bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In silico approach toward the identification of unique peptides from viral protein infection: Application to COVID-19

Benjamin C. Orsburn[1*¥], Conor Jenkins[2], Sierra M. Miller[3], Benjamin A Neely[1], Namandje N Bumpus[4*]

[1]Proteomic und Genomic Sciences, LLC, Baltimore, MD; USA
[2]Hood College Department of Biology, Frederick, MD; USA
[3]Millersville University, Science Lane, Millersville, PA; USA
[4] Department of Medicine, Johns Hopkins University, Baltimore, MD; USA
*Corresponding:
Benjamin C. Orsburn; orsburn@vt.edu
Namandje M. Bumpus; nbumpus@jhmi.edu

Footnote:
[¥]Lead contact

Summary

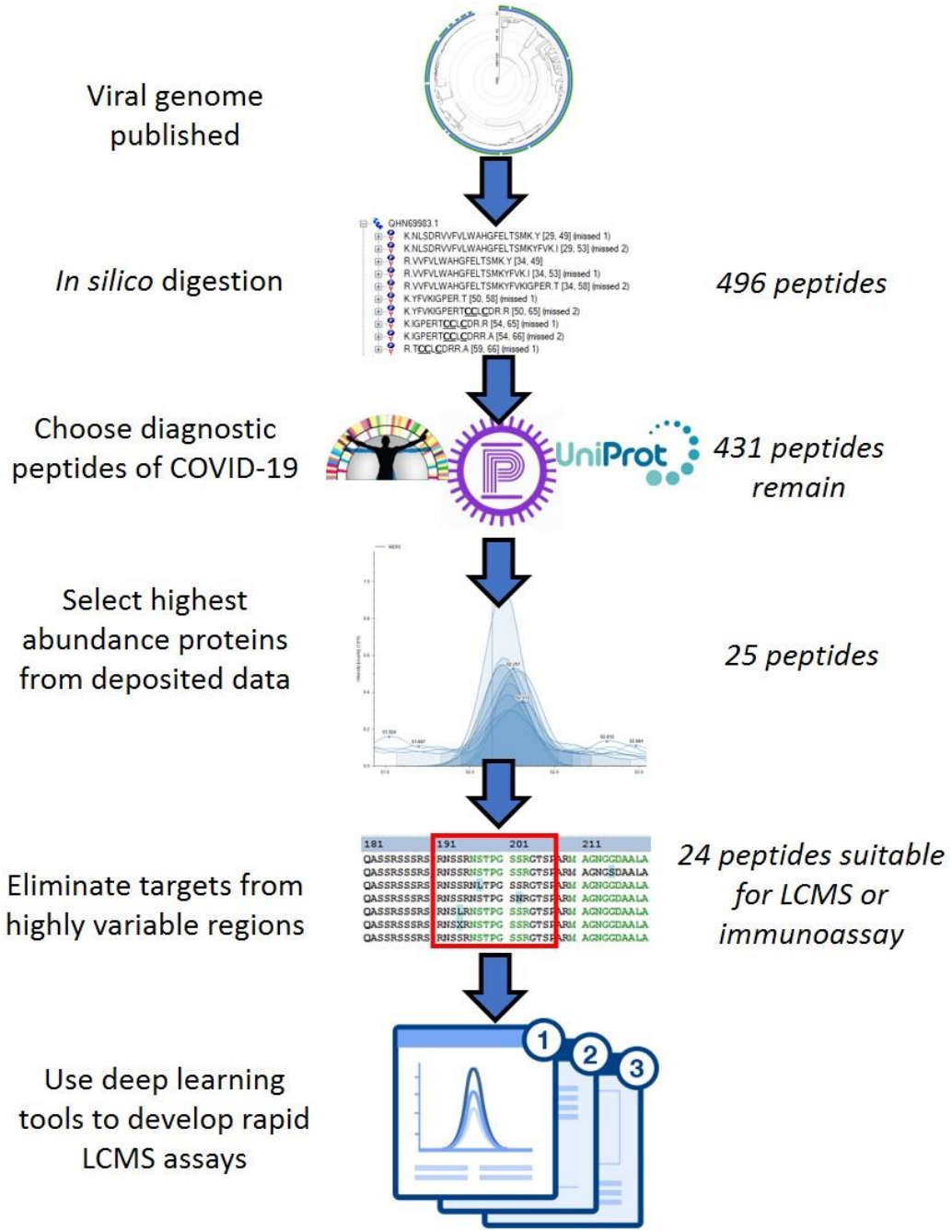
We describe a method for rapid in silico selection of diagnostic peptides from newly described viral pathogens and applied this approach to SARS-CoV-2/COVID-19. This approach is multi-tiered, beginning with compiling the theoretical protein sequences from genomic derived data. In the case of SARS-CoV-2 we begin with 496 peptides that would be produced by proteolytic digestion of the viral proteins. To eliminate peptides that would cause cross-reactivity and false positives we remove peptides from consideration that have sequence homology or similar chemical characteristics using a progressively larger database of background peptides. Using this pipeline, we can remove 47 peptides from consideration as diagnostic due to the presence of peptides derived from the human proteome. To address the complexity of the human microbiome, we describe a method to create a database of all proteins of relevant abundance in the saliva microbiome. By utilizing a protein-based approach to the microbiome we can more accurately identify peptides that will be problematic in COVID-19 studies which removes 12 peptides from consideration. To identify diagnostic peptides, another 7 peptides are flagged for removal following comparison to the proteome backgrounds of viral and bacterial pathogens of similar clinical presentation. By aligning the protein sequences of SARS-CoV-2 field isolates deposited to date we can identify peptides for removal due to their presence in highly variable regions that may lead to false negatives as the pathogen evolves. We provide maps of these regions and highlight 3 peptides that should be avoided as potential diagnostic or vaccine targets. Finally, we leverage publicly deposited proteomics data from human cells infected with SARS-CoV-2, as well as a second study with the closely related MERS-CoV to identify the two proteins of highest abundance in human infections. The resulting final list contains the 24 peptides most unique and diagnostic of SARS-CoV-2 infections. These peptides represent the best targets for the development of antibodies are clinical diagnostics. To demonstrate one application of this we model peptide fragmentation using a deep learning tool to rapidly generate targeted LCMS assays and data processing method for detecting CoVID-19 infected patient samples.

Keywords

SARS-CoV-2, COVID-19, 2019-nCOV, proteomics, LCMS, mass spectrometry, peptides

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Graphical Abstract:



bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The identification of peptides expressed unique to pathogens is required for the development of diagnostic assays as well as vaccine targets. Antibody based techniques such as enzyme linked immunosorbent assay rely on antibodies raised against specific peptide targets. Mass spectrometry (MS) based diagnostic assays typically require many rounds of optimization to identify peptides that are unique in both sequence or in chemical characteristics to distinguish them from the complex host background.[1]

The detection of viral proteins in body fluids can be a rapid and specific diagnostic for infection in severe acute respiratory syndrome (SARS).[2–4] During the 2003 (SARS) outbreak, non-MS based methods of protein detection proved to be more successful[5,6] than LCMS methods.[7–9] Non-MS based methods, such as western blots, enzyme-linked immunosorbent assays (ELISAs), and protein arrays, rely on antibodies for the detection of proteins. Given recent studies concerning high variability in antibody production, LCMS-based methods are an attractive alternative approach for the rapid identification of small molecules, proteins, and peptides in clinical settings where consistency is paramount.[10,11]

In the 15 years since the 2003 SARS outbreak, LCMS technology has experienced a revolution led primarily by increases in the speed, sensitivity, and resolution of MS instruments. Today, protein array and antibody-based methods are falling out of favor in both research and clinical diagnostics, due in large part to the improvements in LCMS technology.[12,13] A review of this growth by Grebe and Singh described a clinical lab with no LCMS systems in 1998 that completed over 2 million individual LCMS clinical assays in 2010.[14] Incremental improvements in rapid sample preparation techniques, chromatography, and data processing have also contributed to the increasing use of LCMS-based clinical testing. A 2013 study demonstrated the level of advance by identifying 4,000 yeast proteins in one hour of LCMS run time, identifying approximately 75 proteins/min at a rate 100 times faster than studies a decade prior.[15]

