P4PP: A Universal Shotgun Proteomics Data Analysis Pipeline for Virus Identification

Authors

Armand Paauw, Evgeni Levin, Ingrid A. I. Voskamp-Visser, Ilka M. F. Marissen, Vincent Ramisse, Marine Eschlimann, Jiří Dresler, Petr Pajer, Christoph Stingl, Hans C. van Leeuwen, Theo M. Luider, and Luc M. Hornstra

Correspondence

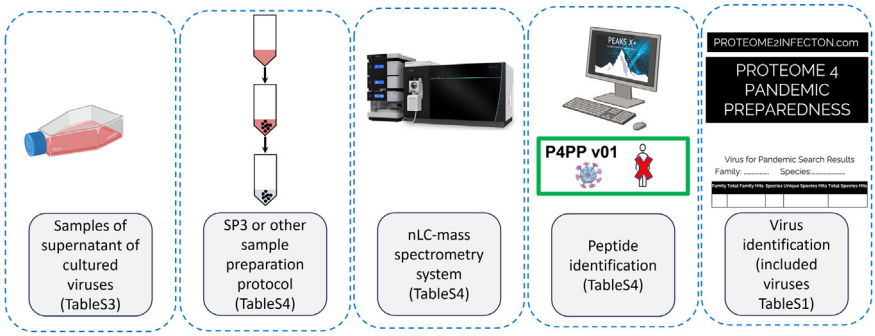
armand.paauw@tno.nl

In Brief

Shotgun proteomics combined with an innovative data analysis pipeline, P4PP, enables the identification of 1896 virus species across 32 families. The application was successfully validated with 174 cell-cultivated virus samples from 9 different MS datasets. P4PP is able to subtype influenza A and SARS-CoV-2. P4PP offers an improved accessibility and simplification of proteomics data analysis and a comprehensive report of identified virus(es). Implementing this capability significantly enhances global preparedness against emerging, mutated, or unexpected viruses worldwide.

Graphical Abstract

Schematic overview of shotgun proteomics-based virus identification of cultivated viruses. Virus cell culture, sample preparation, data acquisition, peptide identification, and virus identification.



Highlights

- A universal shotgun proteomics-based method for agnostic virus identification.
- Simple agnostic method for emerging, unknown or mutated viruses.
- Easy executable pipeline that enables the identification of 1896 virus species.
- Adopting P4PP enhances readiness for newly emerging viruses.
- Extensively validated with nine datasets obtained with different MS protocols.

2025, Mol Cell Proteomics *24*(*7*), 101004

https://doi.org/10.1016/j.mcpro.2025.101004

P4PP: A Universal Shotgun Proteomics Data Analysis Pipeline for Virus Identification

Armand Paauw[1,*], Evgeni Levin[2], Ingrid A. I. Voskamp-Visser[1], Ilka M. F. Marissen[1], Vincent Ramisse[3], Marine Eschlimann[3], Jiří Dresler[4], Petr Pajer[4], Christoph Stingl[5], Hans C. van Leeuwen[1], Theo M. Luider[5], and Luc M. Hornstra[1]

Humans can be infected by a wide variety of virus species. We developed a data analysis approach for shotgun proteomic data to detect these viruses. A proteome for pandemic preparedness (P4PP) pipeline, a corresponding database (P4PP v01), and a web application (P4PP) were constructed. The P4PP pipeline enables the identification of 1896 virus species from the 32 virus families, based on multiple identified discriminatory peptides, in which at least one human infectious virus is described. P4PP was evaluated using different datasets of cell-cultivated viruses, generated at different institutes, measured with different instruments, and prepared with different sample preparation methods. In total, 174 mass spectrometry datasets of 160 and 14 protein trypsin digests of virus-infected and noninfected cell lines were analyzed, respectively. Of the 160 samples, 146 were correctly identified at the species level, and an additional four samples were identified at the family level. In the remaining 10 samples, no virus was detected. However, all these 10 samples tested positive in follow-up samples obtained later in time series were negative samples were measured, indicating that the number of peptides derived from the virus was initially too low in the samples obtained at the start of the experiment. Furthermore, results show that influenza A or severe acute respiratory syndrome coronavirus 2 can be subtyped if enough discriminative peptides of the virus are identified. In the noninfected cell lines, no virus was detected except in one sample where the in that experiment studied virus was detected. Shotgun proteomics, in combination with the developed data analysis approach, can identify all types of virus species after cultivation in a cell line. Implementing this agnostic virus proteome analysis capability in viral diagnostic laboratories has the potential to improve their capabilities to cope with unexpected, mutated, or re-emerging viruses.

Rapid identification of existing and emerging infectious viruses is crucial for therapeutic interventions, prediction of disease progression, and combating outbreaks or epidemics. The coronavirus disease 2019 (COVID-19) crisis and the recent Mpox outbreak highlighted that current biosurveillance methods lack the capability of detection of *a priori* unexpected viruses promptly (1–5). It is plausible that other viruses will emerge and have to be controlled rapidly in the future. Although it is time consuming, virus culturing is an agnostic approach to enrich and then determine the presence of an infectious virus in a sample. During culture, virus growth can often, but not always, be detected by observing cytopathic effects that occur during virus replication (6). Additionally, cell culturing is used in studies to analyze the interaction of viruses with the cells of various organisms (7). Previously, shotgun proteomics was demonstrated as an alternative method for the identification of viruses in complex mixtures. Shotgun proteomics can be used to identify viruses using LC-MS/MS based on identified protein fragments (peptides) (6, 8–12). These studies demonstrated that shotgun proteomics-based approaches have the potential to be used as a complementary method for virus detection and identification. Shotgun proteomics refers to the direct analysis of complex protein mixtures (13). LC-MS/MS is known for successfully identifying all classes of proteins, including those originating from viral capsids and membrane proteins from enveloped viruses. In short, the proteins are subjected to proteolytic cleavage, most often with trypsin, producing smaller peptide sequences. Then, LC-MS/MS determines the unique peptide sequence pattern. The tandem mass spectra results are then compared to a protein database. This allows reference-based identification of peptides in the sample, as well as proteins. Furthermore, the identified peptides (each with a specific amino acid sequence) can be used to identify the virus or organism from which the proteins originate. This interpretation step seems simple but is complicated by several factors such as: the bias in the number of sequenced virus strains per species, inconsistent nomenclature of species names, and extreme high redundancy of some species in public databases. As a major challenge, the occurrence of similar peptide sequences in proteins of different virus species and host cells in which the virus grow, hamper

From the [1]Department of CBRN Protection, Netherlands Organization for Applied Scientific Research TNO, TNO, Rijswijk, the Netherlands; [2]HORAIZON Technology BV., Delft, the Netherlands; [3]DGA CBRN Defence Center, Vert-le-Petit, France; [4]Military Health Institute, Military Medical Agency, Prague, Czech Republic; and [5]Department of Neurology, Erasmus MC, Rotterdam, the Netherlands

*For correspondence: Armand Paauw, armand.paauw@tno.nl.

https://doi.org/10.1016/j.mcpro.2025.101004

quick, and unambiguous identification of the virus with the existing software (10).

To determine the feasibility of shotgun proteomics to identify viruses, we initially developed the proteome2virus application (app) to identify 46 most common human pathogenic viruses in Western Europe (8). This proteome2virus app demonstrated the feasibility of trustworthy identifying viruses in complex samples including clinical fecal samples and cell cultures. The power of the proteome2virus app is the simple and fast data analysis. Each virus identification is based on multiple identified peptides that only perfectly match with protein sequences of the identified virus. Moreover, if certain peptides exclusively match to a particular virus proteome, this virus must be present in the tested sample.

Furthermore, the result of the identified virus(es) is reported clearly, which enables the implementation of the method in clinical diagnostic laboratories.

However, the full potential of the method will be its ability to test for all potential pathogenic viruses at once. Moreover, routinely testing for all potential infectious viruses enables rapid detection of new pandemic threats as they emerge.

The proteomics data generated with shotgun proteomics is an agnostic alternative to test on a myriad of viruses at once. The used settings (data-depended acquisition) during the measurements use only a minimal detection level of an analyte as a selection criterion. Therefore, shotgun proteomics data will theoretically contain peptide signatures of any abundant virus present in the sample. By extending the proteome2virus app, a similar simple, robust, easy executable app is developed that direct tests on the presence of all known (potential) human pathogenic viruses in complex samples. The app proteome for pandemic preparedness (P4PP) enables to test for 1896 virus species at once. Encompassing all recognized human pathogenic virus species and additional virus species from the 32 virus families associated with human pathogens, including the proteomic variations of each species.

Subsequently, to validate the broad scope of the developed app, the proteomes of the supernatant from 174 virus cultures were measured and analyzed using the P4PP pipeline.

EXPERIMENTAL PROCEDURES

MS/MS Datasets

A total of 9 Tandem Mass Spectrometry (MS/MS) datasets were used for this study. Of all 174 samples analyzed in this study, associated data including raw data files, sample information, and supplemented files (PEAKS exported peptide lists in worksheets of file PEAKSexported_peptide_lists.xlsx, supplemented cleaned peptide lists for P4PP analysis in worksheets of pep-lists2P4PP.xlsx and P4PP downloaded peptide lists in worksheets of out-p4pp.xlsx) are described in Supplemental Table S1.

Dataset 1, PXD036663, is the dataset previously generated in house for the validation of proteome2virus containing data of 17 cultured viruses and the supernatant of 2 cell cultures used (negative controls) (8).

