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, machine learning, neural networks

Abstract

A novel coronavirus (SARS-CoV-2) has devastated the globe as a pandemic that has killed more than 1,600,000 people. Widespread vaccination is still uncertain, so many scientific efforts have been directed toward 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, I set out to determine this protease's targets and corresponding potential drug targets. Using a neural network trained on cleavages from 388 coronavirus proteomes with a Matthews correlation coefficient of 0.983, I 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 putative 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, ribosomal stoichiometry and nascent-chain detection and degradation, coagulation, pattern recognition receptors, growth factors, lipoproteins, redox, ubiquitination, and apoptosis. This whole proteome cleavage prediction demonstrates the importance of 3CLpro in expected and nontrivial pathways affecting virulence, lead me to propose more than a dozen potential therapeutic targets against coronaviruses, and should therefore be applied to all viral proteases and subsequently 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 human 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.

Angiotensin-converting enzyme 2 (ACE2), the main 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.[2-5]

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[16-18] 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,[19] 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.[20-23] 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 human immunodeficiency virus (HIV).

PLpro is a highly conserved protein domain that has been shown to determine virulence of coronaviruses[24] and possess deubiquinating and deISGylating activity including cleaving interferon-stimulated gene 15 (ISG15) induced by interferon via the Janus kinases and signal transducer and activator of transcription proteins (JAK-STAT) pathway from ubiquitin-conjugating enzymes and potentially from downstream effectors.[25-29] PLpro deubiquination also prevents activating phosphorylation of interferon regulatory factor 3 (IRF3) and subsequent type-I interferon production,[30, 31] 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.[32, 33]

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.[34-36] 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[37], NF-kappa-B essential modulator (NEMO)[38], the nucleotide-binding oligomerization domain (NOD)-like receptor NLRP12, and TGF-beta activated kinase 1 (TAB1)[39] to modulate interferon signaling. Similar proteases have been studied in the other members of *Nidovirales*[40] and the related *Picornavirales*[41-45], with foot-and-mouth disease virus (FMDV) 3Cpro cleaving histone H3,[46, 47] poliovirus 3Cpro cleaving TFIID and TFIIIC,[48-52] and polio- and rhinovirus but not cardiovirus 3Cpro cleaving microtubule-associated protein 4 (MAP4).[53, 54] These results, however, have not been reproduced for SARS-CoV-2 yet, and STAT2, NEMO, NLRP12, TAB1, H3, TFIIIC, TFIID, and MAP4 are only a few of many cleaved proteins.

The high number of 3CLpro cleavages in coronavirus polyproteins has, however, allowed for sequence logos and resulting sequence rules and training of decision trees and neural networks (NN) for additional cleavage site prediction.[55-60] Notably, Kiemer et al.'s NN[59] based on Blom et al.'s equivalent picornaviral NN[60] was trained on 7 arbitrary coronavirus genomes, totaling 77 cleavages, 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 polymerase I and III subunits (RPA1 and RPC1), eukaryotic translation initiation factor 4 gamma 1 (eIF4G1), the cytoskeletal proteins MAP4 and microtubule-associated protein RP/EB members 1 and 3 (MAPRE1/3), and many members of the ubiquitin pathway (ubiquitin hydrolases USP1/4/5/9X/9Y/13/26 and suppressor of cytokine signaling 6 (SOCS6)).

Additionally, Yang's decision trees[58] were trained on 4 amino acid sliding windows and substitution matrix similarity score-based embeddings, achieved MCCs up to 0.95, but were limited to only 18 coronavirus polyproteins. The embedding-derived non-orthogonality somewhat stabilized the prediction to small changes in sequence assuming the substitution matrix reflects how the cleavages evolve. Decision trees have the benefit of being symbolic and explainable but often predict suboptimally when presented with interpolated or extrapolated inputs, making alternative machine learning techniques more attractive for predicting human protein cleavage prediction. For example, Narayanan et al.[61] and later Singh et al.[62] demonstrated that neural networks outperform decision trees for HIV and hepatitis C virus (HCV) protease cleavage prediction. Additional mixed methods such as Li et al.'s nonlinear dimensionality reduction and subsequent support vector machine (SVM) are able to retain some of the benefits of both linear and nonlinear classifiers.[63] Rognvaldsson et al.[64, 65] argue that nonlinear models including neural networks should not be used for cleavage prediction, however the HIV dataset from Cai et al.[66] that they used and their expanded dataset only included 299 and 746 samples, respectively. Additionally, physiochemical or structural encodings have outperformed one-hot encoding (also called orthogonal encoding) for their small HIV datasets[67] and have moreover eliminated differences between linear and nonlinear classifiers in an equivalent HCV dataset with 891 samples.[68] To my knowledge no one has expanded the 3CLpro cleavage dataset to the point where nonlinearity becomes significant, investigated the entire human proteome for 3CLpro cleavages sites with any method, or 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).[69]

Additional coronavirus polyprotein cleavages were collected from GenBank.[70] Searching for "orf1ab," "pp1ab," and "1ab" within the family *Coronaviridae* returned 388 different, complete polyproteins with 762 different cleavages manually discovered using the Clustal Omega multiple sequence alignment server.[71-73] All 4,268 balanced positive cleavages were used for subsequent classifier training in addition to all other uncleaved coronavirus sequence windows centered at glutamines (17,493) and histidines (11,421), totaling 33,182 samples.

Cleavage Prediction

The NetCorona 1.0 server as in Kiemer et al.'s work[59], my reproductions of their sequence logo-derived rules and NN, and my improved sequence logo-based logistic regression and naïve Bayes classification and NNs were used for prediction of cleavage sites.[74] 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 omitted because although they may be within important localization sequences, their cleavage kinetics are likely significantly retarded by truncation.

Enrichment Analysis

Protein annotation, classification, and enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) 6.8.[75, 76] My training data, prediction methods, and results can be found on GitHub (https://github.com/Luke8472NN/NetProtease).

Results

Here I assumed that SARS-CoV-2 3CLpro is capable of cleaving all aligned cleavages between the four genera of coronaviruses (*Alpha-*, *Beta-*, *Gamma-*, and *Delta-*) because variation in cleavage sequences is greater within polyproteins than between them (Figure 1) no matter the existence of protease/cleavage cophylogeny (Figure 2).[77]

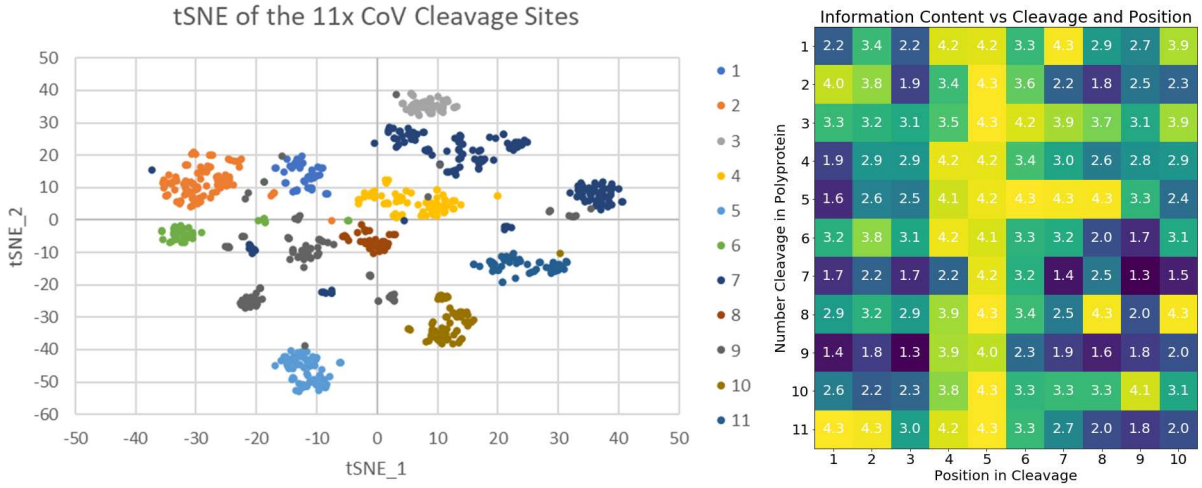

Figure 1: One-hot encoded t-distributed stochastic neighbor embedding (t-SNE)[78] and information content both demonstrate that cleavage variation within genomes is more important than variation between genomes.

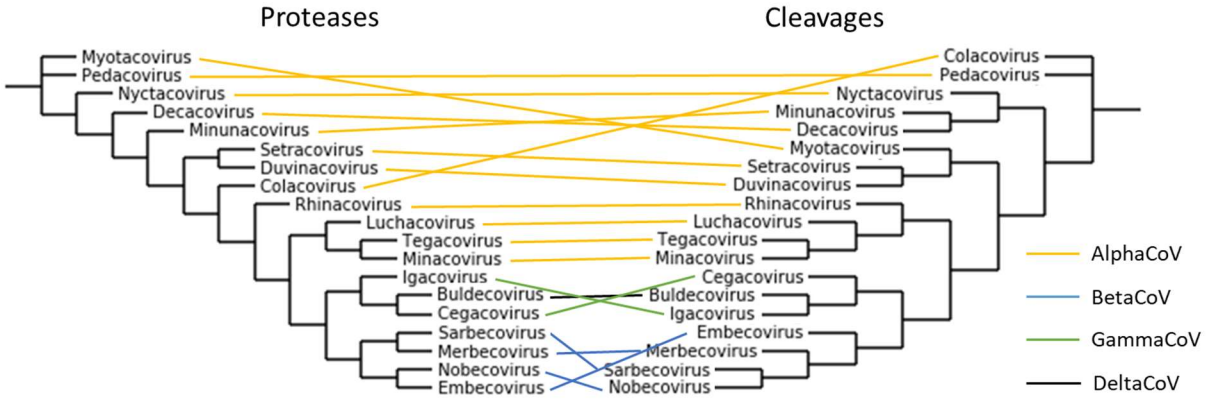

Figure 2: Unscaled subgenera-averaged tanglegram of 3CLpro and respective cleavages based on BLOSUM62 substitution matrix similarity scores with and without default affine gap penalties (opening 10 and extension 0.2).