Targeted peptide-centric assays are advantageous when sensitivity is paramount over the quantity of identified targets. Targeted assays often rely on tandem MS with high speed, but relatively low accuracy, quadrupoles.[16] Quadrupoles can be used to select ions for fragmentation with quantification of fragment ions by other quadrupoles in single reaction monitoring (SRM). They can also be used in conjunction with high resolution systems in single ion monitoring (SIM) and parallel reaction monitoring (PRM). SRM relies on fragmentation which requires *a priori* the mass to charge ratio (m/z) of both the peptide ion of interest and its dominant ions produced during fragmentation. Commonly, the collision energy is optimized for each peptide fragment in order to maximize efficiency and signal. SIM uses higher resolution scans in lieu of fragmentation, which requires *a priori* only an approximate m/z ratio for the peptide. In SIM targeted assays, the peptide ion's exact mass is extracted post-run during data processing. Two studies have shown that SIM scans with ≥60,000 resolution produce selectivity comparable to SRM.[17,18] PRM targeted assays combine fragmentation and high resolution molecule monitoring. In contrast to SIM, isolated ions are subjected to fragmentation. Post-run data processing then calculates quantification from the high-resolution accurate mass of the fragment ions. For proteins with only one available peptide target, PRM is the most selective technique in modern proteomics.[19,20]

Due to the high sensitivity of targeted LCMS assays, background signal is of central concern. Peptides must be selected that are unique in one or more of the following characteristics to be truly diagnostic; m/z, peptide sequence, or hydrophobicity. Due to the complexity of the human background thousands of peptides may be possible that are within 0.1 Da in m/z.[21] In the case of ions with high m/z similarity, fragmentation to reveal peptides with sequences of low homology to the background can be used to easily

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

discern these targets. The elution time of the peptides, once characterized and monitored with internal standards can be utilized to distinguish between even highly similar peptides, including sequences that are chemically modified but on alternative residues.[22]

To address these concerns an increasing number of in silico tools have been developed that can speed the identification of suitable LCMS targets. Tools such as Picky,[23]the Phosphopedia,[22]the Peptide Manager, and Purple[1], and Skyline[24] allow the removal of peptides from theoretical proteome backgrounds or the intelligent selection of peptide targets based on other parameters. Peptide retention times can be calculated based on amino acid hydrophobicity calculations, although the inclusion of internal reference calibration standards such as ProCal are necessary for adjustments to different chromatographic conditions.[25]

In traditional LCMS targeted assay development, once suitable targets are identified, peptides are then synthesized, and ideal fragment ions are selected for use following rounds of selection and optimization. One way to improve development time is by using databases of experimental fragmentation data such as the PeptideAtlas.[26] However, peptide fragmentation follows specific energetic patterns, resulting primarily in fragments caused by separation at the peptide bond. It is therefore possible to create theoretical spectral libraries *in silico* from peptide sequence alone. Theoretical spectral libraries are especially useful when the biological samples are unavailable. New tools that employ deep learning algorithms have been demonstrated to produce theoretical MS2 spectra superior to previous prediction models and, in the absence of true experimental data, are the best resources currently available.[27,28] These deep learning algorithms can learn from vast libraries of experimental data to predict the fragmentation patterns of new peptide sequences that they are given. One such algorithm, PROSIT, uses the vast synthetic human peptide libraries, from the ProteomeTools project[29] for its training dataset. Due to the high quality of the 450,000 synthetic peptides experimentally fragmented in ProteomeTools to date, PROSIT has been demonstrated to create spectral libraries that are, in some cases, superior to experimentally derived in-house spectral libraries.[30]

In this study, we describe a *in silico* approach to identify unique peptides presented during viral infection of human cells. Using a stepwise elimination approach, we outline strategies to begin with a theoretical protein sequence and remove peptides with homology in clinically relevant systems. By removing peptides with homology to human and the relevant human microbiome, we can eliminate many targets that would produce false positive tests for the virus in any healthy person. We further improve on this by removing peptides that could be present in diseases with a similar clinical presentation. By utilizing the data in public repositories, we can determine the protein targets with the highest abundance as well as the highest likelihood of being diagnostic in different assays. We further apply this method toward the analysis of SARS-CoV-2 infected materials by identifying regions of variability in field isolates identified to date and flag these peptides, as diagnostics targeted on these peptides may lead to false negatives. The restricted target list may be utilized toward the generation of immunological based assays such as ELISA or immunoswabs[2] or provide peptides of value as vaccine targets.

To demonstrate an application of this approach we use the peptides identified in this work to create materials to facilitate LCMS studies of SARS-CoV-2 infected tissues. We provide along with this work necessary resources for both DDA and DIA investigation of these materials through the production of spectral libraries and a list of predicted PTMs investigators should consider when analyzing these materials. The peptide spectral libraries are further utilized to create transition lists optimized for

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

hardware from 5 LCMS instrument vendors. All materials and methods described herein are available as supplemental material to this publication. The supplemental material includes protein sequences (FASTA files), predicted PTMs, theoretical MS2 spectral libraries, instrument methods and targeted method data processing templates.

Our work demonstrates not only the feasibility of this approach, but also its ability to rapidly develop methods even in the face of limitation of access to sample experimental data. We use the example of SARS-CoV-2 viral protein detection to underscore the power of today's protein informatics tools in responding to an urgent public health crisis.

Results and Conclusions

Comparison of LCMS to Alternative Pathogen Detection Methods

Genetics based assays that rely on the polymerase chain reaction (PCR) have well established sensitivity due to the ability to amplify nucleotide chains. Primer design must be optimized and may be adversely affected by unknown complex matrices such as in human stool where the true depth of the microbiome may be unknown.[39] PCR-based assays have also been considered cost-prohibitive in some areas of clinical microbiology compared to other assays.[40] A comparison of the relative sensitivity of PCR and protein based assays is challenging due to the differences in nomenclature and in the actual targets being analyzed. A comparison of detection limits of real-time (RT) PCR and other methods for *Giardia* pathogens in human stool determined revealed conflicting results. When compared to a rapid immunoassay for *Giardia lambia*, PCR was found inferior to both a rapid immunoassay and manual microscopic analysis. A later similar work compared assays for *Giardia intestinalis* and concluded that RT-PCR was a superior method to ELISA or microscopic analysis. In addressing the detection of emerging pathogens PCR requires several steps to be deployed *en masse*. First, a genomic sequence must be generated from which primers can be designed. These primers must be evaluated for cross-reactions in clinical samples and these reagents must be generated in bulk, subjected to quality control (QC) procedures and shipped to the site of testing. The lead time for these reagents has been well-publicized in addressing the COVID-19 pandemic, as has the challenges in quality of the primers for an assay deployed by the US CDC in February, 2020.[41] A summary of relevant assays is shown in Table 1. ELISA based assays for SARS-CoV display the least sensitivity of the these assays with detections in nanograms/mL.[3] ImmunoSwab assays first described for SARS-CoV and recently shown to provide rapid utility for SARS-CoV-2 report sensitivity in the former case of 10 picograms/mL.[2] Due to the high similarity of the two platforms we hypothesized that the SARS-CoV-2 assay will be of similar performance. Two examples of optimized peptide pharmacokinetics assays in serum have demonstrated the utility of rapid LCMS based targeted assays that can achieve limits of detection and assay time comparable to current validated assays.

Assay and Reference	**Matrix**	**Detection limit**	**Assay Time**
SARS-CoV Nucleocapsid ImmunoSwab[2]	Nasopharyngeal aspirate medium	10 picogram/mL	45 min
SARS-CoV-2 Nucleocapsid Immunoswab[42]	Nasopharyngeal aspirate/urine	Unstated / based on assay above	10 min
SARS-CoV Nucleocapsid ELISA[4]	Nasopharyngeal aspirate	2.5 nanogram/mL	4-hour EST

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SARS-CoV Nucleocapsid ELISA[4]	Stool	9.0 nanograms/mL	4-hour EST
SARS-CoV Protein Microarray[43]	Human Serum	1:64,000 Dilution of Positive Serum	~2 hours
Inflammatory Mediator Protein Microarray[44]	Human Serum	20 picogram/mL	~3 hours
CDC-SARS-CoV-2[45]	Nasopharyngeal aspirate	3.2 copies/µL	4 hours
SARS-CoV-2 RT-Lamp/CAS12[45]	Nasopharyngeal aspirate	10 copies/µL	45 min
Enkephalin LCMS[46]	Plasma	10pg/mL	6.5 min
Orexin LCMS[47]	CSF	4pg/mL	7 min

Table 1. A comparison of relevant assays and their respective limits of detection.