Dataset 2, PXD054958, contains data of 12 purchased heat inactivated supernatant of virus cultures (Helvetica Health Care Sàrl), the supernatant of an in-house cultured human adenovirus 5 (American Type Culture Collection VR-5) and modified vaccinia virus Ankara (American Type Culture Collection VR1508). The purchased supernatant of the virus cultures and human adenovirus 5 were tested undiluted. Vaccinia virus Ankara was tested, undiluted, 2.5× and 5× diluted supernatant. Human adenovirus 5 and vaccinia virus Ankara were cultured in Vero cell line. Vero cells used for infection were cultured using Dulbecco's Modified Eagle medium containing 5% fetal bovine serum with 1% Pen-Strep solution (Gibco media). Cells were incubated at 37 °C with 5% CO_2. Vero cells were grown until a 70 to 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. Cultures were maintained for 5 days before the supernatant was removed and stored in −80 °C until sample preparation was executed. Before analysis in-house cell cultured viruses were identified by PCR.

Dataset 3, PXD057159, contains the virus cultures of four Poxviridae (Cowpox CPXV-0031/2004 EP-4 lidi, Cowpox CPXV-0023/2004, BR VR302, Camelpox CML-0375/2004, CP-1 and vaccinia strain VACV-0264/2004, Elstree B5), and a negative control (uninfected Vero-E6 cell line). Cultivation details are described (14). Before analysis, cell cultured viruses were identified by PCR.

Dataset 4, PXD057670, comprised supernatant of Vero Cell cultures of three Togaviridae (Rio Negro virus, Sindbis virus and Pixuna virus) and a negative sample. Proteomic samples were prepared from in-house cultured viruses. The viruses were produced on Vero cells in M199 medium (Gibco) supplemented with 2% fetal bovine serum and 1% antibiotic-antifungal at 37 °C under 5% CO_2. Before analysis, cell cultured viruses were identified by PCR or sequencing.

Dataset 5, PXD059217, raw data from Ouwendijk *et al*. (15) in this study herpes simplex virus 1 (HSV-1, International Committee on Taxonomy of Viruses [ICTV]) designation human alphaherpesvirus 1) and varicella-zoster virus (VZV, ICTV designation human alphaherpesvirus 3) were cultured in human epithelial cells to study the proteomes of host and virus during infection (15).

Dataset 6, MSV000080032, contains eight data files, obtained from four uninfected and four human respiratory syncytial virus (RSV)-infected A549 cells. Proteins were enriched using 2D clean-up kit (GE Healthcare) as described. Digestion was done with Lys-C followed by trypsin (16).

Dataset 7, PXD018594, contains 20 data files of SARS-CoV-2 (strain 2019-nCoV/Italy-INMI1 (008N-03893)) infected Vero E6 cells. Cells were infected at two multiplicities of infection (MOIs; 0.01 and 0.001) and harvested at 1, 2, 3, 4, and 7 days post infection (17). Moreover, the strain 2019-nCoV/Italy-INMI1 (Genbank MT066156) is a Pangolin B strain (18).

Dataset 8, PXD034494, proteomic analysis of Mpox virus produced with experimentally infected Vero E6 cells in triplicate (19). For sample preparation, the SP3 protocol was used (20).

Dataset 9, PXD035900, proteomic data from a study in which among others, the global protein abundance of human bronchial epithelial cells (Normal Human Bronchial Epithelial [NHBE] cells) that were infected with A/California/04/2009 H1N1, A/Wyoming/03/2003 H3N2 or A/Vietnam/1203/2004 H5N1 influenza A strains were analyzed (21). Cell were harvested a four time points (3, 6, 12, and 18 h) post infection. From the NHBE cell infected from each time point and each strain the raw data used to study the global protein abundance in IAV-infected human cells was retrieved and analyzed with the developed proteomic pipeline (P4PP).

The mass spectrometry proteomics data generated for this study have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner

repository (22, 23) with the datasets generated in this study, 2, 3, and 4 identifiers PXD54958, PXD057159, and PXD057670, respectively. Datasets 1 (PXD036663) and 5 (PXD059217) are previously generated raw data from this study group. The additional datasets 6, 7, 8, and 9 were downloaded from the proteomeXchange database (22, 23) and MassIVE (24).

Sample Preparation

Sample preparation of samples analyzed in this study. Samples used for the generation of dataset 1 and 2 were prepared as described (8) using the modified SP3 protocol (20). Samples of dataset 4 were also prepared using the SP3 but with modifications. In short, supernatant of virus cultures were diluted in lysis buffer (lithium dodecyl sulfate 1X containing Tris/HCl, Tris base, lithium dodecyl sulfate, glycerol, EDTA, supplemented with 5% beta-mercaptoethanol). The sample was incubated 2 min at 99 °C. Subsequently, the sample was sonicated in an ultrasonic bath for 5 min, followed by cell disruption with a Precellys Evolution instrument. Then, solution of Sera-Magnetic beads was adding (2 μg of Sera-Mag particles were added to the sample). Protein binding to magnetic beads was activated with the addition of CH_3CN. Sera-Magnetic beads were retained in solution by a neodymium magnet while the liquid was removed by pipetting. Next, Sera-Magnetic beads were washed three times with 80% ethanol, using the neodymium magnet. After the third wash, digestion buffer (1 μg/μl Trypsin Gold (Promega) in 50 mM NH_4HCO_3) was added and incubated for 15 min at 50 °C. Digestion reaction was stopped by adding 0.5% TFA final concentration.

Samples for dataset 3 were prepared as described by Pajer *et al.* using the filter-aided sample preparation method (14, 25).

Samples of dataset 5 to 9 were prepared as described in the published papers (15–17, 19, 21).

MS Data Acquisition

The digests of dataset 1 and 2 were analyzed by LC-MS/MS using a nano-liquid chromatography (nano-LC) system (Ultimate 3000; Dionex) coupled to a Orbitrap mass spectrometer (Orbitrap Eclipse, Thermo Fisher Scientific) equipped with a high-field asymmetric ion mobility spectrometer device, as previous described (8).

The digests of dataset 3 were analyzed by LC-MS/MS using a nano-LC system (Ultimate 3000; Dionex) coupled to Orbitrap mass spectrometer (Q-Exactive, Thermo Fisher Scientific) (14).

The digests of dataset 4 were analyzed by LC-MS/MS using a liquid chromatography system (Vanquish, Thermo Fisher Scientific) coupled to an Orbitrap mass spectrometer (Q exactive plus, Thermo Fisher Scientific).

From the samples of dataset 5, ~1 μg of protein sample was measured using a nano-LC system as previous described (15). Moreover, protein digests were loaded onto a C18 column. To measure the HSV-1–containing samples or the negative controls, thereof, the nano-LC was coupled to a nanospray source of an LTQ Orbitrap XL, while for the VZV measurements and its negative controls the nano-LC was coupled to an Orbitrap Fusion Tribrid mass spectrometer (15).

The MS data from eight samples of dataset 6 were generated using a Waters NanoAcquity system (maximum pressure 10,000 psi) interfaced to an LTQ-Orbitrap Elite hybrid mass spectrometer as described in the protein fractionation-free workflow section (16).

The MS data of dataset 7, containing 20 data files of SARS-CoV-2 (2019-nCoV/Italy-INMI1) infected Vero E6 cells, were generated using an ultimate 3000 nano-LC system (Thermo Fisher Scientific) coupled to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) obtained as described (17).

Dataset 8: The obtained peptides after sample preparation of supernatant from Mpox cultures in the Vero 6 cell line were analyzed with an Exploris 480 tandem mass spectrometer coupled to a Vanquish Neo ultra-high performance liquid chromatography (UHPLC) system (19).

Dataset 9: The obtained peptide digest after sample preparation of influenza A–infected NHBE cells were analyzed with an Orbitrap Fusion Tribrid coupled to an Easy-nLC 1000 system (21).

For each dataset, the used UHPLC system and mass spectrometer are described in Supplemental Table S2.

Database P4PP v01 Development

Based on a similar approach for the development of the 46Virus_db used for peptide assignment and the identification of 46 different species using proteome2virus a database (**P4PP v01; P4PP v01.fasta)** was developed containing the proteome of 1896 different virus species (8). Moreover, P4PP v01 contains proteomes from 32 virus families with viruses that are able to infect humans, which are described at ViralZone 27-12-2020 (26). For all virus species described in the ICTV 2020 Master Species List (MSL36) (27) were searched if viral proteins were annotated in the National Center for Biotechnology Information virus database (18). An overview of the included virus families that are present in the database can be found at proteome2pathogen.com in the proteome for pandemic preparedness app after pressing the explore button. Next, by pressing the family, an overview of the virus species present in the database belonging to that family is shown. Viruses included are described in Supplemental Table S3. Virus species are excluded when no or a limited number of proteins (<4 and <35 proteins of RNA and DNA viruses, respectively) from that particular species were available in National Center for Biotechnology Information virus database during construction (January to April 2022). For SARS-CoV-2, 24 proteomes were included (Supplemental Table S4) based on the European Centre for Disease Prevention and Control and WHO guidelines for tracking SARS-CoV-2 variants in March 2022. Influenza A entries contain species variation until March 2022, as well. The proteome of *Homo sapiens* (UP000005640_9606) was used to remove identified peptides that potentially have sequence similarities in human and virus proteomes (28).

Peptide Assignment

The peptide assignment was executed as previously described (8). From each dataset the obtained MS spectra were assigned to peptides using PEAKS X (Bioinformatics Solutions Inc), using database **P4PP v01** and a database of the proteome of *H. sapiens* to assess the identifications that might be seen as background. The peptides were identified using a semispecific trypsin digestion setting. For each dataset the used mass tolerances, maximal missed cleavages, and posttranslational modifications are described in Supplemental Table S2.