Kiemer et al.'s seven genome sequence logo and multilayer perceptron 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.[59] First, logistic regression was performed on the logit of the probability output of the sequence logo (as opposed to Chou et al.'s manual probability cutoff setting by maximizing an unbalanced measure of accuracy[79]) with a nonzero but optimally extremely small pseudocount and returned an MCC of 0.825 with 74.0% recall. Updating the sequence logo with all known cleavages improved its MCC to 0.936 with 94.8% recall (Figure 3). A naïve Bayes classifier was additionally constructed from both the positive and negative sequence logos and slightly improved the MCC to 0.947 with 95.7% recall. Figure 4 demonstrates correlations (represented as the

mutual information variant known as total entropy correlation coefficients or symmetric uncertainties) between positions that are not captured by simple sequence logos and classifiers assuming independence.[80, 81] 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 reaffirming that one-hot encoding performs better than any physiochemical, lower-dimensional inputs when the training set is large enough.

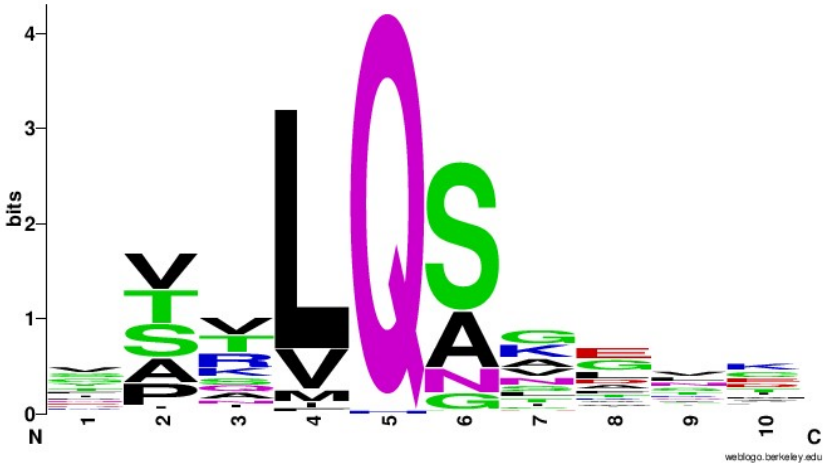

Figure 3: Improved 3CLpro cleavage site sequence logo plotted by WebLogo v2.8.2.[82]

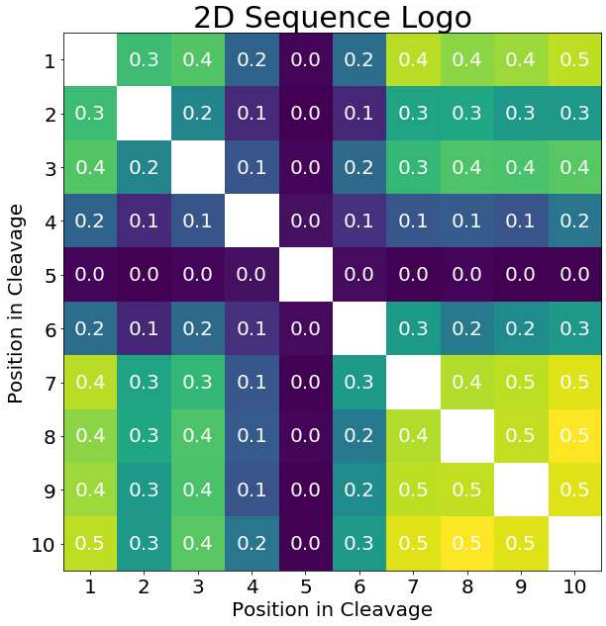

Figure 4: Entropy correlation coefficients (also known as symmetric uncertainties) between positions within the improved sequence logo.

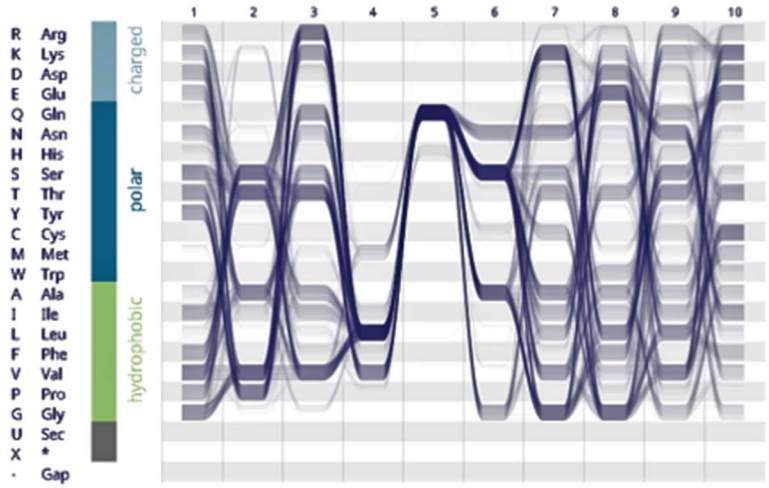

Figure 5: Sequence bundle with charge-polarity-hydrophobicity encoding.[83]

As for my improvements to the NN, note that Kiemer et al.'s MCC of 0.840 is an average from triple cross-validation (CV).[59] 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 cleavages following leucine, methionine, and arginine (VSKLL^AGFKK in APY26744.1 and LVDYM^AGFKK and DAALR^NNELM in ADV71773.1). To my knowledge, synthetic tetra/octapeptides have been cleaved following histidine, phenylalanine, tryptophan, methionine, and possibly proline residues,[56, 84] but only one natural histidine substitution has been documented in HCoV-HKU1[85] and likely does not affect function.[86-88] 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[89]), 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 for all three classification methods, although the optimal and my finally method used 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. Note that Figure 6 displays a curve for an equivalent physiochemical encoding (with input side 40 containing normalized volumes, interface and octanol hydrophobicity scales, and isoelectric points) underperforming when compared to one-hot encoding even at relatively small training sizes. Of these four physiochemical scales, octanol hydrophobicity alone reached a test MCC of 0.959, and, in the order of importance, addition of volume, interface hydrophobicity, and isoelectric point features increased the maximum test MCC to 0.977. Table 1 finally lists the one-hot encoded NN's few incorrectly labeled sequences and their respective sources and scores.

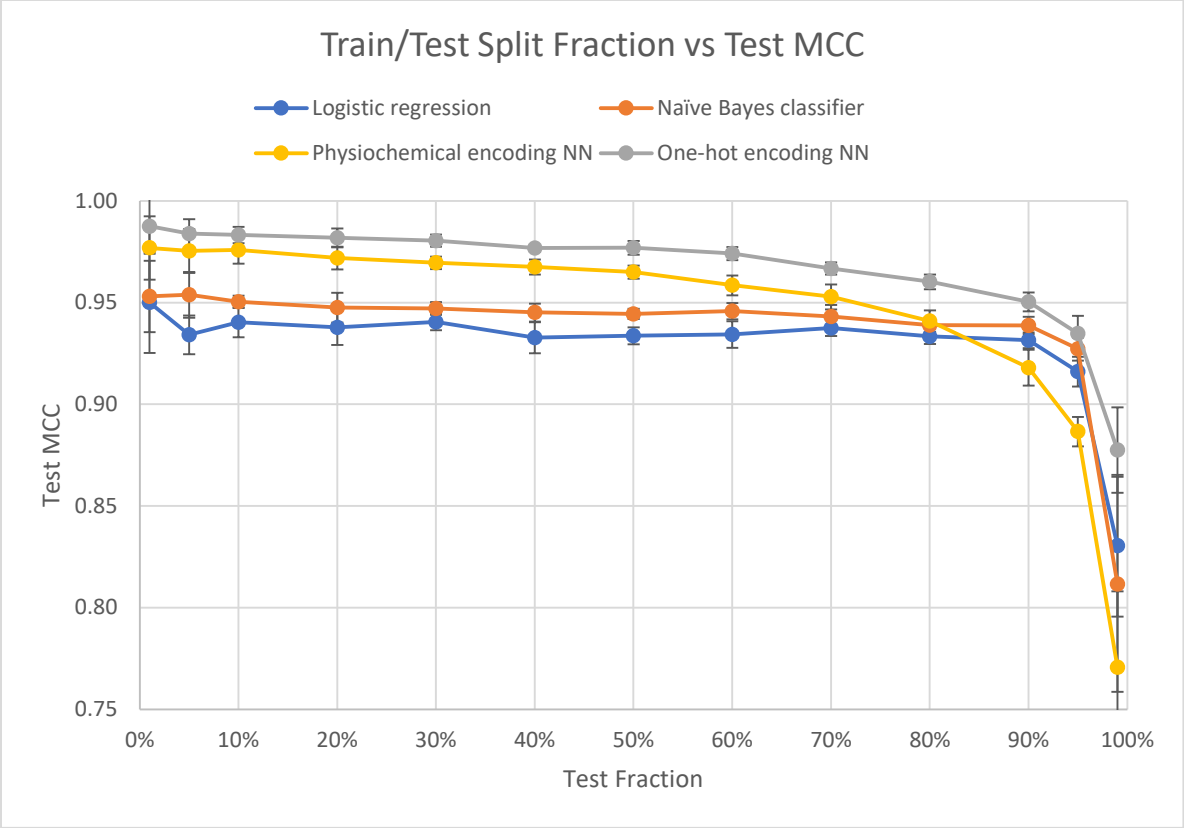

Figure 6: Train/test split fraction vs MCC demonstrating that performance quickly approaches a limit for all classifiers.

Table 1: Only 15 out of 33,182 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 cleaved at least once with a NN score greater than or equal to 0.5. To prove that the 5,887 cleavages 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)[90] 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; my 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.[75, 76] 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 or inclusion of some measure of centrality is required to determine if any insignificantly enriched or depleted pathways are still affected at central nodes (i.e. false negatives). Second, protease-,[56] substrate sequence-,[77, 84, 91-93] substrate truncation-,[94] pH-, temperature-, inhibitor type and concentration-, and time after infection-dependent cleavage kinetics convert this classification problem into a regression problem. Cleavage rates among the 11 cleavages per pp1ab vary by at least 50-fold, so predictions here assume that 3CLpro exists in high enough concentrations and for a long enough time that rate constants do not matter because cleavage reactions are complete. Third, 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 annotations containing proteins with multiple cleavages deviating from Poisson distributions 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 exerted little or no selective pressure before cross-species transmission.