Theoretical peptides

In shotgun proteomics, proteins are first digested into smaller peptide fragments that are more easily detected by the instrument. Given its widespread use, high efficiency and speed of digestion, we chose to develop methods that exclusively use the proteolytic enzyme trypsin, which produces “tryptic peptides." Sequencing grade trypsin exhibits high efficiency cleavage at unmodified (1) arginine and (2) lysine residues unless followed by a proline. Trypsin also has the advantage of leaving a terminal basic residue at the cleavage site, which increases the likelihood of complete fragment ion coverage from the charged terminal.[48] Due to these reasons, trypsin is utilized unless the protein sequence has an abnormally high or low number of lysine or arginine residues. A very high frequency of the residues (such as Lysine-rich proteins) will create very short peptides that could be uninformative for protein identification. A very low frequency of the residues will create very large peptides, or undigested (“intact”) proteins in some cases, that are difficult to detect and fragment. Our theoretical trypsin digest of the 2019-nCoVpFASTA1 database produced tryptic peptides with average lengths of 8 to 18 amino acids. These results indicate that trypsin digestion is an appropriate choice for detection of these viral proteins.

Removal of proteome theoretical background

The recently described program Purple was used for background subtraction of relevant theoretical proteomes using default parameters. Peptides were considered relevant with a sequence between 6 and 24 amino acids in length. Using these parameters, the COVID-19 UniProt FASTA produced 496 unique peptide sequences. In the first stage analysis we eliminated human peptides from further consideration. By subtracting all the human proteome background, we find that two peptide sequences are identical in SARS-CoV-2 and the human sequence. For example, the peptide sequence TLLSLR is found in the Replicase 1a protein of SARS-CoV-2 as well as in 4 human proteins, including Focadhesin (S4R400). These identical peptides should be excluded due from any further analysis due to these identical sequences. Purple was also used to remove peptides of identical amino acid composition with alternate sequence according the to default parameters previously described,[1] removing another 45 peptides from this list.

In the second stage we developed a pipeline to remove proteins derived from the complex human microbiome from consideration. The NIH Microbiome initiative has created vast libraries using genomic sequencing technologies to identify the organisms present in and on *Homo sapiens*.[49,50] These libraries do

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not, however, address protein abundance. In order to create a relevant proteome background for subtraction we performed a meta-analysis of a recent in depth study of the human saliva metaproteome.[38] The raw instrument data was searched using the complete NIH saliva microbiome FASTA database of 536,284 protein sequences derived from 249 organisms as well as the complete UniProt catalog of reviewed bacterial and human protein sequences to date as well as the common list of protein lab contaminants. A new database was created of all protein sequences that contained at least one peptide spectral match (PSM), resulting in 29,816 unique protein sequences. Although it is possible that the proteins provided in the NIH FASTA are possible in the saliva microbiome, without adjusting for protein abundance and removing redundancy for shared proteins between organisms the proteome background may be misconstrued as more complex than it truly is. By removing from consideration, the 29,816 proteins of most relevant abundance we can more accurately measure the proteome background (data not shown.) Subtraction of the human microbiome background removed another 12 peptides that should be excluded in oral diagnostics or in the development of vaccine targets for SARS-CoV-2.

For the third stage subtraction we narrowed the peptide list to those targets that would be diagnostic for SARS-CoV-2 infection but would not cross-react with organisms of similar symptoms or clinical presentation. We created separate protein FASTA databases from UniProt 2020_01 for the viral pathogens, rhinovirus, influenza, pneumoniae, pneumophila and respiratory syncytial virus (RSV) as well as staph aureus and streptococcus species. The subtraction of these proteins from the background removed another six peptides from consideration.

The final stage of proteome background subtraction focused on the removal of peptides that could be derived from other coronavirus strains, including MERS, SARS-CoV and other protein sequences. Unsurprisingly, this proved to be the largest removal step, with 306 peptides subtracted from consideration and only 125 remaining as clearly unique and diagnostic of SARS-CoV-2. These results are available as Supplemental X and visualized in Figure 1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

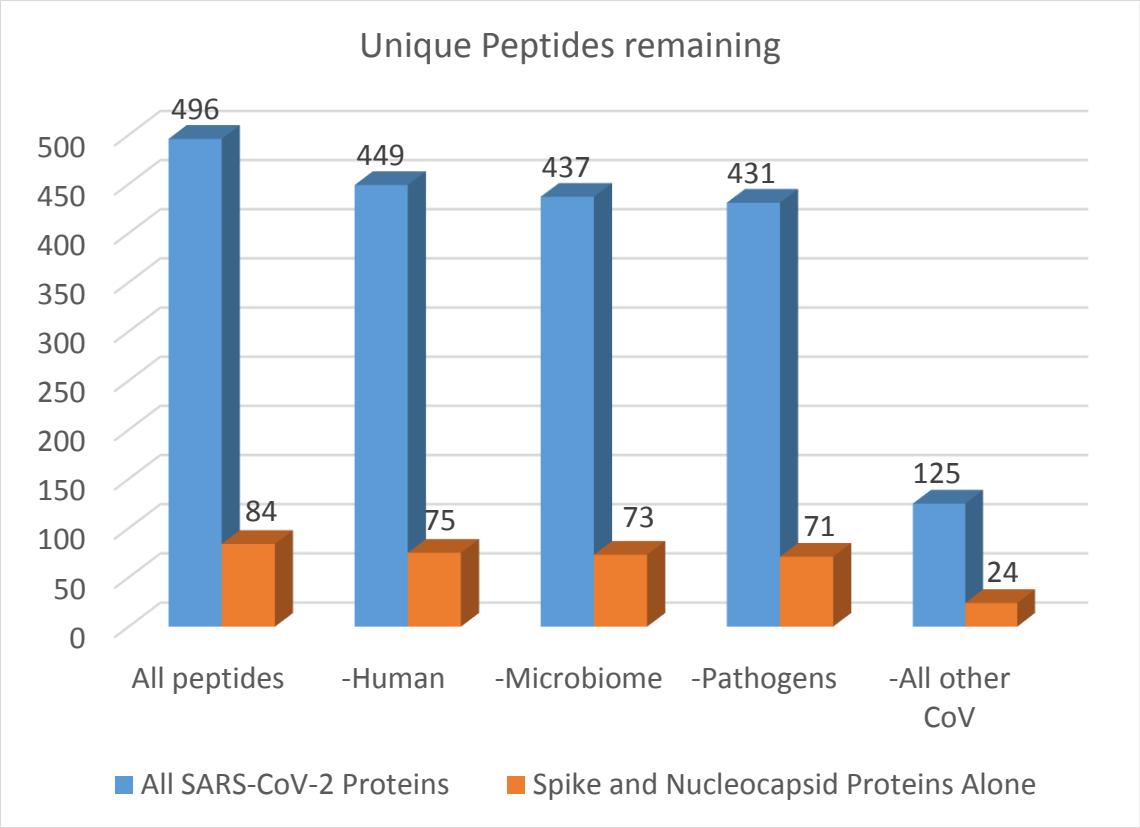


Figure 1. A summary of the peptide numbers as they are removed from an increasingly complex theoretical proteome background

Utilization of Publicly deposited proteomics data to identify ideal protein targets

As shown in Table X, several studies for related coronavirus strains had been deposited in public repositories. In March of 2020, three new studies were released featuring work with SARS-CoV-2. In order to determine if some viral proteins would be of higher relative abundance in human infection, we first began by searching previously released, but unpublished, files from MERS-CoV infection of Calu3 cells. As shown in Figure 2A more peptides were identified from the nucleocapsid protein than any other protein, followed by the Spike protein. When the first SARS-CoV-2 study was released we reprocessed this study and observed the same trend, as shown in Figure 2B. Although not strictly quantitative, counting the number of peptides identified in a global proteomics experiment has been historically used as a metric for approximating relative protein abundance in a sample. These results suggest that the Nucleoprotein and Spike Glycoprotein may be the highest abundance SARS-CoV-2 proteins present in human infections

