Abstract #49 | Poster

In search of lost ORFs: high confidence map of noncanonical Open Reading Frames in lymphoid cells

J. A. Krupka[1, 2], E. Whittle[1, 2], K.H. J. Tan[1, 2], G. Shevchenko[1, 2], R. Jackson[1, 2], C. Gong[1, 2], J. Gao[1, 2], D. J. Hodson[1, 2]

[1] *University of Cambridge, Department of Haematology, Cambridge, United Kingdom*
[2] *University of Cambridge, Cambridge Stem Cell Institute, Cambridge, United Kingdom*

The human genome is annotated with ~20,000 protein-coding genes, historically defined using a minimum length threshold of 100 amino acids. This cut-off reflects computational convention rather than biological boundary. Ribosome profiling(Ribo-seq) has revealed pervasive translation beyond current annotations, identifying thousands of small open reading frames (smORFs) that encode microproteins across tissues and disease contexts. These findings indicate that genomic coding capacity remains incomplete.

Despite repeated reports of widespread smORF translation, there is limited agreement on which smORFs are translated and under what conditions. Up to 75% of reported smORFs were unique to the original study, raising a key question: do these discrepancies reflect technical artefacts or genuine context-dependent biology? To address this, we analysed nearly 4,000 public and patient-derived Ribo-seq datasets. After controlling for technical heterogeneity, we clustered datasets by translation signatures as a proxy for tissue and cell-state specificity, generating a high-confidence catalogue of ~30,000 translated smORFs across diverse contexts.

To infer function, we applied Parsimonious Gene Co-expression Network Analysis(PGCNA) to large expression matrices. In a proof-of-concept analysis of ~500 lymphoid datasets, smORFs co-clustered with core lymphoid programmes, including endoplasmic reticulum homeostasis, B-cell receptor signalling, and activated B-cell(ABC) and germinal centre B-cell GCB) states characteristic of Large B-cell Lymphoma. Integration with immunopeptidomics showed that a subset of smORF-derived peptides preferentially enters MHC I presentation pathways, supporting roles in immune surveillance. Evolutionary analysis distinguished conserved from lineage-specific elements. smORF-encoded microproteins displayed distinct amino acid composition, with upstream ORFs enriched for intrinsic disorder, associated with transient interactions and rapid functional innovation.

To move beyond correlation, we directly interrogate smORF function using genome-scale base-editing screens targeting endogenous loci, enabling precise perturbation without double-strand breaks and systematic identification of smORFs influencing cellular fitness.

Together, these results support a model in which smORF translation reflects structured, context-dependent regulatory programmes rather than stochastic noise - a framework to expand genome annotation and uncover therapeutic targets.