Tissues

As expected in this data, the most significant tissue enrichment of 3CLpro cleavages 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 (vacuolar protein sorting-associated VPS35, eukaryotic translation initiation factor EIF4G1, DNAJ homolog DNAJC13), Huntington's disease (huntingtin), amyotrophic lateral sclerosis (trans-activation response element (TAR) DNA-binding TARDBP), 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 cyclic adenosine monophosphate (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 protein kinase A (PKA) and exchange factor directly activated by cAMP (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,[95] but olfactory receptor neurons are quickly regenerated and sacrifice themselves when infected by influenza A virus.[96] 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 has been shown to be transmitted to the brainstem through the facial and trigeminal nerves.[97]

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. More specifically, cleavage of the cilia-associated proteins nephrocystins 1/2/4/5/6 (NPHP1/2/4/5/6), Bardet-Biedl syndromes proteins 1/9/12 (BBS1/9/12), Alstrom syndrome 1 (ALMS1), coiled-coil and C2 domain-containing protein 2A (CC2D2A), retinitis pigmentosa 1 (RP1), protein fantom (RPGRIP1), tubby-related protein 1 (TULP1), polycystin 1/2, protein kintoun (DNAAF2), dynein axonemal heavy chain 5/11 (DNAH5/11) and intermediate chain 2 (DNAI2), radial spoke head protein 6 homolog A (RSPH6A), and leucine-rich repeat-containing protein 50 (LRRC50) may contribute to dyskinesia and reduced mucociliary escalator effectiveness associated with many respiratory viruses including HCoV-229E and SARS and their resulting bacterial pneumonias.[98, 99] Additionally, cilial dysfunction in olfactory cells in COVID-19 leads to anosmia, although the main reported mechanism is nsp13 (helicase/triphosphatase)-centrosome interaction.[100] 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, nsp13, and spike proteins have so far been shown to interact with the cytoskeleton,[101-103] although many other viruses including influenza A virus,[104] herpes simplex virus, rabies virus, vesicular stomatitis virus, and adeno-associated virus[105] also modulate the cytoskeleton.[106] 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.[107-110]

Modulation of these structural and motor proteins is required for formation of the double-membrane vesicles surrounding replicase complexes[111, 112] 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[113, 114] and contribution to membrane curvature preventing inward nucleocapsid engulfment.[115] Cleavage of retromer component VSP35, ADP-ribosylation factor-binding protein 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 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, separase (the cysteine protease that cleaves cohesin to separate sister chromatids), and topoisomerase III alpha, but not CCCTC-binding factor (CTCF) nor any other topoisomerases 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,[116-120] and some but not all DNA methyltransferases and demethylases are cleaved, further complicating these effects. 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 toward their own open reading frames.[121, 122] 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.[123] Ribosome cleavages are depleted here because they are required 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,[124] indicating that ribosomes that initiate faster than they elongate are preferred because they likely frameshift more frequently, allowing for control of the stoichiometric ratio of pp1a and pp1ab.[125] If slower ribosomes are not directly more likely to frameshift, they are still less likely to participate if frameshift-induced traffic jams, collision-stimulated translation abortion and splitting,[126] and subsequent 60S subunit obstruction sensing and nascent-chain ubiquitylation, which is especially noteworthy because zinc finger 598 (ZNF598), nuclear export mediator factor (NEMF), and listerin E3 ubiquitin ligase 1 (LTN1) are predicted to be cleaved.[127] 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.[128-130] 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).[131] 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.[132] 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.[133] 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 toward 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.[134-136] Telomerase reverse transcriptase (TERT) is also cleaved, but is more frequently reported to be activated by other viral infections and/or promote oncogenesis.[137]

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,[138] is also cleaved, although one of its fragments, lactoferricin, has known antiviral activity.[139] Cleaved PRRs include the toll-like receptors TLR6/8; the C-type lectin receptors CLEC4G/H1/4K/4L/10A/13B/13C/16A; NK cell lectin-like receptors (KLRC4/G1), aspartate/glutamate carrier 1 (ACG1), collectin-7/12, neurocan core protein, FRAS1-related extracellular matrix protein 1 (FREM1), layilin, polycystin 1, E-selectin, and thrombomodulin primarily present on dendritic cells; and the NOD-like receptors NOD2 and NLRP1/2/3/6/10/12/14. Cleaved proteins downstream of these PRRs include receptor-interacting serine-threonine-protein kinase 1/2 (RIP1/2), NF-kappa-B p100 subunit, CASP8 and FADD-like apoptosis regulator (CFLAR), TIR-domain-containing adapter-inducing interferon-beta (TRIF), IRF2, and death domain-associated protein 6 (DAXX), and other relevant downstream pathways similarly include many cleaved proteins: phosphoinositide 3-kinase/protein kinase B (PI3K/AKT) pathway (PIK3CG/D, PIK3R2/5/6, serine/threonine-protein phosphatase 2A (PPP2R1A/2A/2B/2C/2D/3B/5B)(PP2A dephosphorylates AKT), neuronal and inducible nitric oxide synthase (n/iNOS)(where nitric oxide has conflicting effects on viral infections[140, 141]), tuberous sclerosis 1/2 (TSC1/2)); mechanistic target of rapamycin (mTOR) pathway (ribosomal S6 kinase 1/2, sterol regulatory element-binding protein 1 (SREBP1), RB1-inducible coiled-coil protein 1 (RBCC1), regulatory-associated protein of mTOR (RAPTOR), rapamycin-insensitive companion of mTOR (RICTOR), proline-rich 5-like (PRR5L), CAP-Gly domain containing linker protein 1 (CLIP1), forkhead box protein O1/3 (FOXO1/3)); and mitogen-activated protein kinase (MAPK) pathway (MAP4K2/4/5, MAP3K4/5/6/8/13/14, kinase suppressor of RAS 2 (KSR2), MAP2K3/4, MAPK7/13/15, dual specificity protein phosphatase 8 (DUSP8), and mouse double minute 2 homolog (MDM2)). Additional cleaved transcription factors include c-Jun, activating transcription factor 6 (ATF6), the cAMP-responsive element-binding proteins CREB3/5/BP, specificity protein 1 (SP1), octamer transcription factors OCT1/2, the heat schock factors HSF2/2BP/4/X1, RNA polymerase I initiator nucleolar transcription factor 1 (UBTF), RNA polymerase II initiators TFIID and selective factor 1 (SL1) subunits (TATA-binding protein (TBP), TBP-like 2, TBP-associated factor 1C/6/172) and mediator coactivator subunits 1/12L/13/15/17/22/26/28, RNA polymerase III initiators TFIIIB150, TFIIIC, and snRNA-activating protein complex subunit 4 (SNAPC4). No interferons are cleaved likely due to their redundancy, and no interferon receptors are cleaved. The downstream effectors STAT1/2/4; the ISGs guanylate-binding protein 1 (GBP1), interferon alpha-inducible protein 6 (IFI6), membrane spanning 4-domain A4A (MS4A4A), 2'-5'-oligoadenylate synthetase 1 (OAS1), promyelocytic leukemia protein (PML), mitoferrin-2, three prime repair exonuclease 1 (TREX1), and tripartite motif-containing protein 5 (TRIM5); and the tumor necrosis factor (TNF) ligands (TNFSF3/13/18 and ectogysplasin A) 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 (baculoviral IAP repeat-containing proteins BIRC2/3/6) are also cleaved.

Other Pathways and Keywords

Lipoproteins are a depleted keyword, but apolipoproteins APOA-V/B/L1/(a), cholesteryl ester transfer protein (CETP), microsomal triglyceride transfer protein (MTTP), and the lipid transfer receptors LDL-related proteins LRP2/6/12 are all predicted to be cleaved and, other than the proapoptotic APOL1,[142] 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 scavenger receptor SR-B1.[143] 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.[144, 145]

Ubiquitinating and deubiquitinating (DUBs) enzymes are most enriched in the epithelium and the nucleus and include cleaved 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 neural precursor cell expressed developmentally downregulated protein 4 (NEDD4) and the related SMAD ubiquitination regulatory factor 1/2 (SMURF1/2) and HECT, C2, and WW domain containing E3 ubiquitin ligase 1 (HECW1) are, cleaved. NEDD4 has been shown to enhance influenza infectivity by inhibiting interferon-induced transmembrane protein 3 (IFITM3)[146, 147] and Japanese encephalitis virus by inhibiting autophagy,[148] but its ubiquitination of many diverse human viruses promotes their egress. IFITMs generally have antiviral activity (others include HIV-1,[149] dengue virus,[150] and filoviruses[151]), but its use as a treatment for COVID-19 should be carefully considered given its varying effects among other coronaviruses.[152, 153] SARS-CoV-2 has two probable NEDD4 binding sites: the proline-rich, N-terminal PPAY and LPSY[154] 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.[155-157] Further 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.[158, 159]

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.[160-162] SARS nsp13 and nsp14 have been shown to be enhanced by the uncleaved human DDX5 and DDX1, respectively.[163, 164], however multiple proteins interacting with DDX1 (FAM98A) and DDX5 (DHX15, SNW1, MTREX, and HNRNPH1), the retinoic acid-inducible gene I (RIG-I)-associating DDX6, and DDX20 involved in ribosome assembly are predicted to be cleaved, making these effects enigmatic without knowledge of additional interactions with other nsps.

Fibronectin type-III domains are enriched, but fibronectin itself is not cleaved. No cleaved proteins with this domain are directly related to coagulation, but the related 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.[165, 166] Angiotensinogen is, however, unrelated to coagulation and is cleaved far from its N-terminus, so its effects on the renin-angiotensin system remain unknown. The structurally similar alpha-2-macroglobulin has a predicted cleavage outside its protease bait region, however, the addition of a missense mutation Q694S would allow cleavage at the same site as factor XIII without reducing protease trapping ability as much as large deletions.[167, 168] Additional support for this potential exogenous replacement includes presence of serine in the same position in pregnancy zone protein (PZP), which shares 71% identity with alpha-2-macroglobin and contains a neighboring GAG site resembling known PLpro cleavages in its primary bait region. 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 or alpha globulin replacement therapy or treatment with modified small, 3CLpro competitive inhibitors may be a useful treatment for COVID-19.[169]

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.[170] 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.[171-173] 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.[174, 175] 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.[176, 177]

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 dual oxidase 2 (DUOX2)-derived reactive oxygen species (ROS)-induced interferon response.[178] In addition to depleted antioxidant proteins, cleavage of DUOX1/2, NADPH oxidase 5 (NOX5), and xanthine oxidase (XO), the former of which are upregulated in chronic obstructive pulmonary disease (COPD),[179] 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 should be thoroughly evaluated before being recommended as a treatment of COVID-19.

