A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing

Extended Data
Extended Data Figure 1. Clustering analysis of AP-MS dataset reveals biological replicates of individual baits are well correlated. All MS runs were compared and clustered using artMS (David Jimenez-Morales, Alexandre Rosa Campos and John Von Dollen. (2019). artMS: Analytical R tools for Mass Spectrometry. R package version 1.3.9. https://github.com/biodavidjm/artMS). This figure depicts all Pearson’s pairwise correlations between MS runs, and is clustered according to similar correlation patterns.
Extended Data Figure 2. Gene Ontology Biological Process Enrichments for SARS-CoV-2 Host Factors.
We performed GO biological process enrichments (Methods) for the host factors identified as binding to each SARS-CoV-2 viral protein and represent here the top 5 most significant terms for each viral protein.
Extended Data Figure 3. Pfam Protein Families Enrichments for SARS-CoV-2 Host Factors. The enrichment of individual PFAM domains was calculated using a hypergeometric test where success is defined as the number of domains, and the number of trials is the number of individual preys pulled-down with each viral bait. The population values were the numbers of individual PFAM domains in the human proteome. To make sure that the p-values that signify enrichment were meaningful, we only considered PFAM domains that have been pulled-down at least three times with any SARS-CoV-2 protein, and which occur in the human proteome at least five times. Here, we show PFAM domains with the lowest p-value for a given viral bait protein.
Extended Data Figure 4. Lung mRNA expression of the interacting human proteins relative to other proteins. The gene expression in the lung of the interacting proteins was observed to be higher than all other proteins (blue interacting proteins; median=25.52 TPM, grey all other proteins; median=3.198 TPM), p=0.0007).
Extended Data Figure 5. Gene level expression values and specificity for the lung. Scatterplot of the lung mRNA expression (TPM) versus enrichment of lung mRNA expression (lung TPM/median all tissue TPM) for human interacting proteins. Red points denote drug targets that are labelled with their gene names. Points above the horizontal blue line represent interacting proteins that are enriched in lung expression and show how most human interacting proteins tend to be enriched in the lung.
Extended Data Figure 6. Evolutionary Conservation of SARS-CoV-2 Interacting Proteins. We calculated the observed/expected ratio from gnomAD\textsuperscript{1} for loss of function, missense, and synonymous mutations across all RefSeq genes. Compared to other genes (RefSeq Genes, grey), interacting proteins (blue) had lower ratios (indicating stronger intolerance of mutations) of observed/expected missense (median refseq score: 0.49, median interacting score: 0.36, \( p=2.46\times10^{-11} \); t-test), and loss of function (median refseq score: 0.89, median interacting score: 0.85 \( p=9.12\times10^{-7} \); t-test) but no difference in observed/expected synonymous mutations (median refseq score: 1.009 median interacting score: 1.004, \( p=0.48 \); t-test). Collectively, these results indicate that the interacting proteins have reduced genetic variation in human populations.
Extended Data Figure 7. Candidate targets for the viral Nsp5 protease. (a) nsp5 WT and nsp5 C145A (catalytic dead mutant) interactome. (b) Domain maps of HDAC2 and TRMT1 illustrating predicted cleavage sites (using Coronanet 1.0). HDAC: Histone Deacetylase Domain, NLS: Nuclear Localization Sequence, MTS: Mitochondrial Targeting Sequence, SAM-MT: S-adenosylmethionine-Dependent Methyltransferase Domain. (c) Peptide docking of predicted cleavage peptides into the crystal structure of SARS-CoV nsp5.
Consensus analysis of SARS-CoV-2 Orf6 homologs

Reference sequence (1): ref|YP_009724394.1
Identities normalised by aligned length.
Colored by: identity

Extended Data Figure 8. Consensus analysis of SARS-CoV-2 Orf6 homologs. Multiple sequence alignment of 34 homologs of SARS-CoV-2 Orf6. Sequence coverage (cov) and percent identity (pid) shown for each homologous sequence. Amino acids are colored according to identity with SARS-CoV-2 Orf6 (sequence 1 ref|YP_009724394.1). Colors indicate chemical properties of amino acids: small and/or hydrophobic (bright green, A, V, L, I, M, P, G), bulky (dark green, F, Y, W, H), basic (red, R, K), acidic (blue, D, E), neutral (purple, Q, N), polar (light blue, S, T), cysteine (C, yellow). The key methionine M58, and the acidic residues E55, D59, D61 of the putative NUP98-RAE1 binding motif are shown to be highly conserved. Multiple sequence alignment was visualized using the MView web server\(^2\): [https://www.ebi.ac.uk/Tools/msa/mview/](https://www.ebi.ac.uk/Tools/msa/mview/).
Experimental and predicted structural features of BRD2 and Protein E.

Extended Data Figure 9. Experimental and predicted structural features of BRD2 and Protein E. Domain names are indicated for BRD2 along with residue numbers. Areas of protein E where its SARS-CoV homolog has been described by high-resolution structural models (PDB ID: 2MM4) are shown in white blocks and thick black lines above the sequence line. The ribbons indicate alpha helical sections. Protein E is predicted to have two N-linked glycosylation sites (red stars on sequence line). Structural predictions, as generated using PredictProtein\(^4\) are found below the sequence line. The first row indicates secondary structure predictions for alpha helices (red) and beta sheets (blue). The second row beneath the sequence line indicates areas expected to be buried (yellow) and exposed (blue). At the bottom, green bars indicate areas predicted to be disordered.
Extended Data References


