SARS-CoV-2 3CLpro Whole Human Proteome Cleavage Prediction and Enrichment/Depletion Analysis

Lucas Prescott
Correspondence: lskywalker2015@gmail.com
Keywords: SARS-CoV-2, COVID-19, Coronavirus, Protease, Proteomics, 3CLpro, Neural Networks

Abstract

A novel coronavirus (SARS-CoV-2) has devastated the globe as a pandemic that has killed more than 800,000 people. Effective and widespread vaccination is still uncertain, so many scientific efforts have been directed towards discovering antiviral treatments. Many drugs are being investigated to inhibit the coronavirus main protease, 3CLpro, from cleaving its viral polyprotein, but few publications have addressed this protease's interactions with the host proteome or their probable contribution to virulence. Too few host protein cleavages have been experimentally verified to fully understand 3CLpro's global effects on relevant cellular pathways and tissues. Here, we set out to determine this protease's targets and corresponding potential drug targets. Using a neural network trained on coronavirus proteomes with a Matthews correlation coefficient of 0.983, we predict that a large proportion of the human proteome is vulnerable to 3CLpro, with 4,460 out of approximately 20,000 human proteins containing at least one predicted cleavage site. These cleavages are nonrandomly distributed and are enriched in the epithelium along the respiratory tract, brain, testis, plasma, and immune tissues and depleted in olfactory and gustatory receptors despite the prevalence of anosmia and ageusia in COVID-19 patients. Affected cellular pathways include cytoskeleton/motor/cell adhesion proteins, nuclear condensation and other epigenetics, host transcription and RNAi, coagulation, pattern recognition receptors, growth factor, lipoproteins, redox, ubiquitination, and apoptosis. This whole proteome cleavage prediction demonstrates the importance of 3CLpro in expected and nontrivial pathways affecting virulence, lead us to propose more than a dozen potential therapeutic targets against coronaviruses, and should therefore be applied to all viral proteases and experimentally verified.

Introduction

Coronaviruses are enveloped, positive-sense, single-stranded RNA viruses with giant genomes (26-32 kb) that cause diseases in many mammals and birds. Since 2002, three coronavirus outbreaks have occurred: severe acute respiratory syndrome (SARS) in 2002-2004, Middle East respiratory syndrome (MERS) from 2012 to present, and coronavirus disease 2019 (COVID-19) from 2019 to present. The virus that causes the latter disease, SARS-CoV-2, was first thought to directly infect the lower respiratory epithelium and cause pneumonia in susceptible individuals. The most common symptoms include fever, fatigue, nonproductive or productive cough, myalgia, anosmia, ageusia, and shortness of breath. More recently, however, correlations between atypical symptoms (chills, arthralgia, diarrhea, conjunctivitis, headache, dizziness, nausea, severe confusion, stroke, and seizure) and severity of subsequent respiratory symptoms and mortality have motivated researchers to investigate additional tissues that may be infected. One way to explain these symptoms and associated cellular pathways is to review enrichment and depletion in virus-host interaction networks, particularly those including the coronavirus proteases.

ACE2, the receptor for SARS-CoV-1 and -2, has been shown to be less expressed in lung than in many other tissues. Respiratory coronaviruses likely first infect the nasal epithelium and tongue[1] and then work their way down to the lung and/or up through the cribriform plate to the olfactory bulb, through the rhinencephalon, and finally to the brainstem.[2][3][4][5] Additionally, based on ACE2 expression and *in vitro* and *in vivo* models, multiple parts of the gastrointestinal tract (mainly small and large intestine, duodenum, rectum, and esophagus; less appendix and stomach) and accessory organs (mainly gallbladder, pancreas, liver[6][7], salivary gland[8]; less tongue and spleen)[9], kidney,[10] male

and female reproductive tissues,[11][12] heart,[13] immune cells,[14][15] and adipose tissue[16][17][18] may be infectible with corresponding symptoms and comorbidities.

Coronaviruses have two main open reading frames, orf1a and orf1b, separated by a ribosomal frameshift and resulting in two large polyproteins, pp1a and pp1ab, containing proteins including two cysteine proteases, an RNA-dependent RNA polymerase, and other nonstructural proteins (nsp1-16). The main function of these proteases is to cleave the polyproteins into their individual proteins to form the transcription/replication complex, making them excellent targets for antiviral drug development.[19][20][21][22] The papain-like protease (PLpro) and 3 chymotrypsin-like protease (3CLpro) only have 3 and 11 cleavage sites, respectively, in the polyproteins, but it is reasonable to assume that both proteases may cleave host cell proteins to modulate the innate immune response and enhance virulence as in picornaviruses and retroviruses, such as the human immunodeficiency virus (HIV).

PLpro is a highly conserved protein domain that has been shown to determine virulence of coronaviruses[23] and possess deubiquinating and deISGylating activity including cleaving ISG15 induced by interferon via the JAK-STAT pathway from ubiquitin-conjugating enzymes and potentially from downstream effectors.[24][25][26][27][28] PLpro deubiquination also prevents activating phosphorylation of IRF3 and subsequent type-I interferon production,[29][30] however the ubiquitinated leucine in human IRF3 is replaced by a serine in bats likely including *Rhinolophus affinus* (intermediate horseshoe bat), the probable species of origin of SARS-CoV-2.[31][32]

3CLpro is also highly conserved among coronaviruses; SARS-CoV-2 3CLpro is 96.08% and 50.65% identical, respectively, to the SARS- and MERS-CoV homologs, the former with only 12 out of 306 amino acids substituted with all 12 outside the catalytic dyad or surrounding pockets.[33][34][35] Even the most distant porcine deltacoronavirus HKU15 3CLpro shares only 34.97% identity yet is similarly conserved in the these important residues. This conservation indicates that all these proteases are capable of cleaving similar sequences no matter the protease genus of origin. In addition to the 11 sites in the polyproteins, these proteases are known to cleave host proteins including STAT2[36] and NEMO[37] to modulate interferon signaling. Similar proteases have been studied in the order Picornavirales[38] and the family Caliciviridae[39], but results have not been reproduced for SARS-CoV-2 yet, and STAT2 and NEMO are only two of many cleaved proteins. The high number of 3CLpro cut sites in coronavirus polyproteins has, however, allowed for sequence logos and training of a neural network (NN) for additional cleavage site prediction.[40][41] Notably, Kiemer et al.'s NN was trained on 7 arbitrary coronavirus genomes, totaling 77 cut sites, and had a Matthews correlation coefficient (MCC) of 0.84, much higher than the traditional consensus pattern's 0.37 for the same training set size. They predicted cleavage sites in select host proteins, namely the transcription factors CREB-RP, OCT-1, and multiple subunits of TFIID, the innate immune modulators interferon alpha-induced protein 6 (IFI6) and IL-1 receptor-associated kinase 1 (IRAK-1), the epithelial ion channels cystic fibrosis transmembrane conductance regulator (CFTR) and amiloride-sensitive sodium channel subunit delta (SCNN1D), the tumor suppressors p53-binding proteins 1 and 2 (although not p53 itself), RNA polymerases and eukaryotic translation initiation factor 4 gamma 1 (eIF4G1), the cytoskeletal proteins microtubule-associated protein 4 (MAP4) and microtubule-associated protein RP/EB members 1 and 3 (MAPRE1/3), and many members of the ubiquitin pathway (USP1/4/5/9X/9Y/13/26 and SOCS6). To our knowledge no one has investigated the entire human proteome for 3CLpro cleavages sites with current or updated sequence logos or neural networks, let alone performed enrichment analysis and classification of these affected proteins.

Methods

Data Set Preparation

A complete, manually reviewed human proteome containing 20,350 sequences (not including alternative isoforms) was retrieved from UniProt/Swiss-Prot (proteome:up000005640 AND reviewed:yes).[42]

Additional coronavirus polyprotein cleavages were collected from GenBank.[43] Searching for "orf1ab," "pp1ab," and "1ab" within the *Coronaviridae* family returned 388 different, complete polyproteins with 762 different cut sites manually discovered using the Clustal Omega multiple sequence alignment server.[44][45][46]

Cleavage Prediction

The NetCorona 1.0 server as in Kiemer et al.'s work[41], our reproductions of their sequence logo and NN, and our improved sequence logo and NNs were used for prediction of cleavage sites.[47] As recommended to maximize the MCC, only predicted cleavages with NetCorona scores greater than or equal to 0.5 were used. Some predicted cleavage sites were close enough to the N- and C-termini that the nine amino acid window input into the neural network was not filled. These sites with center glutamine residue less than four amino acids from the N-terminus or less than five amino acids from the C-terminus were not omitted because they may be in important localization sequences.

Enrichment Analysis

Protein annotation, classification, and enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) 6.8.[48][49] Our training data, prediction methods, and results can be found on GitHub (https://github.com/Luke8472NN/NetProtease).

Results

Here we assumed that SARS-CoV-2 3CLpro is capable of cleaving all aligned cut sites between the four genera of coronaviruses (*Alpha*-, *Beta*-, *Gamma*-, and *Delta*-), because variation in cleavage sequences is greater within polyproteins than between them and because protease/cleavage cophylogeny only demonstrates a weak association (Figures 1 and 2).

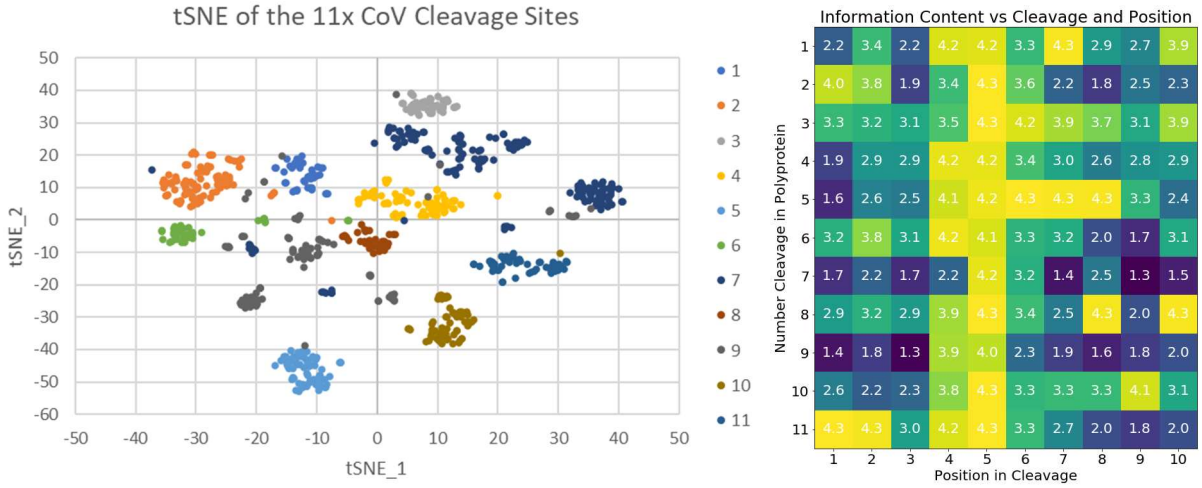

Figure 1: One-hot encoded t-distributed stochastic neighbor embedding (tSNE)[50] and information content both demonstrate that cleavage variation within genomes is more important than variation between genomes.

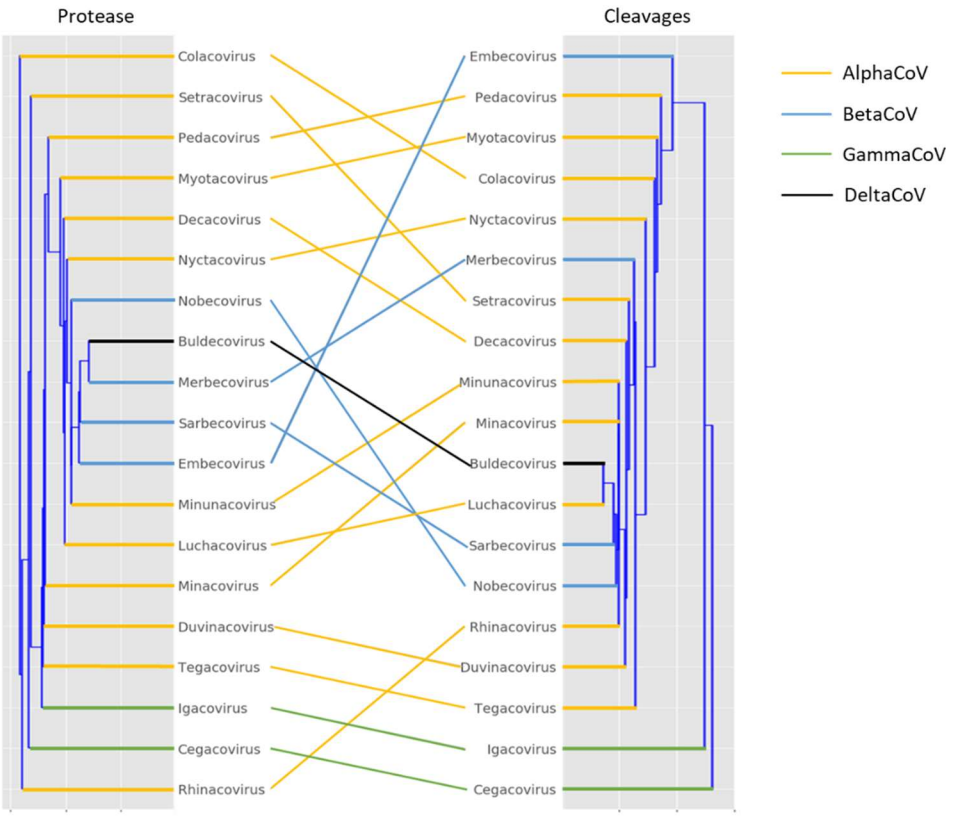

Figure 2: Cophylogeny of 3CLpro and respective cleavages.

Kiemer et al.'s seven genome sequence logo and multilayer perceptron (MLP) structure with each amino acid one-hot encoded as a binary vector of length 20 (an input of 200 bits i.e. linearized 10 amino acids surrounding the cleavage) were both reproduced.[41] Logistic regression was performed on the logarithm of the probability output of the sequence logo with best pseudocount of 2 (out of 20 AAs) and returned an MCC of 0.825 with 74.0% recall. Updating the sequence logo with all known cleavages improved its MCC to 0.927 with 92.5% recall (Figure 3). Figure 4 demonstrates correlations (represented as total information content) between positions that are not captured by simple sequence logos.[51] NNs, however, allow inclusion of 2D and higher-order correlations not easily visualizable and therefore often improve accuracy. Finally, in addition to information content, Figure 5 shows a charge-polarity-hydrophobicity scale with a lack of obvious trend or conservation indicating that one-hot encoding likely performs better than any physiochemical, lower-dimensional inputs.