The impact of post-translational modifications on viral protease cleavage frequency remains uncharacterized. 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 believed. 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 toward 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 modified antiproteases including alpha globulins, 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 IFITMs. Pathway components 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 the training dataset to the whole order *Nidovirales* may provide more diversity to improve classifying methods if additional protease/cleavage coevolution does not invalidate the assumption of cross-reactivity. Issues requiring *in vitro* and *in vivo* 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.

Acknowledgments

I am very grateful for my mother, Victoria Prescott, Esq., and friends who have given me invaluable help and advice throughout my work on this project.

References

1. Lechien JR, Chiesa-Estomba CM, De Siati DR, Horoi M, Le Bon SD, Rodriguez A, 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. 2020 Aug;277(8):2251–61. https://doi.org/10.1007/s00405-020-05965-1 PMID: 32253535
2. Baig AM, Khaleeq A, Ali U, Syeda H. Evidence of the COVID-19 virus targeting the CNS: Tissue distribution, host-virus interaction, and proposed neurotropic mechanisms. ACS Chem Neurosci. 2020 Mar;11(7):995–8. https://doi.org/10.1021/acschemneuro.0c00122 PMID: 32167747
3. Lau KK, Yu WC, Chu CM, Lau ST, Sheng B, Yuen KY. Possible central nervous system infection by SARS coronavirus. Emerg Infect Dis. 2004 Feb;10(2):342–4. https://doi.org/10.3201/eid1002.030638 PMID: 15030709
4. Netland J, Meyerholz DK, Moore S, Cassell M, Perlman S. Severe acute respiratory syndrome coronavirus infection causes neuronal death in the absence of encephalitic in mice transgenic for human ACE2. J Virol. 2008 Aug;82(15):7264–75. https://doi.org/10.1128/JVI.00737-08 PMID: 18495771
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 Mar;2020:1–4. https://doi.org/10.1002/jmv.25728 PMID: 32104915
6. Zhang C, Shi L, Wang FS. Liver injury in COVID-19: management and challenges. Lancet Gastroenterol Hepatol. 2020 May;5(5):428–30. https://doi.org/10.1016/S2468-1253(20)30057-1 PMID: 32145190
7. Chau TN, Lee KC, Yao H, Tsang TY, Chow TC, Yeung YC, et al. SARS-associated viral hepatitis caused by a novel coronavirus: Report of three cases. Hepatology. 2004 Feb;39(2):302–10. https://doi.org/10.1002/hep.20111 PMID: 14767982
8. Liu L, Wei Q, Alvarez X, Wang H, Du Y, Zhu H, 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. 2011 Apr;85(8):4025–30. https://doi.org/10.1128/JVI.02292-10 PMID: 21289121
9. Zhan J, Deng R, Tang J, Zhang B, Tang Y, Wang JK, et al. The spleen as a target in severe acute respiratory syndrome. FASEB J. 2006 Nov;20(13):2321–8. https://doi.org/10.1096/fj.06-6324com PMID: 17077309

10. Naicker S, Yang CW, Hwang SJ, Liu BC, Chen JH, Jha V. The novel coronavirus 2019 epidemic and kidneys. Kidney Int. 2020 May;97:824–8. https://doi.org/10.1016/j.kint.2020.03.001 PMID: 32204907
11. Fan C, Ding Y, Lu WL, Wang J. ACE2 expression in kidney and testes may cause kidney and testis damage after 2019-nCoV infection. medRxiv. 2020 Feb. https://doi.org/10.1101/2020.02.12.20022418
12. Chen H, Guo J, Wang C, Luo F, Yu X, Zhang W, et al. Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. Lancet. 2020 Mar;395(10226):809–15. https://doi.org/10.1016/S0140-6736(20)30360-3 PMID: 32151335
13. Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and the cardiovascular system. Nat Rev Cardiol. 2020 Mar;17:259–60. https://doi.org/10.1038/s41569-020-0360-5 PMID: 32139904
14. Dandekar AA, Perlman S. Immunopathogenesis of coronavirus infections: implications for SARS. Nat Rev Immunol. 2005 Dec;5(12):917–27. https://doi.org/10.1038/nri1732 PMID: 16322745
15. Gu J, Gong E, Zhang B, Zheng J, Gao Z, Zhong Y, et al. Multiple organ infection and the pathogenesis of SARS. J Exp Med. 2005 Aug;202(3):415–24. https://doi.org/10.1084/jem.20050828 PMID: 16043521
16. Jia X, Yin C, Lu S, Chen Y, Liu Q, Bai J, et al. Two things about COVID-19 might need attention. Preprints. 2020 Feb. https://doi.org/10.20944/preprints202002.0315.v1
17. Simonnet A, Chetboun M, Poissy J, Raverdy V, Noulette J, Duhamel A, et al. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) requiring invasive mechanical ventilation. Obesity. 2020 Apr;28(7):1196–9. https://doi.org/10.1002/oby.22831 PMID: 32271993
18. Elliot JG, Donovan GM, Wang KCW, Green FHY, James AL, Noble PB, et al. Fatty airways: Implications for obstructive disease. Eur Respir J. 2019 Dec;54(6):1900857. https://doi.org/10.1183/13993003.00857-2019 PMID: 31624112
19. Ziebuhr J Snijder EJ, Gorbalenya AE. Virus-encoded proteinases and proteolytic processing in the *Nidovirales*. J Gen Virol. 2000 Apr;81(4):853–79. https://doi.org/10.1099/0022-1317-81-4-853 PMID: 10725411
20. Baez-Santos YM, St. John SE, Mesecar AD. The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds. Antiviral Res. 2015 Mar;115:21–38. https://doi.org/10.1016/j.antiviral.2014.12.015 PMID: 25554382
21. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung SH. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: Peptidomimetics and small molecule chemotherapy. J Med Chem. 2016 Feb;59(14):6595–628. https://doi.org/10.1021/acs.jmedchem.5b01461 PMID: 26878082
22. Yang H, Xie W, Xue X, Yang K, Ma J, Liang W, et al. Design of wide-spectrum inhibitors targeting coronavirus main proteases. PLOS Biology. 2005 Nov;3(10):e428. https://doi.org/10.1371/journal.pbio.0030324 PMID: 16128623
23. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: Basis for design of anti-SARS drugs. Science. 2003 Jun;300(5626):1763–7. https://doi.org/10.1126/science.1085658 PMID: 12746549
24. Neimeyer D, Mosbauer K, Klein EM, Sieberg A, Mettelman RC, Mielech AM, et al. The papain-like protease determines a virulence trait that varies among members of the SARS-coronavirus species. PLOS Pathog. 2018 Sep;14(9):e1007296. https://doi.org/10.1371/journal.ppat.1007296 PMID: 30248143

25. Barretto N, Jukneliene D, Ratia K, Chen Z, Mesecar AD, Baker SC. The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquinating activity. J Virol. 2005 Dec;79(24):15189–98. https://doi.org/10.1128/JVI.79.24.15189-15198.2005 PMID: 16306590
26. Yang X, Chen X, Bian G, Tu J, Xing Y, Wang Y, et al. Proteolytic processing, deubiquitinase and interferon antagonist activities of Middle East respiratory syndrome coronavirus papain-like protease. J Gen Virol. 2014 Mar;95(3):614–26. https://doi.org/10.1099/vir.0.059014-0 PMID: 24362959
27. Bailey-Elkin BA, Knaap RCM, Johnson GG, Dalebout TJ, Ninaber DK, van Kasteren PB, 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. 2014 Dec;289:34667–82. https://doi.org/10.1074/jbc.M114.609644 PMID: 25320088
28. Li SW, Lai CC, Ping JF, Tsai FJ, Wan L, Lin YJ, 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. 2011 May;92(5):1127–40. https://doi.org/10.1099/vir.0.028936-0 PMID: 21270289
29. Xing Y, Chen J, Tu J, Zhang B, Chen X, Shi H, 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. 2013 Jul;94(7):1554–67. https://doi.org/10.1099/vir.0.051169-0 PMID: 23596270
30. Matthews K, Schafer A, Pham A, Frieman M. The SARS coronavirus papain like protease can inhibit IRF3 at a post activation step that requires deubiquination activity. Virol J. 2014 Dec;11:209. https://doi.org/10.1186/s12985-014-0209-9 PMID: 25481026
31. Devaraj SG, Wang N, Chen Z, Chen Z, Tseng M, Barretto N, 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. 2007 Nov;282:32208–21. https://doi.org/10.1074/jbc.M704870200 PMID: 17761676
32. Banerjee A, Zhang X, Yip A, Schulz KS, Irving AT, Bowdish D, et al. Positive selection of a serine residue in bat IRF3 confers enhanced antiviral protection. iScience. 2020 Mar;23(7):100958. https://doi.org/10.1016/j.isci.2020.100958 PMID: 32179480
33. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Feb;579:270–3. https://doi.org/10.1038/s41586-020-2012-7 PMID: 32015507
34. Needle D, Lountos GT, Waugh DS. Structures of the Middle East respiratory syndrome coronavirus 3C-like protease reveal insights into substrate specificity. Acta Cryst. 2015 Feb;D71:1102–11. https://doi.org/10.1107/S1399004715003521 PMID: 25945576
35. Xue X, Yu H, Yang H, Xue F, Wu Z, Shen W, et al. Structures of two coronavirus main proteases: Implications for substrate binding and antiviral drug design. J Virol. 2008 Mar;82(5):2515–27. https://doi.org/10.1128/JVI.02114-07 PMID: 18094151
36. Anand K, Palm GJ, Mesters JR, Siddell SG, Ziebuhr J, Hilgenfeld R. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. EMBO J. 2002 Jul;21(13):3213–24. https://doi.org/10.1093/emboj/cdf327 PMID: 12093723
37. Zhu X, Wang D, Zhou J, Pan T, Chen J, Yang Y, et al. Porcine deltacoronavirus nsp5 antagonizes type I interferon signaling by cleaving STAT2. J Virol. 2017 May;91(10):e00003–17. https://doi.org/10.1128/JVI.00003-17 PMID: 28250121
38. Wang, D, Fang L, Shi Y, Zhang H, Gao L, Peng G, et al. Porcine epidemic diarrhea virus 3C-like protease regulates its interferon antagonism by cleaving NEMO. J Virol. 2016 Feb;90(4):2090–101. https://doi.org/10.1128/JVI.02514-15 PMID: 26656704