As previously described, only peptides with a high degree of certainty (false discovery rate [FDR] of ≤0.1%) were used to determine which viruses were present in the original sample (8). 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 and annotations of posttranslational modifications were removed. The obtained cleaned list with the identified amino acids sequences of the peptides can be used for downstream data analysis.

Running the P4PP App

The peptides were processed using the developed web app called P4PP that can be accessed at proteome2pathogen.com. The supplemented file; Instructions-p4pp.pdf, is a tutorial for using the P4PP app in proteome2pathogen.com. The P4PP app uses the same analysis approach as proteome2virus (8). By extending the number of virus families from 10 to 32 and including all species of each family (if enough protein sequences were present) and not only known human

pathogenic viruses the number of virus species that can be identified expanded from 46 to 1896. Furthermore, the P4PP app differs for some virus species identified in proteome2virus. Differences are that the Parechovirus A entry contains strains Parechovirus 1, 3, 4, and 6 which are all separate entries in proteome2virus app. Next, the proteomes of SARS-CoV-1 and SARS-CoV-2 are both severe acute respiratory syndrome-related coronavirus in P4PP species analysis, while these strains are separated entries in proteome2virus app. In proteome2virus, betacoronavirus OC43 is a species entry while it is actually a strain of the virus species betacoronavirus 1. Therefore, the proteome of betacoronavirus OC43 is now added to betacoronavirus 1 entry. Differences are also indicated in Supplemental Table S3.

Next, an extra function is added for SARS and influenza A analysis to enable identification to subspecies level (viral lineage and genotype in case of SARS and influenza A, respectively).

In short, the app replaces all isoleucine (I) amino acid residues in the input peptides with leucine (L) before analysis. Subsequently, duplicate input peptides are removed. The analysis starts with a family-level search followed by a species-level search in each relevant family. In the family-level step, matches between all input peptides and a family-level viral database are searched. 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 **P4PP v01**) and also whether the peptide has a "family discriminative hit" (*i.e.* whether the peptide matches only one viral family in the **P4PP v01** and for that reason is discriminative). Each family with at least three family discriminative hits is considered to be present in the sample and therefore selected for the species-level search. With the species-level search, the identified species is the species within a family with the highest number, but at least 3, species discriminative hits.

Experimental Design and Statistical Rationale

The study utilized a total of 9 MS/MS datasets, encompassing 174 samples. The experimental design focused on ensuring the reliability and validity of the data collected from various virus cultures, the database and the developed data analysis approach.

The data analyzed it this study was generated in six different laboratories, using six different sample preparation protocols and 9 UHPLC-Orbitrap systems using often different measurement strategies to obtain MS data. To test the pipeline with as many varied virus samples, 36 different virus strains originating from 21 different virus species of nine different virus families were tested. Also, replicates and not infected cell cultures were tested abundantly to demonstrate the reproducibility of the data analysis pipeline. Of the 14 negative controls (uninfected cells), one was tested positive for human alphaherpesvirus 1, which was possibly caused by carry over effect during the shotgun proteomics measurements, because, the identified virus is the same virus species that was used in that study.

Furthermore, all data is publicly accessible and the pandemic preparedness app on proteome2pathogen.com is open access making it possible to reanalyze the published results. In this study, 174 samples were analyzed, all 21 different virus species tested were identified correctly, which indicate that the results are robust, reliable, and applicable to broader context.

RESULTS

The use of P4PP to identify viral species was evaluated for different datasets. In total, 174 MS/MS data files were processed using PEAKS X to assign the peptides using the P4PPv01 database followed by the P4PP analysis to detect and identify viral species. Of the 160 virus-infected samples analyzed, 146 were correctly identified at the species level, an additional four samples were identified at the family level. In 10 samples that were infected with a virus, no virus could be detected. These 10 negative samples were all taken shortly after the start of the infection and/or infected with a low MOI. All these 10 samples tested positive in follow-up samples obtained later in time series where negative samples were measured, indicating that the number of peptides derived from the virus was initially too low in the samples obtained at the start of the experiment. Of the 14 negative controls (uninfected cells), one was tested positive for human alphaherpesvirus 1, which was the virus studied (Fig. 1).

In more detail, the samples of dataset 1, with LC-MS/MS data from samples of 17 cultured viruses and two cell cultures were analyzed using the P4PPv01 database and the P4PP app. All 17 viruses were identified at the species level, while in the two negative controls no virus was identified (Supplemental Table S1, Supplemental Dataset 1).

The four influenza A samples identified were also subtyped correctly (twice H1N1 and H3N2).

SARS-CoV-2 strain BetaCoV/Netherlands/01 infected cell cultures with 1.10^9, 1.10^8, and 1.10^7 genome copies per milliliter in the samples were subtyped based on the described Pango lineages as (B.1.351, B.1.621, B.1.1.318, or B), (B.1.1.318, B.1.621, or B), and (B.1.1.318, B.1.1.7, B.1.351, B.1.621, or B), respectively (29). Each isolate in dataset 1 was measured in two (or three in cases of SARS-CoV-2) concentrations, where the second sample was a 10 times dilution of the first. Results show a decrease of identified peptides after 10 times dilution (Fig. 2*A*). Of the different identified virus specific peptides, a large percentage (66%–15.5%) is detected in both samples tested (Fig. 2*B*). In the sample with 10^7, 10^8, and 10^9 genome copies per milliliter SARS-CoV-2 BetaCoV/Netherlands/01 14, 26, and 28 species unique peptides were identified, respectively. In total, 44 different SARS-CoV-2–specific peptides were identified of which 7 were detected in all three samples (Fig. 3) (30).

The samples of dataset 2 containing data obtained from 13 heat inactivated virus cultures with different viruses and three data files of vaccinia lysates in three different concentrations were analyzed with P4PP app. All 13 heat inactivated virus cultures were identified at family level and 11 of these also to the species level (Supplemental Table S1, Supplemental Dataset 2). Moreover, the two influenza A samples were correctly identified to their subtype. The four SARS-CoV-2 strains from four different lineages (alpha B.1.1.7, beta B.1.351, gamma P.1, and delta B.1.617.2) were identified to the correct Pango lineage. The beta strain B.1.351 was identified as a B.1.351 or B.1.621 because 71 peptides were identified that are in the proteome of both lineages. Moreover, B.1.621 arose from the beta lineage, which implies that proteome differences are limited. Of the in total 103 species-specific peptides identified, 20 (19.4%) peptides were identified in all four SARS-CoV-2 strains (Fig. 4).

The undiluted sample of vaccinia (strain) was identified at the species level. The two other samples were identified to their family level (Poxviridae), identification of the species level was

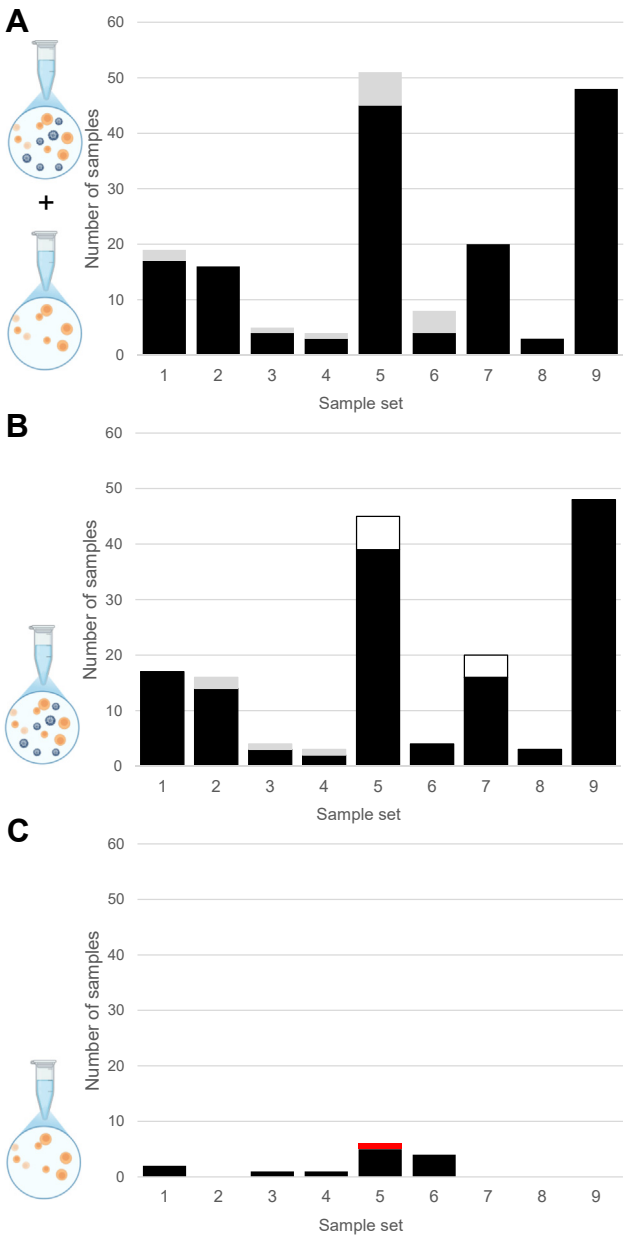


FIG. 1. **Analysis of samples using P4PP application.** *A*, this figure provides an overview of the 174 different sample sets tested. The *black bar* indicates the virus-infected cell culture samples, and the *gray bar* represents the negative controls (noninfected cell cultures). *B*, results of the 160 virus-infected samples. The *black bar* shows the number of correctly identified viruses at the species level, the *gray bar* indicates the number of correctly identified viruses at the family level, and the *white bar* represents samples where no virus was identified. The negative samples were taken shortly after infection or infected with a low MOI. All 10 samples eventually tested positive, suggesting that the viral protein levels were initially below the detection limit of our method and data analysis approach. *C*, the test results for the negative controls (noninfected cell cultures). The *black bar* represents the number of correct results (no virus detected), while the *red bar* indicate a false positive result, likely due to carryover effects. The identified virus in this false positive sample was the same species used in this study. MOI, multiplicity of infection; P4PP, proteome for pandemic preparedness.

not possible because not enough (n = 1) peptides were identified, which were discriminating for one Poxviridae species.