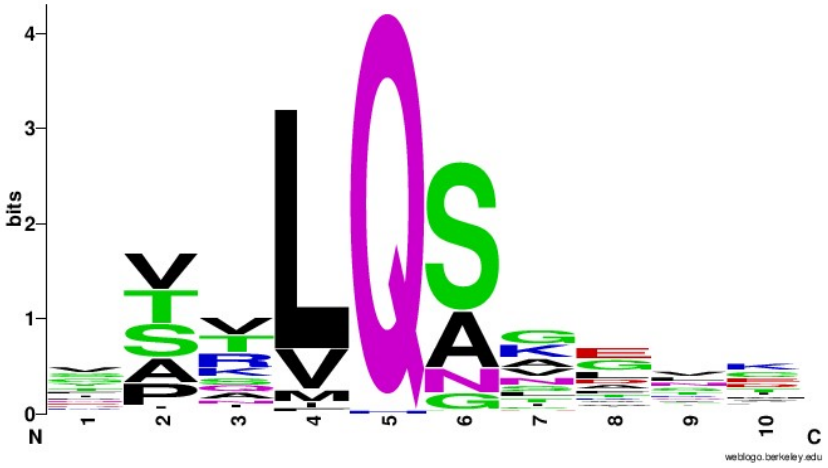

Figure 3: Improved 3CLpro cleavage site sequence logo plotted by WebLogo v2.8.2.[52]

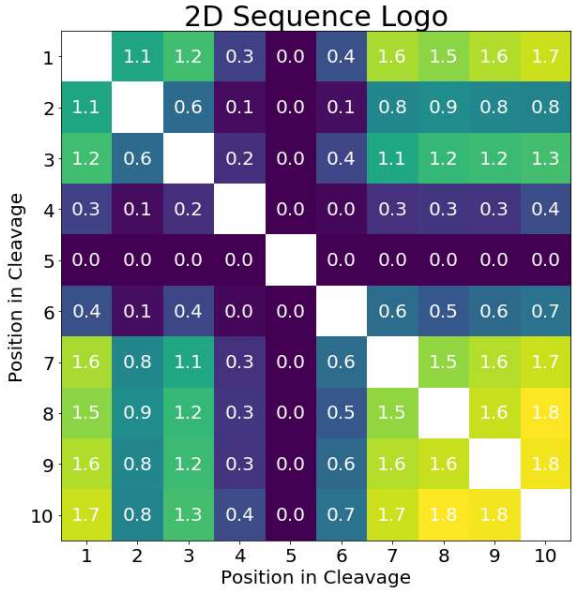

Figure 4: 2D correlations between positions within the improved sequence logo.

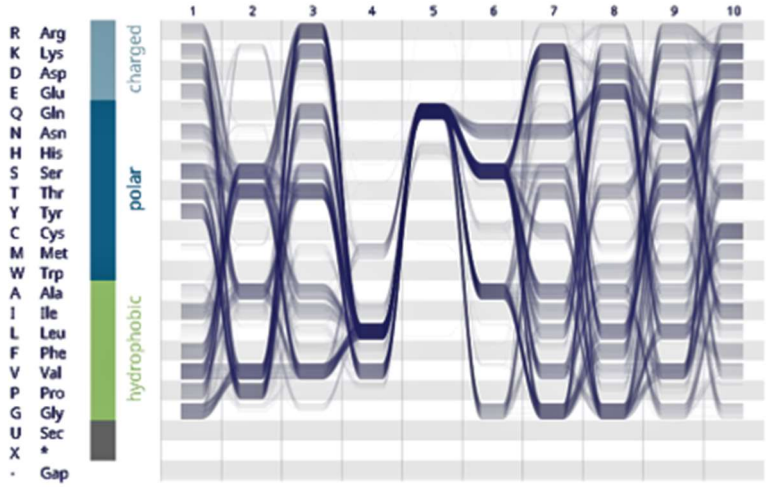

Figure 5: Sequence bundle with charge-polarity-hydrophobicity encoding.[53]

As for our improvements to the NN, note that Kiemer et al.'s MCC of 0.840 is an average from triple cross-validation (CV).[41] Because the known cleavage dataset is small, no data went unused; the three NN output scores were averaged and similarly considered cleavages when greater than 0.5. Applying this average scoring to the entire small and large dataset resulted in single MCCs of 0.946 and 0.849. Retraining the same NN structures (each with one hidden layer with 2 neurons) on the larger dataset resulted in three-average CV and single final MCCs of 0.979 and 0.996, a significant improvement even though the datasets are less balanced. Adding all other histidines (which precede 19/762 different cleavages) as negatives again improved the CV MCC to 0.994 and slightly reduced the final MCC to 0.992. Interestingly, two infectious bronchitis virus (IBV) polyproteins contained cut sites following leucine, methionine, and arginine (VSKLL^AGFKK in APY26744.1 and LVDYM^AGFKK and DAALR^NNELM in ADV71773.1). To our knowledge, only one histidine substitution has been documented[54] and likely does not affect function.[55][56][57] To optimize hyperparameters, the whole dataset was repeatedly split into 80% training/20% testing sets with further splitting of the 80% training set for cross validation. The optimal settings, naive oversampling (within training folds[58]), averaged three-fold cross-validation (on the whole dataset, not just the initial 80%), limited-memory Broyden-Fletcher-Goldfarb-Shanno (lbfgs) solver, hyperbolic tangent activation, 0.00001 regularization, and 1 hidden layer with 10 neurons, had a 20% test set MCC average and standard deviation of 0.983+/-0.003 when split and trained many times. Train/test sets repeatedly split with different ratios in Figure 6 demonstrate that the entire dataset is not required for acceptable performance, although we finally trained three networks on all the data (with three-fold cross-validation), returning a three-average CV and final MCC of 0.983 and 0.998, respectively. Table 1 lists the few incorrectly labeled sequences and their respective sources and scores.

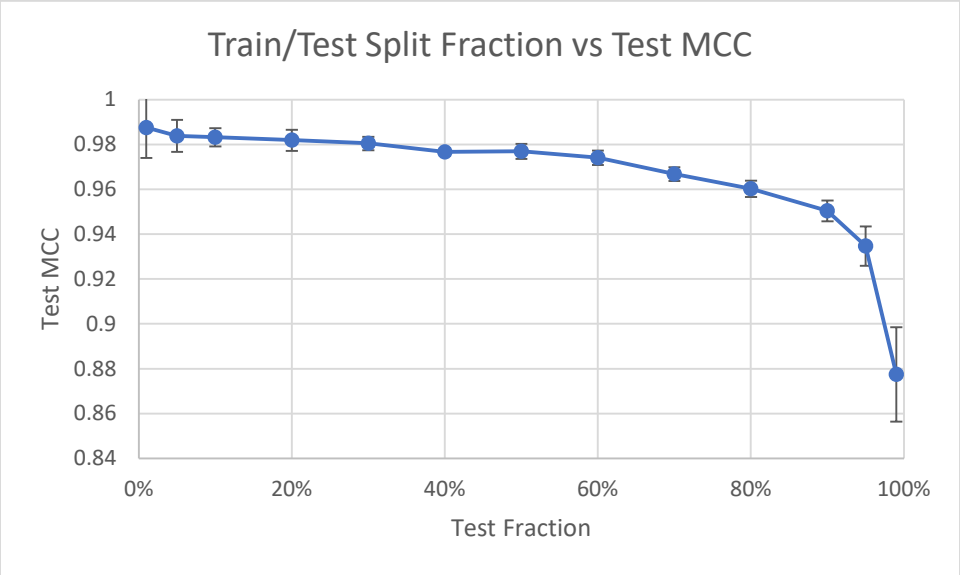

Figure 6: Train/test split fraction vs MCC demonstrating that performance approaches a limit.

Table 1: Only 15 out of 34,500 sequences were incorrectly labeled by the final NN. FN, false negative; FP, false positive.

Error	Genera	Virus	Taxonomy ID	Sequence	Score
FN	Gamma	Beluga whale CoV SW1	NCBI:txid694015	SLELQSVPQN	0.00000
FN	Unclassified	Shrew CoV	NCBI:txid2050019	SYQIQGKDES	0.33024
FN	Unclassified	Shrew CoV	NCBI:txid2050019	YPTLQGQWAP	0.33170
FN	Alpha	Wencheng Sm shrew CoV	NCBI:txid1508228	NNNLQVLERL	0.33329
FN	Unclassified	Guangdong Chinese water skink CoV	NCBI:txid2116470	GVKVQSFKVK	0.37234

FN	Gamma	Canada goose CoV	NCBI:txid2569586	RPTMQFDSYS	0.38064
FN	Beta	SARS	NCBI:txid694009	VAVLQAENVT	0.42198
FP	Beta	CoV BtRI-BetaCoV/SC2018	NCBI:txid2591233	FVRIQSGQTF	0.53847
FP	Alpha	Myotis ricketti CoV SAX2011	NCBI:txid1503289	NKTLHAGILD	0.66122
FP	Beta	HCoV-OC43	NCBI:txid31631	PAALHSKCLT	0.66138
FP	Beta	MERS	NCBI:txid1335626	VIILQATKFT	0.66151
FP	Alpha	Feline CoV	NCBI:txid12663	ETSLQCLIST	0.66504
FP	Alpha	Unclassified Minacovirus Mink/China/1/2016	NCBI:txid2163884	KTKIQAKFGT	0.66668
FP	Beta	MERS	NCBI:txid1335626	FVVLQGKVST	0.71328
FP	Alpha	NL63-related bat CoV	NCBI:txid2501929	NSILQGTSLS	0.99993

Of the 20,350 manually reviewed human proteins, 4,460 were cut at least once with a NN score greater than or equal to 0.5. To prove that the 5,887 cut sites were nonrandomly distributed among human proteins (with a maximum of 25 cleavages in the 5,795 amino acid, RNA splicing regulation nucleoprotein protein AHNAK2), random sequences with weighted amino acid frequencies were checked for cleavages. Cleavages occurred at 1.10% of glutamines (4.77% of amino acids)[59] or every 1,900 amino acids in these random sequences. Most proteins are shorter than this and would, if randomly distributed, follow a Poisson distribution; our data's deviation from this distribution indicates that many cleavages are intentional.

Tissue (UP_TISSUE and UNIGENE_EST_QUARTILE), InterPro, direct Gene Ontology (GO includes cellular compartment (CC), biological process (BP), and molecular function (MF)), Reactome pathways, sequence features, and keywords annotations were all explored in DAVID. Only annotations with Benjamini-Hochberg-corrected p-values less than 0.05 were considered statistically significant, and both enriched and depleted (no cleavages) annotations are listed in Tables S1-S9.

Discussion

Enrichment and depletion analyses are often used to probe the importance of annotations in many disease states, yet quantification is not possible without experimentation. First, if a protein is central to a pathway, a single cleavage may be all that is required to generate equivalent downstream outcomes. Cleaved proproteins such as coagulation factors or complement proteins may even be activated by 3CLpro cleavage. Additional exhaustive analysis is required to determine if any insignificantly enriched or depleted pathways are still affected at central nodes (i.e. false negatives). Second, longer proteins are more likely to be randomly cleaved and may confound conclusions about annotations containing them. Cleavages in longer proteins (e.g. cytoskeletal or cell-cell adhesion components) are no less important than those in shorter sequences, and proteins with multiple cleavages deviating from Poisson distributions within annotations are more likely due to highly conserved sequences than simply protein length. Lastly, convergent evolution within the host may also result in false positives and may be partially avoided by investigating correlations between domains, motifs, repeats, compositionally biased regions, or other sequence or structural similarities and other functional and ontological annotations. Ideally, a negative control proteome from an uninfectable species could prevent false positives, but coronaviruses are extremely zoonotic. Here, depletions in the human proteome are taken to be negative controls. Comparison with a bat proteome with deficiencies in many immune pathways, however, may show which human cleavages are unintentional or exert little or no selective pressure.

Tissues

As expected in our data, the most significant tissue enrichment of 3CLpro cleavages in our data are in the epithelium, but central and peripheral nervous tissues are also affected due to their similar expression and enrichment of complex structural and cell junction proteins. It is noteworthy that major

proteins associated with neurodegenerative disease are also predicted to be cleaved: Alzheimer's disease (amyloid precursor protein (APP), tau protein), Parkinson's disease (VPS35, EIF4G1, DNAJC13), Huntington's disease (huntingtin), amyotrophic lateral sclerosis (TDP-43), and spinocerebellar ataxia type 1 (ataxin-1). Testis has somewhat similar expression to epithelium and brain, highly expresses ACE2, and is enriched in movement/motility- (subset of structural proteins) and meiosis-related (chromosome segregation) proteins, further increasing the likelihood that this tissue is infectible. Spleen, however, does not express much ACE2, and its enrichment is likely due to genes with immune function and mutagenesis sites. Proteins with greater tissue specificity (3rd quartile) show additional enrichments along the respiratory tract (tongue, pharynx, larynx, and trachea), in immune tissues (lymph node and thymus), and in other sensory tissues (eye and ear). Combining tissues, tobacco use disorder is the only significantly enriched disease, but acquired immunodeficiency syndrome (AIDS) and atherosclerosis were surprisingly depleted.

