Journal of Clinical Virology Plus 3 (2023) 100147

Contents lists available at ScienceDirect

Journal of Clinical Virology Plus

journal homepage: www.elsevier.com/locate/jcvp

Proteome2virus: Shotgun mass spectrometry data analysis pipeline for virus identification

Manon Balvers[a], Isabelle F. Gordijn[b], Ingrid A.I. Voskamp-Visser[b], Merel F.A. Schelling[b], Rob Schuurman[c], Esther Heikens[d], Rene Braakman[b], Christoph Stingl[e], Hans C. van Leeuwen[b], Theo M. Luider[e], Lennard J. Dekker[e], Evgeni Levin[a], Armand Paauw[b,*]

[a] HORAIZON Technology BV., Delft, the Netherlands
[b] Department of CBRN Protection, Netherlands Organization for Applied Scientific Research TNO, TNO, P.O. Box 45, 2280 AA Rijswijk, Rijswijk, the Netherlands
[c] Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, the Netherlands
[d] Department of Medical Microbiology, Hospital St Jansdal, Harderwijk, the Netherlands
[e] Department of Neurology, Erasmus MC, Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Diagnosis
Mass spectrometry
Proteome
Peptides
SARS-CoV-2
Virus

ABSTRACT

Objectives: Shotgun proteomics is a generic method enabling detection of multiple viral species in one assay. The reliable and accurate identification of these viral species by analyzing peptides from MS-spectra is a challenging task. The aim of this study was to develop an easy accessible proteome analysis approach for the identification of viruses that cause respiratory and gastrointestinal infections.

Methods: For this purpose, a shotgun proteomics based method and a web application, 'proteome2virus', were developed. Identified peptides were searched in a database comprising proteomic data of 46 viruses known to be infectious to humans.

Results: The method was successfully tested for cultured viruses and eight fecal samples consisting of ten different viral species from seven different virus families, including SARS-CoV-2. The samples were prepared with two different sample preparation methods and were measured with two different mass spectrometers.

Conclusions: The results demonstrate that the developed web application is applicable to different MS data sets, generated from two different instruments, and that with this approach a high variety of clinically relevant viral species can be identified. This emphasizes the potential and feasibility for the diagnosis of a wide range of viruses in clinical samples with a single shotgun proteomics analysis.

1. Introduction

Rapid, sensitive, and specific identification of pathogenic microorganisms is of utmost importance in clinical diagnostics. Mass spectrometry (MS) has been widely used for identification of bacterial pathogens. For the identification of viral pathogens RT-PCR is still the most commonly used method. However, the current status is that there are no routinely used diagnostic assays that are non-targeted based and are able to identify viruses directly from clinical samples. The genetic diversity between viral species makes it impossible to use an amplicon-based strategy as is often used for bacteria [1]. An approach closest to a diagnostic assay that can identify generic viruses directly from clinical samples is metagenomic sequencing (MGS). However, the use of MGS for diagnostics purposes is limited by its technical complexity, relative long runtime and high costs [1,2]. To be prepared for new or altered (due to genomic mutations) infectious viruses a non-targeted based detection method which is able to detect these viruses directly from clinical samples is of great added value [1,2].

In recent years different MS-based proteomics methods have been developed for the detection of viruses as alternative methods to RT-PCR and to provide complementary data. For example, a variety of targeted-based MS methods were published enabling detection of SARS-CoV-2 [3–6], human metapneumoviruses (hMPV), and multiple flavivirus species by analyzing specific protein fragments (peptides) in clinical samples [7,8]. Recently, a multitargeted MS (multiple reaction monitoring (MRM)) method was developed that enables to screen for eight respiratory viruses at once [9].

Shotgun proteomics is non-targeted based and refers to direct analysis of complex protein mixtures using liquid chromatography tandem mass spectrometry (LC-MS/MS) [10]. In short, samples are subjected to proteolytic cleavage, using trypsin to produce small peptide sequences. These peptides are subjected to LC-MS/MS analysis

* Corresponding author.
E-mail address: armand.paauw@tno.nl (A. Paauw).

https://doi.org/10.1016/j.jcvp.2023.100147
Received 29 September 2022; Received in revised form 27 February 2023; Accepted 27 March 2023

in order to measure the masses of the peptides and fragments. The acquired MS and MS/MS-spectra are compared to a protein database, to identify the peptide sequences and subsequently the proteins.

Shotgun proteomics has become an excellent method to study expression of proteins and even protein-protein interactions in life sciences [11]. During the COVID-19 pandemic shotgun proteomics is used in studies to analyze the interaction of SARS-CoV-2 with the human body, varying from studying cellular pathways in cell cultures to determining the immune response in human samples (serum, body secretions) [11]. In addition, shotgun proteomics is used to determine which peptides are potential good targets in MRM assays to detect SARS-CoV-2 in clinical samples [3–6,12]. In contrast to targeted based MS assays, shotgun proteomics is not limited by the fact that only viruses that match the target of the assay can be identified. Furthermore, more viral species can be detected in one assay.

Reliable and accurate identification of viruses by analyzing peptides from MS-spectra is a challenging task. The occurrence of highly similar protein and peptide sequences in different viral species and the high and low redundancy of some of these species and strains in public databases complicate the interpretation of results. Furthermore, extension of these databases with detailed information increases runtime and needs more computer power.

The existing software packages mainly identify proteins in a sample based on known peptide sequences from LC-MS/MS data. An exception is TaxIT, which is an iterative computational pipeline for untargeted strain level identification using MS/MS data [13]. This pipeline is innovative and can be used to identify species to the strain level. Limitations are the relative long runtime, the use of a non-curated database, and the assumption that only one single organism is present in the analyzed sample. Furthermore, significant bio-informatic knowledge is required to perform the analysis with TaxIT.

