Catching the Wave: Detecting Strain-Specific SARS-CoV-2 Peptides in Clinical Samples Collected during Infection Waves from Diverse Geographical Locations

Subina Mehta, Valdemir M Carvalho, Andrew T Rajczewski, Olivier Pible, Björn A Grüning, James E Johnson, Reid Wagner, Jean Armengaud, Timothy J Griffin, Pratik D Jagtap

viruses

MDPI

Article

Catching the Wave: Detecting Strain-Specific SARS-CoV-2 Peptides in Clinical Samples Collected during Infection Waves from Diverse Geographical Locations

Subina Mehta [1], Valdemir M. Carvalho [2], Andrew T. Rajczewski [1], Olivier Pible [3], Björn A. Grüning [4], James E. Johnson [5], Reid Wagner [5], Jean Armengaud [3], Timothy J. Griffin [1,*] and Pratik D. Jagtap [1,*]

[1] Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN 55455, USA
[2] Division of Research and Development, Fleury Group, São Paulo 04344-070, Brazil
[3] Département Médicaments et Technologies pour la Santé (DMTS), Université Paris-Saclay, CEA, INRAE, 30200 Bagnols-sur-Cèze, France
[4] Department of Computer Science, University of Freiburg, 79110 Freiburg, Germany
[5] Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN 55455, USA
* Correspondence: tgriffin@umn.edu (T.J.G.); pjagtap@umn.edu (P.D.J.)

Abstract: The Coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) resulted in a major health crisis worldwide with its continuously emerging new strains, resulting in new viral variants that drive "waves" of infection. PCR or antigen detection assays have been routinely used to detect clinical infections; however, the emergence of these newer strains has presented challenges in detection. One of the alternatives has been to detect and characterize variant-specific peptide sequences from viral proteins using mass spectrometry (MS)-based methods. MS methods can potentially help in both diagnostics and vaccine development by understanding the dynamic changes in the viral proteome associated with specific strains and infection waves. In this study, we developed an accessible, flexible, and shareable bioinformatics workflow that was implemented in the Galaxy Platform to detect variant-specific peptide sequences from MS data derived from the clinical samples. We demonstrated the utility of the workflow by characterizing published clinical data from across the world during various pandemic waves. Our analysis identified six SARS-CoV-2 variant-specific peptides suitable for confident detection by MS in commonly collected clinical samples.

Keywords: SARS-CoV-2; variant detection; strain-specific; mass-spectrometry

Citation: Mehta, S.; Carvalho, V.M.; Rajczewski, A.T.; Pible, O.; Grüning, B.A.; Johnson, J.E.; Wagner, R.; Armengaud, J.; Griffin, T.J.; Jagtap, P.D. Catching the Wave: Detecting Strain-Specific SARS-CoV-2 Peptides in Clinical Samples Collected during Infection Waves from Diverse Geographical Locations. *Viruses* **2022**, *14*, 2205. https://doi.org/10.3390/v14102205