Cleavages are also surprisingly depleted in olfactory and gustatory pathways given the virus' ability to infect related cells and present as anosmia and ageusia. Olfactory receptors are transmembrane rhodopsin-like G protein-coupled receptors that, when bound to an odorant, stimulate production of cAMP via the G protein and adenylate cyclase. The G proteins GNAL and GNAS are not cleaved, and some but not all adenylate cyclases are cleaved, likely resulting in an increase in cAMP. cAMP is mainly used in these cells to open their respective ligand-gated ion channels and cause depolarization, but it is also known to inhibit inflammatory responses through PKA and Epac. Multiple phosphodiesterases (PDEs) that degrade cAMP but not PDE4, the major PDE in inflammatory and immune cells, are cleaved. PDE4 inhibitors have been shown to reduce destructive respiratory syncytial virus-induces inflammation in lung,[60] but olfactory receptor neurons are quickly regenerated and sacrifice themselves when infected by influenza A virus.[61] The depletion in cleavages and resulting increase in cAMP in these neurons is likely to inhibit their programmed cell death long enough for the virus to be transmitted through the glomeruli to mitral cells and the rest of the olfactory bulb. Tongue infection may have similar mechanisms, and herpes simplex virus (HSV) has been shown to be transmitted to the brainstem through the facial and trigeminal nerves.[62]

Gene Ontology

Cleaved proteins are depleted in the extracellular space (except for structural collagen, laminin, and fibronectin mainly associated) and enriched in the cytoplasm and many of its components, indicating that the selective pressure for cleavage is weaker once cells are lysed and the protease is released. In the cytoplasm, the most obviously enriched sets are in the cytoskeleton (microfilament, intermediate filament, microtubule, and spectrin), motor proteins (myosin, kinesin, and dynein), cell adhesion molecules (integrin, immunoglobulin, cadherin, and selectin), and relevant Ras GTPases (Rho, Rab, Ran, Rac, and Arf), particularly in microtubule organizing centers (MTOCs) including centrosomes, an organelle central to pathways in the cell cycle including sister chromatid segregation. Coiled coils account for many of these cleavages and are primarily expressed in corresponding cellular compartments in the epithelium, testis, and brain. Only the coronavirus nsp1 (suppresses host antiviral response), nsp13 (helicase/triphosphatase), and spike proteins have so far been shown to interact with the cytoskeleton,[63][64][65] although many other viruses including influenza A virus,[66] herpes simplex virus (HSV), rabies virus, vesicular stomatitis virus (VSV), and adeno-associated virus (AAV)[67] also modulate the cytoskeleton.[68] In neurons, this allows for axonal and trans-synaptic transport of viruses which can often be inhibited but sometimes exaggerated by cytoskeletal drugs often used in oncology.[69][70][71][72]

Modulation of these structural and motor proteins is required for formation of the double-membrane vesicles surrounding replicase complexes[73][74] and for egress. Similarly required for vesicular transport, the coatomer COPI, clathrin, and caveolae pathways are untouched by 3CLpro other than the muscle-specific cavin-3, but COPII's SEC24A/24B/31A are likely cleaved due to their function in

selecting cargo[75][76] and contribution to membrane curvature preventing nucleocapsid engulfment.[77] Cleavage of retromer (VSP35), GGA (GGA1), and many adaptor protein complexes (AP1B1/G1/G2, AP2A1/B1, AP3B1/B2/D1/M1/M2, and AP5B1/M1) often targeting degradation leaves only the poorly characterized AP4 or other unknown pathways to handle egress. Modulators of any of these vesicle trafficking pathways may be effective treatments for COVID-19.

The nucleus is enriched because its nuclear localization signals and its scaffolding proteins are cleaved. Additionally, many nuclear pore complex proteins and importins/exportins associated with RNA transport are also cleaved. Lamins, which are cleaved by caspases during apoptosis to allow chromosome detachment and condensation, are also cleaved by 3CLpro. Chromatin-remodeling proteins including histone acetyltransferases (HATs) often containing bromodomains, histone deacetylases (HDACs), structural maintenance of chromosomes (SMC) proteins (cohesins and condensins) also containing coiled coils, and separase (the cysteine protease that cleaves cohesin to separate sister chromatids), but not CCCTC-binding factor (CTCF) or topoisomerase II are cleaved, complicating the effects on chromosome condensation and global gene expression. HDAC inhibitors have been shown to decrease or increase virulence depending on the virus.[78][79][80][81][82] Some but not all DNA methyltransferases and demethylases are cleaved, further complicating the global effects on transcription. Viruses benefit from preventing programmed cell death and its corresponding chromosomal compaction in response to viral infection (pyknosis), but they also attempt to reduce host transcription by condensing chromosomes and reroute translation machinery towards their own open reading frames.[83][84] Relatedly, 28S rRNA has been shown to be cleaved by murine coronavirus, and ribosomes with altered activity are likely directed from host to viral RNAs.[85] Ribosome cleavages are depleted here because they are recruited for viral translation, but the few ribosomal proteins that are cleaved (RPL4/10 and RPS3A/19) tend to be more represented in monosomes, not polysomes,[86] indicating that ribosomes that initiate much faster than they elongate are preferred because they likely frameshift more frequently, allowing for control of the stoichiometric ratio of pp1a and pp1ab.[87] Signal recognition particle (SRP) subunits 54/68/72kDa associated with the ribosome are also predicted to be cleaved. SRP, especially the uncleaved SRP9/14kDa heterodimer, encourage translation elongation arrest to allow translocation including transmembrane domain insertion (e.g. coronavirus envelope protein) and has been associated with frameshifts.[88][89][90] In fact, frameshifting is a highly enriched keyword in cleaved proteins mainly due to endogenous retroviral (ERV) elements, some of which can activate an antiviral response via pattern recognition receptors (PRRs).[91] Some also resemble reverse transcriptases and may, like the CRIPSR system in prokaryotes, be capable of copying coronavirus genomic RNA to produce an RNAi response via the similarly cleaved DICER1, AGO1/2, and PIWL1/3.[92] If the latter is true, individuals with distinct ERV alleles and loci may differentially respond to SARS-CoV-2 infection and/or treatment, especially exogenous RNAi. Lastly, ribosomal proteins are also included in the nonsense-mediated decay (NMD) pathway, which is likely depleted in cleavages because NMD has been shown to be a host defense against coronavirus' genomic and subgenomic RNAs' multiple ORFs and large 3' UTRs.[93] It was also shown that the nucleocapsid protein inhibits this degradation but often cannot protect newly synthesized RNAs early in infection. The selective pressure on 3CLpro may be reversed by this nucleocapsid inhibition and the preferential degradation of host mRNAs such that host resources can again be directed towards viral translation.

In addition to affecting large organelles, 3CLpro is predicted to cleave all known components of vault: major vault proteins (MVP), telomerase protein component 1 (TEP1), and poly(ADP-ribose) polymerase 4 (PARP4). Vault function has not been completely described, but it has known interactions with other viruses.[94][95][96] Telomerase reverse transcriptase (TERT) is also cleaved, but is more frequently reported to be activated by other viral infections and/or promote oncogenesis.[97]

Other common viral process proteins are enriched in the epithelium and adaptive immune cells, and those cleaved may affect the heat shock response and other small RNA processing. Lactoferrin, an

antiviral protein that is upregulated in SARS infection,[98] is also cleaved, although one of its fragments, lactoferricin, has known antiviral activity.[99] Cleaved PRRs include the toll-like receptors TLR6/8, the C-type lectin receptors CLEC4G/H1/4K/4L/10A/13B/13C/16A, KLRC4/G1, ACG1, COLEC7/12, CSPG3, FREM1, LAYN, PKD1, SELE, and THBD primarily present on dendritic cells, the NOD-like receptors NOD2 and NLRP1/2/3/6/10/12/14. Cleaved proteins downstream of these PRRs include RIP1/2, NFKB2, CFLAR, TRIF, IRF2, and DAXX, and other relevant downstream pathways similarly include many cleaved proteins: PI3K/AKT (PIK3CG/D, PIK3R2/5/6, PPP2R1A/2A/2B/2C/2D/3B/5B (PP2A dephorphorylates AKT), n/iNOS (where nitric oxide has conflicting effects on viral infections[100][101]), TSC1/2), mTOR (S6K, SRBP1, RBCC1, RAPTOR, RICTOR, PRR5L, FIP200, CLIP1, CLIP-170, KS6B1/2, SREBP1, FOXO1/3), and MAPK (MAP4K2/4/5, MAP3K4/5/6/8/13/14, KSR2, MAP2K3/4, MAPK7/13/15, and DUSP8, MDM2). Additional cleaved transcription factors include c-Jun, ATF6, CREB3/5/BP, SP1, OCT1/2, TFIIIC, TAF6, TBPL2, and HSF2/2BP/4/X1. No interferons are cleaved likely due to their redundancy, and no interferon receptors are cleaved. The downstream effectors STAT1/2/4 and the interferon-stimulated genes (ISGs) GBP1, IFI6, MS4A4A, OAS1, PML, mitoferrin-2, TREX1, and TRIM5 and the TNF ligands (TNFSF3/13/18 and EDA) and receptor TNFRSF21 are, however, also cleaved. Finally, pro-apoptotic protein cleavages exist in the Bcl-2 family (Bcl-rambo) and in caspases (CASP2/5/12), although the anti-apoptotic Bcl-2 protein (Bcl-B) and inhibitors of apoptosis (BIRC2/3/6) are also cleaved.

Other Pathways and Keywords

Lipoproteins are a depleted keyword, but apolipoproteins A-V/B/L1/(a), lipid transfer proteins CETP and MTTP, and receptors, LRP2/6/12 are all predicted to be cleaved and, other than the proapoptotic ApoL1,[102] are associated with chylomicrons, VLDL, and LDL as opposed to HDL, indicating that lipoproteins may contribute to the correlations between COVID-19 symptom severity, dyslipidemia, and cardiovascular disease. It was recently discovered that SARS-CoV-2 spike protein binds cholesterol, allowing for association with and reduced serum concentration of HDL. These findings combined with the 3CLpro cleavages show an opportunity for HDL receptor inhibitor treatment, especially antagonists of the uncleaved SR-B1.[103] Cleavage of the adipokines leptin, leptin receptor, and IL-6 provide a mechanism for COVID-19 comorbidity with obesity independent of lipoproteins and indicate another potential treatment: anti-leptin antibodies.[104][105]

Ubiquitinating and deubiquitinating (DUBs) enzymes are most enriched in the epithelium and the nucleus and include ubiquitin ligase-supporting scaffolding cullins and DUBs such as the ubiquitous proteasomal subunit RPN11 and related lid subunits RPN6/10/12. Ubiquitin itself is not, but NEDD4 and the related SMURF1/2 and HECW1 are, cleaved. NEDD4 has been shown to enhance influenza infectivity by inhibiting IFITM3[106][107] and Japanese encephalitis virus by inhibiting autophagy,[108] but its ubiquitination of many diverse human viruses promotes their egress. SARS-CoV-2 has two probable NEDD4 binding sites: the proline-rich, N-terminal PPAY and LPSY[109] in the spike protein and nsp8, respectively. Although the former sequence is APNY and is likely not ubiquitinated in SARS-CoV, small molecule drugs targeting this interaction or related kinases may be useful treatments for COVID-19 as they have been for other RNA viruses.[110][111][112] Additional research is required to compare these cleavages to the PLpro deubiquitinating activity and the specificity and function of distinct ubiquitin and other ubiquitin-like protein linkage sites.[113][114]

Helicases make up approximately 1% of eukaryotic genes and are enriched in cleavages with many containing RNA-specific DEAD/DEAH boxes. Most viruses except for retroviruses have their own helicase (nsp13 in SARS-CoV-2) and multiple human RNA helicases have been shown to sense viral RNA or enhance viral replication.[115][116][117] SARS nsp13 (helicase) and nsp14 have been shown to be enhanced by the uncleaved human DDX5 and DDX1, respectively.[118][119] Multiple proteins interacting with DDX1 (FAM98A) and DDX5 (DHX15, SNW1, MTREX, and HNRNPH1), the RIG-I-associating DDX6, and DDX20 involved in ribosome assembly are, however, predicted to be cleaved, making these effects complex too.

Fibronectin type-III domains are enriched, but fibronectin itself is not cleaved. No cleaved proteins with this domain are directly related to coagulation, but tissue factor, coagulation factors VIII (antihemophilic factor A glycoprotein, also an acute-phase protein secreted in response to infection), XII (Hageman factor), XIII (fibrin-stabilizing factor transglutaminase), plasmin(ogen), von Willebrand factor, and kininogen-1 are cleaved. Multiple cleaved serpin suicide protease inhibitors (plasminogen activator inhibitor-2, megsin, alpha-1-antitrypsin, and the less relevant angiotensinogen, protein Z-dependent protease inhibitor, leukocyte elastase inhibitor, and heat shock protein 47) are also related to coagulation, potentially increasing both thrombosis and fibrinolysis rates or resulting in dose-dependent effects.[120][121] Most other antiproteases, however, are too small to have many potential cleavage sites even though they are a very important response to respiratory virus infection. Serpin replacement therapy or treatment with modified small, 3CLpro competitive inhibitors may be a useful treatment for COVID-19.[122]

In addition to coagulation factors, the complement system can induce, in addition to many other components of innate immunity, expulsion of neutrophil extracellular traps (NETs) intended to bind and kill pathogens.[123] NETs, however, simultaneously trap platelets expressing tissue factor and contribute to hypercoagulability. The complement pathway is not obviously enriched, but many central proteins (C1/3/4 and factor B) are or have subunits that are cleaved, indicating viral adaptation to the classical, alternative, and likely lectin pathways.[124][125][126] Neutrophilia and NET-associated host damage are known to occur in severe SARS-CoV-2 infection, so inhibitors of the pathway are currently in clinical trials: histone citrullination, neutrophil elastase, and gasdermin D inhibitors to prevent release and DNases to degrade chromatin after release.[127][128] Complement inhibition would likely similarly reduce the risks of hypercoagulability and other immune-mediated inflammation associated with COVID-19, but effects may vary widely between sexes and ages.[129][130]

Redox-active centers including proteins involved in selenocysteine synthesis are additionally depleted in cleavages likely because of their involvement in avoiding cell death and innate immune response. Respiratory viruses differentially modulate redox pathways, balancing lysis-enhanced virion proliferation and DUOX2-derived ROS-induced interferon response.[131] In addition to depleted antioxidant proteins, cleavage of dual oxidases (DUOX1/2), NADPH oxidase 5 (NOX5), and xanthine oxidase (XO), the former of which are upregulated in chronic obstructive pulmonary disease (COPD),[132] indicates that coronaviruses prefer to reduce oxidative stress in infected cells, contrary to most COVID-19 symptoms. Given the diversity of responses to respiratory virus infections, each proposed antioxidant treatment for COVID-19 should be thoroughly evaluated before widespread distribution.

The impact of post-translational modifications on viral protease cleavage frequency have not been well characterized. Glutamine and leucine, the two most important residues in the cleavage sequence logo, are rarely modified, but serine, the next most important residue, is the most frequently phosphorylated amino acid. Analysis of keywords showed enrichment of phosphoproteins and depletion of disulfide crosslinked, lipid-anchored, and other transmembrane proteins.

Lastly, the keywords polymorphism and alternate splicing were enriched, indicating that additional variability between cell lines and between individuals are likely. Once health systems are not so burdened by the quantity of cases and multiple treatments are developed, personalized interventions will likely differ significantly between individuals.