Recently, we developed an easy and open internet application, proteome2pathogen.com, that overcomes these issues. This pipeline assigns identified peptides to bacteria and subsequently identifies the bacteria to the species level [14,15]. The aim of the present study was to develop a similar, easy accessible proteome analysis approach, proteome2virus, for the identification of viruses that cause respiratory and gastrointestinal infections.

2. Materials and methods

2.1. MS-data

Two types of samples were tested. Set number 1 (Table 1) are generated from samples obtained from cultured viruses. SARS-CoV-2, HCoV-OC43 and HCoV-229E were cultured in Vero cell line. Vero cells used for infection were cultured using Dulbecco's Modified Eagle Medium (DMEM) containing 5 % fetal bovine serum (FBS) with 1% Pen Strep solution (Gibco media). Cells were incubated at 37 ˚C with 5% CO_2. Vero cells were grown until a 70–80% confluent layer was formed (daily inspected using microscopy). The supernatant was removed and 5 ml with 1.10^6 genome copies/ml of virus was added. Culture was maintained until Vero cells start floating in the liquid. Supernatant was removed and stored in -70 ˚C until sample preparation was executed. The other viruses used in this study (Influenza A, hMPV, RSV A and RSV B) were cultured as previously described [16]. To determine the concentration of the obtained stocks the titers of cultured viruses were determined with standard rRT-PCRs by using a standard curve [16].

As a negative control supernatant of cell cultures used for virus culturing were applied. Sample preparation was executed with the modified SP3 protocol as described by Hayoun *et al.* [17]. In short, supernatant of the cultured viruses (containing the virus culture medium, DMEM or EMEM (Eagle's minimal essential medium) with 5% FBS, cell debris and virus particles) were diluted in lysis buffer (4% SDS, 100 mM DTT in 100 mM Tris/HCl pH8). HCoV-OC43 and HCoV229E were diluted 10 and 100 times, respectively, while the other samples were diluted to the end concentration based on the number of genome copies /ml determined with rRT-PCR. From the diluted samples, 500 µl was used and subsequent incubated for 30 min at 95˚C. Subsequently, the sample was sonicated in a ultrasonic bath (Crest Ultrasonics) for 5 min, followed by adding 40 µl solution of Sera-Magnetic beads (25 µg/µl of hydrophilic beads and 25 µg/µl of hydrophobic beads). After gently mixing, 500 µl 100% ethanol was added and thereafter incubated for 5 min at 900 rpm on a thermomixer (Thermomixer, Eppendorf). Sera-Magnetic beads were retained in solution by a neodymium magnet (MagRack 6 Cytiva product no. 28-9489-64), while the liquid was removed by pipetting. Next, Sera-Magnetic beads were washed 3 times with 180 µl 80% ethanol whereby the beads were kept in the reaction tube using the neodymium magnet. After removal of the 80% ethanol after the third wash 100 µL digestion buffer (1 µg/µl Trypsin Gold (Promega) in 50 mM ABC-buffer) was added, sonicated 30 s and mixed gently to disperse the beads homogenous in the solution and incubated for 15 min at 50 ˚C. Digestion reaction was stopped by adding 5 µl 10% TFA. Subsequently, the digests were analyzed by LC-MS/MS using a nano-liquid chromatography system (Ultimate 3000; Dionex, Germering, Germany) coupled to a Orbitrap mass spectrometer (Orbitrap Eclipse, Thermo Fisher Scientific, San Jose, USA) equipped with a high-field asymmetric ion mobility spectrometer (FAIMS) device. After preconcentration and washing of the sample on a C18 trap column (5 mm × 300 μm i.d.), peptides were separated on a C18 PepMap column (250 mm × 75 μm internal diameter; Dionex, Amsterdam) using a linear 90 min gradient (3–30% acetonitrile/H20; 0.1% formic acid) at a flow rate of 300 nl/min. The separation of the peptides was monitored by UV detection (absorption at 214 nm). For electrospray ionization we used coated silica nano electrospray emitters (New Objective, Woburn, MA, USA) at a spray voltage of 2.2 kV. FAIMS was setup for consecutively collect ion mobility fractions at compensation voltages (CV) of -45, -60, -75 and -90 V. Blank LC-MS runs were performed between the samples to monitor possible system background. For data dependent acquisition (DDA) measurements, an Orbitrap survey scan (with a resolution of 120,000 and automatic gain control, AGC, set to 400,000) was obtained followed by consecutively isolation (with AGC target set to 10,000), fragmentation (35% normalized HCD collision energy), and ion-trap detection of the peptide precursors identified in the survey scan. Number of collected MS/MS spectra was controlled by limiting the available total duty cycle time to 3.6 s (*top speed mode*) by restricting the maximum cycle times of each FAIMS segment (1 s for CVs ≥ -60 V and 0.8 s for CVs < -60 V).

