Deep Mutational Scanning of the Dark Proteome

David Adams (WTSI) & Alex Bateman (EMBL-EBI)

Millions of variants are known for the human genome sequence and of these many lead to rare genetic diseases or predisposition to common diseases. One of the major challenges to bring precision medicine to fruition is to understand the connection between these DNA variants and their role in human health and disease. Variants that change the amino acid of a protein sequence have been most studied and yet we are still far from having the ability to predict the likely effect of a mutation on a protein's function.

Over the last few years the technology for modifying gene loci via approaches such as deep mutational scanning (1) has been developed now making it feasible to functionally assess virtually every amino acid position in protein coding genes. This technology exploits the fact that CRISPR can precisely engineer into genes all possible variants at a locus and by using DNA sequencing it is possible to analyse pools of cells carrying libraries of variants and thus "readout" their effect on gene function. This technology has been applied to loci such as *BRCA1* to classify variants of unknown significance as either pathogenic or benign, and has significant implications for defining the role of germline variation in disease predisposition. The Adams lab is actively developing and applying this technology to human proteins.

Over the last three decades a large collection of protein domains has been identified, which enable the understanding of common protein functions across proteins. Many known disease mutations fall within these sites and interpretation of their effect on function is somewhat easier to predict. However, once we start to look outside of these regions interpretation of mutations becomes more challenging. Given that more than half of the amino acids in human proteins lie outside of domains there is an important opportunity to apply deep mutational scanning for variant interpretation. The regions outside of domains often fall into the class of disordered regions, such that they do not have a stable structure. In the past decade we have come to learn that these disordered regions are rich in short linear motifs (SLiMs) that are bound by a variety of factors (2). SLiMs are extremely challenging to identify using computational methods and have hence been understudied. SLiMs often mediate low affinity binding to a variety of other proteins, which means that disordered proteins can act as scaffolds for important cellular processes. For example, MDM2 inhibits P53 by binding to a SLiM in its transactivation domain (3). The P53 protein mediates a surprisingly large number (>100) important cellular protein-protein interactions via the SLiMs in its disordered regions.

The Sanger Institute and the EMBL-EBI are engaged in an international effort called the Atlas of Variant Effect (AVE) Consortium, which aims to perform deep mutational scanning across hundreds of disease genes to understand their function. The consortium also includes investigators from the Broad Institute, the University of Toronto, The Walter and Eliza Hall Institute and the University of Washington, Seattle. The AVE Consortium is establishing data standards and methodologies for this project and is also championing open science and innovation.

In this project we aim to select regions of protein coding genes that are conserved but unstructured and to perform mutagenesis on these regions so as to understand all of the amino acids that are important for their function. We expect that this technique will identify hitherto unidentified SLiMs in these regions that will lead to improved understanding of the protein function as well as lead to further specific experiments to dissect the protein interactions. Our focus will be on essential genes and in particular genes implicated in tumorigenesis. This project would be suitable for a computational biologist who is interested in gaining some wet bench skills or a bench biologist who wishes to gain experience in bioinformatics.

References

1 Accurate classification of *BRCA1* variants with saturation genome editing. Findlay *et al*., Nature (2018), 562(7726):217-222.

2 Short linear motifs - ex nihilo evolution of protein regulation. Davey *et al*. Cell Communication and Signaling (2015), 13:43.

3 Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Kussie *et al*. Science (1996) 274(5289):948-953.