Academic Editors: Lars Schaade and Marica Grossegesse

Received: 5 July 2022
Accepted: 5 October 2022
Published: 7 October 2022

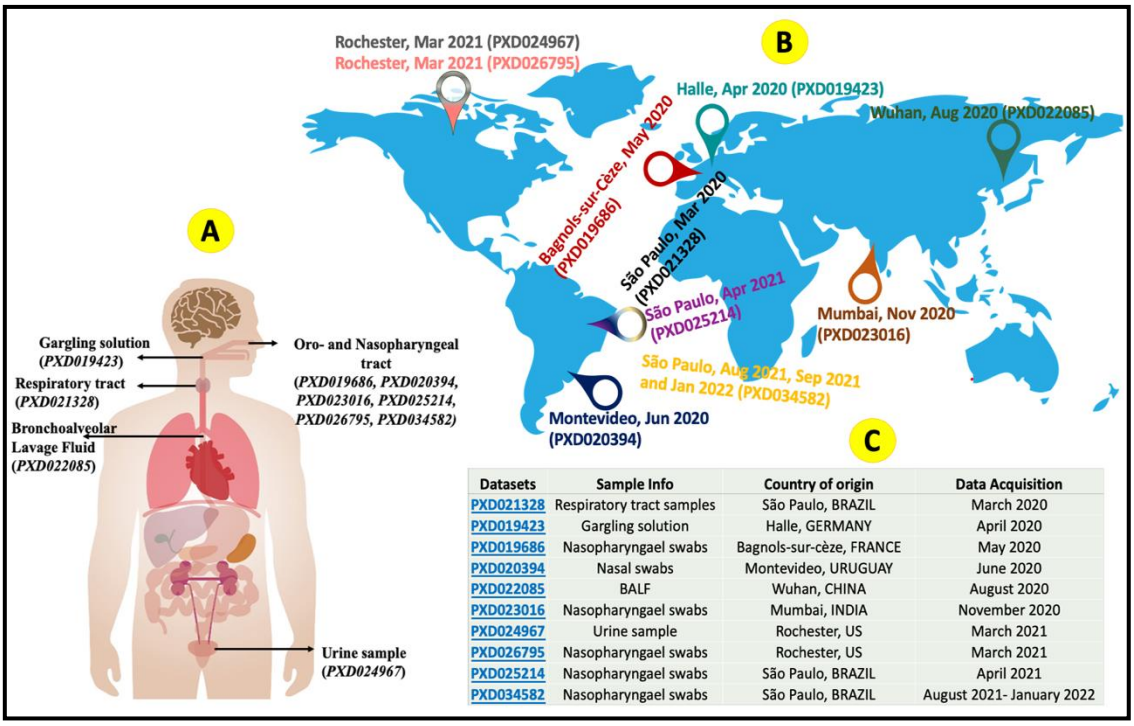
1. Introduction

It has been more than two years since the Coronavirus disease 2019 (COVID-19) outbreak, which has since spread worldwide, resulting in almost 6.3 M deaths [1]. Infected patients have exhibited symptoms that range from asymptomatic to mild fever, cough, and myalgia to severe respiratory distress, organ failure, and death in critical cases. Methods for the detection of viral infection, as well as vaccines and therapeutics, have improved the situation; however, continuous viral mutations, especially in low-vaccinated geographical regions [2], have led to the emergence of new variants at different locations and times across the world. These variants show distinct characteristics with respect to incubation times, infection routes, and severity of the disease [3]. This has posed serious challenges at multiple levels including (a) medical intervention by healthcare workers [4]; (b) detection of new strains carrying new genetic mutations in the population by clinical labs; (c) vaccine efficacy for pharmaceutical companies [5,6], and (d) monitoring of the temporal and geographical course of the pandemic for the scientific community [7]. It

became extremely critical to detect the new strains using molecular techniques to monitor the progression of new waves of the pandemic. The characterization of the viral protein sequences specific to newly defined variants is important, as it directly identifies the structural molecules (nucleocapsid, membrane, and spike proteins) that are recognized by antigen tests and are important molecular targets for vaccine development and other therapeutics [8].

Most commonly, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection is detected using the RT-PCR of patient-derived swab samples, or at-home kits recognizing viral antigens [9]. Despite the effectiveness of these diagnostic tests for the rapid detection of viral infection, other approaches offer a more in-depth characterization of clinical samples [10,11]. Mass spectrometry (MS)-based proteomics provide an orthogonal method to understand the status of infection by directly characterizing the virus-expressed proteins from commonly collected clinical samples (e.g., nasal swabs) [12]. MS-based instrumentation platforms can characterize many clinical samples with high throughput. For example, a turbulent flow chromatography-mass spectrometry (TFC-MS) system method developed by Carvalho and colleagues allows a high-throughput multiplexed analysis of more than 500 samples per day [13]. A customized bioinformatics analysis that allows for the biological interpretation of the complex MS data that are generated is also a requirement to understand the dynamics of viral-protein expression from clinical samples. Fortunately, these bioinformatic tools exist. In our previous studies, we published MS-proteomics-based informatics workflows that can detect and verify SARS-CoV-2 peptides, including those specific to characterized variants [14] and also peptides from potential co-infection pathogens [15] that may be specific to infection waves. These are deployed within the Galaxy bioinformatics ecosystem [16], providing a workbench wherein scientists can share, analyze, and visualize their results in a reproducible manner, carrying out analyses on scalable computer resources accessed through a web browser interface. The ecosystem also provides extensive online and on-demand training material via the Galaxy Training Network [17], including guidance on the usage of the platform for SARS-CoV-2 studies [18].

