Microproteins 2023

Invited speakers titles and abstracts.

Alan Saghatelian
Title: *Discovery of Bioactive Microproteins.*

Abstract: Microproteins are exciting new members of the proteome with at least several thousand members. Now that we've discovered microproteins the next step is to identify which of these are functionally active in the context of cells and tissues. Prior work from the lab has been to look for conserved members of the microproteome to choose microproteins to study. The interactome of these microproteins was then used to identify potential functions for these genes. This approach, while productive at finding bioactive microproteins, was unable to identify microproteins in the context of specific biology. To ameliorate these issues, we will report the use of CRISPR microprotein screens that can identify microproteins with functions in specific pathways, and can uncover conserved as well as lineage specific functional microproteins.

Baldomero M. Olivera
Title: *Venom Peptides from Fish-hunting Cone Snails: General Insights into the Biology of Natural Products*.

Each of the >10,000 species of venomous marine snails in the Superfamily Conoidean has a complex venom with its own distinct complement of ~200 different bioactive components, mostly small, disulfide-rich peptides. A subset of these plays a major role in the prey-capture strategy of that species; I will describe how some of these venom peptides are used by fish-hunting cone snails in diverse prey-capture strategies. The elucidation of the mechanistic basis of the bioactivity of an individual venom peptides demonstrates how Natural Products can be used as pharmacological tools to integrate different biological levels.

Dek Woolfson
Title: *Designing new peptide assemblies for fun and for in-cell applications.*

Peptide design has come of age: it is now possible to generate many stable peptide assemblies from scratch using rational and/or computational approaches. A new challenge for the field is to move past structures offered up by nature and to target the so-called ‘dark matter of protein space’; that is, structures that should be possible in terms of chemistry and physics, but which biology seems to have overlooked or not used prolifically.
I will start this talk by illustrating what is currently possible in this nascent field using de novo designed coiled-coil peptides. These are bundles of 2 or more α helices that wrap around each other in rope-like structures. They are one of the most dominant structures for directing natural protein-protein interactions. Our understanding of coiled coils provides a strong basis for building new peptide assemblies. Indeed, we are nearing the completion of a “toolkit” of de novo coiled coils (Woolfson, J Biol Chem, 2023, DOI: 10.1016/j.jbc.2023.104579). Time allowing, I will show how we are using these peptide assemblies to “seed” de novo computational designs of large single-chain proteins.

Next, I will describe the rational design of a completely new 3-10-helical bundle (Kumar et al., Nature, 2022, DOI:10.1038/s41586-022-04868-x). Finally, I will turn to in-cell applications. I will describe new designs for (i) de novo cell-penetrating peptides and (ii) high-affinity kinesin-binding peptides; and how these can be combined to render peptides that can be delivered exogenously to eukaryotic cells and target subcellular processes, in this case, hijacking active protein motors (Rhys et al., Nature Chem Biol, 2022, DOI: 10.1038/s41589-022-01076-6).

Jean-Philippe Combier
Title: *Complementary peptides represent a credible alternative to agrochemicals by activating translation of targeted proteins.*

The current agriculture main challenge is to maintain food production while facing multiple threats such as increasing world population, temperature increase, lack of agrochemicals due to health issues and uprising of weeds resistant to herbicides. Developing novel, alternative, and safe methods is hence of paramount importance. Here we show that complementary peptides (cPEPs) from any gene can be designed to target specifically plant coding genes. External application of synthetic peptides increases the abundance of the targeted protein, leading to related phenotypes. Moreover, we provide evidence that cPEPs can be powerful tools in agronomy to improve plant traits, such as growth, resistance to pathogen or heat stress, without the needs of genetic approaches. Finally, by combining their activity they can also be used to reduce weed growth.

John Prensner.
Title: *The functional landscape of non-canonical open reading frames in cancer.*

Understanding the functions of non-canonical open reading frames remains one the central bottlenecks in biomedical disease research. We have systematically pursued the role of non-canonical open reading frames in cancer cell survival. We find that cancers broadly use these ORFs to maintain diverse and essential cellular needs. By focusing on medulloblastoma, a deadly childhood brain cancer, we have identified cancer subtype-specific translational control and functional relevance of non-canonical ORFs, including disease mechanisms that provide insight into the pathobiology of this disease. Our work shows non-canonical ORFs to be core mediators of cancer biology and advocate for the investigation of non-canonical ORFs as a next step in target discovery across human malignancies.