651 39. Moustaqil M, Ollivier E, Chiu HP, Van Tol S, Rufolffi-Soto P, Stevens C, et al. SARS-CoV-2 proteases
652 cleave IRF3 and critical modulators of inflammatory pathways (NLRP12 and TAB1): implications for
653 disease presentation across species and the search for reservoir hosts. bioRxiv. 2020 Jun.
654 https://doi.org/10.1101/2020.06.05.135699
655 40. de Vries AAF, Horzinek MC, Rottier PJM, de Groot RJ. The genome organization of the *Nidovirales*:
656 Similarities and differences between arteri-, toro-, and coronaviruses. Sem Virol. 1997 Feb;8(1):33–
657 47. https://doi.org/10.1006/smvy.1997.0104 PMID: 32288441
658 41. Ye S, Xia H, Dong C, Cheng Z, Xia X, Zhang J, et al. Identification and characterization of Iflavirus 3C-
659 like protease processing activities. Virol. 2012 Jul;428(2):136–45.
660 https://doi.org/10.1016/j.virol.2012.04.002 PMID: 22534091
661 42. Kuyumcu-Martinez M, Belliot G, Sosnovtsev SV, Chang KO, Green KY, Lloyd RE. Calicivirus 3C-like
662 proteinase inhibits cellular translation by cleavage of poly(A)-binding protein. J Virol. 2014
663 Aug;78(15):8172–82. https://doi.org/10.1128/JVI.78.15.8172-8182.2004 PMID: 15254188
664 43. Cordingley MG, Callahan PL, Sardana VV, Garsky VM, Colonno RJ. Substrate requirements of human
665 rhinovirus 3C protease for peptide cleavage *in vitro*. J Biol Chem. 1990 Jun;265(16):9062–5. PMID:
666 2160953
667 44. Pallai PV, Burkhardt F, Skoog M, Schreiner K, Bax P, Cohen KA, et al. Cleavage of synthetic peptides
668 by purified poliovirus 3C proteinase. J Biol Chem. 1989 Jun;264(17):9738–41. PMID: 2542331
669 45. Dougherty WG, Semler BL. Expression of virus-encoded proteinases: Functional and structural
670 similarities with cellular enzymes. Microbiol Rev. 1993 Dec;57(4):781–822. PMID: 8302216
671 46. Tesar M, Marquardt O. Foot-and-mouth disease virus protease 3C inhibits cellular transcription and
672 mediated cleavage of histone H3. Virology. 1990 Feb;174(2):364–74. https://doi.org/10.1016/0042-
673 6822(90)90090-e PMID: 2154880
674 47. Falk MM, Grigera PR, Gergmann IE, Zibert A, Multhaup G, Beck E. Foot-and-mouth disease virus
675 protease 3C induces specific proteolytic cleavage of host cell histone H3. J Virol. 1990
676 Feb;64(2):748–56. https://doi.org/10.1128/JVI.64.2.748-756.1990 PMID: 2153239
677 48. Kliewer S, Dasgupta A. An RNA polymerase II transcription factor inactivated in poliovirus-infected
678 cells copurifies with transcription factor TFIID. Mol Cell Biol. 1988 Aug;8(8):3175–82.
679 https://doi.org/10.1128/mcb.8.8.3175 PMID: 2850483
680 49. Clark ME, Dasgupta A. A transcriptionally active form of TFIIIC is modified in poliovirus-infected HeLa
681 cells. Mol Cell Biol. 1990 Oct;10(10):5106–13. https://doi.org/10.1128/mcb.10.10.5106 PMID:
682 2204807
683 50. Clark ME, Hammerle T, Wimmer E, Dasgupta A. Poliovirus proteinase 3C converts an active form of
684 transcription factor IIIC to an inactive form: A mechanism for inhibition of host cell polymerase III
685 transcription by poliovirus. EMBO J. 1991 Oct;10(10):2941–7. https://doi.org/10.1002/j.1460-
686 2075.1991.tb07844.x PMID: 1915271
687 51. Clark ME, Lieberman PM, Berk AJ, Dasgupta A. Direct cleavage of human TATA-binding protein by
688 poliovirus protease 3C *in vivo* and *in vitro*. Mol Cell Bio. 1993 Feb;13(2):1232–7.
689 https://doi.org/10.1128/mcb.13.2.1232 PMID: 8380894
690 52. Shen Y, Igo M, Yalamanchili P, Berk AJ, Dasgupta A. DNA binding domain and subunit interactions of
691 transcription factor IIIC revealed by dissection with poliovirus 3C protease. Mol Cell Biol. 1996
692 Aug;16(8):4163–71. https://doi.org/10.1128/mcb.16.8.4163 PMID: 8754815
693 53. Joachims M, Etchison D. Poliovirus infection results in structural alteration of a microtubule-
694 associated protein. J Virol. 1992 Oct;66(10):5797–804. https://doi.org/10.1128/JVI.66.10.5797-
695 5804.1992 PMID: 1326643
696 54. Joachims M, Harris KS, Etchison D. Poliovirus protease 3C mediates cleavage of microtubule-
697 associated protein 4. Virology. 1995 Aug;211(2):451–61. https://doi.org/10.1006/viro.1995.1427
698 PMID: 7645249

55. Thiel V, Ivanov KA, Putics A, Hertzig T, Schelle B, Bayer S, et al. Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol. 2003 Sep;84(9):2305–15. https://doi.org/10.1099/vir.0.19424-0 PMID: 12917450
56. Chuck CP, Chow HF, Wan DCC, Wong KB. Profiling of substrate specificities of 3C-like proteases from group 1, 2a, 2b, and 3 coronaviruses. PLOS One. 2011 Nov;6(11):e27228. https://doi.org/10.1371/journal.pone.0027228 PMID: 22073294
57. Gao F, Ou HY, Chen LL, Zheng WX, Zhang CT. Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. FEBS Lett. 2003 Oct;553(3):451–6. https://doi.org/10.1016/s0014-5793(03)01091-3 PMID: 14572668
58. Yang ZR. Mining SARS-CoV protease cleavage data using non-orthogonal decision trees: a novel method for decisive template selection. Bioinformatics. 2005 Jun;21(11):2644–50. https://doi.org/10.1093/bioinformatics/bti404 PMID: 15797903
59. Kiemer L, Lund O, Brunak S, Blom N. Coronavirus 3CLpro proteinase cleavage sites: Possible relevance to SARS virus pathology. BMC Bioinform. 2004 Jun;5:72. https://doi.org/10.1186/1471-2105-5-72 PMID: 15180906
60. Blom N, Hansen J, Blaas D, Brunak S. Cleavage site analysis in picornaviral polyproteins: Discovering cellular targets by neural networks. Protein Science. 1996 Sep;5(11):2203–16. https://doi.org/10.1002/pro.5560051107 PMID: 8931139
61. Narayanan A, Wu X, Yang ZR. Mining viral protease data to extract cleavage knowledge. Bioinformatics. 2002 Jul;18(1):S5–13. https://doi.org/10.1093/bioinformatics/18.suppl_1.s5 PMID: 12169525
62. Singh O, Su ECY. Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physiochemical features. BMC Bioinform. 2016 Dec;17(Suppl 17):478. https://doi.org/10.1186/s12859-016-1337-6 PMID: 28155640
63. Li X, Hu H, Shu L. Predicting human immunodeficiency virus protease cleavage sites in nonlinear projection space. Mol Cell Biochem. 2010 Jan;339(1–2):127–33. https://doi.org/10.1007/s11010-009-0376-y PMID: 20054614
64. Rognvaldsson T, You L. Why neural networks should not be used for HIV-1 protease cleavage site prediction. Bioinformatics. 2004 Feb;20(11):1702–9. https://doi.org/10.1093/bioinformatics/bth144 PMID: 14988129
65. You L, Garwicz D, Rognvaldsson T. Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. J Virol. 2005 Oct;79(19):12477–86. https://doi.org/10.1128/JVI.79.19.12477–12486.2005 PMID: 16160175
66. Cai YD, Chou KC. Artificial neural network model for predicting HIV protease cleavage sites in protein. Adv Eng Softw. 1998 Mar;29(2):119–28. https://doi.org/10.1016/S0965-9978(98)00046-5
67. Manning T, Walsh P. The importance of physiochemical characteristics and nonlinear classifiers in determining HIV-1 protease specificity. Bioengineered. 2016 Mar–Apr;7(2):65–78. https://doi.org/10.1080/21655979.2016.1149271 PMID: 27212259
68. Chown H. A comparison of machine learning algorithms for the prediction of hepatitis C NS3 protease cleavage sites. J Proteom Bioinform. 2019 Jul;12(5):89–95. https://doi.org/10.35248/0974-276X.19.12.501
69. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan;47(D1):D505–15. https://doi.org/10.1093/nar/gky1049 PMID: 30395287
70. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J. GenBank. Nucleic Acids Res. 2017 Jan;45(D1):D37–42. https://doi.org/10.1093/nar/gkw1070 PMID: 27899564
71. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct;7:539. https://doi.org/10.1038/msb.2011.75 PMID: 21988835

72. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. 2010 Jul;38(2):W695–9. https://doi.org/10.1093/nar/gkq313 PMID: 20439314
73. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis tool web services from the EMBL-EBI. Nucleic Acids Res. 2013 Jul;41(W1):W597–600. https://doi.org/10.1093/nar/gkt376 PMID: 23671338
74. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011 Oct;12:2825–30. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
75. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009 Jan;4(1):44–57. https://doi.org/10.1038/nprot.2008.211 PMID: 19131956
76. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan;37(1):1–13. https://doi.org/10.1093/nar/gkn923 PMID: 19033363
77. Wu A, Wang Y, Zeng C, Huang X, Xu S, Su C, et al. Prediction and biochemical analysis of putative cleavage sites of the 3C-like protease of Middle East respiratory syndrome coronavirus. Virus Res. 2015 Oct;208:56–65. https://doi.org/10.1016/j.virusres.2015.05.018 PMID: 26036787
78. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008 Nov;9:2579–605. http://www.jmlr.org/papers/v9/vandermaaten08a.html
79. Chou KC, Zhang CT, Kezdy FJ. A vector projection approach to predicting HIV protease cleavage sites in proteins. Proteins. 1993 Jun;16(2):195–204. https://doi.org/10.1002/prot.340160206 PMID: 8332607
80. Bindewald E, Schneider TD, Shapiro BA. CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. Nucleic Acids Res. 2006 Jul;34:W405–11. https://doi.org/10.1093/nar/gkl269 PMID: 16845037
81. Maes F, Vandermeulen D, Suetens P. Medical image registration using mutual information. Proc IEEE. 2003 Oct;91(10):1699–722. https://doi.org/10.1109/JPROC.2003.817864
82. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004 Jun;14:1188–90. https://doi.org/10.1101/gr.849004 PMID: 15173120
83. Kultys M, Nicholas L, Schwarz R, Goldman N, King J. Sequence Bundles: a novel method for visualizing, discovering and exploring sequence motifs. BMC Proc. 2014 Aug;8(Suppl 2):S8. https://doi.org/10.1186/1753-6561-8-S2-S8 PMID: 25237395
84. Goetz DH, Choe Y, Hansell E, Chen YT, McDowell M, Jonsson CB, et al. Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the SARS coronavirus. Biochemistry. 2007 Jul;46(30):8744–52. https://doi.org/10.1021/bi0621415 PMID: 17605471
85. Woo PCY, Huang Y, Lau SKP, Tsoi HW, Yuen KY. *In silico* analysis of ORF1ab in coronavirus HKU1 genome reveals a unique putative cleavage site of coronavirus HKU1 3C-like protease. Microbiol Immunol. 2005 Oct;49(10):899–908. https://doi.org/10.1111/j.1348-0421.2005.tb03681.x PMID: 16237267
86. Ma Y, Wu Y, Shaw N, Gao Y, Wang J, Sun Y, et al. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. PNAS. 2015 Jul;112(30):9436–41. https://doi.org/10.1073/pnas.1508686112 PMID: 26159422
87. Neuman BW, Chamberlain P, Bowden F, Joseph J. Atlas of coronavirus replicase structure. Virus Res. 2014 Dec;194:49–66. https://doi.org/10.1016/j.virusres.2013.12.004 PMID: 24355834
88. Fang SG, Shen H, Wang J, Tay FPL, Liu DX. Proteolytic processing of polyproteins 1a and 1ab between non-structural proteins 10 and 11/12 of *Coronavirus* infectious bronchitis virus is dispensable for

794 viral replication in cultured cells. Virol. 2008 Sep;379(2):175–80.
795 https://doi.org/10.1016/j.virol.2008.06.038 PMID: 18678384
796 89. Santos MS, Soares JP, Abreu PH, Araujo H, Santos J. Cross-validation for imbalances datasets:
797 Avoiding overoptimistic and overfitting approaches. IEEE Comp Intell Mag. 2018 Nov;13(4):59–76.
798 https://doi.org/10.1109/MCI.2018.2866730 PMID: 27295638
799 90. Kozlowski LP. Proteome-pI: proteome isoelectric point database. Nucleic Acids Res. 2017
800 Jan;45(D1):D1112–6. https://doi.org/10.1093/nar/gkw978 PMID: 27789699
801 91. Fan K, Wei P, Feng Q, Chen S, Huang C, Ma L, et al. Biosynthesis, purification, and substrate
802 specificity of severe acute respiratory syndrome coronavirus 3C-like proteinase. J Biol Chem. 2004
803 Jan;279(3):1637–42. https://doi.org/10.1074/jbc.M310875200 PMID: 14561748
804 92. Hegyi A, Ziebuhr J. Conservation of substrate specificities among coronavirus main proteases. J Gen
805 Virol. 2002 Mar;83(3):595–99. https://doi.org/10.1099/0022-1317-83-3-595 PMID: 11842254
806 93. Grum-Tokars V, Ratia K, Begaye A, Baker SC, Mesecar AD. Evaluating the 3C-like protease activity of
807 SARS-coronavirus: Recommendations for standardized assays for drug discovery. Virus Res. 2008
808 Apr;133(1):63–73. https://doi.org/10.1016/j.virusres.2007.02.015 PMID: 17397958
809 94. Fan K, Ma L, Han X, Liang H, Wei P, Liu Y, et al. The substrate specificity of SARS coronavirus 3C-like
810 proteinase. Biochem Biophys Res Common. 2005 Apr;329(3):934–40.
811 https://doi.org/10.1016/j.bbrc.2005.02.061 PMID: 15752746
812 95. Ikemura T, Schwarze J, Makela M, Kanehiro A, Joetham A, Ohmori K, et al. Type 4 phosphodiesterase
813 inhibitors attenuate respiratory syncytial virus-induced airway hyper-responsiveness and lung
814 eosinophilia. J Pharmacol Exp Ther. 2000 Aug;294(2):701–6.
815 https://pubmed.ncbi.nlm.nih.gov/10900250/ PMID: 10900250
816 96. Mori I, Goshima F, Imai Y, Kohsaka S, Sugiyama T, Yoshida T, et al. Olfactory receptor neurons
817 prevent dissemination of neurovirulent influenza A virus into the brain by undergoing virus-induced
818 apoptosis. J Gen Virol. 2002 Sep;83(9):2109–16. https://doi.org/10.1099/0022-1317-83-9-2109
819 PMID: 12185263
820 97. Thomander L, Aldskogius H, Vahlne A, Kristensson K, Thomas E. Invasion of cranial nerves and brain
821 stem by herpes simplex virus inoculated into the mouse tongue. Ann Ontol Rhinol Laryngol. 1988
822 Sep;97(5):554–8. https://doi.org/10.1177/000348948809700525 PMID: 2845851
823 98. Chilvers MA, McKean M, Rutman A, Myint BS, Silverman M, O'Callaghan C. The effects of
824 coronavirus of human nasal ciliated respiratory epithelium. Eur Respir J. 2001 Dec;18(6):965–70.
825 https://doi.org/10.1183/09031936.01.00093001 PMID: 11829103
826 99. Kuek LE, Lee RJ. First contact: the role of respiratory cilia in host-pathogen interactions in the
827 airways. Am J Physiol Lung Cell Mol Physiol. 2020 Oct;319(4):L603–19.
828 https://doi.org/10.1152/ajplung.00283.2020 PMID: 32783615
829 100. Li W, Li M, Ou G. COVID-19, cilia, and smell. FEBS J. 2020 Sep;287(17):3672–6.
830 https://doi.org/10.1111/febs.15491 PMID: 32692465
831 101. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein
832 interaction map reveals targets for drug repurposing. Nature. 2020 Jul;583:459–68.
833 https://doi.org/10.1038/s41586-020-2286-9 PMID: 32353859
834 102. Lv X, Li Z, Guan J, Hu S, Zhang J, Lan Y, et al. Porcine hemagluttinating encephalomyelitis virus
835 activation of the integrin α5β1-FAK-cofilin pathway causes cytoskeletal rearrangement to promote
836 its invasion of N2a cells. J Virol. 2019 Mar;93(5):e01736–18. https://doi.org/10.1128/JVI.01736-18
837 PMID: 30541856
838 103. Rudiger AT, Mayrhofer P, Ma-Lauer Y, Pohlentz G, Muthing J, von Brunn A, et al. Tubulins
839 interact with porcine and human S proteins of the genus *Alphacoronavirus* and support successful
840 assembly and release of infectious viral particles. Virol. 2016 Oct;497:185–97.
841 https://doi.org/10.1016/j.virol.2016.07.022 PMID: 27479465