In this study, our MS proteomics-based Galaxy workflows identified six SARS-CoV-2 variant-specific peptides from published clinical samples. This was achieved by extending our previously published workflows by analyzing 12 previously generated and published MS-based proteomics datasets from a variety of geographical areas and infection wave timelines. These datasets cover a timeline from March 2020 to January 2022 and cover seven countries and three continents (Figure 1, Supplementary Data 1—Table S1). We leveraged the flexibility of workflows in Galaxy to match the peptide mass spectra acquired by tandem mass spectrometry (MS/MS) within these published datasets and against an updated SARS-CoV-2 protein sequence database, including sequences specific to variants classified by the World Health Organization (WHO) [19]. Our discovery workflow first detects SARS-CoV-2 peptides of specific sequences from the clinical samples. Identified peptides from all the datasets are combined and re-evaluated using the PepQuery [20] search engine to verify the presence of these peptide spectrum matches (PSMs) before validating their spectral quality by visual inspection. After the diligent evaluation and confirmation of spectral quality, the resultant peptides are aligned against the wild-type SARS-CoV-2 proteome to identify peptide sequences specific to WHO-classified variants and their associated phylogenetic lineages. Our results from the re-analysis of these datasets provided a demonstration of the power of this approach, characterizing protein sequence changes and verifying MS-detectable peptides that follow temporal and geographic dynamics of SARS-CoV-2 infection waves from diverse clinical sample types. In addition, our bioinformatics tools and workflows that generate these results are well-documented and freely and easily accessed, providing a means for others to utilize this approach in the characterization of SARS-CoV-2 samples or potential studies of other viral infections.



Datasets	Sample Info	Country of origin	Data Acquisition
PXD021328	Respiratory tract samples	São Paulo, BRAZIL	March 2020
PXD019423	Gargling solution	Halle, GERMANY	April 2020
PXD019686	Nasopharyngael swabs	Bagnols-sur-cèze, FRANCE	May 2020
PXD020394	Nasal swabs	Montevideo, URUGUAY	June 2020
PXD022085	BALF	Wuhan, CHINA	August 2020
PXD023016	Nasopharyngael swabs	Mumbai, INDIA	November 2020
PXD024967	Urine sample	Rochester, US	March 2021
PXD026795	Nasopharyngael swabs	Rochester, US	March 2021
PXD025214	Nasopharyngael swabs	São Paulo, BRAZIL	April 2021
PXD034582	Nasopharyngael swabs	São Paulo, BRAZIL	August 2021- January 2022

Figure 1. Publicly available clinical tandem mass spectrometry (MS/MS) datasets from ProteomeXchange were used for the variant detection study. (**A**) Samples came from different parts of the human body. (**B**) Samples and generated datasets were obtained at different timepoints and locations, following the geographical and temporal dynamics of the infection waves. (**C**) Table summarizing the ProteomeXchange accession numbers and geographical and temporal information associated with each dataset.

2. Materials and Methods

2.1. Clinical Datasets

Twelve clinical MS datasets (Figure 1) available via the ProteomeXchange consortium were used to detect variant peptides and proteins. For our timeline and geographical evaluation of the COVID-19 infection, we chose five nasopharyngeal swab datasets (PXD019686 [21], PXD023016 [22], PXD034582 (August, September, and January)); a gargling sample dataset (PXD019423) [23]; a nasal swab (PXD020394); an upper respiratory tract sample (PXD021328) [13]; a BALF sample (PXD022085) and a urine sample (PXD024967) [24], all collected at different times and locations. Note that the datasets collected in Sao Paulo (PXD021328 and PXD034583) were pooled clinical samples, wherein each raw file contained data from two patients. Supplementary Data 1—Table S1 also provides more details on these previously published datasets.