Jonathan M. Mudge
Title: *The annotation of microproteins in GENCODE.*

The description of protein sequences remains a core component of the Ensembl-GENCODE human (and mouse) reference gene annotation project. Indeed, we place special importance on ensuring the accuracy of coding sequence annotations, as these regions are central to the majority of scientific and clinical workflows that utilise our resources. Nonetheless, after 20+ years of effort, our annotation of the proteome remains a work in progress. In recent years, GENCODE

has actively focused on the discovery of missing - i.e. 'non-canonical' - proteins, an ongoing drive that integrates multiple strands of analysis and includes the involvement of numerous associated projects and institutions. I will present an overview of these contemporary discovery efforts, which incorporate evolutionary metrics developed by GENCODE alongside the collaborative analysis of experimental datasets including Ribo-seq, proteomics and immunopeptidomics. This integrated approach has led to the annotation of numerous proteins, the vast majority of which are microproteins. Nonetheless, GENCODE remains conservative in protein annotation, a position largely explained by the burden of responsibility to provide high-confidence annotations. Thus, while we have led community efforts to produce a reference catalog of 7,264 non-canonical human Ribo-seq ORFs, very few have as yet become reclassified as protein-coding. This position does not reflect scepticism on our part, rather an appreciation of the knowledge gaps that exist in the field and how such outstanding biological questions undermine efforts to create future-proof annotations. For example, how can we gain further confidence in the in vivo existence of a predicted microprotein? How can the observation that few Ribo-seq ORFs display protein-level constraint be reconciled with a function-based model of protein annotation? How should we classify peptide products that may be solely expressed under aberrant conditions? How can we distinguish translation events that have alternative modes of function to protein production, e.g. uORFs? In this way, it becomes apparent that our drive is not simply a question of identifying missing proteins, rather an attempt to rationalize a modern, broader view of translation.

Julie L. Aspden

Title: *Translation and functional characterisation of novel ORFs from 'long non-coding' RNAs during neuronal differentiation.*

Long non-coding RNA (lncRNAs) are enriched in the nervous system, with ~40% of human lncRNAs specifically expressed in the brain. LncRNAs are less conserved, yet more tissue and developmental-stage specific than mRNAs. Biological functions have only been determined for a relatively small number of lncRNAs, but several have been found to play key roles in development and differentiation. ~54% of lncRNAs are present in the cytoplasm, where many associate with ribosomal complexes. Ribosome profiling, in a range of organisms, has revealed that many lncRNAs can actually be translated.

We developed Poly-Ribo-Seq to distinguish genuine translation events, lncRNAs that are bound by multiple ribosomes, and therefore actively translated, from non-specific background signal (Aspden et al eLife 2014). Poly-Ribo-Seq on a human neuroblastoma cell line, SH-SY5Y, in immature and differentiated states discovered the translation of 45 translated small ORFs (smORFs) in lncRNAs (Douka et al RNA 2021). These lncRNA-smORFs exhibited comparative translational efficiencies to canonical protein-coding ORFs. 39% of the translated lncRNAs are dynamically expressed during human brain development, and 67% are associated with cancers of the CNS.

To dissect the function of micropeptides translated from these neuronal lncRNAs, 13 candidates were selected based on their evolutionary conservation and association with development or disease. FLAG-tagged reporter assays demonstrate that 11/13 smORFs produce stable micropeptides with 6/11 displaying distinct subcellular localisations indicative of function. siRNA knockdown of 4/6 translated lncRNAs resulted in dysregulation of neuronal

differentiation. CRISPR mutants indicate that the phenotype of one translated lncRNA, LIPT2-AS1 is exclusively due to the loss of its translation product.
3/4 of these translated lncRNAs, which contribute to the regulation of neuronal differentiation in SH-SY5Y cells, show dynamic expression in cortical organoids during early development. Preliminary analysis of organoid polysomal fractions indicates that they are also translated in cortical organoids.
Overall, these data highlight the importance of translated lncRNAs in neuronal differentiation and the potential for lncRNAs as a source of novel neuro-peptides.

Petra Van Damme
Title: *Riboproteogenomics to uncover small protein landscapes.*

By monitoring translation (initiation) at a genome-wide level, revolutionary riboproteogenomics has been a game changer for genome (re)annotation which triggered a new flow of information on putative, protein-coding regions [1–3]. Among them, an important part is represented by short open reading frames (sORFs), in general defined as genes encoding proteins (equal to or) shorter than 100 amino acids in length that go under the name of sORF-encoded polypeptides (SEPs). SEPs are largely understudied viewing their under-annotation in genomes both directly and indirectly resulting from their small size and other peculiar biochemical features [3]. In particular, this specific protein class has been linked to low protein abundance and stability, and especially in case of bacterial SEPs, an increased hydrophobicity because of the high incidence of transmembrane properties, all representing aspects that hamper standard protein detection methods [3], which in turn explains the lack of expression validation and functional information on SEPs.
To contribute to the closure of this knowledge gap, we developed a dedicated work-flow integrating annotation, experimental expression validation, subcellular localization determination and interactome characterization of putative, novel bacterial sORFs and their according SEPs [3-4]. Making use of the model bacterial pathogen Salmonella Typhimurium, we set off to functionally characterize the protein products of these fascinating novel genomic elements in the context of (basic) bacterial physiology as well as infection biology. By doing so, we will participate in the broadening of general knowledge on bacterial SEP biology and expand the technical toolkit facilitating small protein biology discoveries.