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

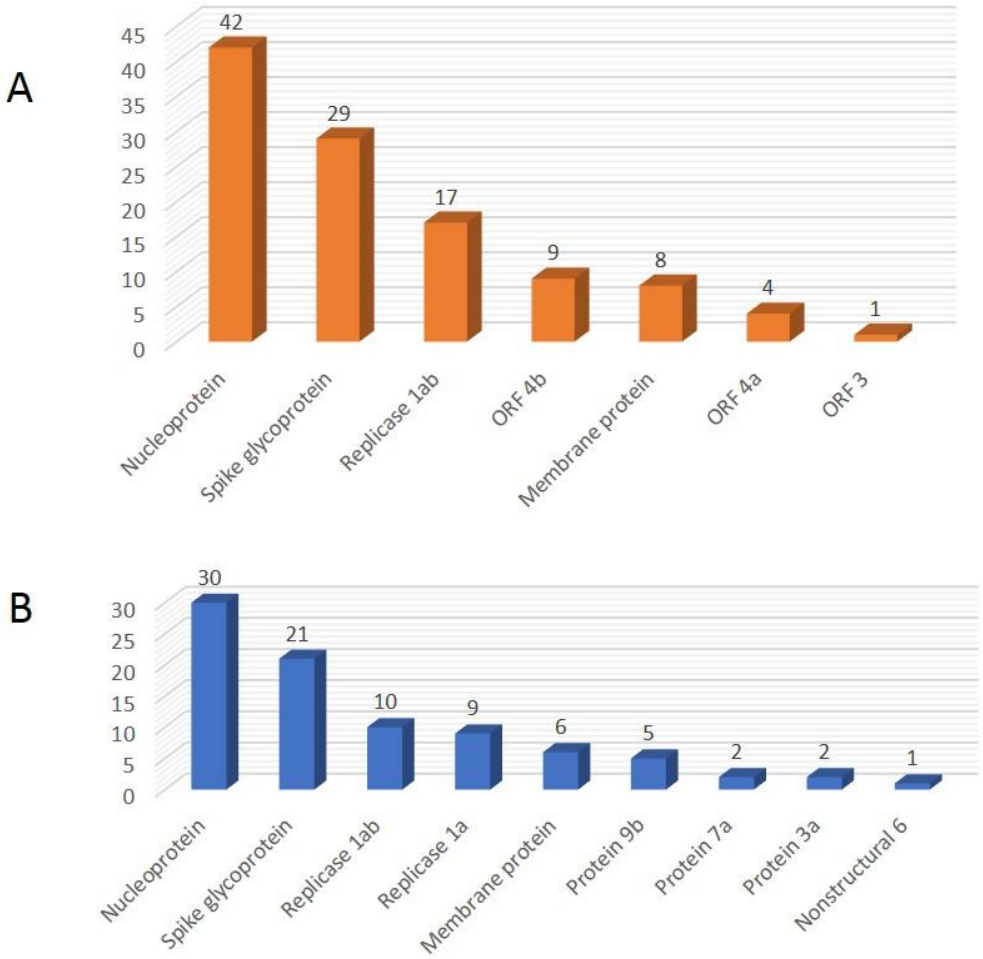


Figure 2A. The number of unique peptides identified in a metanalysis of MERS-CoV infection for each protein in human cells. **2B**. The same analysis performed on the first proteomic study released on SARS-CoV infected human samples.

Identification of peptides most likely to be affected by evolutionary pressures

To identify peptides likely to make poor targets due to genomic variability, we performed a 2-stage approach. The first consisted of aligning and visualizing protein sequences being deposited in NCBI MetAlign and AliView software in R. These tools allow the color coding and rapid flagging of regions of alternative protein sequences in those aligned. A visualization of one such region from the nucleocapsid protein is shown in Figure 3A. To verify these results, we processed proteomics data derived from cloned nucleocapsid proteins utilizing every unique full length nucleocapsid protein sequence available in UniProt. Figure 3B mirrors the results from the visualization analysis. We identify one peptide region in the nucleocapsid and two in the spike glycoprotein that may provide false positives for diagnostic assays.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

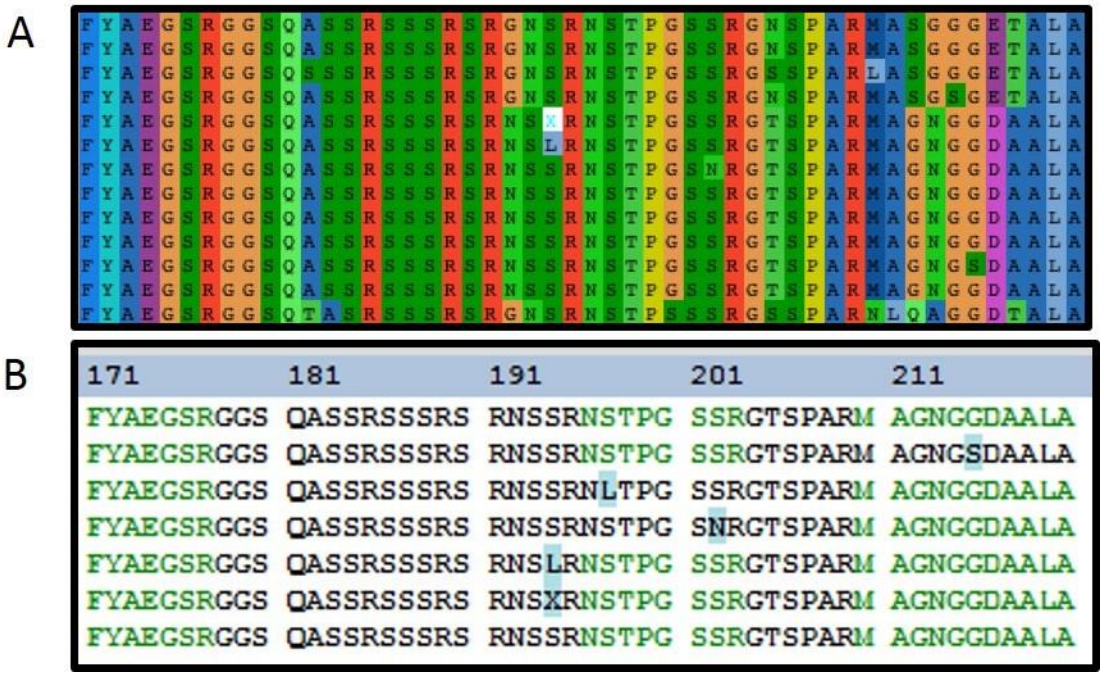


Figure 3A. AliView output predicting protein variability in a region of the nucleocapid protein sequence. **3B**. Comparison of all available annotated NCBI protein sequences to recently deposited proteomics data that reflects the variability in this viral protein region.

Generation of Spectral Libraries and Experimental Methods

In the absence of any experimentally derived spectra from SARS-CoV-2 proteins we used the Prosit deep learning tool to generate theoretical spectra for each peptide. Figure 4 is an example of a nucleocapsid peptide predicted by Prosit on January 27th compared to the first experimental data available for a cloned protein from this virus. The top panel is the experimental spectra and the bottom is a mirror plot showing the predicted fragments and their expected relative intensities. Prosit accurately predicted the absence of fragments y14 and y15 as well as the presence of only b ions central to the peptide sequence. Furthermore, the abundance predictions allow for the selection of the most appropriate ions for MS/MS assays such as PRM and MRM. As one utility of the methods described herein, we used the Skyline software to create transition lists optimized specifically for each of the Skyline-compatible triple quadrupole instruments. While minor modifications are required for SCIEX, Agilent, and Shimadzu instruments, Waters Xevo and Thermo instruments use identical parameters for transition list design. Most modern triple quadrupole instruments are capable of 500 SRMS/sec and fully permit the use of 2,000 transition lists, as provided here. For older instruments that lack this scan speed or that require higher dwell times, the transition lists included in the supplemental methods may be reduced by the end user accordingly.

PRM methods monitor multiple transitions simultaneously but at a time cost. The highest scan speed currently available in Orbitrap instruments is 48 scans per second and is only available on the Exploris 480 system (data not shown). In order to achieve maximum sensitivity, higher fill times are often required for these instruments. We chose to utilize three peptides/protein for these methods. Alternative peptides

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

can be selected from the Skyline files provided (Supplemental Information) or by selecting peptide mass targets from the other transition lists. A summary of the resources provided in this work are available in the Supplemental Information.

