudogenes**

3. On **mRNAs** (near canonical ORFs)
   - Upstream ORFs (uORFs)
   - Downstream ORFs (dORFs) (rare)

True IncRNAs often have regulatory functions (transcription, heterochromatin...)

Pseudogenes usually arise from the duplication of a gene, followed by the accumulation of damaging mutations in one copy.
**uORFs** sometimes **compete with** and inhibit translation of the canonical ORF

- Classic example:
  - The **yeast Gcn4 gene** has 4 uORFs that normally inhibit its translation
  - In conditions of starvation, the ribosome skips the uORFs
    → Gcn4 is translated instead
  - The **amino acid sequences of the translated Gcn4 uORFs are irrelevant**
    → do not code for functional proteins
LETTER

The translation of non-canonical open reading frames controls mucosal immunity

Ruaidhrí Jackson¹, Lina Kroehling¹, Alexandra Khitun², Will Bailis¹, Abigail Jarret¹, Autumn G. York¹, Omair M. Khan¹, J. Richard Brewer¹, Mathias H. Skadow¹, Coco Duizer¹, Christian C. D. Harman¹, Lelina Chang¹, Piotr Bielecki¹, Angel G. Solis¹, Holly R. Steach¹, Sarah Slavoff²,³,⁴ & Richard A. Flavell¹,⁵,*

Nature, 2018
The authors use mouse models of colitis (e.g. colon infection with *Salmonella typhimurium*) to study the mucosal immune system.

While **RNA-seq** offers a global view of **transcription**, the authors wanted to acquire a global view of **translation** in their colitis model.

They used two complementary strategies to identify RNAs that are being translated:

1. **RiboTag RNA-seq**
2. **Ribosome profiling**
**RiboTag RNA-seq** employs Cre mouse lines to enrich for mRNAs from a specific cell type

If Cre is expressed, the **RPL22 ribosomal protein** is altered: The original exon 4 is excised, and a **HA-(hemagglutinin)-tagged** version of exon 4 is transcribed instead.

---

A transgenic mouse expresses Cre recombinase only in a cell type of interest, e.g. LysM-Cre mice in bone-marrow derived macrophages.
RiboTag RNA-seq uses **anti-HA antibodies** to select for ribosome-bound mRNAs from a specific cell type. 

Sanz et al, PNAS, 2009

---

**Diagram:**

- **Tissue** is homogenized.
- **Anti-HA antibodies** coupled to magnetic beads bind to HA-tagged ribosome-bound mRNAs.
- **Elution** releases the bound mRNAs.

**Techniques:**

- **RNA-seq**
- **qPCR**
Ribosome profiling = ribosome footprinting = Ribo-Seq allows for the specific identification of only translated sections of mRNA

Ribosomes are immobilized using cycloheximide (a toxin)

After RNAse treatment, only ribosome-protected fragments remain (24 – 31 nucleotides)
Note: RiboTag RNA-seq (or similar systems) can be combined with ribosome profiling into one workflow (but this was not done by the authors).
With RiboTag RNA-seq, they found many differentially expressed ribosome-associated transcripts that mapped to non-coding genes!

Bone marrow derived-macrophages were generated from RiboTag\textsuperscript{LysM} mice and stimulated with 1 ng/ml bacterial lipopolysaccharide (LPS) for 6 or 24 hours \textit{in vitro}.

Fig. 1a

\begin{itemize}
  \item Upregulation with LPS
  \item Downregulation with LPS
\end{itemize}
With RiboTag RNA-seq, they found many differentially expressed ribosome-associated transcripts that mapped to non-coding genes!

Fig. 1b

10,971 ribosome-associated transcripts

90.1%

9.9%

Fig. 1c

1,096 ribosome-associated transcripts

Pseudogene: 30.3%

IncRNA: 51.3%

Other RNA: 18.4%
Comparison with paired RNA-seq data indicates that **one third** of expressed **IncRNAs** associate with ribosomes.
Two top upregulated ribosome-associated “IncRNAs” were examined in more detail:

*Aw112010*

*Gm13822*
Bone-marrow derived macrophages were stimulated with LPS or infected with \textit{Salmonella typhimurium}.

Ribosome-bound mRNA was measured by qPCR.
The two ribosome-bound RNAs were \textit{induced in colonic macrophages in vivo} 24h after infection with \textit{S. typhimurium}.

Colon samples were washed, homogenized, and incubated overnight with HA beads.
Next, the authors used **ribosome profiling** to corroborate that IncRNAs are truly being translated. They made heavy use of bioinformatics scores and algorithms to decide, for each ORF, whether translation was taking place.

(Ribosome footprints can be artefacts, e.g. represent noise or protection by non-ribosome RNA-binding proteins.)
A high Percentage of Maximum Entropy (PME) value (a homogenous footprints profile) indicates translation

Protein-coding gene

Homogeneous spread of reads indicates translation

Non-coding small nucleolar RNA

Single Ribo-Seq peak, very inhomogenous

Many lncRNAs had high PME values.
A high **ribosome release score** (RRS) – the ratio of footprints in the **coding region vs. 3’ UTR** – also indicates translation.

**Guttman et al, Cell, 2013**
Three-nucleotide periodicity is also a strong indicator of translation!

- Ribosome profiling offers single-nucleotide resolution.
- The ribosome moves in 3-nucleotide jumps.
- Footprints of a given size often have the same offset to the P site (11 nt for a 24 nt footprint)
  - → can be aligned
  - → typical pattern in coding sequences
The authors used two tools (RibORF and RiboScan) plus a ribosome release score ≥ 7 to identify 96 translated IncRNAs.

Aw112010 was among them.

