**Computational annotation of understudied proteins coexpression network analysis of The Cancer Genome Atlas (TCGA) data**

**Background**

The human genome contains nearly 30,000 genes, out of which about 6,000 are estimated to express proteins with the ability to bind drug-like small molecules. This set of genes is known as the druggable genome. To date, roughly 5% to 10% of the druggable genome is targeted by FDA-approved drugs, while a significant portion of the remaining genes are understudied, and their functions poorly understood1.

Much of biomedical research and the development of therapeutics is focused on a small fraction of the human genome, ignoring many disease-relevant proteins and the associated scientific and commercial opportunities. The National Institutes of Health (NIH) Illuminating the Druggable Genome programme aims to catalyse research around understudied targets2.

**Overview**

In order to characterize the biological function of the understudied elements of the druggable genome, we present a computational method to identify associations between genes and biological terms by performing coexpression network analysis across the 40+ RNA-seq datasets available in The Cancer Genome Atlas (TCGA) database.

The following is a brief summary of the approach used in the analysis. For each tumor type available in TCGA, the following steps are performed:

1. Step 1 – Identify clusters of coexpressed genes by performing network analysis of RNA-seq data.
2. Step 2 – Biologically characterize these gene clusters by performing enrichment analysis.
3. Step 3 – Associate each understudied protein to the biological term(s) enriched in the cluster it belongs to.

The associations between genes and biological terms are integrated across all datasets. This results in a matrix of counts, which indicates the number of tumor types in which a gene is associated to a cluster enriched for a specific biological term. For example: RPL7, a ribosomal protein, may belong to a cluster of genes enriched for the protein translation term in 22 out of 30 tumor types.

While not providing conclusive evidence regarding a gene’s biological function, this ‘guilt-by-association’ method could provide useful insight in better understanding its role in tumor.

**Results**

**Clustering**

The first part of the results section contains an overview of the clustering results in the Mesothelioma (TCGA-MESO) dataset. The following plots display a summary of the identification of modules of co-expressed genes in the Mesothelioma (TCGA-MESO) dataset.

*Figure 1 | Gene clustering dendrogram and module assignments from TCGA Mesothelioma RNA-seq dataset*.

*Figure 2 | Summary of the top significantly enriched biological terms for selected gene modules. The analysis reveals the presence of clusters of genes involved in biological functions widely recognized to be hallmarks of cancer*3*, including enabling metastasis through extracellular matrix organization (yellow module), sustaining proliferative signaling by inducing cell cycle (red), and inducing angiogenesis (green). The analysis also reveals a module enriched for terms related to immune response (greenyellow), potentially indicating the presence of infiltrating immune cells.*

**Integration**

The second part of the section contains information integrated across all tumor types.

*Figure 3 | Summary of the biological terms most reproducibly associated with the FAM26F gene.* FAM26F is a protein-coding gene, which encodes for a pore-forming subunit of a voltage-gated ion channel. This gene is annotated with a single term in the Gene Ontology database, cation channel activity (GO:0005261). The analysis reveals the gene associated to clusters of genes involved in immune response and leukocyte activation across over 20 tumor types.

A summary of the most relevant results of the analysis can be interactively browsed below:

*Figure 4 | Summary of the most relevant associations between Gene Ontology terms and understudied genes*. The heatmap’s rows represent genes, columns represent Gene Ontology terms, values indicate the number of tumor types where the gene belongs to a cluster enriched for the corresponding term.

**References**

1. Illuminating the Druggable Genome - Overview | NIH Common Fund. Available at: https://commonfund.nih.gov/idg/overview. (Accessed: 6th March 2017)

2. The IDG Knowledge Management Center. Poster : Unexplored opportunities in the druggable human genome. Available at: http://www.nature.com/nrd/posters/druggablegenome/index.html. (Accessed: 6th March 2017)

3. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).