842 104. Ohman T, Rintahaka J, Kalkkinen N, Matikainen S, Nyman TA. Actin and RIG-I/MAVS signaling
843 components translocate to mitochondria upon influenza A virus infection of human primary
844 macrophages. J Immunol. 2009 May;182(9):5682–92. https://doi.org/10.4049/jimmunol.0803093
845 PMID: 19380815
846 105. Dohner K, Sodeik B. The role of the cytoskeleton during viral infection. Curr Top Microbiol. 2005
847 Feb;285:67–108. https://doi.org/10.1007/3-540-26764-6_3 PMID: 15609501
848 106. Naghavi MH, Walsh D. Microtubule regulation and function during virus infection. J Virol. 2017
849 Aug;91(16):e00538–17. https://doi.org/10.1128/JVI.00538-17 PMID: 28615197
850 107. Kristensson K, Lyche E, Roytta M, Svennerholm B, Vahlne A. Neuritic transport of herpes simplex
851 virus in rat sensory neurons in vitro. Effects of substances interacting with microtubular function and
852 axonal flow [nocodazole, taxol, and erythron-9-3-(2-hydroxynonyl)adenine]. J Gen Virol. 1986
853 Sep;67:2023–8. https://doi.org/10.1099/0022-1317-67-9-2023 PMID: 2427647
854 108. Solov'eva MF, Krispin TI, Shloma DV, Kalmakhelidze KA, Balandin IG. Effect of inhibitors that
855 destroy cytoskeleton structures on the antiviral and antiproliferative activity of interferons. Vopr
856 Virusol. 1988 May–Jun;33(3):309–14. https://pubmed.ncbi.nlm.nih.gov/2459850/ PMID: 2459850
857 109. Yi F, Guo J, Dabbagh D, Spear M, He S, Kehn-Hall K, et al. Discovery of novel small-molecule
858 inhibitors of LIM domain kinase for inhibiting HIV-1. J Virol. 2017 Apr;91(13):e02418–16.
859 https://doi.org/10.1128/JVI.02418-16 PMID: 28381571
860 110. Campbell EM, Nunez R, Hope TJ. Disruption of the actin cytoskeleton can complement the ability
861 of Nef to enhance HIV-1 infectivity. J Virol. 2004 Jun;78(11):5745–55.
862 https://doi.org/10.1128/JVI.78.11.5745-5755.2004 PMID: 15140972
863 111. Wolff G, Melia CE, Snijder EJ, Barcena M. Double-membrane vesicles as platforms for viral
864 replication. Trends in Microbiol. 2020 Jun;1839. https://doi.org/10.1016/j.tim.2020.05.009 PMID:
865 32536523
866 112. Neuman BW, Angelini MM, Buchmeier MJ. Does form meet function in the coronavirus
867 replicative organelle? Trends in Microbiol. 2014 Nov;22(11):642–7.
868 https://doi.org/10.1016/j.tim.2014.06.003 PMID: 25037114
869 113. Miller E, Antonny B, Hamamoto S, Schekman R. Cargo selection into COPII vesicles is driven by
870 the Sec24p subunit. EMBO J. 2002 Nov;21(22):6105–13. https://doi.org/10.1093/emboj/cdf605
871 PMID: 12426382
872 114. Mancias JD, Goldberg J. Structural basis of cargo membrane protein discrimination by the
873 human COPII coat machinery. EMBO J. 2008 Oct;27(21):2918–28.
874 https://doi.org/10.1038/emboj.2008.208 PMID: 18843296
875 115. Stagg SM, Gurkan C, Fowler DM, LaPointe P, Foss TR, Potter CS, et al. Structure of the Sec13/31
876 COPII coat cage. Nature. 2006 Jan;439:234–8. https://doi.org/10.1038/nature04339 PMID:
877 16407955
878 116. Nagesh PT, Husain M. Influenza A virus dysregulates host histone deacetylase 1 that inhibits viral
879 infection in lung epithelial cells. J Virol. 2016 Apr;90(6):4614–25. https://doi.org/10.1128/JVI.00126-
880 16 PMID: 26912629
881 117. Chen H, Qian Y, Chen X, Ruan Z, Ye Y, Chen H, et al. HDAC6 restricts influenza A virus by
882 deacetylation of the RNA polymerase PA subunit. J Virol. 2019 Feb;93(4):e01896–18.
883 https://doi.org/10.1128/JVI.01896-18 PMID: 30518648
884 118. Shulak L, Beljanski V, Chiang C, Dutta SM, Van Grevenynghe J, Belgnaou SM, et al. Histone
885 deacetylase inhibitors potentiate vesicular stomatitis virus oncolysis in prostate cancer cells by
886 modulating NF-κB-dependent autophagy. J Virol. 2014 Feb;88(5):2927–40.
887 https://doi.org/10.1128/JVI.03406-13 PMID: 24371063

888 119. Feng Q, Su Z, Song S, Xu H, Zhang B, Yi L, et al. Histone deacetylase inhibitors suppress RSV
889 infection and alleviate virus-induced airway inflammation. Int J Mol Med. 2016 Jul;38(3):812–22.
890 https://doi.org/10.3892/ijmm.2016.2691 PMID: 27460781
891 120. Mosley AJ, Meekings KN, McCarthy C, Shepherd D, Cerundolo V, Mazitschek R, et al. Histone
892 deacetylase inhibitors increase virus gene expression but decrease CG8+ cell antiviral function in
893 HTLV-1 infection. Blood. 2006 Dec;108(12):3801–7. https://doi.org/10.1182/blood-2006-03-013235
894 PMID: 16912225
895 121. Kaminskyy V, Zhivotovsky B. To kill of be killed: how viruses interact with the cell death
896 machinery. J Intern Med. 2010 May;267:473–82. https://doi.org/10.1111/j.1365-2796.2010.02222.x
897 PMID: 20433575
898 122. Spencer CA, Kruhlak MJ, Jenkins HL, Sun X, Bazett-Jones DP. Mitotic transcription repression *in*
899 *vivo* in the absence of nucleosomal chromatin condensation. J Cell Bio. 2000 Jul;150(1):13–26.
900 https://doi.org/10.1083/jcb.150.1.13 PMID: 10893252
901 123. Banerjee S, An S, Zhou A, Silverman RH, Makino S. RNase L-independent specific 28S rRNA
902 cleavage in murine coronavirus-infected cells. J Virol. 2000 Oct;74(19):8793–802.
903 https://doi.org/10.1128/jvi.74.19.8793-8802.2000 PMID: 10982321
904 124. Slavov N, Semrau S, Airoldi E, Budnik B, van Oudenaarden A. Differential stoichiometry among
905 core ribosomal proteins. Cell Rep. 2015 Nov;13(5):865–73.
906 https://doi.org/10.1016/j.celrep.2015.09.056 PMID: 26565899
907 125. Plant EP, Rakauskaite R, Taylor DR, Dinman JD. Achieving a golden mean: Mechanisms by which
908 coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. J Virol. 2019
909 Apr;84(9):4330–40. https://doi.org/10.1128/JVI.02480-09 PMID: 20164235
910 126. Park H, Subramaniam AR. Inverted translational control of eukaryotic gene expression by
911 ribosome collisions. PLOS Biol. 2019 Sep;17(9):e3000396.
912 https://doi.org/10.1371/journal.pbio.3000396 PMID: 31532761
913 127. Joazeiro CAP. Mechanisms and functions of ribosome-associated protein quality control. Nat Rev
914 Mol Cell Biol. 2019 Jun;20(6):368–83. https://doi.org/10.1038/s41580-019-0118-2 PMID: 30940912
915 128. Siegal V, Walter P. Elongation arrest is not a prerequisite for secretory protein translocation
916 across the microsomal membrane. J Cell Biol. 1985 Jun;100(6):1913–21.
917 https://doi.org/10.1083/jcb.100.6.1913 PMID: 2581979
918 129. Rottier P, Armstrong J, Meyer DI. Signal recognition particle-dependent insertion of coronavirus
919 E1, an intracellular membrane glycoprotein. J Biol Chem. 1985 Aug;260(8):4648–52.
920 https://pubmed.ncbi.nlm.nih.gov/2985561/ PMID: 2985561
921 130. Young JC, Andrews DW. The signal recognition particle receptor alpha subunit assembles co-
922 translationally on the endoplasmic reticulum membrane during an mRNA-encoding translation
923 pause *in vitro*. EMBO J. 1996 Jan;15(1):172–81. https://doi.org/10.1002/j.1460-2075.1996.tb00345.x
924 PMID: 25855820
925 131. Grandi N, Tramontano E. Human endogenous retroviruses are ancient acquired elements still
926 shaping innate immune responses. Front Immunol. 2019 Sep;9(2039):1–16.
927 https://doi.org/10.3389/fimmu.2018.02039 PMID: 30250470
928 132. Roy M, Viginier B, Saint-Michel E, Arnaud F, Ratinier M, Fablet M. Viral infection impacts
929 transposable element transcript amounts in *Drosophila*. PNAS. 2020 Jun;117(22):12249–57.
930 https://doi.org/10.1073/pnas.2006106117 PMID: 32434916
931 133. Wada M, Lokugamage KG, Nakagawa K, Narayanan K, Makino S. Interplay between coronavirus,
932 a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway. PNAS. 2018
933 Oct;115(43):e10157–66. https://doi.org/10.1073/pnas.1811675115 PMID: 30297408
934 134. Wang W, Xiog L, Wang P, Wang F, Ma Q. Major vault protein plays important roles in viral
935 infection. IUBMB Life. 2020 Apr;74(4):624–31. https://doi.org/10.1002/iub.2200 PMID: 31769934

135. Steiner E, Holzmann K, Pirker C, Elbling L, Micksche M, Sutterluty F, et al. The major vault protein is responsive to and interferes with interferon-γ-mediated STAT1 signals. J Cell Sci. 2006 Oct;119:459–69. https://doi.org/10.1242/jcs.02773 PMID: 16418217

136. Li F, Chen Y, Zhang Z, Ouyang J, Wang Y, Yan R, 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. 2015 Dec;43(21):10321–37. https://doi.org/10.1093/nar/gkv1078 PMID: 26490959

137. Bellon M, Nicot C. Regulation of telomerase and telomeres: Human tumor viruses take control. J Natl Cancer Inst. 2008 Jan;100(2):98–108. https://doi.org/10.1093/jnci/djm269 PMID: 18182620

138. Reghunathan R, Jayapal M, Hsu LY, Chng HH, Tai D, Leung BP, et al. Expression profile of immune response genes in patients with severe acute respiratory syndrome. BMC Immunol. 2005 Jan;6:2. https://doi.org/10.1186/1471-2172-6-2 PMID: 15655079

139. Berlutti F, Pantanella F, Natalizi T, Frioni A, Paesano R, Polimeni A, et al. Antiviral properties of lactoferrin—A natural immunity molecule. Molecules. 2011 Aug;16(8):6992–7018. https://doi.org/10.3390/molecules16086992 PMID: 21847071

140. Adusumilli NC, Zhang D, Friedman JM, Friedman AJ. Harnessing nitric oxide for preventing, limiting and treating the severe pulmonary consequences of COVID-19. Nitric Oxide. 2020 Oct;103:4–8. https://doi.org/10.1016/j.niox.2020.07.003 PMID: 32681986

141. Perrone LA, Belser JA, Wadford DA, Katz JM, Tumpey TM. Inducible nitric oxide contributes to viral pathogenesis following highly pathogenic influenza virus infection in mice. J Infect Dis. 2013 May;207(10):1576–84. https://doi.org/10.1093/infdis/jit062 PMID: 23420903

142. Wan G, Zhaorigetu S, Liu Z, Kaini R, Jiang Z, Hu CAA. Apolipoprotein L1, a novel Bcl-2 homology domain 3-only lipid-binding protein, induces autophagic cell death. J Biol Chem. 2008 Aug;282(31):21540–9. https://doi.org/10.1074/jbc.M800214200 PMID: 18505729

143. Peng Y, Wan L, Fan C, Zhang P, Wang X, Sun J, et al. Cholesterol metabolism—Impact for SARS-CoV-2 infection prognosis, entry, and antiviral therapies. medRxiv. 2020 Apr. https://doi.org/10.1101/2020.04.16.20068528