Conclusion

Many expected and novel protein annotations were discovered to be enriched and depleted in cleavages, indicating that 3CLpro is a much more important virulence factor than previously thought. 3CLpro cleavages are enriched in the epithelium (especially along the respiratory tract), brain, testis, plasma, and immune tissues and depleted in olfactory and gustatory receptors. Affected pathways with

discussed connections to viral infections include cytoskeleton/motor/cell adhesion proteins, nuclear condensation and other epigenetics, host transcription and RNAi, coagulation, pattern recognition receptors, growth factor, lipoprotein, redox, ubiquitination, apoptosis. These pathways point towards many potential therapeutic mechanisms to combat COVID-19: cytoskeletal drugs frequently used against cancer, modulators of ribosomal stoichiometry to enrich monosomes, upregulation of DICER1 and AGO1/2, exogenous lactoferrin and small, modified antiproteases, upregulation of serpins potentially via dietary antioxidants, complement inhibition, reduction of LDL and inhibition of HDL receptor (e.g. by antagonizing SR-B1), anti-leptin antibodies, and downregulating NEDD4 or related kinases and upregulating IFITM3. Pathways with more complex disruption that may also deliver therapeutic targets but require elucidating experimental results include PDEs, histone acetylation, nitric oxide, and vesicle coatomers. It is also worth further investigating how 3CLpro contributes if at all to the correlations between obesity and severity of infection or to viral induction of autoimmune and potentially oncological conditions.

Expansion of our dataset to the whole order *Nidovirales* may provide more diversity to improve classifying methods if protease/cleavage cophylogeny does not invalidate the assumption of cross-reactivity. Additional issues requiring experimentation include characterization of cleavage kinetics, any functional differences between proteases, the molecular effects of post-translation modifications, the individual and population effects of polymorphisms in cleavage sequences on susceptibility to or severity of infection. Even though many caveats exist without experimentation, similar prediction, enrichment/depletion analysis, and therapeutic target identification should be performed for every other viral protease.

References

1. Lechien JR, Chiesa-Estomba CM, De Siati DR, et al. Olfactory and gustatory dysfunction as a clinical presentation of mild to moderate forms of COVID-19: A multicenter European study. *Eur Arch Otorhinolaryngol*, 277(8):2251-61 (Aug 2020). https://doi.org/10.1007/s00405-020-05965-1
2. Baig AM, Khaleeq A, Ali U, et al. Evidence of the COVID-19 virus targeting the CNS: Tissue distribution, host-virus interaction, and proposed neurotropic mechanisms. *ACS Chem Neurosci*, 11(7):995-8 (Mar 2020). https://doi.org/10.1021/acschemneuro.0c00122
3. Lau KK, Yu WC, Chu CM, et al. Possible central nervous system infection by SARS coronavirus. *Emerg Infect Dis*, 10(2):342-4 (Feb 2004). https://doi.org/10.3201/eid1002.030638
4. Netland J, Meyerholz DK, Moore S, et al. Severe acute respiratory syndrome coronavirus infection causes neuronal death in the absence of encephalitic in mice transgenic for human ACE2. *J Virol*, 82(15):7264-75 (Aug 2008). https://doi.org/10.1128/JVI.00737-08
5. Li YC, Bai WZ, Hashikawa T. The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *J Med Virol*, 2020:1-4 (Mar 2020). https://doi.org/10.1002/jmv.25728
6. Zhang C, Shi L, Wang FS. Liver injury in COVID-19: management and challenges. *Lancet Gastroenterol Hepatol*, 5(5):428-30 (May 2020). https://doi.org/10.1016/S2468-1253(20)30057-1
7. Chau TN, Lee KC, Yao H, et al. SARS-associated viral hepatitis caused by a novel coronavirus: Report of three cases. *Hepatology*, 39(2):302-10 (Feb 2004). https://doi.org/10.1002/hep.20111
8. Liu L, Wei Q, Alvarez X, et al. Epithelial cells lining salivary gland ducts are early target cells of severe acute respiratory syndrome coronavirus infection in the upper respiratory tract of rhesus macaques. *J Virol*, 85(8):4025-30 (Apr 2011). https://doi.org/10.1128/JVI.02292-10
9. Zhan J, Deng R, Tang J, et al. The spleen as a target in severe acute respiratory syndrome. *FASEB J*, 20(13):2321-8 (Nov 2006). https://doi.org/10.1096/fj.06-6324com
10. Naicker S, Yang CW, Hwang SJ, et al. The novel coronavirus 2019 epidemic and kidneys. *Kidney Int*, 97:824-8 (May 2020). https://doi.org/10.1016/j.kint.2020.03.001

11. Fan C, Ding Y, Lu WL, et al. ACE2 expression in kidney and testes may cause kidney and testis damage after 2019-nCoV infection. *medRxiv* (Feb 2020). https://doi.org/10.1101/2020.02.12.20022418
12. Chen H, Guo J, Wang C, et al. Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. *Lancet*, 395(10226):809-15 (Mar 2020). https://doi.org/10.1016/S0140-6736(20)30360-3
13. Zheng YY, Ma YT, Zhang JY, et al. COVID-19 and the cardiovascular system. *Nat Rev Cardiol*, 17:259-60 (Mar 2020). https://doi.org/10.1038/s41569-020-0360-5
14. Dandekar AA, Perlman S. Immunopathogenesis of coronavirus infections: implications for SARS. *Nat Rev Immunol*, 5(12):917-927 (Dec 2005). https://doi.org/10.1038/nri1732
15. Gu J, Gong E, Zhang B, et al. Multiple organ infection and the pathogenesis of SARS. *J Exp Med*, 202(3):415-24 (Aug 2005). https://doi.org/10.1084/jem.20050828
16. Jia X, Yin C, Lu S, et al. Two things about COVID-19 might need attention. *Preprints*, 2020020315 (Feb 2020). https://doi.org/10.20944/preprints202002.0315.v1
17. Simonnet A, Chetboun M, Poissy J, et al. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) requiring invasive mechanical ventilation. *Obesity*, 28(7):1196-9 (Apr 2020). https://doi.org/10.1002/oby.22831
18. Elliot JG, Donovan GM, Wang KCW, et al. Fatty airways: Implications for obstructive disease. *Eur Respir J*, 56(2) (Oct 2019). https://doi.org/10.1183/13993003.00857-2019
19. Baez-Santos YM, St. John SE, Mesecar AD. The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds. *Antiviral Res*, 115:21-38 (Mar 2015). https://doi.org/10.1016/j.antiviral.2014.12.015
20. Pillaiyar T, Manickam M, Namasivayam V, et al. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: Peptidomimetics and small molecule chemotherapy. *J Med Chem*, 59(14):6595-628 (Feb 2016). https://doi.org/10.1021/acs.jmedchem.5b01461
21. Yang H, Xie W, Xue X, et al. Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS Biology*, 3(10):e428 (Nov 2005). https://doi.org/10.1371/journal.pbio.0030324
22. Anand K, Ziebuhr J, Wadhwani P, et al. Coronavirus main proteinase (3CLpro) structure: Basis for design of anti-SARS drugs. *Science*, 300(5626):1763-7 (Jun 2003). https://doi.org/10.1126/science.1085658
23. Neimeyer D, Mosbauer K, Klein EM, et al. The papain-like protease determines a virulence trait that varies among members of the SARS-coronavirus species. *PLoS Pathog*, 14(9):e1007296 (Sep 2018). https://doi.org/10.1371/journal.ppat.1007296
24. Barretto N, Jukneliene D, Ratia K, et al. The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquinating activity. *J Virol*, 79(24):15189-98 (Dec 2005). https://doi.org/10.1128/JVI.79.24.15189-15198.2005
25. Yang X, Chen X, Bian G, et al. Proteolytic processing, deubiquitinase and interferon antagonist activities of Middle East respiratory syndrome coronavirus papain-like protease. *J Gen Virol*, 95(3):614-26 (Mar 2014). https://doi.org/10.1099/vir.0.059014-0
26. Bailey-Elkin BA, Knaap RCM, Johnson GG, et al. Crystal structure of the Middle East respiratory syndrome coronavirus (MERS-CoV) papain-like protease bound to ubiquitin facilitates targeted disruption of deubiquinating activity to demonstrate its role in innate immune suppression. *J Biol Chem*, 289:34667-82 (Dec 2014). https://doi.org/10.1074/jbc.M114.609644
27. Li SW, Lai CC, Ping JF, et al. Severe acute respiratory syndrome coronavirus papain-like protease suppressed alpha interferon-induced responses through downregulation of extracellular signal-

regulated kinase 1-mediated signalling pathways. *J Gen Virol*, 92(5):1127-40 (May 2011). https://doi.org/10.1099/vir.0.028936-0
28. Xing Y, Chen J, Tu J, et al. The papain-like protease of porcine epidemic diarrhea virus negatively regulates type I interferon pathway by acting as a viral deubiquitinase. *J Gen Virol*, 94(7):1554-67 (Jul 2013). https://doi.org/10.1099/vir.0.051169-0
29. Matthews K, Schafer A, Pham A, et al. The SARS coronavirus papain like protease can inhibit IRF3 at a post activation step that requires deubiquination activity. *Virol J*, 11:209 (Dec 2014). https://doi.org/10.1186/s12985-014-0209-9
30. Devaraj SG, Wang N, Chen Z, et al. Regulation of IRF-3-dependent innate immunity by the papain-like protease domain of the severe acute respiratory syndrome coronavirus. *J Biol Chem*, 282:32208-21 (Nov 2007). https://doi.org/10.1074/jbc.M704870200
31. Banerjee A, Zhang X, Yip A, et al. Positive selection of a serine residue in bat IRF3 confers enhanced antiviral protection. *iScience*, 23(7):100958 (Mar 2020). https://doi.org/10.1016/j.isci.2020.100958
32. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature,* 579:270-3 (Feb 2020). https://doi.org/10.1038/s41586-020-2012-7
33. Needle D, Lountos GT, Waugh DS. Structures of the Middle East respiratory syndrome coronavirus 3C-like protease reveal insights into substrate specificity. *Acta Cryst*, D71:1102-11 (Feb 2015). https://doi.org/10.1107/S1399004715003521
34. Xue X, Yu H, Yang H, et al. Structures of two coronavirus main proteases: Implications for substrate binding and antiviral drug design. *J Virol*, 82(5):2515-27 (Mar 2008). https://doi.org/10.1128/JVI.02114-07
35. Anand K, Palm GJ, Mesters JR, et al. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J,* 21(13):3213-24 (Jul 2002). https://doi.org/10.1093/emboj/cdf327
36. Zhu X, Wang D, Zhou J, et al. Porcine deltacoronavirus nsp5 antagonizes type I interferon signaling by cleaving STAT2. *J Virol*, 91(10):e00003-17 (May 2017). https://doi.org/10.1128/JVI.00003-17
37. Wang, D, Fang L, Shi Y, et al. Porcine epidemic diarrhea virus 3C-like protease regulates its interferon antagonism by cleaving NEMO. *J Virol*, 90(4):2090-2101 (Feb 2016). https://doi.org/https://jvi.asm.org/content/90/4/2090
38. Ye S, Xia H, Dong C, et al. Identification and characterization of Iflavirus 3C-like protease processing activities. *Virol*, 428(2):136-45 (Jul 2012). https://doi.org/10.1016/j.virol.2012.04.002
39. Kuyumcu-Martinez M, Belliot G, Sosnovtsev SV, et al. Calicivirus 3C-like proteinase inhibits cellular translation by cleavage of poly(A)-binding protein. *J Virol*, 78(15):8172-82 (Aug 2014). https://doi.org/10.1128/JVI.78.15.8172-8182.2004
40. Chuck CP, Chow HF, Wan DCC, et al. Profiling of substrate specificities of 3C-like proteases from group 1, 2a, 2b, and 3 coronaviruses. *PLoS One*, 6(11):e27228 (Nov 2011). https://doi.org/10.1371/journal.pone.0027228
41. Kiemer L, Lund O, Brunak S, et al. Coronavirus 3CLpro proteinase cleavage sites: Possible relevance to SARS virus pathology. *BMC Bioinform*, 5:72 (Jun 2004). https://doi.org/10.1186/1471-2105-5-72
42. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47(D1):D505-15 (Jan 2019). https://doi.org/10.1093/nar/gky1049
43. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*, 45(D1):D37-42 (Jan 2017). https://doi.org/10.1093/nar/gkw1070
44. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol,* 7:539 (Oct 2011). https://doi.org/10.1038/msb.2011.75
45. Goujon M, McWilliam H, Li W, et al. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*, 38(2):W695-9 (Jul 2010). https://doi.org/10.1093/nar/gkq313

46. McWilliam H, Li W, Uludag M, et al. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res*, 41(W1):W597-600 (Jul 2013). https://doi.org/10.1093/nar/gkt376
47. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12:2825-30 (Oct 2011). http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
48. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44-57 (Jan 2009). https://doi.org/10.1038/nprot.2008.211
49. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res,* 37(1):1-13 (Jan 2009). https://doi.org/10.1093/nar/gkn923
50. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*, 9:2579-605 (Nov 2008). http://www.jmlr.org/papers/v9/vandermaaten08a.html
51. Bindewald E, Schneider TD, Shapiro BA. CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic Acids Res*, 34:W405-11 (Jul 2006). https://doi.org/10.1093/nar/gkl269
52. Crooks GE, Hon G, Chandonia JM, et al. WebLogo: A sequence logo generator. *Genome Res*, 14:1188-90 (Jun 2004). https://doi.org/10.1101/gr.849004
53. Kultys M, Nicholas L, Schwarz R, et al. Sequence Bundles: a novel method for visualizing, discovering and exploring sequence motifs. *BMC Proc*, 8(Suppl 2):S8 (Aug 2014). https://doi.org/10.1186/1753-6561-8-S2-S8
54. Woo PCY, Huang Y, Lau SKP, et al. *In silico* analysis of ORF1ab in coronavirus HKU1 genome reveals a unique putative cleavage site of coronavirus HKU1 3C-like protease. *Microbiol Immunol*, 49(10):899-908 (Oct 2005). https://doi.org/10.1111/j.1348-0421.2005.tb03681.x
55. Ma Y, Wu Y, Shaw N, et al. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *PNAS*, 112(30):9436-41 (Jul 2015). https://doi.org/10.1073/pnas.1508686112
56. Neuman BW, Chamberlain P, Bowden F, et al. Atlas of coronavirus replicase structure. *Virus Res*, 194:49-66 (Dec 2014). https://doi.org/10.1016/j.virusres.2013.12.004
57. Fang SG, Shen H, Wang J, et al. Proteolytic processing of polyproteins 1a and 1ab between non-structural proteins 10 and 11/12 of *Coronavirus* infectious bronchitis virus is dispensable for viral replication in cultured cells. *Virol*, 379(2):175-80 (Sep 2008). https://doi.org/10.1016/j.virol.2008.06.038
58. Santos MS, Soares JP, Abreu PH, et al. Cross-validation for imbalances datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Comp Intell Mag*, 13(4):59-76 (Nov 2018). https://doi.org/10.1109/MCI.2018.2866730
59. Kozlowski LP. Proteome-pI: proteome isoelectric point database. *Nucleic Acids Res*, 45(Database issue):D1112-6 (Jan 2017). https://doi.org/10.1093/nar/gkw978
60. Ikemura T, Schwarze J, Makela M, et al. Type 4 phosphodiesterase inhibitors attenuate respiratory syncytial virus-induced airway hyper-responsiveness and lung eosinophilia. *J Pharmacol Exp Ther*, 294(2):701-6 (Aug 2000). https://pubmed.ncbi.nlm.nih.gov/10900250/
61. Mori I, Goshima F, Imai Y, et al. Olfactory receptor neurons prevent dissemination of neurovirulent influenza A virus into the brain by undergoing virus-induced apoptosis. *J Gen Virol*, 83(9):2109-16 (Sep 2002). https://doi.org/10.1099/0022-1317-83-9-2109
62. Thomander L, Aldskogius H, Vahlne A, et al. Invasion of cranial nerves and brain stem by herpes simplex virus inoculated into the mouse tongue. *Ann Ontol Rhinol Laryngol*, 97(5):554-8 (Sep 1988). https://doi.org/10.1177/000348948809700525
63. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* (Apr 2020). https://doi.org/10.1101/2020.03.22.002386