Set number 2 (Table 2) contains MS data generated from fecal samples of children with a known gastrointestinal virus infection. Each fecal sample was diluted in an approximal 1:1 ratio with 100 mM ammonium bicarbonate (pH 8), resulting in a total volume of 100 µl per sample. The samples were thoroughly mixed and then centrifuged for 5 min at 10,000 rpm. Subsequently, the supernatant was used in the following steps of the sample preparation. To 100 µl supernatant, 100 µl 0.2% RapiGest (Waters Corporation, Milford, CT, USA) in 100 mM ammonium bicarbonate (pH 8) was added (final concentration of RapiGest as advised by manufacturer is 0.1%). Subsequently, the sample was sonicated in an ultrasonic bath (Branson 2510). After heating the sample for 5 min at 95 ˚C, the sample was cooled down and centrifuged for 5 min at 10,000 rpm (Eppendorf minispin). Five µl 200 mM DTT (DL-dithiothreitol, Sigma-Aldrich) was added to a final concentration of 5 mM. Next, the sample was incubated for 30 minutes at 60 ˚C, subsequently cooled down, and then centrifuged for 5 min at 10,000 rpm. Five µl 600 nM iodoacetamide (IAM) was added (final concentration of IAM 15 mM). Once again, the sample was incubated for 30 min at 60 ˚C after which the reactionmixture was quenched by adding 5 µl 700 nM cysteine (final concentration of cysteine 17.5 mM). Next, 5 µl trypsin (20 ng/µl in 50 mM acetic acid) was added (~pH8) followed by incubating the sample overnight at 37 ˚C. Subsequently, 5 µl 20% trifluoroacetic acid (TFA) was added to the sample and the sample was incubated for 30 min at 37 ˚C to stop the trypsin enzymatic reaction and to break down RapiGest components. The sample was centrifuged

Table 1
Viruses identified with proteome2virus after PEAKS X and LC-MS/MS analysis of (diluted) virus cultures. Samples measured with Orbitrap Eclipse lined up with a FAIMS.

Sample	Titer[a]	Number of presented peptides (unique[b])	Identified virus family	Identified virus	Unique peptide number of virus species identified	Total peptide number of virus species identified
SARS-CoV-2	10^9	1273 (1212)	*Coronaviridae*	SARS-CoV-2	30	40
SARS-CoV-2	10^8	1352 (1295)	*Coronaviridae*	SARS-CoV-2	28	36
SARS-CoV-2	10^7	427 (415)	*Coronaviridae*	SARS-CoV-2	13	16
HCoV-OC43	n.d[c]	8262 (8030)	*Coronaviridae*	Betacoronavirus OC43	93	94
HCoV-OC43	n.d[d]	7842 (7649)	*Coronaviridae*	Betacoronavirus OC43	89	90
HCoV-229E	n.d[c]	5180 (5041)	*Coronaviridae*	*Human coronavirus 229E*	42	42
HCoV-229E	n.d[d]	4315 (4201)	*Coronaviridae*	*Human coronavirus 229E*	39	42
Influenza H1N1	10^{10}	13562 (13041)	*Orthomyxoviridae*	*Influenza A virus*	236	237
Influenza H1N1	10^9	1392 (1325)	*Orthomyxoviridae*	*Influenza A virus*	54	54
Influenza H3N2	10^{10}	16473 (15739)	*Orthomyxoviridae*	*Influenza A virus*	122	122
Influenza H3N2	10^9	2451 (2363)	*Orthomyxoviridae*	*Influenza A virus*	34	34
RSV A	10^9	7571 (7355)	*Pneumoviridae*	*Human orthopneumovirus*	134	134
RSV A	10^8	1264 (1223)	*Pneumoviridae*	*Human orthopneumovirus*	43	43
RSV B	10^{10}	27765 (26331)	*Pneumoviridae*	*Human orthopneumovirus*	159	159
RSV B	10^9	9838 (9453)	*Pneumoviridae*	*Human orthopneumovirus*	79	79
hMPV	10^{10}	21326 (20570)	*Pneumoviridae*	*Human metapneumovirus*	125	127
hMPV	10^9	3989 (3856)	*Pneumoviridae*	*Human metapneumovirus*	45	45
Hep2 cell line supernatant	-	32070 (31033)	No virus detected			
LLC-MK2 cell line supernatant	-	9084 (8864)	No virus detected			

[a] Number of genome copies in sample.
[b] number of unique peptides after isoleucine (I) amino acid residues in the input peptides are replaced with leucine (L) and subsequent removal of all duplicates.
[c] Not determined. Tenfold diluted cell culture.
[d] Not determined. Hundred fold diluted cell culture.

Table 2
Viruses identified with proteome2virus after PEAKS X and LC-MS/MS analysis of fecal samples. Samples measured with Orbitrap Q Exactive Plus.

Virus	Ct-value[a]	Number of presented peptides (unique[b])	Identified virus family	Identified virus	Unique peptide number of virus species identified	Total peptide number of virus species identified
Astrovirus	18	1078 (1037)	*Astroviridae*	*Mamastrovirus 1*	7	7
Norovirus	19	379 (369)	*Caliciviridae*	*Norwalk virus*	6	6
Astrovirus	18	1922 (1848)	*Astroviridae*	*Mamastrovirus 1*	6	6
Rotavirus	18	1331 (1283)	*Reoviridae*	*Rotavirus A*	221	221
Adenovirus 40 or 41	13	1691 (1644)	*Adenoviridae*	*Human mastadenovirus F*	34	96
Rotavirus	22	323 (314)	*Reoviridae*	*Rotavirus A*	8	8
Norovirus	18	2413 (2351)	*Caliciviridae*	*Norwalk virus*	22	22
Adenovirus 40 or 41	14	1403 (1332)	*Adenoviridae*	*Human mastadenovirus F*	33	123

[a] Number of cycles for the PCR reaction in which positive signal is higher than the threshold signal. The lower the Ct-value the higher number of genome copies of the targeted nucleic acid sequences.
[b] Number of unique peptides after isoleucine (I) amino acid residues in the input peptides are replaced with leucine (L) and subsequent removal of all duplicates.

for 5 min at 10,000 rpm. The supernatant was cleaned and concentrated using a C18 column (Thermo Fisher Scientific).

