***ESPOD Project proposal:* "From proteoforms to pathogenesis: The systematic identification and functional analysis of proteoforms in a malaria parasite"**

Supervisors: Dr. Juan Antonio Vizcaíno (EMBL-EBI) and Dr. Oliver Billker (Sanger Institute)

Introduction

The term proteoform[1] represents the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts and post-translational modifications (PTMs). In proteogenomics studies, mass spectrometry (MS) proteomics data is combined with genomics and/or transcriptomics information, typically by using sequence databases generated from DNA sequencing efforts, RNA-Seq experiments, or Ribo-Seq approaches, among others. The promise of these approaches is that peptides can be used to detect events such as gene variants, novel splice junctions, small Open Reading Frames (ORFs), or pseudogenes, among others. In addition, expression of long non-coding RNAs (lncRNA) can also be investigated.

In malaria parasites, alternative splicing events and lncRNAs have been identified and some have been implicated in virulence gene regulation[2-4], but no comprehensive proteogenomics study has so far been performed in any *Plasmodium* species to predict proteoforms and lncRNAs systematically across parasite species and life cycle stages. At the Sanger Institute, the *Plasmodium* genetic modification project, *Plasmo*GEM, has created a vector resource and genome scale functional screening platform for *Plasmodium berghei*, a malaria parasite infecting rodents[5], but currently available reagents do not distinguish between proteoforms, and lncRNAs are not included. We therefore here propose to combine proteogenomics and phosphoproteomics approaches to study proteoforms in the malaria parasite and prioritise them for validation. We will then use the power of the *Plasmo*GEM screening platform to build proteoform-specific reagents and systematically screen proteoforms for roles in parasite growth and pathogenesis *in vivo*.

Objectives

1. To identify proteoforms and functional lncRNAs across *Plasmodium* species and life cycle stages using a wide range of suitable omics datasets deposited in public databases.
2. Validate proteoforms and lncRNAs for *P. berghei* using new proteomics, phosphoproteomics or transcriptomics data as required, and prioritise them for functional validation.
3. Use the existing *Plasmo*GEM DNA engineering pipeline to generate proteoform-specific reagents for functional validation both at scale and though in-depth analysis of selected genes.

Available datasets, analysis pipelines and molecular tools

Public MS proteomics datasets are increasingly used for performing proteogenomics studies, as it has been demonstrated for human, mouse and rat, among other species[6]. At the moment of writing, in the PRIDE database at EMBL-EBI there are 15 relevant datasets publicly available (http://bit.ly/2nilua8). Additionally, around 10 datasets are private. Existing public datasets in other resources will be investigated as well.

In addition, several deep proteomics datasets are being generated at present in house at the Sanger Institute in efforts to create global protein interaction maps and understand signal transduction pathways. Samples include extracts from schizonts and gametocytes of *P. berghei*, schizonts of *P. falciparum* and *P. knowlesi,* and a range of phosphoproteomics datasets from different *Plasmodium* species and life stages.

Work plan

A) Proteogenomics study. We will follow a comprehensive proteogenomics approach to study: (i) non-canonical proteins encoded by *Plasmodium*, involving protein splice isoforms, protein variants, small Open Reading Frames (sORFs); and (ii) the expression of lncRNAs. We will perform a multi-stage analysis approach, using different sources of information: e.g. updated *Plasmodium* protein sequence databases, ESTs (Expression Sequence Tag), RNA-Seq studies, and sequence information available in lncRNA resources, among others. Existing proteogenomics data analysis pipelines available in house will be further developed using existing state-of-the-art and free-to-use (ideally open source) software (e.g. the OpenMS framework).

B) Phosphoproteomics study. A quantitative proteomics analysis will be performed using the total and phospho-proteome in the datasets generated in house. The availability of both portions of the proteome will enable the elucidation of those proteins (and potentially specific proteoforms) where the abundance and phosphorylation levels change unambiguously as a result of a particular biological condition, enabling the identification of key players in the mechanism of action of the parasite. At present, we are starting to develop state-of-the-art and free-to-use phoshoproteomics analysis pipelines based on the OpenMS framework, that will also be further developed and benchmarked.

C) Detection of additional proteoforms. The unassigned mass spectra after the above-described analyses (in A and B), will be clustered using the new version of the in-house PRIDE Cluster spectrum cluster algorithm[7]. By doing this, we will target those peptides that are abundantly found across *Plasmodium* samples but that have not been identified and quantified before. The resulting set of spectral clusters will be subjected to a less conventional analysis pipeline involving *de-novo* sequencing and spectral searches (using software such as SpectraST and PepNovo). This will enable an even higher coverage of the *Plasmodium* proteome, detecting novel proteoforms, including additional protein variants and other potentially biologically-relevant PTMs.

D) Expanding the *Plasmo*GEM vector resource for the systematic functional analysis of proteoforms. Data from the above bioinformatics analysis will be fed into available vector design software, and proteoform/transcript specific reagents for *P. berghei* genome engineering will be generated using available molecular biology pipelines for the production of genetic modification vectors. Where necessary, deletion vectors will be combined with proteoform-specific complementation sequences on *Plasmodium* artificial chromosomes, an approach recently developed by the Billker lab (unpublished). Reagents will be barcoded for inclusion in genetic screens and single cell phenotyping approaches currently under development (manuscript submitted, https://doi.org/10.1101/105015) will be used for functional analysis, as appropriate. A range of scalable phenotypic assays are currently available for asexual parasite growth in blood, tissue tropism, sexual differentiation, mosquito transmission and infection of the liver. These will be used to search for functional implications of observed proteoforms, some of which will be selected for further in depth analysis and publication.

References

1. Smith, L.M., Kelleher, N.L. & Consortium for Top Down Proteomics. *Nat Methods* **10**, 186-187 (2013).
2. Broadbent, K.M. et al. *BMC Genomics* **16**, 454 (2015).
3. Zhu, L. et al. *Sci Rep* **6**, 20498 (2016).
4. Gabriel, H.B. et al. *Sci Rep* **5**, 18429 (2015).
5. Gomes, A.R. et al. *Cell Host Microbe* **17**, 404-413 (2015).
6. Martens, L. & Vizcaíno, J.A. *Trends Biochem Sci, in press* (2017).
7. Griss, J. et al. *Nat Methods* **13**, 651-656 (2016).