Finding evidence of *de novo* proteins

Background: New protein coding genes frequently evolve from existing protein coding genes, via various mechanisms. Recent studies have shown that protein coding genes can also emerge "*de novo*" from genomic regions that did not previously encode any gene. Existence of such *de novo* genes has been established based on gene annotation, presence of a transcript (RNA-seq) and ribosome association (Ribo-seq). However, except for a few specific cases, a direct evidence of the protein is lacking for the majority of *de novo* genes. Mass spectrometry is a highly sensitive and high throughput technique to quantify the abundance of proteins. A large amount of mass spectrometry data is publicly available, which can be mined to find evidence for *de novo* proteins.

Objective: The aim of this project is to analyse mass spectrometry data to find and quantify *de novo* proteins. To this end, the student will first download publicly available data from PRIDE database. Then they will build an updated protein sequence database (fasta) using known protein sequences as well as newly discovered *de novo* protein sequences. Finally, they will map mass spectrometry data to this database using the program MS-GF+.

Methods: Mass spectrometry data analysis

Requirements:

- Basic knowledge or interest to work with linux command line, and simple programming
- Interest in protein computational biology
- Ability to work independently

Supervision: Dr. Bharat Ravi, b.ravi@uni-muenster.de, Molecular Evolution and Bioinformatics Group (https://bornberglab.org/)

Relevant literature:

- Heames B, Schmitz J, Bornberg-Bauer E. *A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila.* J Mol Evol. 2020 May;88(4):382-398. doi: 10.1007/s00239-020-09939-z.
- Van Oss SB, Carvunis AR. *De novo gene birth*. PLoS Genet. 2019 May 23;15(5):e1008160. doi: 10.1371/journal.pgen.1008160.
- Kim S, Pevzner PA. *MS-GF+ makes progress towards a universal database search tool for proteomics.* Nat Commun. 2014 Oct 31;5:5277. doi: 10.1038/ncomms6277.