LC-MS/MS measurements of fecal samples were conducted on a nano LC (Ultimate 3000; Thermo Fisher Scientific, Germering, Germany) coupled to Orbitrap Q Exactive Plus mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Samples were loaded (0.5–5 µl) onto a trap column (Acclaim PepMap; 5 mm length × 300 μm inner diameter, ID; Thermo) and desalted and preconcentrated for 10 min using 0.1% aqueous TFA at a flowrate of 20 µl/min. Next, the trap column was switched in-line with the analytical column (Acclaim PepMap, 250 mm × 75 µm ID, 3 µm particle size; Thermo) and peptides were eluted and separated using the following binary gradient: first a 60 min linear gradient from 4 to 25% solvent B followed by a 30 min gradient from 25 to 50% solvent B, where solvent A consisted of 2% acetonitrile and 0.1% formic acid (rest water), and solvent B consisted of 80% acetonitrile and 0.08% formic acid. LC chromatograms were monitored by an UV detector (absorption at 214 nm). For electrospray ionization we used coated silica nano electro-spray emitters (New Objective, Woburn, MA, USA) at a spray voltage of 1.7 kV. All samples were measured with a data dependent acquisition method using survey scans with a resolution of 70,000 (AGC target = 10^6) followed by up to 12 MS/MS scans (NCE = 28%, resolution = 17,500, 60 ms maximal ion injection time and 50,000 AGC target).

2.2. Protein database development

Our goal was to develop a rapid virus peptide analysis application with a high reliability for the identification of human pathogenic viruses. In this section, development of a fasta protein database (46Virus_db_v01) for the identification of 46 different viral species is described. Included are gastrointestinal and respiratory viruses that cause infections in humans, and human *Herpesviridae*. In total 46 virus species of 10 different viridae families, that are able to cause an infection in humans were included in the 46Virus_db_v01. Proteins were extracted from the NCBI virus database [18]. If available the proteomes of reference genomes were extracted. Otherwise, annotated proteins based on nucleotide completeness or, if these were not available, either partial and complete sequences were extracted (TableS1). For *Influenza A*, subtypes were included most relevant to human infections (H1N1, H3N2, H5N1, H7N9, H9N2 and H10N8). From H1N1 and H3N2, proteomes

of 13 and 7 strains were selected (TableS2), respectively. For H5N1, H7N9, H9N2 and H10N8, annotated proteins based on nucleotide completeness were selected. For SARS-CoV-2, 14 proteomes that contain all known variations (as of May 19, 2021) and originated from different locations were included (TableS2).

The database 46Virus_db_v01 contains viruses of *Adenoviridae, Astroviridae, Caliciviridae, Coronaviridae, Herpesviridae, Orthomyxoviridae, Paramyxoviridae, Picornaviridae, Pneumoviridae* and *Reoviridae* (supplement file 46Virus_db.fasta). In the 46Virus_db_v01 all isoleucine (I) amino acid residues are replaced with leucine (L) because of the isobaric nature of leucine and isoleucine.

As a contaminant database, the proteome of *Homo sapiens* (UP000005640_9606) was used to remove identified peptides that potential have sequence similarities in human and virus proteomes [19].

2.3. Peptide assignment

From each sample the obtained MS spectra were assigned to peptides using PEAKS X (Bioinformatics Solutions Inc., Waterloo, Canada), the 46virus_db_v01 and as a contaminant database the proteome of *Homo sapiens*. The peptides were identified using a semi-specific trypsin digestion setting. For sample Set 1, oxidation of methionine moieties was set as a variable modification and the mass tolerances for 'De Novo' and Database searches were used with the following settings; precursor ion mass tolerance 10 ppm and the product ion mass tolerance at 0.5 Da. For sample Set 2 the same settings were used only the precursor ion mass tolerance was set on 20 ppm and the fixed modification: carbamidomethylation of cysteines was added.

For all peptide assignments, only peptides with a high degree of certainty (false discovery rate [FDR] of $\leq$ 0.1%) were used to determine which viruses were present in the original sample. The obtained peptides were exported in comma-separated values (.csv) file format from PEAKS X. The list of peptide sequences was extracted (Column 'peptide' from peptide.csv), the header of the table removed and annotations of PTMs (oxidation of methionine and carbamidomethylation of cysteine) were removed. The obtained cleaned list with the identified amino-acids sequences of the peptides can be used for downstream data analysis.

2.4. Peptide-based VIRUS detection engine analysis

The developed application proteome2virus processes a list of peptides (which can be in a .csv or .txt format and independent of the used software application for peptide assignment). The cleaned peptide lists (see supplementary data for cleaned peptide lists generated in this study) were processed through a web application (app) developed by our team that identifies the virus based on discriminatory peptides (proteome2virus).

To support also peptide lists generated with other software applications, the app replaces all isoleucine (I) amino acid residues in the input peptides with leucine (L) before analysis. Duplicate input peptides, including duplicates caused by the "I-to-L" replacement, are removed. Subsequently, the analysis follows a two-step workflow: it starts with a family-level search followed by a species-level search in each relevant family. Both steps use a fast string searching algorithm (Ag) to search for matches between a query peptide and the 46Virus_db_v01 database [20].

In more detail, in the family-level step, the app searches for matches between all input peptides and a family-level viral database, which contains the proteomes of the included viruses at family level. For each peptide with a match, it determines for which family the peptide has a "hit" (i.e. a perfect match for a reference sequence of a virus family in 46Virus_db_v01) and also whether the peptide has a "family discriminative hit" (i.e. whether the peptide matches only one viral family in the 46Virus_db_v01 and for that reason is discriminative). Each family with at least 3 family discriminative hits is considered to be present in the sample and therefore selected for the species-level search. At this species level, a search is performed for each selected family separately. All peptides previously identified in that family are searched in a species-level database, which contains the proteomes of all included species belonging to this family, to determine "hits" and "species discriminative hits" (i.e. whether the peptide matches exactly one viral species within this selected family).