64. Lv X, Li Z, Guan J, et al. Porcine hemagluttinating encephalomyelitis virus activation of the integrin α5β1-FAK-cofilin pathway causes cytoskeletal rearrangement to promote its invasion of N2a cells. *J Virol*, 93(5):e01736-18 (Mar 2019). https://doi.org/10.1128/JVI.01736-18
65. Rudiger AT, Mayrhofer P, Ma-Lauer Y, et al. Tubulins interact with porcine and human S proteins of the genus *Alphacoronavirus* and support successful assembly and release of infectious viral particles. *Virol*, 497:185-97 (Oct 2016). https://doi.org/10.1016/j.virol.2016.07.022
66. Ohman T, Rintahaka J, Kalkkinen N, et al. Actin and RIG-I/MAVS signaling components translocate to mitochondria upon influenza A virus infection of human primary macrophages. *J Immunol*, 182(9):5682-92 (May 2009). https://doi.org/10.4049/jimmunol.0803093
67. Dohner K, Sodeik B. The role of the cytoskeleton during viral infection. *Curr Top Microbiol*, 285:67-108 (Feb 2005). https://doi.org/10.1007/3-540-26764-6_3
68. Naghavi MH, Walsh D. Microtubule regulation and function during virus infection. *J Virol*, 91(16):e00538-17 (Aug 2017). https://doi.org/10.1128/JVI.00538-17
69. Kristensson K, Lyche E, Roytta M, et al. Neuritic transport of herpes simplex virus in rat sensory neurons *in vitro*. Effects of substances interacting with microtubular function and axonal flow [nocodazole, taxol, and erythron-9-3-(2-hydroxynonyl)adenine]. *J Gen Virol*, 67:2023-8 (Sep 1986). https://doi.org/10.1099/0022-1317-67-9-2023
70. Solov'eva MF, Krispin TI, Shloma DV, et al. Effect of inhibitors that destroy cytoskeleton structures on the antiviral and antiproliferative activity of interferons. *Vopr Virusol* 33(3):309-14 (May-Jun 1988). https://pubmed.ncbi.nlm.nih.gov/2459850/
71. Yi F, Guo J, Dabbagh D, et al. Discovery of novel small-molecule inhibitors of LIM domain kinase for inhibiting HIV-1. *J Virol*, 91(13):e02418-16 (Apr 2017). https://doi.org/10.1128/JVI.02418-16
72. Campbell EM, Nunez R, Hope TJ. Disruption of the actin cytoskeleton can complement the ability of Nef to enhance HIV-1 infectivity. *J Virol*, 78(11): 5745-55 (Jun 2004). https://doi.org/10.1128/JVI.78.11.5745-5755.2004
73. Wolff G, Melia CE, Snijder EJ, et al. Double-membrane vesicles as platforms for viral replication. *Trends in Microbiol,* 1839 (Jun 2020). https://doi.org/10.1016/j.tim.2020.05.009
74. Neuman BW, Angelini MM, Buchmeier MJ. Does form meet function in the coronavirus replicative organelle? *Trends in Microbio*l, 22(11):642-7 (Nov 2014). https://doi.org/10.1016/j.tim.2014.06.003
75. Miller E, Antonny B, Hamamoto S, et al. Cargo selection into COPII vesicles is driven by the Sec24p subunit. *EMBO J*, 21(22):6105-13 (Nov 2002). https://doi.org/10.1093/emboj/cdf605
76. Mancias JD, Goldberg J. Structural basis of cargo membrane protein discrimination by the human COPII coat machinery. *EMBO J*, 27(21):2918-28 (Oct 2008). https://doi.org/10.1038/emboj.2008.208
77. Stagg SM, Gurkan C, Fowler DM, et al. Structure of the Sec13/31 COPII coat cage. *Nature*, 439:234-8 (Jan 2006). https://doi.org/10.1038/nature04339
78. Nagesh PT, Husain M. Influenza A virus dysregulates host histone deacetylase 1 that inhibits viral infection in lung epithelial cells. *J Virol*, 90(6):4614-25 (Apr 2016). https://doi.org/10.1128/JVI.00126-16
79. Chen H, Qian Y, Chen X, et al. HDAC6 restricts influenza A virus by deacetylation of the RNA polymerase PA subunit. *J Virol*, 93(4):e01896-18 (Feb 2019). https://doi.org/10.1128/JVI.01896-18
80. Shulak L, Beljanski V, Chiang C, et al. Histone deacetylase inhibitors potentiate vesicular stomatitis virus oncolysis in prostate cancer cells by modulating NF-κB-dependent autophagy. *J Virol*, 88(5):2927-40 (Feb 2014). https://doi.org/10.1128/JVI.03406-13
81. Feng Q, Su Z, Song S, et al. Histone deacetylase inhibitors suppress RSV infection and alleviate virus-induced airway inflammation. *Int J Mol Med*, 38(3):812-22 (Jul 2016). https://doi.org/10.3892/ijmm.2016.2691

82. Mosley AJ, Meekings KN, McCarthy C, et al. Histone deacetylase inhibitors increase virus gene expression but decrease CG8+ cell antiviral function in HTLV-1 infection. *Blood*, 108(12):3801-7 (Dec 2006). https://doi.org/10.1182/blood-2006-03-013235
83. Kaminskyy V, Zhivotovsky B. To kill of be killed: how viruses interact with the cell death machinery. *J Intern Med*, 267:473-82 (May 2010). https://doi.org/10.1111/j.1365-2796.2010.02222.x
84. Spencer CA, Kruhlak MJ, Jenkins HL, et al. Mitotic transcription repression *in vivo* in the absence of nucleosomal chromatin condensation. *J Cell Bio*, 150(1):13-26 (Jul 2000). https://doi.org/10.1083/jcb.150.1.13
85. Banerjee S, An S, Zhou A, et al. RNase L-independent specific 28S rRNA cleavage in murine coronavirus-infected cells. *J Virol*, 74(19):8793-802 (Oct 2000). https://doi.org/10.1128/jvi.74.19.8793-8802.2000
86. Slavov N, Semrau S, Airoldi E, et al. Differential stoichiometry among core ribosomal proteins. *Cell Rep*, 13(5):865-73 (Nov 2015). https://doi.org/10.1016/j.celrep.2015.09.056
87. Plant EP, Rakauskaite R, Taylor DR, et al. Achieving a golden mean: Mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. *J Virol*, 84(9):4330-40 (Apr 2019). https://doi.org/10.1128/JVI.02480-09
88. Siegal V, Walter P. Elongation arrest is not a prerequisite for secretory protein translocation across the microsomal membrane. *J Cell Biol*, 100(6):1913-1921 (Jun 1985). https://doi.org/10.1083/jcb.100.6.1913
89. Rottier P, Armstrong J, Meyer DI. Signal recognition particle-dependent insertion of coronavirus E1, an intracellular membrane glycoprotein. *J Biol Chem*, 260(8):4648-52 (Aug 1985). https://pubmed.ncbi.nlm.nih.gov/2985561/
90. Young JC, Andrews DW. The signal recognition particle receptor alpha subunit assembles co-translationally on the endoplasmic reticulum membrane during an mRNA-encoding translation pause *in vitro*. *EMBO J*, 15(1):172-81 (Jan 1996). https://doi.org/10.1002/j.1460-2075.1996.tb00345.x
91. Grandi N, Tramontano E. Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses. *Front Immunol*, 9(2039):1-16 (Sep 2019). https://doi.org/10.3389/fimmu.2018.02039
92. Roy M, Viginier B, Saint-Michel E, et al. Viral infection impacts transposable element transcript amounts in *Drosophila*. *PNAS*, 117(22):12249-57 (Jun 2020). https://doi.org/10.1073/pnas.2006106117
93. Wada M, Lokugamage KG, Nakagawa K, et al. Interplay between coronavirus, a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway. *PNAS*, 115(43):e10157-66 (Oct 2018). https://doi.org/10.1073/pnas.1811675115
94. Wang W, Xiog L, Wang P, et al. Major vault protein plays important roles in viral infection. *IUBMB Life*, 74(4):624-31 (Apr 2020). https://doi.org/10.1002/iub.2200
95. Steiner E, Holzmann K, Pirker C, et al. The major vault protein is responsive to and interferes with interferon-γ-mediated STAT1 signals. *J Cell Sci*, 119:459-69 (Oct 2006). https://doi.org/10.1242/jcs.02773
96. Li F, Chen Y, Zhang Z, et al. Robust expression of vault RNAs induced by influenza A virus plays a critical role in suppression of PKR-mediated innate immunity. *Nucleic Acids Res*, 43(21):10321-37 (Dec 2015). https://doi.org/10.1093/nar/gkv1078
97. Bellon M, Nicot C. Regulation of telomerase and telomeres: Human tumor viruses take control. *J Natl Cancer Inst*, 100(2):98-108 (Jan 2008). https://doi.org/10.1093/jnci/djm269
98. Reghunathan R, Jayapal M, Hsu LY, et al. Expression profile of immune response genes in patients with severe acute respiratory syndrome. *BMC Immunol*, 6:2 (Jan 2005). https://doi.org/10.1186/1471-2172-6-2

99. Berlutti F, Pantanella F, Natalizi T, et al. Antiviral properties of lactoferrin—A natural immunity molecule. *Molecules*, 16(8):6992-7018 (Aug 2011). https://doi.org/10.3390/molecules16086992
100. Adusumilli NC, Zhang D, Friedman JM, et al. Harnessing nitric oxide for preventing, limiting and treating the severe pulmonary consequences of COVID-19. *Nitric Oxide*, 103:4-8 (Oct 2020). https://doi.org/10.1016/j.niox.2020.07.003
101. Perrone LA, Belser JA, Wadford DA, et al. Inducible nitric oxide contributes to viral pathogenesis following highly pathogenic influenza virus infection in mice. *J Infect Dis*, 207(10):1576-84 (May 2013). https://doi.org/10.1093/infdis/jit062
102. Wan G, Zhaorigetu S, Liu Z, et al. Apolipoprotein L1, a novel Bcl-2 homology domain 3-only lipid-binding protein, induces autophagic cell death. *J Biol Chem*, 282(31):21540-9 (Aug 2008). https://doi.org/10.1074/jbc.M800214200
103. Peng Y, Wan L, Fan C, et al. Cholesterol metabolism—Impact for SARS-CoV-2 infection prognosis, entry, and antiviral therapies. *medrXiv* (Apr 2020). https://doi.org/10.1101/2020.04.16.20068528
104. Rebello CJ, Kirwan JP, Greenway FL. Obesity, the most common comorbidity in SARS-CoV-2: is leptin the link? *Int J Obes (Lond)*, (Jul 2020). https://doi.org/10.1038/s41366-020-0640-5
105. Zhang AJX, To KKW, Li Can, et al. Leptin mediates the pathogenesis of severe 2009 pandemic influenza A (H1N1) infection associated with cytokine dysregulation in mice with diet-induced obesity. *J Infect Dis*, 207(8):1270-80 (Apr 2013). https://doi.org/10.1093/infdis/jit031
106. Chesarino NM, McMichael TM, Yount JS. E3 ubiquitin ligase NEDD4 promotes influenza virus infection by decreasing levels of the antiviral protein IFITM3. *PLoS Pathog*, 11(8):e1005095 (Aug 2015). https://doi.org/10.1371/journal.ppat.1005095
107. Shi G, Ozog S, Torbett BE, et al. mTOR inhibitors lower an intrinsic barrier to virus infection mediated by IFITM3. *PNAS*, 115(43):e10069-78 (Oct 2018). https://doi.org/10.1073/pnas.1811892115
108. Xu Q, Zhu N, Chen S, et al. E3 ubiquitin ligase Nedd4 promotes Japanese encephalitis virus replication by suppressing autophagy in human neuroblastoma cells. *Sci Rep*, 7:45375 (Mar 2017). https://doi.org/10.1038/srep45375
109. Yang B, Kumar S. Nedd4 and Nedd4-2: closely related ubiquitin-protein ligases with distinct physiological functions. *Cell Death Differ*, 17:68-77 (Jun 2009). https://doi.org/10.1038/cdd.2009.84
110. Han Z, Lu J, Liu Y, et al. Small-molecule probes targeting the viral PPxY-host Nedd4 interface block egress of a broad range of RNA viruses. *J Virol*, 88(13):7294-306 (Apr 2014). https://doi.org/10.1128/JVI.00591-14
111. An H, Krist DT, Statsyuk AV. Crosstalk between kinases and Nedd4 family ubiquitin ligases. *Mol BioSyst*, 10:1643-57 (Jan 2014). https://doi.org/10.1039/C3MB70572B
112. Maaroufi H. SARS-CoV-2 encodes a PPxY late domain motif that is known to enhance budding and spread in enveloped RNA viruses. *biorXiv* (Apr 2020). https://doi.org/10.1101/2020.04.20.052217
113. Isaacson MK, Ploegh HL. Ubiquitination, ubiquitin-like modifiers, and deubiquitination in viral infection. *Cell Host Microbe*, 5:559-70 (Jun 2009). https://doi.org/10.1016/j.chom.2009.05.012
114. Zinngrebe J, Montinaro A, Peltzer N, et al. Ubiquitin in the immune system. *EMBO Rep*, 15(1):28-45 (Nov 2013). https://doi.org/10.1002/embr.201338025
115. Steimer L, Klostermeier D. RNA helicases in infection and disease. *RNA Biol*, 9(6):751-771 (Jun 2012). https://doi.org/10.4161/rna.20090
116. Sharma A, Boris-Lawrie K. Determination of host RNA helicases activity in viral replication. *Methods Enzymol*, 511:405-35 (Jun 2012). https://doi.org/10.1016/B978-0-12-396546-2.00019-X
117. Umate P, Tuteja N, Tuteja R. Genome-wide comprehensive analysis of human helicases. *Commun Integr Biol*, 4(1):118-37 (Jan-Feb 2011). https://doi.org/10.4161/cib.4.1.13844