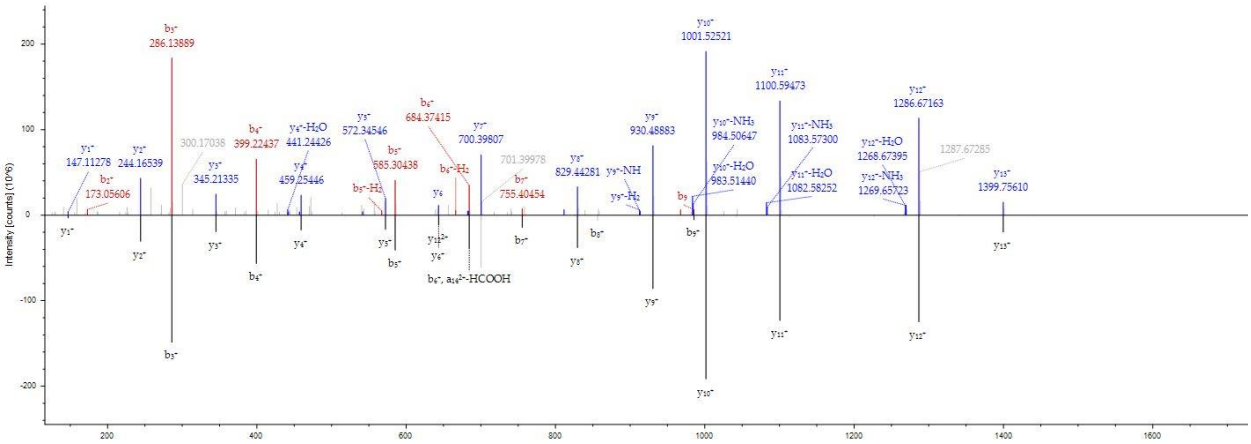


Figure 4. A mirror plot demonstrating the Prosit derived MS/MS spectra (bottom half) against the first known fragmentation pattern for this SARS-CoV-2 nucleocapsid peptide in the literature (top half)

Resources for untargeted proteomics methods

In the course of this work, we created the minimal resources required for untargeted shotgun analysis of SARS-CoV-2 infected materials. These resources are made available in the supplemental material including the full Prosit Spectral libraries, FASTA files and predicted post-translational modifications. The Prosit spectral libraries (Supplemental Material) enable the interrogation of DIA data and may be used for DDA experiments that employ tools such as the MSPepSearch (NIST).[51] DDA data requires only the protein FASTA file and a list of PTMs that may be present in the sample. Our analysis using ModPred predicted 17 possible PTMs (Supplemental Table 1). Amidation was the most frequent predicted PTM, but there is no known biological mechanism that we could derive from a survey of the literature. Palmitoylation, the second most frequent predicted PTM, is a well characterized viral PTM with critical functions in human immunodeficiency virus (HIV), human herpes virus (HHV), and influenza virus infectivity.[52–54] During the assembly of this manuscript, proteomics data became available from the infection of *Chlorocebus sabaeus* (Green monkey) cells with SARS-CoV-2. Multiple phosphorylation sites observed on the nucleocapsid protein from this study were accurately predicted as high confidence sites by the ModPred software, including S176 and S205, further strengthening our confidence in this program for this function (Supplemental Table 1).

Description	Date of Deposit	Repository and ID
SAR-CoV-2 Infection in *Chlorocebus sabaeus cells*	3/25/2020	Zenodo 10.5281/zenodo.3722590
SARS-CoV-2 cloned proteins protein-protein interactions	3/23/2020	ProteomeXchangePXD018117
SARS-CoV-2 infection with tandem mass tag calibrator labeling	3/11/2020	ProteomeXchangePXD017710

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Calu-3 proteome response to an infectious clone of MERS-CoV	6/12/2015	ProteomeXchangePXD002358
Calu-3 metabolome response to an infectious clone of MERS-CoV	6/12/2015	ProteomeXchange PXD002359
Calu-3 lipidome response to an infectious clone of MERS-CoV	6/12/2015	ProteomeXchange PXD002360
Calu-3 proteome response to an infectious clone of H7N9	6/17/2015	ProteomeXchange PXD002385
Calu-3 metabolome response to an infectious clone of H7N9	6/17/2015	ProteomeXchange PXD002362

Table 2. Publicly available data relevant to SARS-CoV-2 Studies

LCMS Data for SARS-CoV-2

The COVID-19 global health crisis has permitted an unprecedented degree of scientific sharing and progress. Prior to March 11, 2020 no proteomics data existed in any public repository for SARS-CoV-2 infected materials. Table 2 is a summary of the LCMS data available as of the date of this writing for both SARS-CoV-2 and the most closely related coronavirus strains as recently described.[55]

Summary

Using *in silico* methods, we have developed methods for the rapid selection of unique viral peptides for diagnostic assay and vaccine development, using the example of SARS-CoV-2. *In vitro* validation of these peptides and methods is required and outside the scope of this work given our current lack of access to such samples. To demonstrate the utility of our pipeline, we have provided the minimum materials for data processing for both DDA and DIA untargeted proteomics methods with FASTA databases, spectral libraries and by predicting relevant PTMs for consideration. To broaden the number of labs that can apply our methods, we optimized run parameters for widely used LCMS systems compatible with Skyline, representing instruments from five companies. We will continue to refine these resources and post updates to these methods to LCMSmethods.org and invite researchers anywhere in the world to contact us for assistance in further optimization to address this emerging threat.

Acknowledgements

We thank Dr. Matthew Monroe at Pacific Northwest National Laboratory (PNNL) for assistance curating unpublished data in the ProteomeXchange/MASSIVE public repository.

Author Contributions

Conceptualization, B.C.O and N.M.B.; Methodology, B.C.O. and C.J.; Investigation, S.M.M, C.J., B.A.N.; Writing-Original Draft, B.C.O, C.J. and N.M.B.; Writing-Review and Editing, B.C.O, B.A.M and N.M.B., Resources, C.J. and S.A.M.; Supervision, N.M.B. and B.C.O.

Declaration of Interests

The authors declare no competing interests

Methods Details

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Coronavirus FASTA databases

At the date of this writing, only theoretical protein sequences for SARS-CoV-2, are available. These sequences are being acquired and annotated and the result of translation of genomic sequence information. All sequences in this study were obtained from NCBI accession: txd2697049, https://www.ncbi.nlm.nih.gov/protein/?term=txid2697049). Using Proteome Discoverer 2.4 (Thermo), the protein sequences were combined into a single protein FASTA database (2019-nCOVpFASTA1; Supplemental Information), and added to human proteome sequences (UniProt SwissProt Human database; downloaded 2/15/2020) to produce a database including both human and COVID-19 protein sequences (Human_plus_2019-nCOVpFASTA2; Supplemental Information). The UniProt 2020_01 pre-released protein FASTA for SARS-CoV-2 was downloaded on 3/15/2020 is also provided here. (UniProt_202001_prerelease_COVID19. Fasta; Supplemental material)

Publicly available proteomics data from human samples infected with other coronaviruses

Publicly available experiments on other coronavirus experiments were found by searching the ProteomeXchange Consortium web interface (http://www.proteomexchange.org/).[31,32] Clarification of the identity of unpublished data from Pacific Northwest National Laboratory (PNNL) was provided by Dr. Michael Monroe. SARS-CoV-2 labeled proteomics data was made available ahead of publication[33] as ProteomeXchange PXD017710. SARS-CoV-2 protein-protein interaction RAW files were downloaded from ProteomeXchange PXD018117.[34]

Creation of peptide spectral libraries exclusive to SARS-CoV-2 with Prosit deep learning

The 2019-nCOVpFASTA1 (Supplemental Information) was converted to PROSIT peptide format with the EncyclopeDIA[35] software, resulting in an *in silico* peptide digestion (parameters: Charge Range = 2-3; Max missed cleavage = 1, m/z range = 400-1,500; default NCE = 30eV; default charge = +2). PROSIT peptide fragmentation prediction libraries were generated using the PROSIT online web portal (https://www.proteomicsdb.org/prosit/) with the spectral library modeling interface (options: prediction model = Prosit_2019_intensity; iRT prediction model = Prosit_2019_irt). The libraries were exported in NIST MSP text format for use in Skyline.[24,36]

PRM and SRM method development

For SRM transitions, the 2019-nCOVpFASTA1 (Supplemental Information) was imported into Skyline v20.1.0.28 (University of Washington) along with the PROSIT tryptic peptide spectral library. Peptide settings and transitions were optimized within Skyline to reflect the vendor optimization requirements. For Agilent systems, the 20ms default dwell time was selected for the transition settings. For SCIEX instruments, the same dwell time was utilized as well as automatic optimization of the declustering potential and compensation voltage from the transition settings menu. For Waters, Thermo, and Shimadzu systems, no further settings were required for transition list generation. All transition lists were exported as unscheduled 15min methods. For PRM methods, three peptides were selected for each protein due to the increased time per scan relative to SRM methods. Instrument-specific Skyline files are included in the Supplemental information as described in Table 1.