In the end, the app reports results, which contain for each identified family one species together with the unique hits (species discriminative peptides) and the total (all peptides identified in the identified species) number of peptide hits for this species. A file can be downloaded, which contains the peptides that have a perfect match for a reference sequence of the final identified virus or viruses. In this table it is also indicated whether the peptide is unique for the identified virus.

The final identified species is the species within a family with the highest number, but at least 3, unique peptides. If there is a draw within a family based on the highest number of unique peptides, the species is selected with the highest number of total peptides. If there are multiple species with similar number of unique and total number of peptides, all species will be reported. If no species with at least 3 species discriminative peptides is detected within a family, it is determined whether there is a species with at least 3 total peptide hits. When at least 3 total peptide hits are identified in a species of this family only the family name will be reported and the species name will be depicted as 'Uncertain' because discrimination of the detected virus to the species level is too uncertain based on these results. The returned total number of hits is the number of total peptides that was identified at the family level, while the number of discriminative hits is undefined (NA). In case in the first family-level step at least 3 family-discriminative peptides were identified, but at the second species-level step no species was present with at least 3 total peptide hits, the identified family is considered to be a false positive hit and is excluded from the report. The app was written in R (version 4.2.0), using the "shiny" package for web development [21].

3. Results

Virus infections in humans are caused by numerous viral species. To be able to screen in one assay for a wide selection of common viral species that cause respiratory or gastrointestinal tract infections, shotgun MS methods were developed. For respiratory and gastrointestinal tract samples the protocols followed are summarized in Fig. 1 and Fig. 2, respectively.

In the undiluted samples >10,000 peptides can be identified (Table 1). To interpret the identity of a virus quickly and accurately, the list of peptides can be processed by proteome2virus. After uploading the file (.csv or .txt) with peptides to the app, the file is processed as shown in Fig. 3, which generates a report on the screen within reasonable analysis time. For example, the proteome analysis of a Hep2 cell line supernatant generated the largest list of peptides 32,070 (31,033 unique after I to L replacement) in this study, and was analyzed by proteome2virus in seven minutes. The highest number of peptides was detected in a sample from an undiluted RSV B virus culture in which 27,765 (26,331 unique after I to L replacement) peptides were identified in 5.81 min. In samples with < 500 peptides, the identification time was less than 10 s.

3.1. Identification of cultured viruses

All cultured viruses were analyzed on an Orbitrap Eclipse mass spectrometer with FAIMS and sample preparation was executed with the SP3 protocol [17]. Peptide lists were generated from the MS-data using PEAKS X. The cleaned peptide lists were uploaded in the proteome2virus application (proteome2virus) (Fig. 1). Subsequently, all the tested viruses SARS-CoV-2, HCoV-OC43, HCoV-229E, *Influenza A* (H1N1 and H3N2), RSV A, RSV B, and hMPV were correctly identified based on 13 to 236 species discriminating peptides (Table 1). The outputted peptide lists of proteome2virus from all results in this study are described in

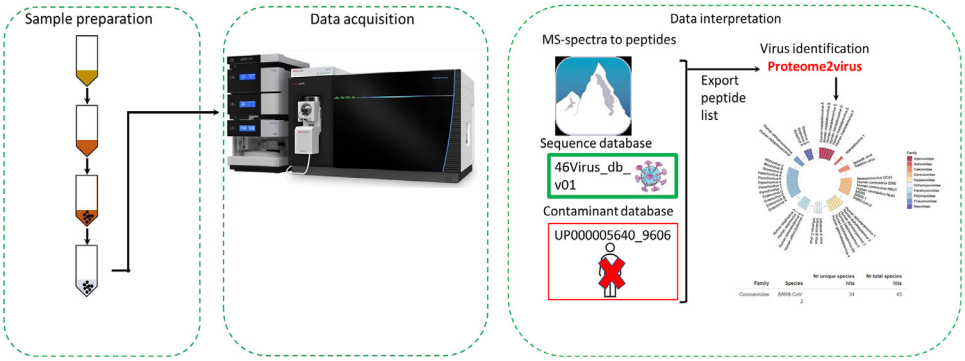

Fig. 1. Schematic overview of the liquid chromatography-tandem mass spectrometry-based method for identifying viruses in Sample set 1. Sample preparation: SP3 protocol, Data acquisition with Orbitrap Eclipse lined up with a FAIMS and Data interpretation; identification of virus peptides in PEAKS X and peptide based virus identification using proteome2virus.

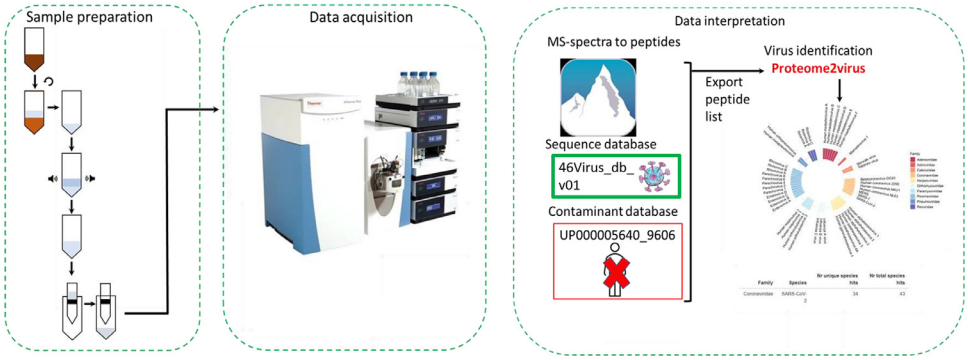

Fig. 2. Schematic overview of the liquid chromatography-tandem mass spectrometry-based method for identifying viruses in Sample set 2. Sample preparation: A traditional in dilution digestion protocol as described in material and methods, Data acquisition Orbitrap Q Exactive Plus and Data interpretation; identification of virus peptides in PEAKS X and peptide based virus identification using proteome2virus.