The samples of dataset 3 containing data obtained from five samples (four cultured Poxviridae and a negative control (supernatant of noninfected Vero6 cells), 2 Cowpox viruses and one Camelpox virus were correctly identified at the species level. Moreover, 46 (29.7%) of the identified Cowpox-specific peptides were identified in both strains analyzed. The vaccinia sample was identified to its family level (Poxviridae) but identification at species level was not possible because not enough (n = 2) discriminating peptides were identified. In the negative sample, no virus was identified (Supplemental Table S1, Supplemental Dataset 3).

In dataset 4, Sindbis virus and Pixuna virus were correctly identified at the species level based on 27 and 21 discriminating peptides, respectively. The third Togaviridae containing sample was also identified correctly, although with uncertainty at the species level. Moreover, this sample containing a Rio Negro virus was identified to its family level based on 15 discriminating peptides and at the species level <3 peptides. However, a reliable identification was not possible with P4PP. All 15 identified peptides on family level matched the Venezuelan equine encephalitis virus (VEEV) entry, as well (Table 1). The Vero cell line, which was used as a negative control tested negative for all 1896 virus species in the P4PP database.

Dataset 5 raw data from Ouwendijk *et al.* (15) in this study HSV-1 and VZV were cultured in human epithelial cells to study the proteomes of host and virus during infection (15). Dataset 5 contains 51 data files; 21, 24, and 6 data obtained from HSV-1, VZV-infected cells and not infected cell cultures, respectively (Supplemental Table S1, Supplemental Dataset 5).

From the HSV-1–infected cell cultures, samples were taken at 0, 2, 4, 6, 8, 10, and 12 h post infection. The experiment was executed in 3-fold. In the 21 samples, 4607 to 5852 unique peptides were identified. In the samples taken at 0 and after 2 h no virus was identified (Supplemental Table S1, Supplemental Dataset 5). In all other samples, HSV-1 was identified. The identified number of HSV-1–derived peptides increased overtime, consistent with the increase in the number of HSV-1 proteins as previously described (Fig. 5) (15).

Of the three negative controls during HSV-1 experiments, one sample generated a positive result for human alphaherpesvirus 1 based on three unique peptides, while in the other two no virus was identified.

From the VZV-infected cell cultures, samples were taken at 0, 3, 6, 9, 12, 15, 18, and 24 h post infection. The experiment was executed in three fold. In the 24 samples, 4992 to 6282 unique peptides were identified. In all samples, VZV was identified (Supplemental Table S1, Supplemental Dataset 5).

In these time series, the temporal correlation between incubation time and identified peptides was not clearly visible, which is in line with the results described in the original study

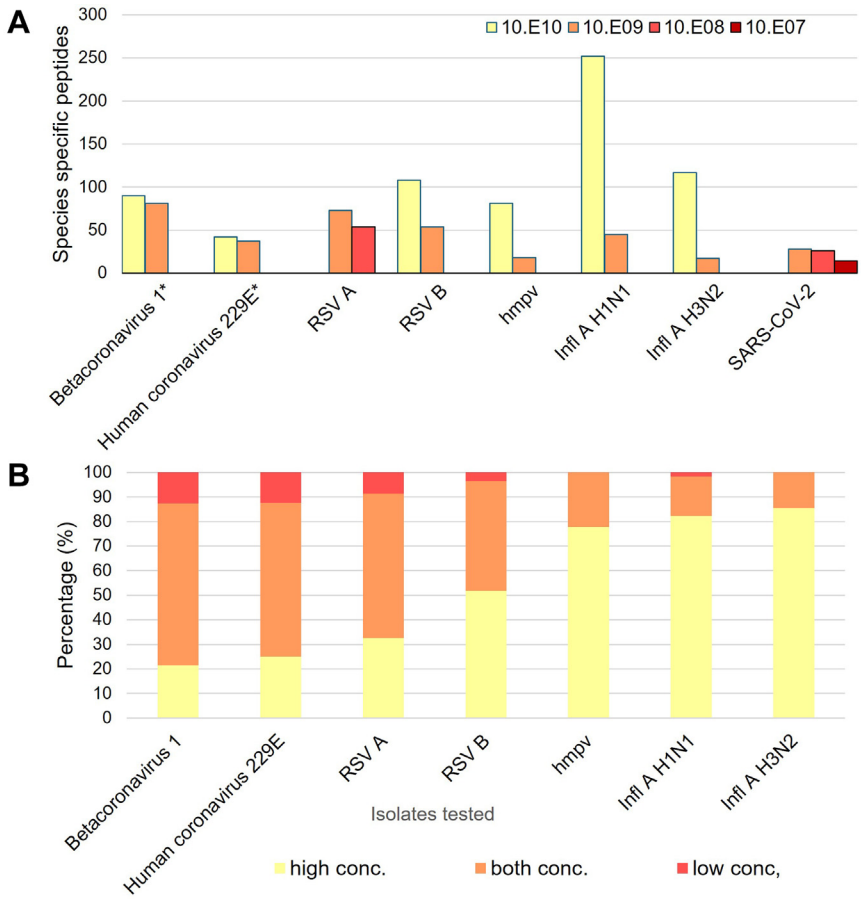


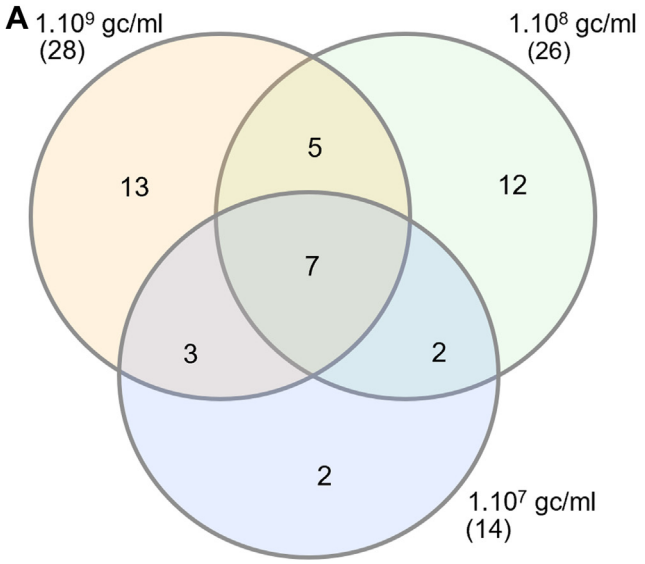
FIG. 2. **Number of peptide identifications and distribution across sample concentrations.** *A*, number of identified peptides in isolates tested in different concentrations. *Estimated concentrations, number of genome copies per milliliter was not determined. *B*, distribution of the species-specific peptides identified in samples. *Yellow* percentage of species-specific peptides identified in the highest concentration tested only. *Red* of species-specific peptides identified in the lowest concentration tested only and in *orange* the percentage of peptides identified in both samples.

(Fig. 6) (15). In that study, it was also noted that in the samples of 0 h post infection already 32 VZV proteins were detected (15). The differences between VZV and HSV-1 were explained by the ratio between infective and noninfective virus particles. The ratio of noninfectious and infectious particles (or plaque-forming units) of VZV-infected cells is 40,000:1, while for HSV-1–infected cells this is 10:1 (15, 31, 32). The three negative controls during VZV experiments tested negative.

Dataset 6 MSV000080032 contains eight data files, four uninfected, and 4 RSV-infected A549 cells (16). Four biological replicates were investigated. Therefore, samples for proteomic analysis were taken 24 h post infection. In the eight samples, 29,543 to 30,900 unique peptides were identified. In total, 52,026 different peptides were identified. In the four RSV-infected A540 cells, RSV was identified based on 124 to 145 unique peptides, while in the four uninfected A549 cells no virus was identified (Supplemental Table S1, Supplemental Dataset 6).

In total, in the four samples, 177 different species discriminative peptides were identified. Thereof, 100 (56.5%) RSV-derived peptides were identified in all four samples analyzed. Next, two to six and two to seven species discriminative peptides were identified in three and two of the four replicates tested, respectively. In each sample species, discriminative peptides (2, 8, 12 and 14) were detected which were only found in one of the four tested replicates (Fig. 7*A*) (30). Next, of the in total 258 different family discriminative peptides, 138 (53.5%) were identified in all four samples (Fig. 7*B*). Next, 4 to 12 and 2 to 14 family discriminative peptides were identified in three and two of the four replicates tested, respectively. In each sample, family discriminative peptides (4, 10, 17 and 20) were detected, which were only found in one of the four tested replicates (Fig. 7*B*).