55% used non-canonical start codons

73% were smaller than 100 amino acids
Ribosome profiling revealed \textit{Aw112010} as the top differentially translated gene upregulated after LPS stimulation of wild-type bone-marrow derived macrophages.
Does Aw112010 really produce a protein?

No antibodies for Aw112010 exist, so an epitope-tagged Aw112010^{HA} knock-in mouse was generated using CRISPR-Cas9.

Mass spectrometry also confirmed expression of the protein.

Aw112010-HA protein is induced by LPS stimulation
To abolish translation of Aw112010 and prove its functional relevance, the authors created Aw112010Stop mice.

Frameshifting stop insert
(14 nt)
Aw112010\textsuperscript{Stop} mice developed more severe infectious colitis
Aw112010\textsuperscript{Stop} had a higher bacterial load of \textit{S. typhimurium}
What is Aw112010’s mechanism of action?

Phagocytosis, phagosome acidification, intracellular killing, and pyroptosis were unaltered in Aw112010\textsuperscript{Stop} macrophages.

However, production of **IL-12 and IL-6** was impaired! (IL-10 was unaltered)

The cytokine IL-12 is crucial for defense against salmonella.
Objection: This still does not prove that the protein product of Aw112010 accounts for the phenotype of Aw112010\textsuperscript{Stop} mice.

The lncRNA itself might perform the function.

Indeed, the authors found that the altered Aw112010\textsuperscript{Stop} transcript is subject to nonsense-mediated decay (NMD), which leads to rapid destruction of the RNA.
Nonsense-mediated decay is a mechanism of the cell to protect against **nonsense mutations** (premature stop codons)

- After splicing, exon junctional complexes (EJC) are placed on exon-exon junctions.
- In the first ever round of translation, these are removed.
- In subsequent rounds of translation, encountering an EJC triggers nonsense-mediated decay.
The authors created a very different version of Aw112010, where the nucleic acid sequence is heavily mutated, but the amino acid sequence remains the same.

The predicted RNA secondary structures are very different.
Re-introduction of both the wild-type and heavily mutated \textit{Aw112010} into \textit{Aw112010}^{\text{Stop}} mice rescued IL-12 production

Plasmids expressing \textit{Aw112010} or the empty vector (EV) were delivered to bone marrow-derived macrophages by electroporation.
Conclusions

The translation of non-canonical open reading frames controls mucosal immunity

- Profiling translation in specific situations (e.g. infection, LPS stimulation) can lead to the discovery of previously unknown, functional proteins.

- Various bioinformatics tools exist for re-analyzing Ribosome-Seq data to identify translated ORFs de novo.

- Some open questions remain:
  - What are the functions of the other translated lncORFs?
  - What role does Aw112010 play in humans?
Pervasive functional translation of noncanonical human open reading frames

Jin Chen¹,², Andreas-David Brunner³, J. Zachery Cogan¹,², James K. Nuñez¹,², Alexander P. Fields¹,²⁎, Britt Adamson¹,²†, Daniel N. Itzhak⁴, Jason Y. Li⁴, Matthias Mann³,⁵, Manuel D. Leonetti⁴, Jonathan S. Weissman¹,²‡

Science, March 2020
The authors aimed to obtain a global view of functional non-canonical ORFs.

1. First, they generated a large Ribo-Seq dataset, and used ORF-RATER to identify potential ORFs.

2. A specialized CRISPR-ko library was then constructed to target 2353 non-canonical ORFs.
   → A pooled screen identified >500 ORFs whose knockout caused a fitness defect.

3. Selected hits were validated.
Analysis of Ribo-Seq data using ORF-Finder identifies 38% new ORFs

Data from human induced pluripotent stem cells (iPSCs), iPSC-derived cardiomyocytes, and human foreskin fibroblasts (HFFs) were pooled.
HLA-I peptidomics (mass spectrometry of peptides eluted from HLA-I) confirms 240 non-canonical peptides.
Design of a CRISPRko library targeting non-canonical ORFs

GuideScan was used to choose the sgRNAs

12044 ORFs
Select high confidence ORFs identified from ribosome profiling (ORF-RATER score > 0.8, length ≥ 10 aa)

2596 ORFs
Design sgRNAs targeting each ORF

Select top 10 sgRNAs based on on-target scores

Filter sgRNAs with off-targets

2353 ORFs

large-scale DNA oligo array synthesis

Plasmid sgRNA library cloning

Lentiviral sgRNA library construction

Lentiviral infection
doublings

Cas9 expressing cells
sgRNA-expressing cells

Endpoint

Next-generation sequencing
The endpoints included cell fitness/growth, as well as transcriptional changes (Perturb-Seq).

**Growth**

$\log_2 \text{sgRNA enrichment} \over \text{cell doublings} = \text{growth phenotype (γ)}$

**Single-cell RNA-seq (Perturb-Seq)**

- $\downarrow$ sgRNA identity
- $\uparrow$ Transcriptome
> 500 ORFs were found to influence fitness in iPSCs!
Guides targeting the ORFs showed much higher fitness effects than control guides.

Predicted efficiency scores were the same.
Selected upstream ORFs were confirmed by ectopically expressing a transcript that encodes only the uORF peptide.

The uCDS (with synonymous mutations to prevent targeting by the sgRNA) was co-delivered with the sgRNA in a lentiviral vector to HEK293 cells.
Example: Overexpression of the MIEF1 uORF increased mitochondrial *fission*, whereas its knockout increased *fusion*.

MIEF1 = Mitochondrial Elongation Factor 1
Conclusions

- A CRISPR screen demonstrates that hundreds of non-canonical ORFs have significant fitness effects.
  - ORFs that are dispensible for cell growth or survival, but have specific other functions, are not detected.

- Many upstream ORFs encode for functional proteins.
  - These are sometimes related to the function of canonical ORF.
Thank you for your attention!