144. Rebello CJ, Kirwan JP, Greenway FL. Obesity, the most common comorbidity in SARS-CoV-2: is leptin the link? Int J Obes (Lond). 2020 Jul;44:1810–7. https://doi.org/10.1038/s41366-020-0640-5 PMID: 32647360

145. Zhang AJX, To KKW, Li C, Lau CCY, Poon VKM, Chan CCS, 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. 2013 Apr;207(8):1270–80. https://doi.org/10.1093/infdis/jit031 PMID: 23325916

146. Chesarino NM, McMichael TM, Yount JS. E3 ubiquitin ligase NEDD4 promotes influenza virus infection by decreasing levels of the antiviral protein IFITM3. PLOS Pathog. 2015 Aug;11(8):e1005095. https://doi.org/10.1371/journal.ppat.1005095 PMID: 26263374

147. Shi G, Ozog S, Torbett BE, Compton AA. mTOR inhibitors lower an intrinsic barrier to virus infection mediated by IFITM3. PNAS. 2018 Oct;115(43):e10069–78. https://doi.org/10.1073/pnas.1811892115 PMID: 30301809

148. Xu Q, Zhu N, Chen S, Zhao P, Ren H, Zhu S, et al. E3 ubiquitin ligase Nedd4 promotes Japanese encephalitis virus replication by suppressing autophagy in human neuroblastoma cells. Sci Rep. 2017 Mar;7:45375. https://doi.org/10.1038/srep45375 PMID: 28349961

149. Yu J, Li M, Wilkins J, Ding S, Swartz TH, Esposito AM, et al. IFITM proteins restrict HIV-1 infection by antagonizing the envelope glycoprotein. Cell Rep. 2015 Oct;13:145–56. http://doi.org/10.1016/j.celrep.2015.08.055 PMID: 26387945

150. Zhu X, He Z, Yuan J, Wen W, Huang X, Hu Y, et al. IFITM3-containing exosome as a novel mediator for anti-viral response in dengue virus infection. Cell Microbiol. 17(1):105–18. https://doi.org/10.1111/cmi.12339 PMID: 25131332

151. Huang IC, Bailey CC, Weyer JL, Radoshitzky SR, Becker MM, Chiang JJ, et al. Distinct patterns of IFITM-mediated restriction of filoviruses, SARS coronavirus, and influenza A virus. PLOS Pathog. 2011 Jan;7(1):e1001258. https://doi.org/10.1371/journal.ppat.1001258 PMID: 21253575

152. Zhao X, Sehgal M, Hou Z, Cheng J, Shu S, Wu S, et al. Identification of residues controlling restriction versus enhancing activities of IFITM proteins on entry of human coronaviruses. J Virol. 2018 Mar;92(6):e01535–17. https://doi.org/10.1128/JVI.01535-17 PMID: 29263263

153. Hachim MY, Al Heialy S, Hachim IY, Halwani R, Senok AC, Maghazachi AA, et al. Interferon-induced transmembrane protein (IFITM3) is upregulated explicitly in SARS-CoV-2 infected lung epithelial cells. Front Immunol. 2020 Jun;11:1372. https://doi.org/10.3389/fimmu.2020.01372 PMID: 32595654

154. Yang B, Kumar S. Nedd4 and Nedd4-2: closely related ubiquitin-protein ligases with distinct physiological functions. Cell Death Differ. 2009 Jun;17:68–77. https://doi.org/10.1038/cdd.2009.84 PMID: 19557014

155. Han Z, Lu J, Liu Y, Davis B, Lee MS, Olson MA, et al. Small-molecule probes targeting the viral PPxY-host Nedd4 interface block egress of a broad range of RNA viruses. J Virol. 2014 Apr;88(13):7294–306. https://doi.org/10.1128/JVI.00591-14 PMID: 24741084

156. An H, Krist DT, Statsyuk AV. Crosstalk between kinases and Nedd4 family ubiquitin ligases. Mol BioSyst. 2014 Jan;10:1643–57. https://doi.org/10.1039/C3MB70572B PMID: 24457516

157. Maaroufi H. SARS-CoV-2 encodes a PPxY late domain motif that is known to enhance budding and spread in enveloped RNA viruses. bioRxiv. 2020 Apr. https://doi.org/10.1101/2020.04.20.052217

158. Isaacson MK, Ploegh HL. Ubiquitination, ubiquitin-like modifiers, and deubiquitination in viral infection. Cell Host Microbe. 2009 Jun;5:559–70. https://doi.org/10.1016/j.chom.2009.05.012 PMID: 19527883

159. Zinngrebe J, Montinaro A, Peltzer N, Walczak H. Ubiquitin in the immune system. EMBO Rep. 2013 Nov;15(1):28–45. https://doi.org/10.1002/embr.201338025 PMID: 24375678

160. Steimer L, Klostermeier D. RNA helicases in infection and disease. RNA Biol. 2012 Jun;9(6):751–771. https://doi.org/10.4161/rna.20090 PMID: 22699555

161. Sharma A, Boris-Lawrie K. Determination of host RNA helicases activity in viral replication. Methods Enzymol. 2012 Jun;511:405–35. https://doi.org/10.1016/B978-0-12-396546-2.00019-X PMID: 22713331

162. Umate P, Tuteja N, Tuteja R. Genome-wide comprehensive analysis of human helicases. Commun Integr Biol. 2011 Jan–Feb;4(1):118–37. https://doi.org/110.4161/cib.4.1.13844 PMID: 21509200

163. Xu L, Khadijah S, Fang S, Wang L, Tay FPL, Liu DX. The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. J Virol. 2010 Sep;84(17):8571–83. https://doi.org/10.1128/JVI.00392-10 PMID: 20573827

164. Chem JY, Chen WN, Poon KMV, Zheng BJ, Lin X, Wang YX, et al. Interaction between SARS-CoV helicase and a multifunctional cellular protein (Ddx5) revealed by yeast and mammalian cell two-hybrid systems. Arch Virol. 2009 Feb;154(3):507–12. https://doi.org/10.1007/s00705-009-0323-y PMID: 19224332

165. Spiezia L, Boscolo A, Poletto F, Cerruti L, Tiberio I, Campello E, et al. COVID-19-related severe hypercoagulability in patients admitted to intensive care unit for acute respiratory failure. Thromb Haemost. 2020 Jun;120(6);998–1000. https://doi.org/10.1055/s-0040-1710018 PMID: 32316063

166. Ji HL, Zhao R, Matalon S, Matthay MA. Elevated plasmin(ogen) as a common risk factor for COVID-19 susceptibility. Physiol Rev. 2020 Jul;100(3):1065–75. https://doi.org/10.1152/physrev.00013.2020 PMID: 32216698

167. Sottrup-Jensen L, Sand O, Kristensen L, Fey GH. The α-macroglobulin bait region: Sequence diversity and localization of cleavage sites for proteinases in five mammalian α-macroglobulins. J Biol Chem. 1989 Sep;264(27):15781–9. PMID: 2476433

168. Gettins PGW, Hahn KH, Crews BC. α2-macroglobulin bait region variants: A role for the bait region in tetramer formation. J Biol Chem. 1995 Jun;270(23):14160–7. https://doi.org/10.1074/jbc.270.23.14160 PMID: 7539801

169. 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. 2015 Apr;308:L1189–201. https://doi.org/10.1152/ajplung.00028.2015 PMID: 25888573

170. De Bont CM, Boelens WC, Pruijn GCM. NETosis, complement, and coagulation: a triangular relationship. Cell Mol Immunol. 2019 Jan;16(1):19–27. https://doi.org/10.1038/s41423-018-0024-0 PMID: 29572545

171. Noris M, Benigni A, Remuzzi G. The case of complement activation in COVID-19 multiorgan impact. Kidney Intl. 2020 Aug;98(2):314–22. https://doi.org/10.1016/j.kint.2020.05.013 PMID: 32461141

172. Agrawal P, Nawadkar R, Ojha H, Kumar J, Sahu A. Complement evasion strategies of viruses: An overview. Front Microbiol. 2017 Jun;8:1117. https://doi.org/10.3389/fmicb.2017.01117 PMID: 28670306

173. Eddie Ip WK, Chan KH, Law HKW, Tso GHW, Kong EKP, Wong WHS, et al. Mannose binding lectin in severe acute respiratory syndrome coronavirus infection. J Infect Dis. 2005 May;191(10):1697–704. https://doi.org/10.1086/429631 PMID: 15838797

174. Narasuraju T, Tang BM, Herrmann M, Muller S, Chow VTK, Radic M. Neutrophilia and NETopathy as key pathologic drivers of progressive impairment in patients with COVID-19. Front Pharmacol. 2020 Jun;11:870. https://doi.org/10.3389/fphar.2020.00870 PMID: 32581816

175. Zou Y, Yalavarthi S, Shi H, Gockman K, Zuo M, Madison JA, et al. Neutrophil extracellular traps in COVID-19. JCI Insight. 2020 Apr;5(11):e138999. https://doi.org/10.1172/jci.insight.138999 PMID: 32329756

176. Stahel PF, Barnum SR. Complement inhibition in coronavirus disease (COVID)-19: A neglected therapeutic option. Front Immunol. 2020 Jul;11:1661. https://doi.org/10.3389/fimmu.2020.01661 PMID: 32733489

177. De Costa MG, Poppelaars F, van Kooten C, Mollnes TE, Tedesco F, Wurzner R, et al. Age and sex-associated changes of complement activity and complement levels in healthy Caucasian population. Front Immunol. 2018 Nov;9:2664. https://doi.org/10.3389/fimmu.2018.02664 PMID: 30515158

178. Khomich OA, Kochetkov SN, Bartosch B, Ivanov AV. Redox biology of respiratory viral infections. Viruses. 2018 Aug;10(8):392. https://doi.org/10.3390/v10080392 PMID: 30049972

179. Schneider D, Ganesan S, Comstock AT, Meldrum CA, Mahidhara R, Goldsmith AM, et al. Increased cytokine response of rhinovirus-infected airway epithelial cells in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2010 Aug;182(3):332–40. https://doi.org/10.1164/rccm.200911-1673OC PMID: 20395558

Supplementary Information

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				

1085

1086 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				

1089