118. Xu L, Khadijah S, Fang S, et al. The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. *J Virol*, 84(17):8571-83 (Sep 2010). https://doi.org/10.1128/JVI.00392-10
119. Chem JY, Chen WN, Poon KMV, et al. Interaction between SARS-CoV helicase and a multifunctional cellular protein (Ddx5) revealed by yeast and mammalian cell two-hybrid systems. *Arch Virol*, 154(3):507-12 (Feb 2009). https://doi.org/10.1007/s00705-009-0323-y
120. Spiezia L, Boscolo A, Poletto F, et al. COVID-19-related severe hypercoagulability in patients admitted to intensive care unit for acute respiratory failure. *Thromb Haemost* (Jun 2020). https://doi.org/10.1055/s-0040-1710018
121. Ji HL, Zhao R, Matalon S, et al. Elevated plasmin(ogen) as a common risk factor for COVID-19 susceptibility. *Physiol Rev*, 100(3):1065-75 (Jul 2020). https://doi.org/10.1152/physrev.00013.2020
122. Meyer M, Jaspers I. Respiratory protease/antiprotease balance determined susceptibility to viral infection and can be modified by nutritional antioxidants. *Am J Physiol Lung Cell Mol Physiol*, 308:L1189-201 (Apr 2015). https://doi.org/10.1152/ajplung.00028.2015
123. De Bont CM, Boelens WC, Pruijn GCM. NETosis, complement, and coagulation: a triangular relationship. *Cell Mol Immunol*, 16(1):19-27 (Jan 2019). https://doi.org/10.1038/s41423-018-0024-0
124. Noris M, Benigni A, Remuzzi G. The case of complement activation in COVID-19 multiorgan impact. *Kidney Intl* 0:0 (May 2020). https://doi.org/10.1016/j.kint.2020.05.013
125. Agrawal P, Nawadkar R, Ojha H, et al. Complement evasion strategies of viruses: An overview. *Front Microbiol*, 8:1117 (Jun 2017). https://doi.org/10.3389/fmicb.2017.01117
126. Eddie Ip WK, Chan KH, Law HKW, et al. Mannose binding lectin in severe acute respiratory syndrome coronavirus infection. *J Infect Dis*, 191(10):1697-704 (May 2005). https://doi.org/10.1086/429631
127. Narasuraju T, Tang BM, Herrmann M, et al. Neutrophilia and NETopathy as key pathologic drivers of progressive impairment in patients with COVID-19. *Front Pharmacol*, 11:870 (Jun 2020). https://doi.org/10.3389/fphar.2020.00870
128. Zou Y, Yalavarthi S, Shi H, et al. Neutrophil extracellular traps in COVID-19. *JCI Insight*, 5(11):e138999 (Apr 2020). https://doi.org/10.1172/jci.insight.138999
129. Stahel PF, Barnum SR. Complement inhibition in coronavirus disease (COVID)-19: A neglected therapeutic option. *Front Immunol*, 11:1661 (Jul 2020). https://doi.org/10.3389/fimmu.2020.01661
130. De Costa MG, Poppelaars F, van Kooten C, et al. Age and sex-associated changes of complement activity and complement levels in healthy Caucasian population. Front Immunol, 9:2664 (Nov 2018). https://doi.org/10.3389/fimmu.2018.02664
131. Khomich OA, Kochetkov SN, Bartosch B, et al. Redox biology of respiratory viral infections. *Viruses*, 10(8):392 (Aug 2018). https://doi.org/10.3390/v10080392
132. Schneider D, Ganesan S, Comstock AT, et al. Increased cytokine response of rhinovirus-infected airway epithelial cells in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 182(3):332-40 (Aug 2010). https://doi.org/10.1164/rccm.200911-1673OC

Supplementary Data

Table S1: Significant UP_TISSUE enrichments and depletions.

Enriched				Depleted			
Tissue	**FE**	**p-value**	**Benjamini**	**Tissue**	**FE**	**p-value**	**Benjamini**
Plasma	1.56	1.40E-06	1.64E-04	Hair root	1.30	1.22E-05	3.95E-03
Fetal kidney	1.55	1.58E-04	8.14E-03	Umbilical cord blood	1.17	6.56E-09	4.24E-06
Hepatoma	1.50	8.76E-05	5.83E-03	Cajal-Retzius cell	1.16	1.69E-05	3.63E-03
Epithelium	1.44	1.50E-37	7.01E-35				
Amygdala	1.32	3.39E-05	2.63E-03				
Teratocarcinoma	1.31	1.62E-04	7.55E-03				
Spleen	1.26	3.12E-05	2.91E-03				
Testis	1.21	1.22E-17	1.90E-15				
Brain	1.18	1.80E-30	4.21E-28				

Table S2: Significant UNIGENE_EST_QUARTILE enrichments and depletions.

Enriched				Depleted			
Tissue	**FE**	**p-value**	**Benjamini**	**Tissue**	**FE**	**p-value**	**Benjamini**
larynx_normal_3rd	1.35	1.50E-26	1.14E-24	salivary gland_normal_3rd	1.05	2.77E-06	2.10E-04
oral tumor_disease_3rd	1.35	8.45E-24	1.60E-22	neonate (< 4 weeks old)_development_3rd	1.03	7.94E-04	1.50E-02
pharynx_normal_3rd	1.33	2.62E-14	1.99E-13	non-glioma_disease_3rd	1.03	1.49E-04	5.65E-03
laryngeal cancer_disease_3rd	1.33	2.42E-24	6.14E-23	bone marrow_normal_3rd	1.03	5.42E-04	1.36E-02
tongue_normal_3rd	1.31	6.07E-26	2.31E-24	heart_normal_3rd	1.02	9.88E-04	1.49E-02
thyroid_normal_3rd	1.26	3.36E-19	5.11E-18	skin_normal_3rd	1.02	1.59E-03	1.99E-02
trachea_normal_3rd	1.26	3.56E-16	3.66E-15				
pharyngeal tumor_disease_3rd	1.25	1.93E-09	1.22E-08				
thyroid tumor_disease_3rd	1.23	1.11E-14	9.37E-14				
mammary gland_normal_3rd	1.22	8.77E-18	1.11E-16				
colorectal tumor_disease_3rd	1.22	1.07E-14	1.01E-13				
breast (mammary gland) cancer_disease_3rd	1.19	3.06E-11	2.11E-10				
adipose tissue_normal_3rd	1.16	4.51E-07	2.28E-06				
colon_normal_3rd	1.15	2.51E-09	1.47E-08				
uterine tumor_disease_3rd	1.13	5.07E-07	2.41E-06				
eye_normal_3rd	1.13	2.10E-08	1.14E-07				
muscle_normal_3rd	1.12	7.95E-06	3.36E-05				
lymph node_normal_3rd	1.12	4.02E-06	1.80E-05				
thymus_normal_3rd	1.11	2.26E-04	9.02E-04				
ear_normal_3rd	1.09	6.52E-03	2.05E-02				
pituitary gland_normal_3rd	1.09	4.07E-03	1.34E-02				
connective tissue_normal_3rd	1.08	8.06E-03	2.43E-02				
chondrosarcoma_disease_3rd	1.08	1.63E-03	5.90E-03				
testis_normal_3rd	1.07	6.92E-04	2.63E-03				

Table S3: Significant InterPro enrichments and depletions.

Enriched				Depleted			
Pfam	**FE**	**p-value**	**Benjamini**	**Pfam**	**FE**	**p-value**	**Benjamini**
Dynein heavy chain	4.50	1.95E-09	6.57E-07	High sulphur keratin-associated protein	1.29	1.64E-05	1.81E-02
Dynein heavy chain domain	4.50	1.95E-09	6.57E-07	Small GTP-binding protein domain	1.21	1.20E-07	2.35E-04
Dynein heavy chain, coiled coil stalk	4.50	1.95E-09	6.57E-07	Thioredoxin-like fold	1.20	5.50E-06	7.13E-03
Dynein heavy chain, domain-2	4.50	1.95E-09	6.57E-07	Olfactory receptor	1.20	1.35E-17	3.51E-14
Dynein heavy chain, P-loop containing D4 domain	4.50	8.26E-09	2.35E-06	GPCR, rhodopsin-like, 7TM	1.18	1.72E-22	1.35E-18
ATPase, dynein-related, AAA domain	4.50	3.48E-08	7.60E-06	G protein-coupled receptor, rhodopsin-like	1.17	8.42E-22	3.29E-18
Peptidase A2A, retrovirus RVP subgroup	4.50	1.04E-05	9.89E-04	Krueppel-associated box	1.13	6.74E-07	1.05E-03
Dynein heavy chain, domain-1	4.50	4.24E-05	3.41E-03				

Retroviral nucleocapsid protein Gag	4.50	4.24E-05	3.41E-03				
Beta-retroviral matrix, N-terminal	4.50	4.24E-05	3.41E-03				
PH-BEACH domain	4.50	1.71E-04	1.08E-02				
Na/K/Cl co-transporter superfamily	4.50	6.77E-04	3.62E-02				
Peptidase A2A, retrovirus, catalytic	4.09	4.59E-05	3.54E-03				
Retrovirus capsid, N-terminal core	4.05	1.70E-04	1.10E-02				
Myosin-like IQ motif-containing domain	4.03	1.64E-08	4.33E-06				
BEACH domain	4.00	6.20E-04	3.37E-02				
Retroviral envelope protein	4.00	6.20E-04	3.37E-02				
MyTH4 domain	4.00	6.20E-04	3.37E-02				
Myosin, N-terminal, SH3-like	3.90	3.24E-06	3.54E-04				
Myosin tail	3.79	2.35E-07	3.63E-05				
Spectrin/alpha-actinin	3.57	1.18E-09	4.38E-07				
Spectrin repeat	3.42	1.55E-07	2.61E-05				
Laminin, N-terminal	3.38	9.01E-05	6.53E-03				
Myosin head, motor domain	3.04	7.82E-09	2.42E-06				
Mitochondrial carrier protein	3.00	6.15E-06	6.16E-04				
Peptidase aspartic, active site	3.00	1.02E-04	6.99E-03				
G-patch domain	2.96	1.09E-06	1.39E-04				
SNF2-related	2.96	1.09E-06	1.39E-04				
Actinin-type, actin-binding, conserved site	2.94	7.14E-05	5.28E-03				
Cadherin, N-terminal	2.91	1.63E-12	6.71E-10				
Arf GTPase activating protein	2.85	8.68E-06	8.46E-04				
IQ motif, EF-hand binding site	2.85	1.34E-15	9.88E-13				
Mitochondrial carrier domain	2.63	5.71E-08	1.11E-05				
Mitochondrial substrate/solute carrier	2.63	5.71E-08	1.11E-05				
Bromodomain, conserved site	2.60	4.26E-04	2.40E-02				
Cadherin conserved site	2.59	1.97E-15	1.23E-12				
HECT	2.57	2.86E-04	1.67E-02				
Cadherin	2.52	7.39E-15	3.45E-12				
Cadherin-like	2.51	5.04E-15	2.65E-12				
Aspartic peptidase	2.50	7.04E-04	3.66E-02				
Laminin G domain	2.48	1.38E-07	2.44E-05				
Forkhead-associated (FHA) domain	2.43	9.60E-05	6.69E-03				
Kinesin, motor region, conserved site	2.42	4.36E-05	3.43E-03				
Dbl homology (DH) domain	2.41	3.34E-08	7.75E-06				
Bromodomain	2.41	2.94E-05	2.48E-03				
Kinesin, motor domain	2.40	1.99E-05	1.71E-03				
Armadillo-like helical	2.38	1.70E-23	3.15E-20				
Calponin homology domain	2.31	9.23E-08	1.71E-05				
Ubiquitin-associated/translation elongation factor EF1B, N-terminal, eukaryote	2.25	1.62E-04	1.07E-02				
GPS domain	2.25	7.84E-04	3.90E-02				
Rho GTPase activation protein	2.23	3.32E-08	8.21E-06				
EGF-like, laminin	2.19	7.78E-04	3.93E-02				
Zinc finger, PHD-finger	2.17	1.09E-06	1.44E-04				
Rho GTPase-activating protein domain	2.15	1.08E-05	1.00E-03				
Band 4.1 domain	2.12	2.33E-04	1.45E-02				
FERM domain	2.12	2.33E-04	1.45E-02				
FERM central domain	2.12	2.33E-04	1.45E-02				
Armadillo-type fold	2.09	7.25E-24	2.69E-20				
Pleckstrin homology domain	2.04	2.71E-17	2.51E-14				
WW domain	2.04	4.62E-04	2.56E-02				
Zinc finger, PHD-type	2.02	4.39E-06	4.65E-04				
Zinc finger, PHD-type, conserved site	2.02	9.39E-05	6.67E-03				
Zinc finger, FYVE/PHD-type	1.95	4.92E-08	1.01E-05				
Helicase, superfamily 1/2, ATP-binding domain	1.92	3.08E-06	3.45E-04				
Collagen triple helix repeat	1.92	7.03E-05	5.31E-03				
Pleckstrin homology-like domain	1.92	2.16E-21	2.67E-18				
Intermediate filament protein, conserved site	1.90	7.00E-04	3.69E-02				
Helicase, C-terminal	1.88	1.12E-05	1.01E-03				
Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2	1.86	2.52E-04	1.54E-02				
AAA+ ATPase domain	1.85	1.25E-06	1.55E-04				
PDZ domain	1.83	4.30E-07	6.14E-05				
Immunoglobulin I-set	1.82	1.99E-06	2.38E-04				
Galactose-binding domain-like	1.81	4.25E-04	2.43E-02				

Intermediate filament protein	1.79	8.40E-04	4.12E-02				
von Willebrand factor, type A	1.76	2.70E-04	1.60E-02				
Immunoglobulin E-set	1.73	2.56E-04	1.54E-02				
Src homology-3 domain	1.71	1.58E-07	2.55E-05				
Fibronectin, type III	1.63	5.57E-06	5.73E-04				
Ankyrin repeat-containing domain	1.60	9.58E-07	1.32E-04				
Ankyrin repeat	1.59	2.11E-06	2.44E-04				
Epidermal growth factor-like domain	1.56	1.78E-05	1.57E-03				
Immunoglobulin subtype 2	1.51	3.77E-05	3.10E-03				
EGF-like, conserved site	1.47	7.18E-04	3.68E-02				
P-loop containing nucleoside triphosphate hydrolase	1.33	2.45E-07	3.64E-05				
Immunoglobulin subtype	1.32	1.42E-04	9.53E-03				

Table S4: Significant GO CC enrichments and depletions.