In the dataset 7, Vero E6 cells were infected at two MOIs; 0.01 and 0.001 and harvested at 1, 2, 3, 4, and 7 days post infection (17). Experiments were biological replicated, which resulted in 20 data files (Supplemental Table S1,



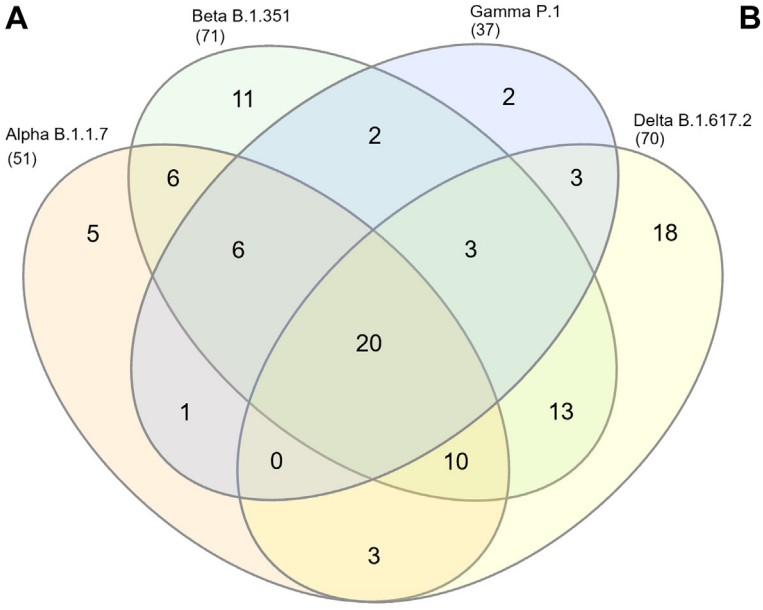
B

Repeatedly identified peptides*
LTFGGPSDSTGSNQNGER
ADETQALPQR
NPANNAALVLQLPQGTTLPK
VAGDSGFAAYSR
KKADETQALPQR
ASANLAATK
KADETQALPQR

FIG. 3. **SARS-CoV-2 identified peptides of SARS-CoV-2 strain BetaCoV/Netherlands/01 infected cell cultures at varying viral loads.** *A*, The Venn diagram shows the overlap of peptides identified from SARS-CoV-2 strain BetaCoV/Netherlands/01 in cell cultures infected with 1.10^9, 1.10^8, and 1.10^7 gc/ml. Between brackets the number of specific identified peptides in each isolate. *B*, the sequences of the seven peptides derived from SARS-CoV-2 identified in all three concentrations analyzed. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Supplemental Dataset 7). In the 20 samples, 3789 to 8854 unique peptides were identified. At the first time points (J1) in the supernatant of VeroE6-infected cells with 0.01 and 0.001 MOI, no virus was identified with the P4PP data analysis approach. In the samples taken at time points J2 to J7, SARS-CoV-2 was identified.

The number of identified SARS-CoV-2–derived peptides on species and family level at various time points corresponds to the measured protein concentrations over time, as previously described (Fig. 8) (17). With an increasing number of identified peptides the number of potential subtypes decreased. However, even with 60 different peptides identified of SARS-CoV-2



B

Repeatedly identified peptides
MAGNGGDAALALLLLDRLNQLESK
LTFGGPSDSTGSNQNGER
RGPEQTQGNFGDQELLRQGTDYKHWPQLAQFAPSASAFFGMSR
LTFGGPSDSTGSNQNGERSGAR
NAPRLTFGGPSDSTGSNQNGERSGAR
TATKAYNVTQAFGR
DHLGTRNPANNAALVLQLPQGTTLPK
YRLGNYKLNTDHSSSSDNLALLVQ
VAGDSGFAAYSR
NAPRLTFGGPSDSTGSNQNGER
VYSTGSNVFQTR
RPQGLPNNTASWFTALTQHGK
GPEQTQGNFGDQELLR
AYNVTQAFGR
GGDGKMKDLSPR
RGPEQTQGNFGDQELLR
HLDAYKTFPPTEPK
AYNVTQAFGRRGPEQTQGNFGDQELLR
NPANNAALVLQLPQGTTLPKGFYAEGSR
QKRTATKAYNVTQAFGR

FIG. 4. **Relationships of the identified peptides of the four SARS-CoV-2 strains analyzed in dataset 2.** *A*, Venn diagram of SARS-CoV-2 strains (Alpha B.1.1.7, Beta B.1.351, Gamma P.1 and Delta B.1.617.2). Between the brackets the number of species-specific peptides identified of the tested isolate. *B*, the sequences of the 20 peptides identified that were identified in all four SARS-CoV-2 strains. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

TABLE 1
Togaviridae-specific peptides identified in sample virus1 of dataset 4 with a match (1) in the proteomes of Togaviridae virus species or not (0)

Identified peptides of Togaviridae	Rio Negro virus	Venezuelan equine encephalitis virus	Mosso das Pedras virus	Seven other Togaviridae[a]
AGFLHVPYTQAPSGFAQWVK	1	1	0	0
AGFLHVPYTQAPSGFAQWVKDKPPSLK	1	1	0	0
FEHATTMPNQVGMPFNTLVNRPGYAPLALSVTPLK	1	1	0	0
FEHATTMPNQVGMPFNTLVNRPGYAPLALSVTPLKVK	1	1	0	0
GSYVEMHLPGSEVDSSLLSMSGNAVK	1	1	0	0
KLVQYAGEVYNYDFPEYGAGHAGAFGDLQAR	1	1	0	0
LQTSSQYGLDPSGTVKGR	1	1	0	0
LVPTLNLEYLTCHYK	1	1	1	0
MVAGPLSTAWTPFDR	1	1	0	0
SLDDLFKEYKLTKPYMATCAR	1	1	0	0
TALSVVTWNEKGVTVK	1	1	0	0
TSQPCHLVDGHGYFLLAR	1	1	1	0
TTSSPDVYANTNLVLQRPK	1	1	0	0
VSDTPTLSTAECTLNECVYSSDFGGLATVK	1	1	0	0
VVALVLGGANEGSR[a]	1	1	1	1

[a]peptide VVALVLGGANEGSR matches also with the proteomes of Bebaru virus, Eastern equine encephalitis virus, Highlands J virus, Semliki Forest virus, Tonate virus, and Una virus.

the exact subtype could not be determined because it identified an equal number of peptides of SARS-CoV-2 subtype B.1.1.318 and B.

Dataset 8 contained three samples of proteomic data from Mpox virus-infected Vero E6 cells (Supplemental Table S1, Supplemental Dataset 8) (19). In the three samples, 3194 to 3886 unique peptides were identified. In all samples, Mpox virus was identified on family level as a poxviridae based on 529, 513, and 570 peptides and subsequently identified as Mpox virus based on 86, 83, and 89 unique peptides, respectively (Supplemental Table S1, Supplemental Dataset 8). In total, in the three samples, 133 different species discriminative peptides were identified. There off, 66 (58.4%) were identified in all three samples analyzed. Next, 3, 8, and 11 different species discriminative peptides were identified in two of the three replicates. Also, in each sample species discriminative peptides (3, 9 and 13) were detected which are not found in the two other replicates tested (Fig. 9*A*). Next, of the in total 646 different family discriminative peptides 420 (65%) were identified in all three samples. Next, 16, 33, and 44 different family discriminative peptides were identified in two of the three replicates. Also, in each sample, family discriminative peptides (22, 50 and 61) were detected which are not found in the two other replicates tested (Fig. 9*B*).

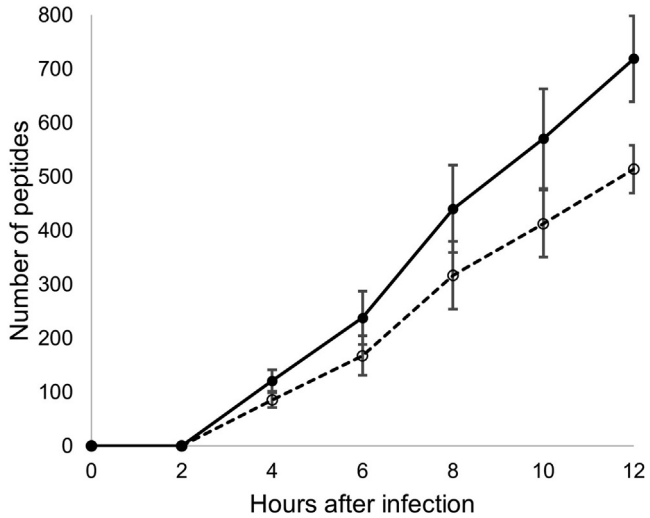


FIG. 5. **Average number of unique peptides identified of human alphaherpesvirus 1 in dataset 5.** On family level (*black line*) and species level (*dotted line*).

Of dataset 9, 48 proteomic data files in which the global protein abundance of NHBE cells that were infected with influenza A H1N1, H3N2, or H5N1 influenza A strains were analyzed (Supplemental Table S1, Supplemental Dataset 9). Cells were harvested at 3, 6, 12, and 18 h post infection. In the 48 samples, 5649 to 12,462 unique peptides were identified. In all 48 samples, influenza A was identified. The subtype was identified correctly in all except 6 samples (Supplemental Table S1, Supplemental Dataset 9). Moreover, in four samples with influenza A H1N1 and in one sample with influenza A H5N1 taken 3 h post infection, the subtype could not be determined because multiple options were possible. In one of the four samples taken 3 h post infection with the influenza A H3N2 strain the P4PP app could not discriminate between H1N2 and H3N2 based on the identified influenza A peptides. From 3 to 12 h post infection, there is a clear increase in identified peptides from all three influenza A virus strains, after which the number of identified peptides stabilizes (Fig. 10). The number of identified influenza A–derived peptides on species level shows a similar picture as the Log2 intensity of IAV NP AB detected over the time course of pH1N1, H3N2, and H5N1 IAV infection in NHBE, as previously described (21).