Prediction of PTMs

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PTMs were predicted for the 2019-nCOVpFASTA1 proteins (Supplemental Information) using the ModPred web interface (www.modpred.org; accessed 1/31/2020).[37] All PTMs available at the date of analysis were selected as theoretical sites, and the Basic non evolutionary model was applied. ModPred ranks each PTM as high, medium, or low confidence as previously described.[37] The ModPred web interface can accept a maximum of 5,000 amino acids. In order to analyze YP_009724389.1 the predicted sequence had to be divided into five sequences, using a 100 amino acid overlap to avoid disrupting potential large motifs. The results of ModPred were compiled into a single spreadsheet with all modifications of all confidence levels. A second sheet was created that contained only the high confidence PTMs predicted, as well as a final summary for the counting of predicted high confidence PTM occurrence, provided as Supplemental Material as detailed in Supplemental Table 1.

Evaluation of genomic stability in SARS-COV-2

To ensure that our assay targets did not lie in a mutable portion of the genome, all available SARS-CoV-2 genome constructs (as of March 21st, 2020) were downloaded from NCBI and aligned using MAFFT (sorted fasta output and automatic input detection; ver 7.453) and visualized with AliView (ver 1.26). The protein sequences for the nucleocapsid, polymerase, spike protein, and membrane glycoproteins were collected from NCBI's IPG database. Search terms included: nucleocapsid, N protein, E protein, envelope, polymerase, replicase, rdrp, spike protein, surface protein, S protein. Entries labeled 'hypothetical' were not included. Final alignments visualized included only entries originating directly from the virus (excluding things such as Bat-SARS-CoV, civet, Rhinolophus, etc.). All collected accession numbers can be found in Supplemental File 1. The evaluation of suitable targets was performed by manual review.

Evaluation of Nucleocapsid Protein Stability and Identification of Variable Regions

LCMS RAW instrument files from Gordon *et al.,*[34] were downloaded from ProteomeXchange (Table 2). The files for cloned nucleocapsid protein were processed in PD 2.2 using the default template for Q Exactive Basic ID. The files were searched against the cRAP database, the UniProt 2020_01 FASTA and a compiled database of all full length nucleocapsid protein variants available in NCBI on 3/24/2020. Of the 15 protein sequences available, only 10 were flagged as having unique sequences by PD. The search results were not merged into protein groups in order to facilitate the visualization of protein sequences using the sequence comparison view post processing tool.

Creation of saliva microbiome background interference databases

Files deposited in ProteomeXchange with identifier PXD004319 were utilized to create a saliva microbiome background.[38] The files were searched in Proteome Discoverer 2.2 using the parameters detailed in Supplemental Table 1. The FASTA databases used were the complete Oral Microbiome.FA file from the NIH Microbiome initiative, a custom database of all bacterial species in UniProt 2020_01 as well as the Homo sapiens FASTA from the same, as well as the cRAP database of common lab protein and peptide contaminants. The database marker node was utilized to flag sequences derived from each FASTA. Protein grouping was not employed in this analysis to allow the largest possible sequence variation. All protein sequences with more than one potential peptide spectral match (PSM) were retained and exported. The FASTA database utilities tool in PD 2.2 was used to compile these results into a FASTA that contained only proteins in UniProt standard accession format, entitled Adjusted_Human_Saliva_Microbiome_3_24_20.FASTA (Supplemental Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FASTA databases of pathogens with similar clinical presentation

Protein FASTA databases were downloaded from UniProt 2020_01 to create databases of proteins indicative of pathogens with similar clinical presentation using the following search terms: Coronavirus reviewed [no], Influenza, Middle East Respiratory, Pneumoniae, Respiratory Syncytial Virus, Rhinovirus, Staphylococcus aureus, Streptococcus reviewed [yes].

Automatic removal of interfering peptide targets with Purple

To remove peptides that with homology to those of the COVID-19 theoretical peptides, Purple_portable_0.4.2 for Windows was installed locally. The UniProt 2020_01 release of the SARS-CoV-2 FASTA was placed into the same local file with the FASTA databases from Human and the FASTA databases of pathogens with similar clinical presentation. Two analyses were performed, the first to remove peptides that occur in human alone for use in serum or plasma-based LCMS assays. A second analysis was performed using these in addition to the saliva microbiome peptides for use in saliva or nasopharyngeal sampling methods.

Evaluation of SARS-CoV-2 Proteomics Data for Evaluating In silico targets

Labeled proteomics of SARS-CoV-2 infected human cells were recently deposited as ProteomeXchange PXD017710. To evaluate the relative abundance of proteins during infectivity, the RAW files were reprocessed in PD 2.2, using the default workflow for TMT 10-plex quantification with the FASTA databases for UniProt human and the SARS-CoV-2 FASTA prerelease from UniProt 2020_01. Protein and peptide assemblies were manually reviewed to cross-reference suitable peptide targets.

Data and Code Availability

All data described in this manuscript, including protein databases, instrument methods, transition lists and Skyline templates are included in the Supplemental Information as detailed in Supplemental Table 1.

(1) Lechner, J.; Hartkopf, F.; Hiort, P.; Nitsche, A.; Grossegesse, M.; Doellinger, J.; Renard, B. Y.; Muth, T. Purple: A Computational Workflow for Strategic Selection of Peptides for Viral Diagnostics Using MS-Based Targeted Proteomics. *Viruses* **2019**. https://doi.org/10.3390/v11060536.

(2) Kammila, S.; Das, D.; Bhatnagar, P. K.; Sunwoo, H. H.; Zayas-Zamora, G.; King, M.; Suresh, M. R. A Rapid Point of Care Immunoswab Assay for SARS-CoV Detection. *J. Virol. Methods* **2008**. https://doi.org/10.1016/j.jviromet.2008.05.023.

(3) Cho, S. J.; Woo, H. M.; Kim, K. S.; Oh, J. W.; Jeong, Y. J. Novel System for Detecting SARS Coronavirus Nucleocapsid Protein Using an SsDNA Aptamer. *J. Biosci. Bioeng.* **2011**. https://doi.org/10.1016/j.jbiosc.2011.08.014.

(4) Che, X. Y.; Hao, W.; Wang, Y.; Di, B.; Yin, K.; Xu, Y. C.; Feng, C. Sen; Wan, Z. Y.; Cheng, V. C. C.; Yuen, K. Y. Nucleocapsid Protein as Early Diagnostic Marker for SARS. *Emerg. Infect. Dis.* **2004**. https://doi.org/10.3201/eid1011.040516.

(5) Yip, T. T. C.; Cho, W. C. S.; Cheng, W. W.; Chan, J. W. M.; Ma, V. W. S.; Yip, T. T.; Lau Yip, C. N. B.;

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ngan, R. K. C.; Law, S. C. K. Application of ProteinChip Array Profiling in Serum Biomarker Discovery for Patients Suffering from Severe Acute Respiratory Syndrome. *Methods Mol. Biol.* **2007**. https://doi.org/10.1007/978-1-59745-304-2_20.

(6) Yip, T. T. C.; Chan, J. W. M.; Cho, W. C. S.; Yip, T. T.; Wang, Z.; Kwan, T. L.; Law, S. C. K.; Tsang, D. N. C.; Chan, J. K. C.; Lee, K. C.; et al. Protein Chip Array Profiling Analysis in Patients with Severe Acute Respiratory Syndrome Identified Serum Amyloid A Protein as a Biomarker Potentially Useful in Monitoring the Extent of Pneumonia. *Clin. Chem.* **2005**. https://doi.org/10.1373/clinchem.2004.031229.