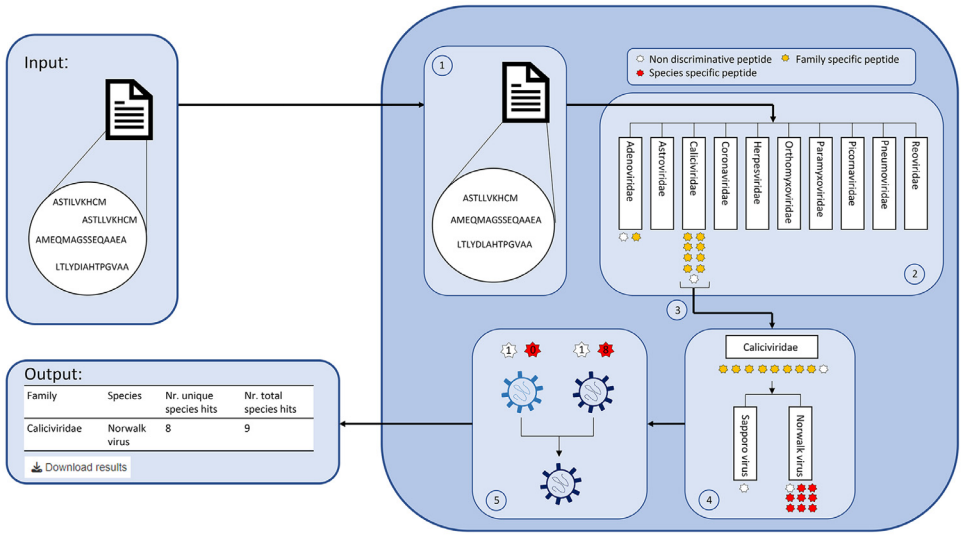

Fig. 3. A file with peptides generated with PEAKS X is uploaded to the proteome2virus app. (1) The peptides are then preprocessed by replacing I to L and removing the duplicate peptides. (2) After this, a family search is performed. In this search, input peptides are searched in proteomes of all families in the database. If a match is identified for a family, the match is determined to be discriminative (only present in this family) or not discriminative for this family. (3) After the family search, all families are selected with at least 3 family discriminative peptides. (4) In each family, a species search is performed. In this species search, all identified peptides for this family are searched in proteomes of the species present in the database that belongs to this family. If a match is identified for a species, it is determined to be discriminative (only present in this species) or not discriminative for this species within the family that is selected. (5) A final species is then selected based on the number of species discriminative peptides (and if necessary also by the number of total identified peptides for this species). The proteome2virus app subsequently generates an output table, which contains the number of species discriminative (unique) peptides and the total number of species identified peptides for each selected family. With a button a table can be downloaded, which contains the peptides that have a perfect match for a reference sequence of the identified virus or viruses. In this table it is also indicated whether the peptide is unique for the identified virus species.

the supplemented file: Peptides lists of identified viruses.xlsx. Discrepancy between species name in the tested sample and the reported virus is caused by difference in the commonly used name in clinical diagnostics and the official name as described by the International Committee on Taxonomy of Viruses (ICTV) [22]. In the supernatant of Hep2 and LLC-MK2 cell cultures that were not infected, no viruses were identified.

3.2. Identification of viruses in fecal clinical samples

To determine whether gastrointestinal viruses could be identified, clinical fecal samples from eight patients with virus infection were analyzed by shotgun proteomics and compared with RT-PCR test results (Table 2, Fig. 2). In all eight samples only one viral species was identified. Again, the discrepancy between species name in the tested sample and the reported virus (see Table 2) is caused by difference in the commonly used name in clinical diagnostics and the official name as described by the ICTV.

Mamastrovirus 1, Norwalk virus, Rotavirus A, and *Human mastadenovirus F* were identified. Each virus was identified twice in the eight different patient samples. The number of species discriminating peptides for each virus varied from 6 to 221.

3.3. Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [23,24] with the dataset identifier PXD36663 and 10.6019/PXD036663. The samples with an overview of the associated data (raw datafiles on PXD036663, supplemented cleaned peptide lists, outputted peptide lists of proteome2virus) are described in TableS3.

4. Discussion

To enable detection of the most commonly occurring respiratory and gastrointestinal viruses in a single assay, a data analysis approach for shotgun mass spectrometry data has been developed. With this developed data analysis pipeline, it is possible to identify 46 viral species within one test. The analysis report provides an easy-to-interpret result, which describes the identified virus (or viruses) and the number of discriminatory peptides on which the identification is based.

Viruses were correctly identified with the developed web application, based on at least three species discriminative peptides. In the samples used, it was unknown which virus was present and because virus species contains multiple proteins the cut-off number was increased to three discriminative peptides in order to reduce the possibility of encountering false positive results. Moreover, in proteomics a cut-off of two or more peptides is often used to identify a protein.

In total ten different viral species were identified from seven different virus families.

Using the constructed database with 46 viral species, PEAKS X, and the developed MS-data analysis approach proteome2virus, viruses can be identified based on shotgun mass spectrometry data. The analyzed data were acquired using two different sample preparations methods and measured on two different Orbitraps (Orbitrap Q Exactive Plus and Orbitrap Eclipse equipped with FAIMS). This implicates that our data analysis pipeline is applicable for different MS data sets.