Enriched				**Depleted**			
CC	**FE**	**p-value**	**Benjamini**	**CC**	**FE**	**p-value**	**Benjamini**
spectrin	4.55	3.94E-05	3.16E-03	ribosome	1.17	2.45E-05	1.76E-03
axonemal dynein complex	4.09	1.58E-04	7.50E-03	mitochondrion	1.05	8.74E-05	4.13E-02
spectrin-associated cytoskeleton	3.98	2.08E-03	4.83E-02	integral component of membrane	1.03	3.75E-08	5.43E-05
nuclear pore nuclear basket	3.79	1.36E-04	7.44E-03				
microtubule plus-end	3.74	2.65E-06	4.62E-04				
myosin filament	3.13	5.56E-04	1.98E-02				
costamere	3.11	1.31E-04	7.60E-03				
dynein complex	3.10	3.13E-05	2.97E-03				
viral capsid	3.03	1.61E-03	3.93E-02				
apicolateral plasma membrane	2.94	1.08E-03	3.38E-02				
nuclear periphery	2.81	4.91E-04	1.88E-02				
viral envelope	2.73	1.30E-03	3.52E-02				
desmosome	2.65	5.75E-04	1.98E-02				
myosin complex	2.46	3.16E-06	4.72E-04				
cell leading edge	2.22	6.89E-04	2.30E-02				
adherens junction	2.09	4.20E-04	1.67E-02				
kinesin complex	2.06	4.00E-04	1.73E-02				
nuclear pore	1.96	1.38E-04	7.17E-03				
basement membrane	1.90	1.58E-04	7.85E-03				
microtubule	1.89	1.55E-14	8.12E-12				
spindle pole	1.88	1.14E-05	1.49E-03				
growth cone	1.84	1.27E-05	1.47E-03				
centriole	1.77	7.59E-05	4.95E-03				
recycling endosome	1.69	4.03E-04	1.67E-02				
cytoskeleton	1.69	2.80E-11	9.74E-09				
PML body	1.67	1.30E-03	3.60E-02				
chromosome	1.65	1.27E-03	3.63E-02				
cell-cell junction	1.64	3.45E-05	3.00E-03				
dendritic spine	1.64	1.94E-03	4.60E-02				
midbody	1.62	5.24E-04	1.94E-02				
synapse	1.61	4.95E-05	3.69E-03				
centrosome	1.59	7.32E-10	1.91E-07				
axon	1.52	1.09E-04	6.70E-03				
microtubule organizing center	1.52	1.43E-03	3.75E-02				
actin cytoskeleton	1.50	1.87E-04	8.45E-03				
cytoplasmic vesicle	1.41	1.20E-03	3.61E-02				
apical plasma membrane	1.36	1.45E-03	3.73E-02				
cell-cell adherens junction	1.34	1.54E-03	3.84E-02				
protein complex	1.30	1.20E-03	3.52E-02				
cytoplasm	1.18	1.67E-15	1.74E-12				
nucleoplasm	1.17	2.69E-07	5.62E-05				
membrane	1.16	2.10E-05	2.19E-03				
cytosol	1.12	6.14E-05	4.27E-03				
nucleus	1.07	8.52E-04	2.74E-02				

Table S5: Significant GO BP enrichments and depletions.

Enriched				**Depleted**			
BP	**FE**	**p-value**	**Benjamini**	**BP**	**FE**	**p-value**	**Benjamini**

tRNA export from nucleus	2.63	3.69E-05	2.72E-02	detection of chemical stimulus involved in sensory perception of smell	1.20	6.97E-18	3.61E-14
microtubule-based movement	2.46	4.95E-10	1.11E-06	detection of chemical stimulus involved in sensory perception	1.25	1.37E-06	4.73E-03
homophilic cell adhesion via plasma membrane adhesion molecules	2.25	3.10E-14	2.09E-10	sensory perception of smell	1.18	1.24E-05	3.15E-02
regulation of Rho protein signal transduction	2.25	1.02E-07	1.38E-04	G-protein coupled receptor signaling pathway	1.15	8.98E-19	9.31E-15
single organismal cell-cell adhesion	1.98	2.05E-06	1.97E-03				
cytoskeleton organization	1.76	1.48E-06	1.66E-03				
regulation of small GTPase mediated signal transduction	1.72	3.44E-05	2.86E-02				
positive regulation of GTPase activity	1.56	4.00E-12	1.35E-08				
cell adhesion	1.52	8.97E-09	1.51E-05				

Table S6: Significant GO MF enrichments and depletions.

Enriched				Depleted			
MF	**FE**	**p-value**	**Benjamini**	**MF**	**FE**	**p-value**	**Benjamini**
microfilament motor activity	3.57	8.54E-07	1.58E-04	odorant binding	1.24	6.66E-06	8.48E-03
structural constituent of nuclear pore	2.97	1.14E-04	1.53E-02	olfactory receptor activity	1.20	7.19E-18	2.76E-14
nuclear localization sequence binding	2.71	7.24E-05	1.13E-02	G-protein coupled receptor activity	1.15	8.55E-15	1.64E-11
microtubule motor activity	2.68	2.40E-12	2.44E-09				
motor activity	2.63	6.78E-10	3.46E-07				
spectrin binding	2.57	4.74E-04	4.96E-02				
Rho guanyl-nucleotide exchange factor activity	2.43	3.59E-09	1.05E-06				
calmodulin binding	2.03	4.30E-12	2.92E-09				
ATPase activity	1.85	1.25E-08	2.55E-06				
microtubule binding	1.84	1.62E-09	5.52E-07				
structural constituent of cytoskeleton	1.74	1.32E-04	1.67E-02				
guanyl-nucleotide exchange factor activity	1.74	8.04E-05	1.16E-02				
actin binding	1.68	8.75E-09	2.23E-06				
GTPase activator activity	1.68	1.09E-08	2.47E-06				
protein kinase binding	1.36	2.03E-04	2.41E-02				
chromatin binding	1.35	3.00E-04	3.35E-02				
ATP binding	1.34	1.15E-12	2.34E-09				
calcium ion binding	1.32	4.90E-06	8.32E-04				
protein binding	1.08	1.08E-09	4.42E-07				

Table S7: Significant Reactome pathway enrichments and depletions.

Enriched				Depleted			
Pathway	**FE**	**p-value**	**Benjamini**	**Pathway**	**FE**	**p-value**	**Benjamini**
Cation-coupled Chloride cotransporters	4.68	5.39E-04	2.80E-02	Peptide chain elongation	1.20	1.05E-04	3.80E-02
Anchoring fibril formation	4.06	2.07E-06	7.64E-04	Viral mRNA Translation	1.20	1.05E-04	3.80E-02
Extracellular matrix organization	3.43	2.00E-04	1.29E-02	Formation of a pool of free 40S subunits	1.20	4.62E-05	2.24E-02
Non-integrin membrane-ECM interactions	3.04	2.05E-08	2.27E-05	G alpha (i) signalling events	1.19	3.90E-10	2.88E-07
Laminin interactions	2.81	2.51E-05	3.08E-03	Olfactory Signaling Pathway	1.19	1.05E-16	1.64E-13
Pre-NOTCH Transcription and Translation	2.74	6.79E-05	7.49E-03				
NS1 Mediated Effects on Host Pathways	2.66	1.27E-05	3.50E-03				
Regulation of Glucokinase by Glucokinase Regulatory Protein	2.57	1.94E-04	1.42E-02				
Nuclear import of Rev protein	2.55	1.26E-04	1.07E-02				

Rev-mediated nuclear export of HIV RNA	2.48	2.03E-04	1.24E-02				
Vpr-mediated nuclear import of PICs	2.41	4.86E-04	2.65E-02				
Nuclear Pore Complex (NPC) Disassembly	2.41	3.16E-04	1.82E-02				
Assembly of collagen fibrils and other multimeric structures	2.24	1.95E-04	1.34E-02				
Collagen biosynthesis and modifying enzymes	2.17	1.43E-05	2.63E-03				
Loss of Nlp from mitotic centrosomes	2.14	1.36E-05	3.00E-03				
SUMOylation of RNA binding proteins	2.09	8.45E-04	3.98E-02				
Anchoring of the basal body to the plasma membrane	2.05	1.24E-06	6.85E-04				
Recruitment of mitotic centrosome proteins and complexes	2.04	2.32E-05	3.66E-03				
Regulation of PLK1 Activity at G2/M Transition	1.97	2.44E-05	3.37E-03				
SUMOylation of DNA damage response and repair proteins	1.95	1.26E-04	1.15E-02				
ECM proteoglycans	1.93	1.84E-04	1.44E-02				
Regulation of HSF1-mediated heat shock response	1.87	7.48E-04	3.69E-02				
Rho GTPase cycle	1.74	8.86E-05	8.87E-03				

Table S8: Significant sequence feature enrichments and depletions.

Enriched				**Depleted**			
Seq Feature	**FE**	**p-value**	**Benjamini**	**Seq Feature**	**FE**	**p-value**	**Benjamini**
region of interest:AAA 4	4.67	2.17E-08	1.07E-05	disulfide bond	1.06	6.61E-11	1.42E-06
repeat:Spectrin 5	4.67	2.17E-08	1.07E-05	transmembrane region	1.03	3.45E-06	3.64E-02
region of interest:AAA 3	4.67	2.17E-08	1.07E-05				
region of interest:Stem	4.67	2.17E-08	1.07E-05				
region of interest:Stalk	4.67	2.17E-08	1.07E-05				
region of interest:AAA 2	4.67	2.17E-08	1.07E-05				
region of interest:AAA 1	4.67	2.17E-08	1.07E-05				
region of interest:AAA 5	4.67	2.17E-08	1.07E-05				
repeat:Spectrin 7	4.67	9.45E-08	3.41E-05				
region of interest:AAA 6	4.67	9.45E-08	3.41E-05				
repeat:Spectrin 8	4.67	9.45E-08	3.41E-05				
repeat:Spectrin 6	4.67	9.45E-08	3.41E-05				
repeat:Spectrin 9	4.67	9.45E-08	3.41E-05				
repeat:Spectrin 17	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 13	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 16	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 15	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 11	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 10	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 14	4.67	4.09E-07	1.32E-04				
repeat:Spectrin 12	4.67	4.09E-07	1.32E-04				
region of interest:5 X 4 AA repeats of P-X-X-P	4.67	3.18E-05	4.80E-03				
domain:Chromo 2	4.67	3.18E-05	4.80E-03				
repeat:Spectrin 18	4.67	3.18E-05	4.80E-03				
repeat:Spectrin 20	4.67	1.33E-04	1.39E-02				
domain:BEACH	4.67	1.33E-04	1.39E-02				
repeat:Spectrin 19	4.67	1.33E-04	1.39E-02				
repeat:Spectrin 21	4.67	5.46E-04	4.36E-02				
domain:Laminin EGF-like 10	4.25	3.34E-05	4.96E-03				
domain:Laminin EGF-like 8	4.25	3.34E-05	4.96E-03				
domain:Laminin EGF-like 9	4.21	1.29E-04	1.36E-02				
domain:Laminin G-like 5	4.16	4.86E-04	4.03E-02				
domain:Laminin EGF-like 11	4.16	4.86E-04	4.03E-02				
repeat:PXXP 4	4.05	2.13E-06	5.14E-04				
repeat:PXXP 3	4.05	2.13E-06	5.14E-04				
repeat:PXXP 2	4.05	2.13E-06	5.14E-04				
repeat:PXXP 5	4.05	2.13E-06	5.14E-04				
repeat:PXXP 1	4.05	2.13E-06	5.14E-04				