(7) Ren, Y.; He, Q. Y.; Fan, J.; Jones, B.; Zhou, Y.; Xie, Y.; Cheung, C. Y.; Wu, A.; Chiu, J. F.; Peiris, J. S. M.; et al. The Use of Proteomics in the Discovery of Serum Biomarkers from Patients with Severe Acute Respiratory Syndrome. *Proteomics* **2004**. https://doi.org/10.1002/pmic.200400897.

(8) Zhang, L.; Zhang, Z. P.; Zhang, X. E.; Lin, F. S.; Ge, F. Quantitative Proteomics Analysis Reveals BAG3 as a Potential Target To Suppress Severe Acute Respiratory Syndrome Coronavirus Replication. *J. Virol.* **2010**. https://doi.org/10.1128/jvi.00213-10.

(9) Ying, W.; Hao, Y.; Zhang, Y.; Peng, W.; Qin, E.; Cai, Y.; Wei, K.; Wang, J.; Chang, G.; Sun, W.; et al. Proteomic Analysis on Structural Proteins of Severe Acute Respiratory Syndrome Coronavirus. In *Proteomics*; 2004. https://doi.org/10.1002/pmic.200300676.

(10) Voskuil, J. L. A. Commercial Antibodies and Their Validation. *F1000Research* **2014**. https://doi.org/10.12688/f1000research.4966.2.

(11) Voskuil, J. L. A. The Challenges with the Validation of Research Antibodies. *F1000Research* **2017**. https://doi.org/10.12688/f1000research.10851.1.

(12) Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M. Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst.* **2016**. https://doi.org/10.1016/j.cels.2016.02.015.

(13) Grebe, S. K. G.; Singh, R. J. Clinical Peptide and Protein Quantification by Mass Spectrometry (MS). *TrAC - Trends in Analytical Chemistry*. 2016. https://doi.org/10.1016/j.trac.2016.01.026.

(14) Grebe, S. K. G.; Singh, R. J. LC-MS/MS in the Clinical Laboratory - Where to from Here? *Clin. Biochem. Rev.* **2011**.

(15) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **2014**. https://doi.org/10.1074/mcp.M113.034769.

(16) Pino, L. K.; Searle, B. C.; Yang, H.-Y.; Hoofnagle, A. N.; Noble, W. S.; MacCoss, M. J. Matrix-Matched Calibration Curves for Assessing Analytical Figures of Merit in Quantitative Proteomics. *J. Proteome Res.* **2020**. https://doi.org/10.1021/acs.jproteome.9b00666.

(17) Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B. Targeted Proteomic Quantification on Quadrupole-Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **2012**. https://doi.org/10.1074/mcp.O112.019802.

(18) Higgs, R. E.; Butler, J. P.; Han, B.; Knierman, M. D. Quantitative Proteomics via High Resolution MS Quantification: Capabilities and Limitations. *Int. J. Proteomics* **2013**. https://doi.org/10.1155/2013/674282.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(19) Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol. Cell. Proteomics* **2012**. https://doi.org/10.1074/mcp.O112.020131.

(20) Gallien, S.; Bourmaud, A.; Kim, S. Y.; Domon, B. Technical Considerations for Large-Scale Parallel Reaction Monitoring Analysis. *J. Proteomics* **2014**. https://doi.org/10.1016/j.jprot.2013.10.029.

(21) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; et al. How Many Human Proteoforms Are There? *Nature Chemical Biology*. 2018. https://doi.org/10.1038/nchembio.2576.

(22) Lawrence, R. T.; Searle, B. C.; Llovet, A.; Villén, J. Plug-and-Play Analysis of the Human Phosphoproteome by Targeted High-Resolution Mass Spectrometry. *Nat. Methods* **2016**. https://doi.org/10.1038/nmeth.3811.

(23) Zauber, H.; Kirchner, M.; Selbach, M. Picky: A Simple Online PRM and SRM Method Designer for Targeted Proteomics. *Nature Methods*. 2018. https://doi.org/10.1038/nmeth.4607.

(24) Schilling, B.; Rardin, M. J.; MacLean, B. X.; Zawadzka, A. M.; Frewen, B. E.; Cusack, M. P.; Sorensen, D. J.; Bereman, M. S.; Jing, E.; Wu, C. C.; et al. Platform-Independent and Label-Free Quantitation of Proteomic Data Using MS1 Extracted Ion Chromatograms in Skyline. *Mol. Cell. Proteomics* **2012**. https://doi.org/10.1074/mcp.M112.017707.

(25) Zolg, D. P.; Wilhelm, M.; Yu, P.; Knaute, T.; Zerweck, J.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration. *Proteomics* **2017**. https://doi.org/10.1002/pmic.201700263.

(26) Deutsch, E. W. The PeptideAtlas Project. *Methods Mol. Biol.* **2010**. https://doi.org/10.1093/nar/gkj040.

(27) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotechnol.* **2004**. https://doi.org/10.1038/nbt930.

(28) Yang, Y.; Liu, X.; Shen, C.; Lin, Y.; Yang, P.; Qiao, L. In Silico Spectral Libraries by Deep Learning Facilitate Data-Independent Acquisition Proteomics. *Nat. Commun.* **2020**. https://doi.org/10.1038/s41467-019-13866-z.

(29) Zolg, D. P.; Wilhelm, M.; Schmidt, T.; Médard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. ProteomeTools: Systematic Characterization of 21 Post-Translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides. *Mol. Cell. Proteomics* **2018**. https://doi.org/10.1074/mcp.tir118.000783.

(30) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; et al. Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nat. Methods* **2019**. https://doi.org/10.1038/s41592-019-0426-7.

(31) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nature Biotechnology*. 2014. https://doi.org/10.1038/nbt.2839.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(32) Ternent, T.; Csordas, A.; Qi, D.; Gómez-Baena, G.; Beynon, R. J.; Jones, A. R.; Hermjakob, H.; Vizcaíno, J. A. How to Submit MS Proteomics Data to ProteomeXchange via the PRIDE Database. *Proteomics* **2014**. https://doi.org/10.1002/pmic.201400120.

(33) Klann, K.; Koch, B.; Krause, D. SARS-CoV-2 Infected Host Cell Proteomics Reveal Potential Therapy Targets. *Preprint* **2020**. https://doi.org/10.21203/rs.3.rs-17218/v1.

(34) Gordon, D. E.; Jang, G. M.; Bouhaddou, M.; Xu, J.; Obernier, K.; O'Meara, M. J.; Guo, J. Z.; Swaney, D. L.; Tummino, T. A.; Huettenhain, R.; et al. A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv* **2020**. https://doi.org/10.1101/2020.03.22.002386.

(35) Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villén, J.; MacCoss, M. J. Chromatogram Libraries Improve Peptide Detection and Quantification by Data Independent Acquisition Mass Spectrometry. *Nat. Commun.* **2018**. https://doi.org/10.1038/s41467-018-07454-w.

(36) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**. https://doi.org/10.1093/bioinformatics/btq054.

(37) Pejaver, V.; Hsu, W. L.; Xin, F.; Dunker, A. K.; Uversky, V. N.; Radivojac, P. The Structural and Functional Signatures of Proteins That Undergo Multiple Events of Post-Translational Modification. *Protein Sci.* **2014**. https://doi.org/10.1002/pro.2494.

(38) Belstrøm, D.; Jersie-Christensen, R. R.; Lyon, D.; Damgaard, C.; Jensen, L. J.; Holmstrup, P.; Olsen, J. V. Metaproteomics of Saliva Identifies Human Protein Markers Specific for Individuals with Periodontitis and Dental Caries Compared to Orally Healthy Controls. *PeerJ* **2016**. https://doi.org/10.7717/peerj.2433.

(39) Schuurman, T.; Lankamp, P.; van Belkum, A.; Kooistra-Smid, M.; van Zwet, A. Comparison of Microscopy, Real-Time PCR and a Rapid Immunoassay for the Detection of Giardia Lamblia in Human Stool Specimens. *Clin. Microbiol. Infect.* **2007**. https://doi.org/10.1111/j.1469-0691.2007.01836.x.

(40) Yang, S.; Rothman, R. E. PCR-Based Diagnostics for Infectious Diseases: Uses, Limitations, and Future Applications in Acute-Care Settings. *Lancet Infectious Diseases*. 2004. https://doi.org/10.1016/S1473-3099(04)01044-8.