Our study results show that with shotgun LC-MS/MS and the developed data analysis pipeline, it is possible to screen for different viral species simultaneously. Another advantage of our method is its simplicity, which ensures that all viruses (DNA or RNA viruses, enveloped or non-enveloped viruses) can be detected with the same procedure and reagents. Subsequently, the identification of a virus is highly reliable because the identification is based on at least three but most often more unique peptide sequences. The database can easily be extended with other virus genomes.

There are some drawbacks to shotgun mass spectrometry which need to be further studied before this method can be used as a routine virus diagnostic test in clinical laboratories. The relative low throughput and sensitivity, and the risk of carryover should be addressed. Concerning the risk of carryover, Foundraine *et al.* demonstrated that by using an Evosep One LC system, which uses a disposable StageTip for sample loading, there was no carryover between the tested samples [25]. Another practical approach is to re-analyze a sample if it detects the same viral species as in the previous sample [6]. The throughput is mainly limited by the serial measurement of samples in contrast to RT-PCR. Throughput, can be increased by optimizing elution gradients and/or the data acquisition method of the mass spectrometer. Recently, scanning SWATH was introduced as a new acquisition method that enables the measurement of several hundreds of proteomes per day on a single LC–MS instrument [26,27]. So far, the sensitivity of shotgun mass spectrometry is relatively low, which limits its clinical applicability for certain infections or certain samples like cerebrospinal fluid, where viral loads are often low. However, in acute phase samples or samples from immunocompromised patients, and in certain infections like gastrointestinal infections, viral loads are usually high. Moreover, in immunocompromised patients the infectious agent is more often unpredictable. In these cases shotgun mass spectrometry as non-targeted based method could already be of additional value despite the relative low sensitivity. Another advantage of the developed shotgun MS-method compared to PCR or antibody based detection methods, is that this-method is less sensitive for genetic mutations, and less dependent on specific reagents. Therefore, MS-based proteomic applications may play a role in detecting new virus variants in patients as previously suggested [11].

Despite the fact that the presented data were acquired with conventional nano-LC settings and not in a high-throughput manner, the method can be applied in critical, rare or complex situations to complement clinical diagnostics.

In summary, we have developed an open access, simplified data analysis pipeline for the diagnosis of most common causes of respiratory and gastrointestinal infections, using shotgun LC-MS/MS. Results indicate that the causative agent of viral diseases can be elucidated using shotgun mass spectrometry in combination with an easily accessible data analysis approach which is feasible in a meaningful time frame and can assist in the screening of a suspected viral infection.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Manon Balvers: Data curation, Formal analysis, Writing – original draft, Validation, Writing – review & editing. **Isabelle F. Gordijn:** Conceptualization, Investigation, Writing – review & editing. **Ingrid A.I. Voskamp-Visser:** Conceptualization, Investigation, Writing – review & editing. **Merel F.A. Schelling:** Conceptualization, Investigation, Writing – review & editing. **Rob Schuurman:** Conceptualization, Validation, Writing – review & editing. **Esther Heikens:** Conceptualization, Supervision, Writing – review & editing. **Rene Braakman:** Conceptualization, Data curation, Investigation, Supervision, Writing – review & editing. **Christoph Stingl:** Data curation, Investigation, Validation, Writing – review & editing. **Hans C. van Leeuwen:** Data curation, Supervision, Validation, Writing – review & editing. **Theo M. Luider:** Supervision, Validation, Writing – review & editing. **Lennard J. Dekker:** Data curation, Investigation, Validation, Writing – review & editing. **Evgeni Levin:** Conceptualization, Data curation, Methodology, Software, Visualization, Supervision, Validation, Writing – review & editing. **Armand Paauw:** Conceptualization, Data curation, Formal analysis, Writing – original draft, Methodology, Software, Visualization, Supervision, Validation, Writing – review & editing.

Acknowledgments

We would like to thank Francine Bruggeman for technical assistance and Inge D. Wijnberg and Hugo-Jan Jansen for their comments and critical reading of the manuscript.

This work was supported by the Dutch Ministry of Defence [grant number V2207] and by the Dutch Ministry of Economic Affairs through the Early Research Program funded project Pandemic Preparedness. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jcvp.2023.100147.

References

[1] T.N. Wylie, K.M. Wylie, B.N. Herter, G.A. Storch, Enhanced virome sequencing using targeted sequence capture, Genome Res. 25 (2015) 1910–1920.

[2] S.A. Jansen, W. Nijhuis, H.L. Leavis, A. Riezebos-Brilman, C.A. Lindemans, R. Schuurman, Broad virus detection and variant discovery in fecal samples of hematopoietic transplant recipients using targeted sequence capture metagenomics, Front. Microbiol. 11 (2020) 560179.