repeat:Spectrin 4	4.01	9.75E-09	5.08E-06				
domain:Cadherin 9	3.96	2.98E-05	4.65E-03				
domain:Cadherin 8	3.96	2.98E-05	4.65E-03				
domain:Laminin EGF-like 6	3.96	2.98E-05	4.65E-03				
repeat:ANK 16	3.96	2.98E-05	4.65E-03				
repeat:ANK 18	3.90	1.08E-04	1.19E-02				
domain:Laminin EGF-like 7	3.90	1.08E-04	1.19E-02				
repeat:ANK 19	3.90	1.08E-04	1.19E-02				
repeat:ANK 17	3.90	1.08E-04	1.19E-02				
repeat:Spectrin 3	3.82	3.43E-08	1.53E-05				
repeat:ANK 20	3.82	3.83E-04	3.27E-02				
repeat:ANK 21	3.82	3.83E-04	3.27E-02				
domain:IPT/TIG 1	3.82	3.83E-04	3.27E-02				
repeat:Spectrin 1	3.78	2.29E-09	1.34E-06				
repeat:Spectrin 2	3.78	2.29E-09	1.34E-06				
domain:Cadherin 7	3.69	1.58E-06	3.91E-04				
domain:IPT/TIG 2	3.60	2.84E-04	2.56E-02				
domain:Laminin EGF-like 4	3.60	2.84E-04	2.56E-02				
domain:IPT/TIG 3	3.60	2.84E-04	2.56E-02				
domain:Actin-binding	3.51	4.23E-06	8.83E-04				
domain:Laminin N-terminal	3.51	6.26E-05	7.70E-03				
repeat:ANK 14	3.43	2.04E-04	1.93E-02				
repeat:ANK 13	3.43	2.04E-04	1.93E-02				
repeat:ANK 15	3.43	2.04E-04	1.93E-02				
domain:Laminin G-like 4	3.43	2.04E-04	1.93E-02				
region of interest:Actin-binding	3.38	5.06E-08	1.98E-05				
domain:Importin N-terminal	3.34	6.44E-04	4.92E-02				
domain:IQ 3	3.19	2.25E-05	3.63E-03				
short sequence motif:LXXLL motif 2	3.12	6.78E-05	8.12E-03				
short sequence motif:LXXLL motif 1	3.12	6.78E-05	8.12E-03				
region of interest:Triple-helical region	3.05	4.61E-05	6.35E-03				
domain:Cadherin 6	3.04	6.71E-17	1.30E-13				
domain:Laminin G-like 3	3.04	1.99E-04	1.93E-02				
domain:IQ 1	2.97	1.18E-06	3.07E-04				
domain:IQ 2	2.97	1.18E-06	3.07E-04				
domain:Laminin G-like 2	2.96	4.98E-06	1.02E-03				
domain:Laminin G-like 1	2.96	4.98E-06	1.02E-03				
domain:Arf-GAP	2.96	4.98E-06	1.02E-03				
domain:Myosin head-like	2.91	5.47E-07	1.66E-04				
repeat:Solcar 3	2.84	6.81E-09	3.76E-06				
zinc finger region:PHD-type 2	2.80	2.62E-05	4.15E-03				
domain:CH 2	2.77	1.07E-04	1.19E-02				
domain:CH 1	2.77	1.07E-04	1.19E-02				
short sequence motif:DEAH box	2.72	8.44E-07	2.40E-04				
domain:Cadherin 4	2.71	1.29E-16	1.16E-13				
domain:Cadherin 3	2.71	1.29E-16	1.16E-13				
zinc finger region:PHD-type 1	2.69	3.06E-05	4.69E-03				
repeat:Solcar 1	2.68	4.20E-08	1.71E-05				
repeat:Solcar 2	2.68	4.20E-08	1.71E-05				
domain:ABC transporter 2	2.67	2.02E-05	3.44E-03				
domain:ABC transporter 1	2.67	2.02E-05	3.44E-03				
domain:HECT	2.67	1.85E-04	1.83E-02				
domain:Cadherin 1	2.64	9.71E-16	8.53E-13				
domain:Cadherin 2	2.64	9.71E-16	8.53E-13				
domain:Cadherin 5	2.63	3.25E-14	2.54E-11				
domain:Laminin EGF-like 2	2.60	4.72E-04	3.95E-02				
domain:PH 1	2.58	2.24E-05	3.69E-03				
domain:Bromo	2.58	3.07E-04	2.71E-02				
domain:SH3 2	2.57	1.48E-05	2.61E-03				
compositionally biased region:Gln-rich	2.55	9.21E-19	1.44E-15				
domain:Kinesin-motor	2.55	6.40E-06	1.25E-03				
domain:FHA	2.54	8.57E-05	1.00E-02				
domain:PH 2	2.53	5.62E-05	7.20E-03				
domain:SH3 1	2.52	3.68E-05	5.30E-03				
nucleotide phosphate-binding region:ATP 2	2.46	8.99E-05	1.02E-02				

nucleotide phosphate-binding region:ATP 1	2.46	8.99E-05	1.02E-02				
domain:DH	2.41	1.71E-07	5.94E-05				
repeat:ANK 9	2.39	9.20E-05	1.03E-02				
domain:CH	2.34	1.40E-04	1.44E-02				
domain:GPS	2.34	4.98E-04	4.05E-02				
domain:IQ	2.30	4.99E-06	9.96E-04				
repeat:ANK 8	2.24	2.04E-04	1.95E-02				
domain:Fibronectin type-III 4	2.23	1.10E-05	1.99E-03				
domain:Rho-GAP	2.20	7.10E-06	1.36E-03				
repeat:HEAT 1	2.18	5.46E-05	7.09E-03				
repeat:HEAT 2	2.18	5.46E-05	7.09E-03				
domain:Fibronectin type-III 3	2.17	8.68E-07	2.40E-04				
compositionally biased region:Poly-Lys	2.16	7.38E-16	7.29E-13				
domain:Fibronectin type-III 5	2.13	6.29E-04	4.92E-02				
compositionally biased region:Poly-Ser	2.12	3.65E-34	1.14E-30				
compositionally biased region:Thr-rich	2.11	3.92E-04	3.32E-02				
domain:FERM	2.10	5.87E-04	4.64E-02				
compositionally biased region:Ser-rich	2.09	3.83E-28	8.98E-25				
region of interest:Tail	2.01	3.44E-05	5.03E-03				
repeat:ANK 7	2.00	2.85E-04	2.54E-02				
repeat:LRR 11	1.99	3.60E-06	7.85E-04				
domain:PDZ	1.94	2.56E-06	5.72E-04				
region of interest:Head	1.94	1.34E-04	1.39E-02				
compositionally biased region:His-rich	1.94	3.16E-04	2.76E-02				
repeat:LRR 10	1.93	2.34E-06	5.36E-04				
domain:Fibronectin type-III 2	1.93	3.96E-07	1.33E-04				
repeat:LRR 13	1.93	1.96E-04	1.92E-02				
domain:Fibronectin type-III 1	1.92	5.27E-07	1.65E-04				
domain:PH	1.91	1.00E-11	6.71E-09				
compositionally biased region:Poly-Leu	1.90	6.97E-07	2.04E-04				
region of interest:Rod	1.87	4.88E-04	4.01E-02				
domain:Ig-like C2-type 4	1.85	6.30E-04	4.89E-02				
repeat:ANK 6	1.84	8.53E-05	1.01E-02				
domain:EGF-like 2	1.84	2.19E-04	2.03E-02				
domain:Helicase C-terminal	1.82	6.14E-05	7.66E-03				
short sequence motif:Cell attachment site	1.80	3.56E-04	3.08E-02				
compositionally biased region:Poly-Asp	1.79	5.06E-04	4.08E-02				
repeat:LRR 9	1.78	1.76E-05	3.06E-03				
repeat:LRR 12	1.77	6.37E-04	4.90E-02				
domain:Helicase ATP-binding	1.76	1.12E-04	1.21E-02				
domain:Ig-like C2-type 3	1.75	4.96E-05	6.63E-03				
domain:EGF-like 1	1.75	8.83E-05	1.02E-02				
repeat:ANK 5	1.74	2.21E-05	3.70E-03				
compositionally biased region:Poly-Glu	1.74	2.09E-17	2.81E-14				
repeat:ANK 1	1.72	4.13E-08	1.76E-05				
repeat:ANK 2	1.72	5.14E-08	1.93E-05				
repeat:LRR 8	1.71	3.77E-05	5.35E-03				
repeat:ANK 4	1.70	9.53E-06	1.75E-03				
repeat:LRR 6	1.70	9.62E-07	2.58E-04				
repeat:LRR 7	1.70	7.51E-06	1.41E-03				
repeat:ANK 3	1.69	1.34E-06	3.39E-04				
compositionally biased region:Glu-rich	1.66	2.78E-08	1.30E-05				
repeat:TPR 3	1.65	1.17E-04	1.26E-02				
repeat:LRR 5	1.64	2.15E-06	5.05E-04				
domain:SH3	1.63	5.30E-05	6.99E-03				
compositionally biased region:Poly-Gln	1.62	1.83E-04	1.83E-02				
compositionally biased region:Pro-rich	1.62	1.41E-22	2.64E-19				
compositionally biased region:Poly-Arg	1.59	5.91E-05	7.47E-03				

compositionally biased region:Poly-Pro	1.59	2.28E-09	1.42E-06				
domain:Ig-like C2-type 1	1.53	2.23E-04	2.05E-02				
domain:Ig-like C2-type 2	1.52	2.60E-04	2.37E-02				
repeat:LRR 4	1.50	6.72E-05	8.16E-03				
compositionally biased region:Poly-Ala	1.46	3.77E-06	8.04E-04				
nucleotide phosphate-binding region:ATP	1.46	1.51E-13	1.09E-10				
repeat:LRR 1	1.46	4.23E-05	5.90E-03				
repeat:LRR 2	1.45	4.84E-05	6.56E-03				
compositionally biased region:Poly-Gly	1.44	1.58E-04	1.60E-02				
repeat:LRR 3	1.42	2.09E-04	1.96E-02				
splice variant	1.30	1.82E-68	1.71E-64				
sequence variant	1.18	2.22E-68	1.04E-64				

Table S9: Significant keyword enrichments and depletions.

Enriched				**Depleted**			
Keyword	**FE**	**p-value**	**Benjamini**	**Keyword**	**FE**	**p-value**	**Benjamini**
Ribosomal frameshifting	3.61	5.90E-06	1.19E-04	Redox-active center	1.26	5.44E-04	2.96E-02
Thick filament	3.55	2.00E-05	3.39E-04	Antibiotic	1.20	6.55E-04	3.30E-02
Dynein	3.10	2.15E-07	4.98E-06	Olfaction	1.19	3.87E-17	9.25E-15
Aspartyl protease	3.10	7.24E-05	1.08E-03	Ribosomal protein	1.19	1.68E-07	1.72E-05
Viral envelope protein	2.93	6.00E-04	6.46E-03	G-protein coupled receptor	1.14	1.37E-17	4.92E-15
Laminin EGF-like domain	2.70	5.03E-05	8.07E-04	Transducer	1.14	2.56E-18	1.84E-15
Bromodomain	2.62	9.69E-06	1.89E-04	Sensory transduction	1.13	5.84E-11	1.05E-08
Autism	2.60	7.71E-04	7.88E-03	Palmitate	1.12	2.64E-05	1.72E-03
Motor protein	2.58	1.93E-17	1.72E-15	Ribonucleoprotein	1.10	5.39E-04	3.17E-02
Transposable element	2.56	3.29E-04	3.81E-03	Lipoprotein	1.10	3.30E-09	4.73E-07
Basement membrane	2.56	1.63E-05	2.83E-04	Receptor	1.06	2.31E-06	1.65E-04
ERV	2.49	8.08E-04	8.13E-03	Mitochondrion	1.05	7.81E-04	3.67E-02
Myosin	2.46	3.35E-06	6.99E-05	Disulfide bond	1.04	2.70E-07	2.15E-05
Calmodulin-binding	2.26	3.48E-13	1.67E-11	Transmembrane	1.03	9.41E-08	1.12E-05
Triplet repeat expansion	2.17	6.00E-03	4.77E-02	Transmembrane helix	1.03	1.93E-07	1.73E-05
Autism spectrum disorder	2.09	2.28E-03	1.89E-02				
Guanine-nucleotide releasing factor	2.09	3.82E-10	1.50E-08				
Nuclear pore complex	2.04	8.79E-04	8.70E-03				
Autocatalytic cleavage	2.04	3.70E-04	4.21E-03				
Microtubule	2.01	6.19E-16	4.63E-14				
Intermediate filament	1.93	1.51E-04	2.09E-03				
Actin-binding	1.90	7.05E-13	3.15E-11				
GTPase activation	1.86	6.04E-09	1.99E-07				
Hydroxylation	1.81	1.98E-04	2.58E-03				
Collagen	1.80	1.78E-04	2.42E-03				
SH3 domain	1.79	1.88E-08	5.35E-07				
Cell adhesion	1.78	4.28E-17	3.35E-15				
Tight junction	1.73	1.40E-03	1.27E-02				
Helicase	1.72	3.28E-05	5.40E-04				
ANK repeat	1.67	8.97E-08	2.25E-06				
mRNA transport	1.67	1.03E-03	9.84E-03				
Coiled coil	1.67	5.05E-87	1.58E-84				
Calcium transport	1.66	1.80E-03	1.56E-02				
Cytoskeleton	1.66	1.22E-29	1.53E-27				
Nucleotidyltransferase	1.65	6.37E-03	4.88E-02				
Chromosomal rearrangement	1.63	1.49E-08	4.44E-07				
Chromatin regulator	1.62	2.39E-07	5.33E-06				
TPR repeat	1.61	1.10E-04	1.56E-03				
Extracellular matrix	1.60	1.85E-06	4.00E-05				
Cell projection	1.60	6.71E-16	4.17E-14				
Ciliopathy	1.59	1.40E-03	1.28E-02				
Cilium biogenesis/degradation	1.57	7.04E-04	7.45E-03				
Biological rhythms	1.56	2.23E-03	1.87E-02				
Cilium	1.55	9.77E-05	1.42E-03				
Endocytosis	1.54	3.96E-03	3.22E-02				
Mental retardation	1.51	1.20E-05	2.28E-04				
Cell junction	1.47	7.52E-10	2.62E-08				

EGF-like domain	1.44	5.73E-04	6.39E-03				
Proto-oncogene	1.44	7.35E-04	7.64E-03				
Calcium	1.40	7.26E-10	2.67E-08				
ATP-binding	1.40	7.04E-15	3.65E-13				
DNA repair	1.40	5.87E-04	6.43E-03				
DNA damage	1.39	2.25E-04	2.75E-03				
Deafness	1.38	6.01E-03	4.73E-02				
Mitosis	1.38	1.60E-03	1.42E-02				
Activator	1.33	1.20E-05	2.21E-04				
Phosphoprotein	1.33	1.13E-88	7.06E-86				
Isopeptide bond	1.33	9.52E-09	2.98E-07				
Ubl conjugation	1.32	2.48E-12	1.04E-10				
Repressor	1.32	5.82E-05	9.11E-04				
Methylation	1.32	1.42E-07	3.43E-06				
Protein transport	1.31	6.82E-05	1.04E-03				
Cell division	1.31	1.77E-03	1.55E-02				
Disease mutation	1.29	2.84E-15	1.64E-13				
Cell cycle	1.28	1.87E-04	2.49E-03				
Immunoglobulin domain	1.26	1.38E-03	1.28E-02				
Nucleotide-binding	1.24	4.55E-08	1.19E-06				
Cytoplasm	1.24	4.82E-23	5.03E-21				
Differentiation	1.23	1.83E-03	1.56E-02				
Alternative splicing	1.22	7.27E-68	1.14E-65				
Polymorphism	1.20	1.83E-76	3.83E-74				
Developmental protein	1.20	9.18E-04	8.95E-03				
Zinc-finger	1.19	1.58E-05	2.82E-04				
Transport	1.15	2.25E-04	2.81E-03				
Transcription regulation	1.13	2.64E-04	3.18E-03				
Transcription	1.13	2.22E-04	2.83E-03				
Nucleus	1.12	4.01E-08	1.09E-06				
Zinc	1.12	1.07E-03	1.01E-02				
Metal-binding	1.10	3.11E-04	3.67E-03				
Acetylation	1.08	6.24E-03	4.84E-02				