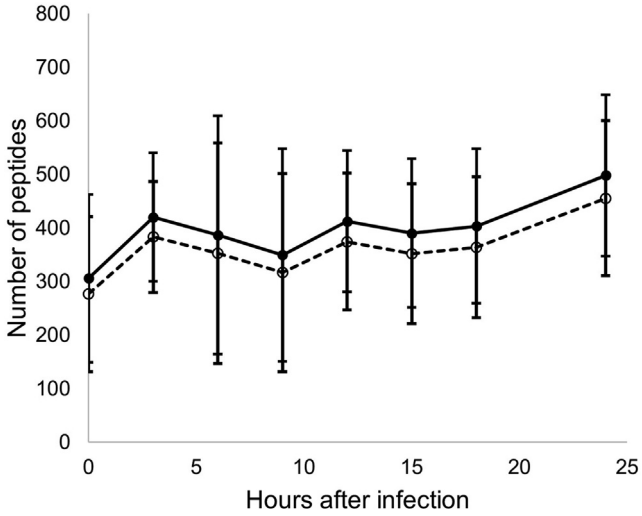


FIG. 6. **Average number of unique peptides identified of human alphaherpesvirus 3 in dataset 5.** On family level (*black line*) and species level (*dotted line*).

Finally, the distribution of the number of species false positive peptides in each sample was analyzed. In 102 (58.6%) samples, no false positive peptides were identified. In 41 (58.6%), 17 (23.6%), 5 (2.9%), 4 (2.3%), 2 (1.1%), 1 (0.6%), 1 (0.6%), and 1 (0.6%), sample(s) 1, 2, 3, 4, 5, 6, 8, and 11 false positive peptides were identified, respectively (Fig. 11). Next, the effect on the results, if the majority rule was not implemented was analyzed, for each sample. Only, in 2 (1.1%) samples a second species would have been identified. In both cases, Bovine orthopneumovirus would have been identified as a second species based on three species-specific peptides identified in samples infected with human orthopneumovirus. The 3 bovine orthopneumovirus specific peptides were identified in samples with in total four and five false positive peptides. Moreover, the other peptides in these samples were specific for Avian metapneumovirus and Murine orthopneumovirus.

DISCUSSION

In this work, we have shown that the P4PP software app is likely able to detect and identify 1896 virus species, including known virus variants within these species, from 32 different virus families. The virus families were selected because they were known to contain human pathogenic viruses (26). The P4PP app enables the use of shotgun proteomics-based analytics as an easy to execute virus detection and identification assay without the need to take potential virus species-specific properties into account (*e.g.* DNA or RNA virus), select a test panel based on clinical symptoms or to be afraid of mismatches to virus specific reagents (*e.g.* primers or antibodies in a PCR or immune-based detection assay based technology, respectively).

Influenza A Subtyping

The cultured influenza A in PXD03663, PXD054958, and PXD035900 could be subtyped if enough discriminating peptides were identified. Using this approach, influenza A subtypes H1N1, H3N2, and H5N1 were correctly identified. The capability to screen for subtypes enables detecting also nonendemic influenza A virus subtypes, if they emerge and immediately indicate which subtype it concerns.

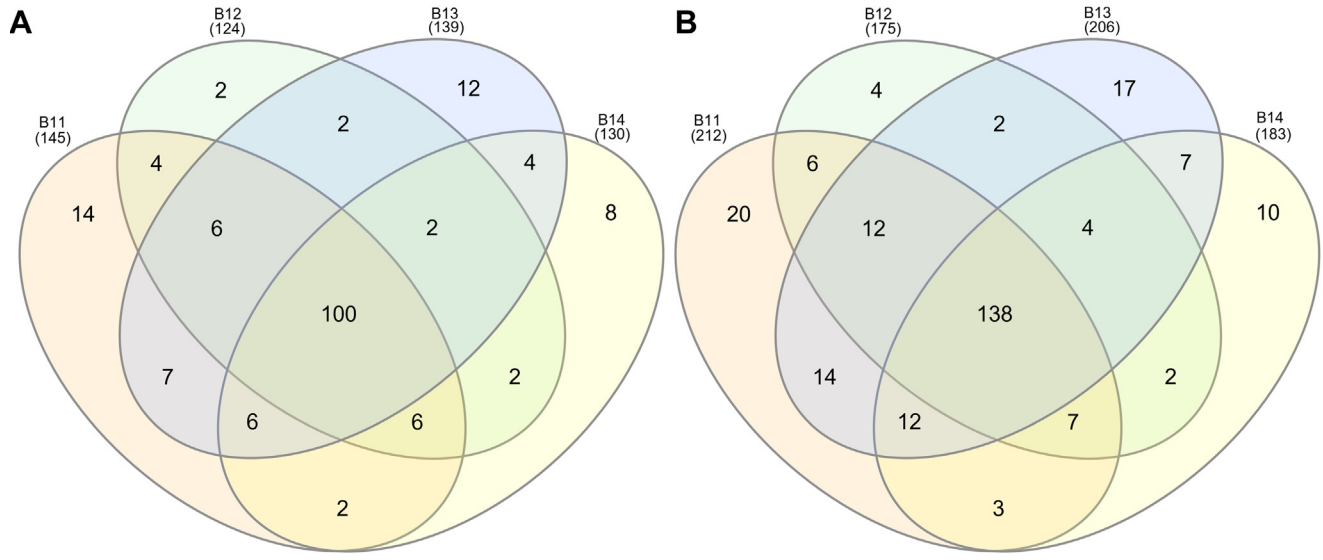


FIG. 7. **Venn diagram of the number of peptides identified in the four analyzed samples (B11, B12, B12, and B14) of RSV and infected A540 cells of MSV000080032.** *A*, number of species discriminative peptides of each sample. *B*, number of family discriminative peptides of each sample. Between brackets the number of species unique peptides identified in the sample. RSV, respiratory syncytial virus.

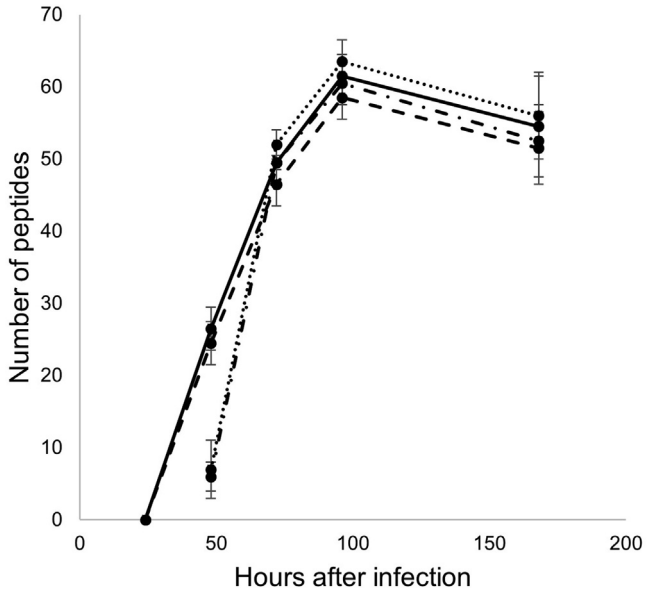


FIG. 8. **Increase of the average number of unique peptides identified of SARS-CoV-2 in dataset 7.** Infected Vero E6 cells with MOI 0.01 on family level (*black line*) and species level (*large dotted line*) and MOI 0.001 on family discriminative peptides (*small dotted line*) and species level (*irregular line*). MOI, multiplicity of infection. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

SARS-CoV-2 Subtyping

The database contained the proteomes of SARS-CoV-2 isolates until March 2022. All analyzed SARS-CoV-2 proteomes were obtained from isolates before that date. This study shows that our approach can be used for subtyping SARS-CoV-2 if enough SARS-CoV-2–derived peptides are identified. However, to be able to subtype new variants circulating the **P4PP v01** database and P4PP software analysis database should be updated frequently and it should be noted that subtyping with new isolates is not useful until the database is supplemented with known circulating SARS-CoV-2 subtypes.

The analysis of Togaviridae strains showed species identification is complicated.

Rio Negro virus and Pixuna virus are both species within the VEEV IV group, which also includes other species such as Everglades virus, Tonate virus, Mucambo virus, Mosso das Pedras virus, Cabassou virus, and six different VEEV species (33). In the P4PP analysis and P4PP v01 database, VEEV is listed as a single entry. However, the six different VEEV species are not genetically most related to each other but are distributed among other species within the VEEV IV group. Therefore, a high number of species-specific peptides are required to accurately identify Togaviridae at the species level. The limited number of peptides identified for Rio Negro virus may explain why it was not as well identified compared to Pixuna virus (15 *versus* 35 peptides). Of the 35 peptides identified for Pixuna virus, 11 also matched the proteome of the identified VEEV entry (data not shown).