(41) Sheridan, C. Coronavirus and the Race to Distribute Reliable Diagnostics. *Nat. Biotechnol.* **2020**. https://doi.org/10.1038/d41587-020-00002-2.

(42) Diao, B.; Wen, K.; Chen, J.; Liu, Y.; Yuan, Z.; Han, C.; Chen, J.; Pan, Y.; Chen, L.; Dan, Y.; et al. Diagnosis of Acute Respiratory Syndrome Coronavirus 2 Infection by Detection of Nucleocapsid Protein. *medRxiv* **2020**. https://doi.org/10.1101/2020.03.07.20032524.

(43) Zhu, H.; Hu, S.; Jona, G.; Zhu, X.; Kreiswirth, N.; Willey, B. M.; Mazzulli, T.; Liu, G.; Song, Q.; Chen, P.; et al. Severe Acute Respiratory Syndrome Diagnostics Using a Coronavirus Protein Microarray. *Proc. Natl. Acad. Sci. U. S. A.* **2006**. https://doi.org/10.1073/pnas.0510921103.

(44) Selvarajah, S.; Negm, O. H.; Hamed, M. R.; Tubby, C.; Todd, I.; Tighe, P. J.; Harrison, T.; Fairclough, L. C. Development and Validation of Protein Microarray Technology for Simultaneous

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Inflammatory Mediator Detection in Human Sera. *Mediators Inflamm.* **2014**. https://doi.org/10.1155/2014/820304.

(45) Broughton, J. P.; Deng, X.; Yu, G.; Fasching, C. L.; Singh, J.; Streithorst, J.; Granados, A.; Sotomayor-Gonzalez, A.; Zorn, K.; Gopez, A.; et al. Rapid Detection of 2019 Novel Coronavirus SARS-CoV-2 Using a CRISPR-Based DETECTR Lateral Flow Assay. *medRxiv* **2020**. https://doi.org/10.1101/2020.03.06.20032334.

(46) Ozalp, A.; Barroso, B.; Meijer, J.; van den Beld, C. Determination of Methionine-Enkephalin and Leucine-Enkephalin by LC-MS in Human Plasma: Study of Pre-Analytical Stability. *Anal. Biochem.* **2018**. https://doi.org/10.1016/j.ab.2018.07.001.

(47) Hirtz, C.; Vialaret, J.; Gabelle, A.; Nowak, N.; Dauvilliers, Y.; Lehmann, S. From Radioimmunoassay to Mass Spectrometry: A New Method to Quantify Orexin-A (Hypocretin-1) in Cerebrospinal Fluid. *Sci. Rep.* **2016**. https://doi.org/10.1038/srep25162.

(48) Kolsrud, H.; Malerod, H.; Ray, S.; Reubsaet, L.; Lundanes, E.; Greibrokk, T. A Critical Review of Trypsin Digestion for LC-MS Based Proteomics. In *Integrative Proteomics*; 2012. https://doi.org/10.5772/29326.

(49) Methé, B. A.; Nelson, K. E.; Pop, M.; Creasy, H. H.; Giglio, M. G.; Huttenhower, C.; Gevers, D.; Petrosino, J. F.; Abubucker, S.; Badger, J. H.; et al. A Framework for Human Microbiome Research. *Nature* **2012**. https://doi.org/10.1038/nature11209.

(50) Peterson, J.; Garges, S.; Giovanni, M.; McInnes, P.; Wang, L.; Schloss, J. A.; Bonazzi, V.; McEwen, J. E.; Wetterstrand, K. A.; Deal, C.; et al. The NIH Human Microbiome Project. *Genome Res.* **2009**. https://doi.org/10.1101/gr.096651.109.

(51) Zhang, Z.; Burke, M.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Markey, S. P.; Yu, W.; Chaerkady, R.; Hess, S.; Stein, S. E. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *J. Proteome Res.* **2018**. https://doi.org/10.1021/acs.jproteome.7b00614.

(52) Serwa, R. A.; Abaitua, F.; Krause, E.; Tate, E. W.; O'Hare, P. Systems Analysis of Protein Fatty Acylation in Herpes Simplex Virus-Infected Cells Using Chemical Proteomics. *Chem. Biol.* **2015**. https://doi.org/10.1016/j.chembiol.2015.06.024.

(53) Veit, M. Palmitoylation of Virus Proteins. *Biol. Cell* **2012**. https://doi.org/10.1111/boc.201200006.

(54) Veit, M.; Serebryakova, M. V.; Kordyukova, L. V. Palmitoylation of Influenza Virus Proteins. In *Biochemical Society Transactions*; 2013. https://doi.org/10.1042/BST20120210.

(55) Zhang, C.; Zheng, W.; Huang, X.; Bell, E. W.; Zhou, X.; Zhang, Y. Protein Structure and Sequence Re-Analysis of 2019-NCoV Genome Refutes Snakes as Its Intermediate Host or the Unique Similarity between Its Spike Protein Insertions and HIV-1. *J. Proteome Res.* **2020**. https://doi.org/10.1021/acs.jproteome.0c00129.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

List of Supplemental Materials:

Supplemental Table 1: LCMS resources provided in Zip File with this Submission

Supplemental Table 2: Peptide lists remaining following subtraction of each proteome background

Supplemental File 1: Alignments of Nucleocapsid and Spike Protein Variants and analysis details

Supplemental Material 1: Zip file containing all resources detailed in Supplemental Table 1

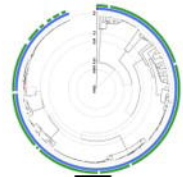
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Table 1. LCMS resources provided as supplemental material with this submission.

Proteomics Databases	**Description**
2019-nCoVpFASTA1	FASTA file of COVID-19 Protein Sequences
Human_plus_2019-nCoVpFASTA2	Human UniProt FASTA plus COV-19 Proteins above
2019-nCoV_Trypsin_Prosit_Spectral_Library-MSP-format.zip	Prosit derived spectral library of COV-19 peptides
Adjusted Oral Microbiome FASTA	FASTA database generated from meta-analysis of saliva metaproteomics samples
Instrument-Specific Run Parameters	
SRM Transition lists	
Thermo TSQ Quantum/Vantage	Formatted for Thermo/Finnigan Legacy TSQ
Thermo TSQ Altis/Quantis/Fortis	Formatted for current generation Thermo TSQ
SCIEX Triple Quadrupole Systems	Contains specific SCIEX parameters as described in methods
Agilent all models	Formatted for Agilent
Waters XEVO	Formatted for Waters and Waters XEVO
Shimadzu all models	Transitions in .txt format as required by vendor interfaced
PRM Window List for 3 peptides/protein	Monoisotopic mass for 3 peptides for building PRM methods
Example PRM methods for Thermo Q Exactive/Exploris/Fusion	Premade .meth files with all parameters for these systems
Data Processing	
PRM Example Processing Skyline Templates	Skyline PRM format .sky file
SRM Example Processing Skyline Templates	SRM format .sky files with 10 peptides/protein
SIM Example Processing Skyline Template	SIM Templates for Orbitrap or TOF systems
PD settings for processing	Excel sheet with all PD processing details
Post Translational Modification Predictions	
ModPred All Predicted Modifications Excel Sheet	A compiled list of all PTMs from ModPred output
ModPred All High confidence Predicted Modifications	A simplified sheet of all high confidence PTMs
ModPred Summary and counts of high confidence PTMs	A combined sheet for counting high confidence PTMs

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.980383; this version posted April 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Viral genome published



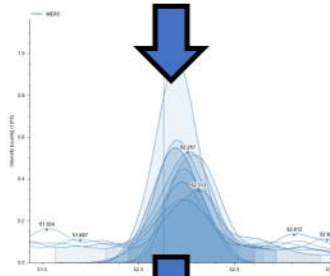
In silico digestion

555 peptides

Choose diagnostic peptides of COVID-19

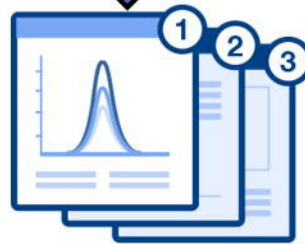
110 peptides remain

Select highest abundance proteins from deposited data



23 peptides remain

Eliminate targets from highly variable regions

21 peptides suitable for LCMS or immunoassay

Use deep learning tools to develop rapid LCMS assays