Polly Hsu
Title: *Prevalent Unannotated ORFs Revealed by Improved Super-Resolution Ribosome Profiling.*

A crucial step in functional genomics is identifying Open Reading Frames (ORFs) that are associated with various biological functions. To systematically and robustly identify these ORFs, we improved Ribo-seq to achieve both a strong three-nucleotide periodicity and high read coverage in Arabidopsis. From the improved dataset, we discovered ~7700 unconventional translation events, including upstream ORFs (uORFs) and downstream ORFs (dORFs) on annotated protein-coding genes, as well as small ORFs (sORFs) on annotated non-coding genes such as microRNAs, trans-acting small interfering RNAs (tasi-RNAs), and pseudogenes. Some

of the novel translation events were validated by proteomics data, suggesting that their protein products are stable in plants. In contrast, we observed that approximately one hundred annotated coding genes appear non-coding in our experimental conditions. We specifically focused on uORFs as they are common negative translational regulators and emerging targets of plant engineering. We found that numerous key components of circadian clock and phytohormone pathways possess actively translated uORFs, highlighting the potential roles of uORFs in plant physiology and development. Among different categories of uORFs, strong ribosome stalling was observed on Conserved Peptide uORFs (CPuORFs) and minimum uORFs (i.e., AUG-stop), implying significant repression of these uORFs. Finally, we found that genes containing translated uORFs are generally larger and have more conserved domains – features commonly observed in evolutionarily older genes. In summary, our study demonstrates that improved super-resolution ribosome profiling can efficiently identify widespread non-canonical ORFs, verify the translation of annotated coding genes, and shed light on translational regulation in plants.

Renaud Vincentelli
Title: *Combining High-Throughput protein purification and quantitative interactomics to generate the genome wide scale affinity mapping of the human PDZome*.

The HTP protein production and interaction facility of the AFMB offers custom and automated protocols (in 96/384 format) covering all the stages of cloning, protein production (in E. coli), purification (µg to tens of milligrams) as well as in vitro study of protein-protein, protein-peptide or protein-DNA interactions. The facility purified more than 10.000 proteins for research teams, French (>10 ANR) or international networks (>10 EU grants).
We developed and validated protein purification protocols at a pace of >1.000 cultures and purifications per week (1, 2). This is used to either improve the soluble level of difficult proteins (3) or to purify protein librairies such as transcription factors (4), the 5.000 disulfide rich animal toxins of the EU VENOMICS project (5), hundreds of CAZYmes (6) or the full repertoire of the 266 human PDZ domains (7).
To characterize proteins at a pace and scale that is compatible with hundreds of proteins in micrograms, several custom-made protocols have been developed such as a new in vitro HTP protein- DNA interaction assay (HTP SELEX (4)) and in vitro HTP quantitative protein-protein and protein- peptide interaction assays (HTP Hold-up (7)) able to determine thousands of affinities per day. Using holdup, we published the biggest quantitative dataset for PDZ-pbm interactions (8) and were the only ones to identify the human PDZ binders of SARS-CoV-2 (9). The most recent developments of the holdup will open the way to the systematic deciphering of the Full human- PDZ-pbm quantitative interactome and could be adapted to many more interactomes.

Sebastiaan van Heesch
Title: *VERY small human microproteins – Is there a reasonable lower size cut-off?*

Microproteins, encoded by short open reading frames (ORFs), are small by definition. Still, arbitrary size cutoffs are used for coding region annotation, whereas (very) short length does not necessarily prohibit function: bioactive peptides cleaved from longer precursor proteins can

function as peptide hormones, neuropeptides or immunomodulators; and can be synthetically produced as pharmacological drugs. To date, we have sparse or no endogenous evidence of human peptides that are born small, i.e., translated independently from short non-canonical ORFs instead of being processed from longer precursors.

Therefore, we set out to detect and characterize microproteins of 15 aa and smaller. Using stringent cutoffs, we reproducibly detected 221 highly translated short ORFs in human tissues. Most of these were upstream ORFs that—in contrast to most newly found ORFs—were conserved across mammals and translated at on average 3-fold greater levels than the main coding sequence located within the same mRNA template. For several candidates we provide initial protein-level evidence using targeted mass spectrometry and reprocessed HLA immunopeptidomics data.

To explore the putative functions of these very small proteins, we synthesized all 221 candidates on triplicate cellulose membranes and sought for protein interaction partners using PRISMA: a protein interaction screen on peptide matrix. PRISMA defined hundreds of unique interactomes and clustered them into functional groups, providing hints about possible roles in human cells. For instance, we could validate predicted involvement of several very small microproteins in translational regulation and endocytosis.

Our findings demonstrate that very small ORFs in the < 15 aa size range should not be overlooked, but instead be incorporated in reference annotations and future study designs so that their functional roles can be explored systematically across mammals.