We obtained additional data from other studies in which the expression of virus and host proteins were analyzed thoroughly with shotgun proteomics (15–17, 19, 21). The raw data were analyzed with the P4PP data analysis pipeline to test the compatibility with MS data obtained with other sample preparation protocols, liquid chromatography equipment, mass spectrometers, and/or data acquisition programs. In a study of one (or a few) virus strain(s), the use of a database with only the proteome of the studied viruses and the host cell for data

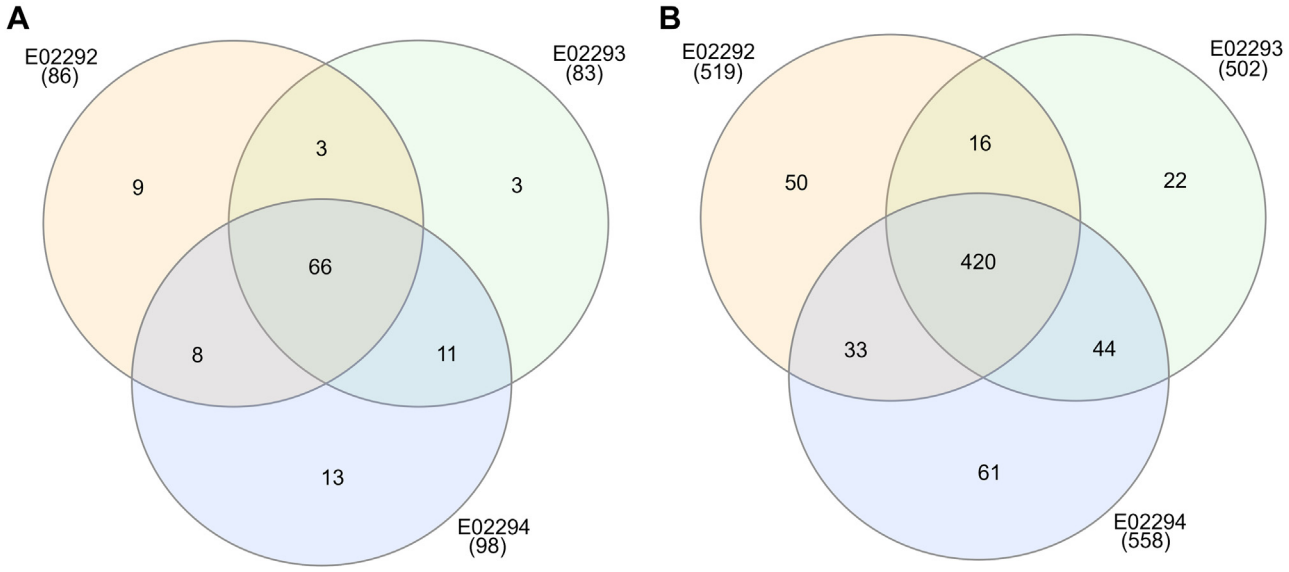


FIG. 9. **Venn diagram of the number of peptides identified in the three analyzed samples (E02292, E02293, and E02294) of Mpox of PXD034494.** *A*, number of species discriminative peptides of each sample. *B*, number of family discriminative peptides of each sample. Between brackets the number of species unique peptides in the sample.

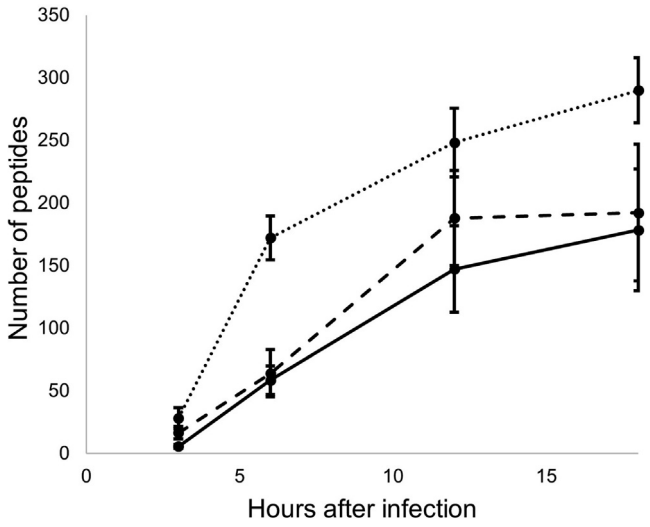


FIG. 10. **Increase in the average number of unique peptides identified in influenza A strains.** H1N1 (*black line*), H3N2 (*large dotted line*), and H5N1 (*small dotted line*) grown in human bronchial epithelial cells (NHBE cells) overtime. NHBE, Normal Human Bronchial Epithelial.

analysis is usual. Therefore, unexpected contaminant coinfecting viruses will not be noticed. Our results showed that none of the datasets used in this study contained a contaminant coinfection, indicating the high quality of the used datasets. Furthermore, it partially explains the differences between the results of previous studies and those presented in this study. Moreover, in the RSV study (Supplemental Dataset 6) now 52,026 different peptides were identified, which is a 23% increase of identified peptides compared to original publication (16). This increase could be attributed to the use of different databases as well as the advancements in proteomics data analysis over the past decade. In dataset 7 and 8, in the original studies, 94 SARS-CoV-2 and 680 Mpox-derived peptides identified were identified, respectively, while in this study, 90 SARS-CoV-2 and 646 Mpox-derived peptides were identified (17, 19). Given the extensive data analysis, the

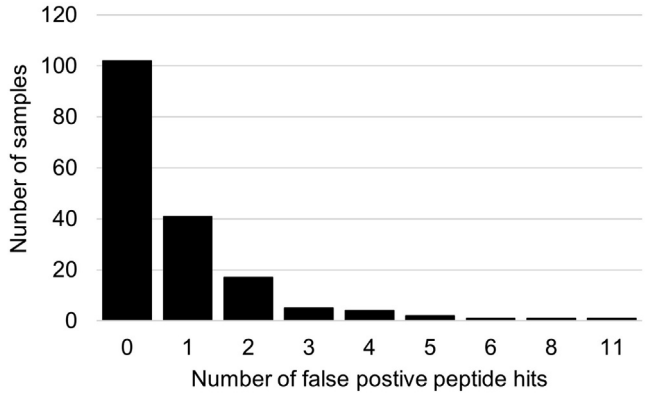


FIG. 11. **Distribution of the number of species false positive peptides in each sample.**

variety of software programs, the differences in database sizes, and the stricter FDR applied in this study, the observed difference is minimal. Theoretical, multiple virus species from different virus families in a sample can be identified with P4PP but this is not tested in this study. In this study, cell lines were infected with one virus species therefore no second virus from another virus family is identified.

The robustness of our proteome analysis approach was demonstrated by the different datasets tested. Except for the putative cross-contamination, no false positive identifications were observed. Furthermore, the majority rule in species identification was applied in only 1.1% of the samples. The number of false positive peptide identifications was very limited due to the use of a stringent FDR of $\leq$0.1%.

For the generation of the MS data in the different datasets (including the datasets generated in this study), data were generated in six different laboratories, using six different sample preparation protocols and 9 UHPLC-Orbitrap systems using often different measurement strategies to obtain MS data (Supplemental Table S2). The SP3 protocol, used in four of the datasets analyzed, is a rapid easy-to-execute method, which yields enough peptides to identify the cultivated virus (8, 19).

Results from this work suggest that the developed proteome data analysis approach is a promising method for the detection and identification of viruses in cell culture samples, or to control the identity of grown viruses in case of a potential contamination with another virus.

In this study, 21 different virus species were identified without false positive identifications, demonstrating the robustness of the developed MS data analysis approach, which also indicate that it is likely that other virus species included in the database can be identified, as well. However, confirmation of the results is recommended until the app is validated more thoroughly.

The P4PP v01 contains proteomes from 32 virus families with viruses that are able to infect humans. If a new pathogenic virus emerges from another virus family, the virus will not be detected until the app and database are updated. Clearly, the detection of unexpected virus species is a trade-off between routine practicality, required information level, costs, speed, and comprehensiveness. Recent studies showed the potential of shotgun proteomics for detection of most virus species at once as well (9–11). In these studies, the used data analysis approaches require complex bioinformatic analysis expertise and aim to strain level identification of the virus, which can be used when other diagnostic methods fail.

The developed P4PP app enables to screen for pathogenic viruses to the species level for all species included (except for influenza A and SARS-CoV-2, which can be subtyped). Pathogenic viruses are identified based on multiple peptides to the family level and subsequently to the species level. Therefore, P4PP will identify a new virus variant to its family and/or species level and not miss

emerging virus strain variants. Moreover, in routine clinical diagnostics the first priority is to determine the cause of an infection because this will generate a specific medical intervention. If there is a suspected clinical, a public health urgency or a scientific interest, the same MS data can be analyzed to obtain strain information with recently described methods (9–11).

In routine diagnostic laboratories, P4PP can be implemented and executed. The ability to test directly on all known possible pathogenic viruses species generates several advantages that targeted diagnostics lack. New, (re-)emerging viruses or viruses that jump from wildlife and/or livestock can be detected faster, it increases the number of viruses that are identified in one measurement, it generates an objective overview of endemic viruses and the method is not depending on pathogen specific materials (primers, antibodies, and crRNAs).

Metagenomic sequencing (MGS) is also an option to identify cultivated viruses. However, the simplicity of the sample preparation and the robustness of different data acquisition methods in combination with the developed P4PP data analysis approach generates a direct actionable result makes shotgun proteomics a potentially faster but at least an equivalent method to MGS. The study of Ye *et al.* showed that a virus can be cultured in a cell line and subsequently identified based on identified peptides with LC-MS/MS (6). The advantages of a proteomics-based analysis approach on viruses cultured in cells compared to MGS are that it is possible to screen for DNA and RNA viruses simultaneously, an easy data collection and a shorter data analysis time (6). Another possible advantage of a cell culturing step in the analysis of in this case (waste) water samples is that it demonstrates the viability of the identified virus (6). In case the method is used for biosurveillance purposes, an identified virus directly indicates whether a water source (pond, waste, lake, sea) contains a virus that may pose a risk to our society.

Incidental findings of an infectious agent in a clinical sample with shotgun proteomics can raise ethical challenges similar as previous described for shotgun metagenomics (34). Results can be suppressed by clinical virologist if these virus species are not requested to be tested on or alternatively the software is adjusted to analyze only a subset of virus species or virus families that are of interest. However, these limitations also limit the preparedness for unexpected emerging viruses.

Recently, Mpox and SARS-CoV-2 pandemics have shown that testing for the most common infectious viruses may neglect the detection of (re-)emerging uncommon virus species. Implementing the developed virus proteome analysis capability in viral diagnostic laboratories will improve the ability to cope with unexpected, mutated or re-emerging viruses. Potentially, P4PP can be used as an additional complementary tool in the toolbox for diagnosing causes of infectious disease outbreaks or in cases where a virus can be cultivated but standard tests to identify the virus remain negative.