[3] J. Saadi, S. Oueslati, L. Bellanger, F. Gallais, L. Dortet, A.M. Roque-Afonso, et al., Quantitative assessment of SARS-CoV-2 virus in nasopharyngeal swabs stored in transport medium by a straightforward LC-MS/MS assay targeting nucleocapsid, membrane, and spike proteins, J. Proteome Res. 20 (2021) 1434–1443.
[4] C. Ihling, D. Tanzler, S. Hagemann, A. Kehlen, S. Huttelmaier, C. Arlt, et al., Mass Spectrometric identification of SARS-CoV-2 proteins from gargle solution samples of COVID-19 patients, J. Proteome Res. 19 (2020) 4389–4392.
[5] D. Gouveia, G. Miotello, F. Gallais, J.C. Gaillard, S. Debroas, L. Bellanger, et al., Proteotyping SARS-CoV-2 virus from nasopharyngeal dwabs: A Proof-of-Concept focused on a 3 min Mass Spectrometry window, J. Proteome Res. 19 (2020) 4407–4416.
[6] K.H.M. Cardozo, A. Lebkuchen, G.G. Okai, R.A. Schuch, L.G. Vianam, A.N. Olive, C.D.S. Lazari, et al., Fast and low-cost detection of SARS-CoV-2 peptides by tandem mass spectrometry in clinical samples, Eur. PMC (2020), doi:10.21203/Rs.3.Rs-28883/V1. PPR:PPR163832.
[7] S. Wee, A. Alli-Shaik, R. Kek, H.L.F. Swa, W.P. Tien, V.W. Lim, et al., Multiplex targeted mass spectrometry assay for one-shot flavivirus diagnosis, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 6754–6759.
[8] M.W. Foster, G.C. Gerhardt, L. Robitaille, P.L. Plante, G. Boivin, J. Corbeil, et al., Targeted proteomics of human metapneumovirus in clinical samples and viral cultures, Anal. Chem. (2015).
[9] C. Hodgkins, L.K. Buckton, G.J. Walker, B. Crossett, S.J. Cordwell, A.R. Horvath, et al., Simultaneous monitoring of eight human respiratory viruses including SARS–CoV-2 using liquid chromatography-tandem mass spectrometry, Sci. Rep. 12 (2022) 13392,022-16250-y.
[10] A. Garcia, Two-dimensional gel electrophoresis in platelet proteomics research, Methods Mol. Med. 139 (2007) 339–353.
[11] J. Zecha, C. Lee, F.P. Bayer, C. Meng, V. Grass, J. Zerweck, et al., Data, reagents, assays and merits of proteomics for SARS-CoV-2 research and testing, Mol. Cell Proteom. 19 (2020) 1503–1522.
[12] S. Renuse, P.M. Vanderboom, A.D. Maus, J.V. Kemp, K.M. Gurtner, A.K. Madugundu, et al., A mass spectrometry-based targeted assay for detection of SARS-CoV-2 antigen from clinical specimens, EBioMedicine 69 (2021) 103465.
[13] M. Kuhring, J. Doellinger, A. Nitsche, T. Muth, R. By, TaxIt: An iterative computational pipeline for untargeted strain-level identification using MS/MS spectra from pathogenic single-organism samples, J. Proteome Res. 19 (2020) 2501–2510.
[14] E.M. Berendsen, E. Levin, R. Braakman, D.V. der Riet-van Oeveren, N.J. Sedee, A. Paauw, Identification of microorganisms grown in blood culture flasks using liquid chromatography-tandem mass spectrometry, Futur. Microbiol. 12 (2017) 1135–1145.
[15] E.M. Berendsen, E. Levin, R. Braakman, A. Prodan, H.C. van Leeuwen, A. Paauw, Untargeted accurate identification of highly pathogenic bacteria directly from blood culture flasks, Int. J. Med. Microbiol. 310 (2020) 151376.
[16] J.A. Majchrzykiewicz-Koehorst, E. Heikens, H. Trip, A.G. Hulst, A.L. de Jong, M.C. Viveen, et al., Rapid and generic identification of influenza A and other respiratory viruses with mass spectrometry, J. Virol. Methods 213 (2015) 75–83.
[17] K. Hayoun, D. Gouveia, L. Grenga, O. Pible, J. Armengaud, B. Alpha-Bazin, Evaluation of sample preparation methods for fast proteotyping of microorganisms by tandem mass spectrometry, Front. Microbiol. 10 (2019) 1985.
[18] E.L. Hatcher, S.A. Zhdanov, Y. Bao, O. Blinkova, E.P. Nawrocki, Y. Ostapchuck, et al., Virus variation resource - improved response to emergent viral outbreaks, Nucleic Acids Res. 45 (2017) D482–D490.
[19] UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (2021) D480-9.
[20] Boyer R.S., Moore J.S. A fast string searching algorithm. Communications of the ACM. New York, NY, USA: Association for Computing Machinery 1977;20,762–72.
[21] Chang W., Cheng J., Allire J.J., Xie Y., McPherson J. Shiny: web application framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny. 2017.
[22] F.M. Zerbini, S.G. Siddell, A.R. Mushegian, P.J. Walker, E.J. Lefkowitz, E.M. Adriaenssens, et al., Differentiating between viruses and virus species by writing their names correctly, Arch. Virol. (2022).
[23] E.W. Deutsch, N. Bandeira, V. Sharma, Y. Perez-Riverol, J.J. Carver, D.J. Kundu, et al., The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics, Nucleic Acids Res. 48 (2020) D1145-52.
[24] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, et al., The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucleic Acids Res. 47 (2019) D442–D450.
[25] D.E. Foudraine, L.J.M. Dekker, N. Strepis, M.L. Bexkens, C.H.W. Klaassen, T.M. Luider, et al., Accurate detection of the four most prevalent carbapenemases in E. coli and K. pneumoniae by high-resolution mass spectrometry, Front. Microbiol. 10 (2019) 2760.
[26] C.B. Messner, V. Demichev, N. Bloomfield, J.S.L. Yu, M. White, M. Kreidl, et al., Ultra-fast proteomics with Scanning SWATH, Nat. Biotechnol. 39 (2021) 846–854.
[27] C.B. Messner, V. Demichev, D. Wendisch, L. Michalick, M. White, A. Freiwald, et al., Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection, Cell Syst. 11 (2020) 11,24.e4.