DATA AVAILABILITY

All the mass spectrometry proteomics data generated for this study have been deposited to the ProteomeXchange Consortium as described in Experimental Procedures paragraph MS/MS datasets.

Supplemental Data—This article contains supplemental data (8, 14–17, 19, 21).

Acknowledgments—We would like to thank Inge D. Wijnberg and Hugo-Jan Jansen (Ministry of Defence, the Netherlands) for the comments and critical reading of the manuscript and Clotilde Favino, Mathieu Giraud, and Valérie Lafontaine (DGA, France) for generating the data of dataset 4, analysis, extraction, and virus culture, respectively.

Author Contributions—A. P., E. L., I. A. I. V.-V., I. M. F. M., V. R., M. E., J. D., P. P., C. S., H. C. v. L., T. M. L., and L. M. H. writing–review and editing; A. P., H. C. v. L., and L. M. H. writing–original draft; A. P. visualization; A. P., E. L., I. A. I. V.-V., I. M. F. M., V. R., M. E., J. D., P. P., C. S., H. C. v. L., and T. M. L. validation; A. P., E. L., and L. M. H. supervision; A. P. and E. L. software; A. P., E. L., and M. E. methodology; A. P., E. L., I. A. I. V.-V., I. M. F. M., V. R., M. E., J. D., P. P., C. S., H. C. v. L., T. M. L., and L. M. H. investigation; A. P., I. M. F. M., J. D., and C. S. formal analysis; A. P., E. L., I. A. I. V.-V., I. M. F. M., V. R., M. E., J. D., and C. S. data curation; A. P., E. L., I. A. I. V.-V., V. R., J. D., and L. M. H. conceptualization; L. M. H. project administration; L. M. H. funding acquisition.

Funding and Additional Information—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.

Conflict of Interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: App, Application; FDR, False discovery rate; HSV-1, Herpes simplex virus 1; HTBE, Human bronchial epithelial; ICTV, International Committee on Taxonomy of Viruses; MGS, Meta genomic sequencing; MOI, Multiplicity of infection; NHBE, Normal Human Bronchial Epithelial; P4PP, Proteome for pandemic preparedness; RSV, Respiratory syncytial virus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; UHPLC, ultra-high performance liquid chromatography; VEEV, Venezuelan equine encephalitis virus; VZV, Varicella-zoster virus.

Received January 10, 2025, and in revised form, May 22, 2025
Published, MCPRO Papers in Press, May 29, 2025, https://doi.org/10.1016/j.mcpro.2025.101004

REFERENCES

1. Minhaj, F. S., Ogale, Y. P., Whitehill, F., Schultz, J., Foote, M., Davidson, W., *et al*., Monkeypox Response Team, 2. (2022) Monkeypox outbreak - nine states, may 2022. *MMWR Morb. Mortal. Wkly. Rep.* **71**, 764–769
2. Perez Duque, M., Ribeiro, S., Martins, J. V., Casaca, P., Leite, P. P., Tavares, M., *et al*. (2022) Ongoing monkeypox virus outbreak, Portugal, 29 april to 23 may 2022. *Euro Surveill.* **27**, 2200424
3. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., *et al*., China Novel Coronavirus Investigating and Research Team. (2020) A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733
4. Wang, C., Horby, P. W., Hayden, F. G., and Gao, G. F. (2020) A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473
5. Colella, J. P., Bates, J., Burneo, S. F., Camacho, M. A., Carrion Bonilla, C., Constable, I., *et al*. (2021) Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. *PLoS Pathog.* **17**, e1009583
6. Ye, Y., Zhao, L., Imperiale, M. J., and Wigginton, K. R. (2019) Integrated cell culture-mass spectrometry method for infectious human virus monitoring. *Environ. Sci. Technol. Lett.* **6**, 407–412
7. Zecha, J., Lee, C., Bayer, F. P., Meng, C., Grass, V., Zerweck, J., *et al*. (2020) Data, reagents, assays and merits of proteomics for SARS-CoV-2 research and testing. *Mol. Cell. Proteomics* **19**, 1503–1522
8. Balvers, M., Gordijn, I. F., Voskamp-Visser, I., Schelling, M. F. A., Schuurman, R., Heikens, E., *et al*. (2023) Proteome2virus: shotgun mass spectrometry data analysis pipeline for virus identification. *J. Clin. Virol. Plus* **3**, 100147
9. Kuhring, M., Doellinger, J., Nitsche, A., Muth, T., and Renard, B. Y. (2020) TaxIt: an iterative computational pipeline for untargeted strain-level identification using MS/MS spectra from pathogenic single-organism samples. *J. Proteome Res.* **19**, 2501–2510
10. Holstein, T., Kistner, F., Martens, L., and Muth, T. (2023) PepGM: a probabilistic graphical model for taxonomic inference of viral proteome samples with associated confidence scores. *Bioinformatics* **39**, btad289
11. Lozano, C., Pible, O., Eschlimann, M., Giraud, M., Debroas, S., Gaillard, J., *et al*. (2024) Universal identification of pathogenic viruses by liquid chromatography coupled to tandem mass spectrometry proteotyping. *Mol. Cell. Proteomics* **23**, 100822
12. Majchrzykiewicz-Koehorst, J. A., Heikens, E., Trip, H., Hulst, A. G., de Jong, A. L., Viveen, M. C., *et al*. (2015) Rapid and generic identification of influenza A and other respiratory viruses with mass spectrometry. *J. Virol. Methods* **213**, 75–83
13. Garcia, A. (2007) Two-dimensional gel electrophoresis in platelet proteomics research. *Methods Mol. Med.* **139**, 339–353
14. Pajer, P., Dresler, J., Kabíckova, H., Písa, L., Aganov, P., Fucik, K., *et al*. (2017) Characterization of two historic smallpox specimens from a Czech museum. *Viruses* **9**, 200
15. Ouwendijk, W. J. D., Dekker, L. J. M., van den Ham, H., Lenac Rovis, T., Haefner, E. S., Jonjic, S., *et al*. (2020) Analysis of virus and host proteomes during productive HSV-1 and VZV infection in human epithelial cells. *Front. Microbiol.* **11**, 1179
16. Dave, K. A., Norris, E. L., Bukreyev, A. A., Headlam, M. J., Buchholz, U. J., Singh, T., *et al*. (2014) A comprehensive proteomic view of responses of A549 type II alveolar epithelial cells to human respiratory syncytial virus infection. *Mol. Cell. Proteomics* **13**, 3250–3269
17. Grenga, L., Gallais, F., Pible, O., Gaillard, J., Gouveia, D., Batina, H., *et al*. (2020) Shotgun proteomics analysis of SARS-CoV-2-infected cells and how it can optimize whole viral particle antigen production for vaccines. *Emerg. Microbes Infect.* **9**, 1712–1721
18. Hatcher, E. L., Zhdanov, S. A., Bao, Y., Blinkova, O., Nawrocki, E. P., Ostapchuck, Y., *et al*. (2017) Virus variation resource - improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45**, D482–D490
19. Lozano, C., Grenga, L., Gallais, F., Miotello, G., Bellanger, L., and Armengaud, J. (2023) Mass spectrometry detection of monkeypox virus: comprehensive coverage for ranking the most responsive peptide markers. *Proteomics* **23**, e2200253
20. Hayoun, K., Gouveia, D., Grenga, L., Pible, O., Armengaud, J., and Alpha-Bazin, B. (2019) Evaluation of sample preparation methods for fast proteotyping of microorganisms by tandem mass spectrometry. *Front. Microbiol.* **10**, 1985
21. Haas, K. M., McGregor, M. J., Bouhaddou, M., Polacco, B. J., Kim, E., Nguyen, T. T., *et al*. (2023) Proteomic and genetic analyses of influenza A viruses identify pan-viral host targets. *Nat. Commun.* **14**, 6030
22. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., *et al*. (2019) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450
23. Deutsch, E. W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J. J., Kundu, D. J., *et al*. (2020) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* **48**, D1145–D1152
24. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., and Bandeira, N. (2018) Assembling the community-scale discoverable human proteome. *Cell. Syst.* **7**, 412–421.e5
25. Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362
26. Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., *et al*. (2011) ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**, 576
27. Siddell, S. G., Smith, D. B., Adriaenssens, E., Alfenas-Zerbini, P., Dutilh, B. E., Garcia, M. L., *et al*. (2023) Virus taxonomy and the role of the international committee on taxonomy of viruses (ICTV). *J. Gen. Virol.* **104**, 001840
28. UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489
29. Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., *et al*. (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407
30. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015) InteractiVenn: a web-based tool for the analysis of sets through venn diagrams. *BMC Bioinformatics* **16**, 169
31. Watson, D. H., Russel, W. C., and Wildy, P. (1963) Electron microscopic particle counts on herpes virus using the phosphotungstate negative staining technique. *Virology* **19**, 250–260
32. Carpenter, J. E., Henderson, E. P., and Grose, C. (2009) Enumeration of an extremely high particle-to-PFU ratio for varicella-zoster virus. *J. Virol.* **83**, 6917–6921
33. Chen, R., Mukhopadhyay, S., Merits, A., Bolling, B., Nasar, F., Coffey, L. L., *et al*., Ictv Report Consortium. (2018) ICTV virus taxonomy profile: togaviridae. *J. Gen. Virol.* **99**, 761–762
34. Houldcroft, C. J., Beale, M. A., and Breuer, J. (2017) Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* **15**, 183–192