2.2. Discovery Workflow

Our previous study [14] used two workflows: a) discovery workflow for COVID-19 peptide detection and b) verification workflow for confirmation using the PepQuery tool (Galaxy Version 1.6.2+galaxy1) [20]. The discovery workflow (Figure 2A) used several sequence database search algorithms, such as X! tandem, MSGF+, and OMSSA within SearchGUI (Galaxy Version 3.3.10.1) [25]/PeptideShaker (Galaxy Version 1.16.36.3) and Andromeda [26] within MaxQuant (Galaxy Version 1.6.17.0+galaxy3) [27] to detect PSMs, peptides, and infer proteins at a 1% global False Discovery Rate (FDR). The COVID-19 protein sequence database used for matching peptide MS/MS data consisted of the nucleocapsid, membrane, and spike protein mutations from the variants of concern (B.1.1.7, B.1.351, P.1, B.1.671.2, AY.4, AY.4.2, XE, B.1.1.529, BA.1, BA.2, BA.3, BA.4). Along with variant structural proteins, we added all the proteins from the wild-type strain

(EPI_ISL_402124—dated 30 December 2019), and sequences were obtained from the GISAID database (https://www.gisaid.org/, last accessed on 3 February 2022). As the datasets were from clinical samples, we added human proteins and common contaminants to the COVID-19 protein sequence database.

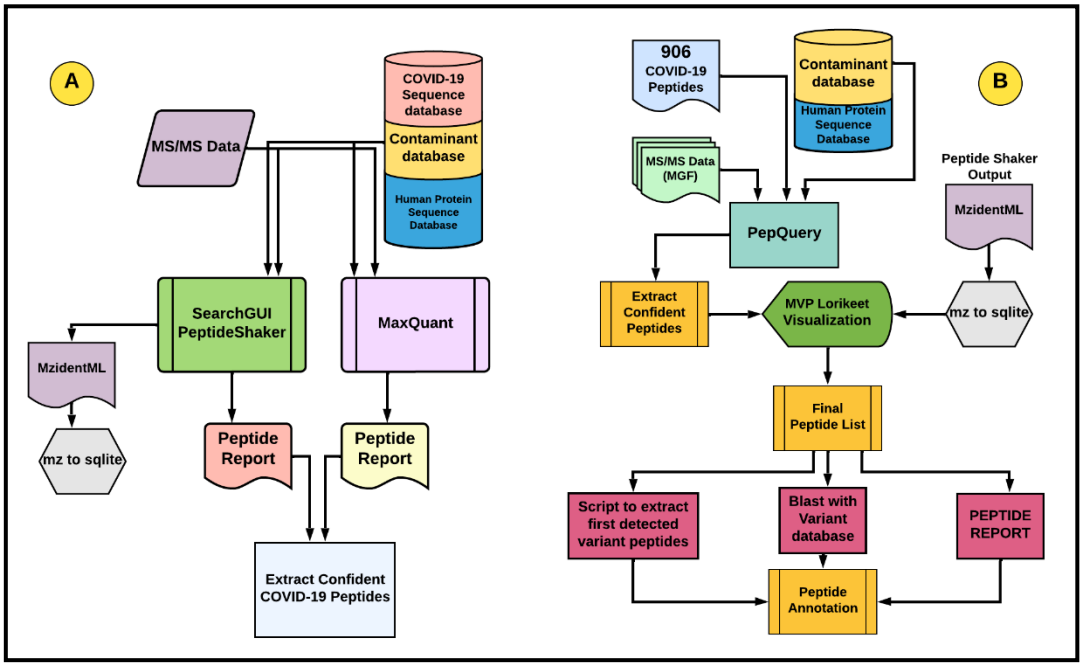


Figure 2. Galaxy-based workflows were used to identify and verify SARS-CoV-2 peptides from MS-based clinical datasets. (**A**) **Discovery workflow:** MS/MS spectra from clinical datasets were searched against a database consisting of SARS-CoV-2 structural protein sequences from SARS-CoV-2 variants, wild-type virus protein sequence, common contaminant sequences, and human protein sequences using SearchGUI/Peptide Shaker and MaxQuant. The PSM output was filtered to extract confident matches to COVID-19 peptides with sequences unique to viral variants. mzidentML generated from SearchGUI-Peptide shaker was used for spectral quality analysis via Lorikeet viewer. (**B**) **Verification Workflow:** A peptide panel of 906 SARS-CoV-2 peptides (theoretical and empirically detected peptides obtained from in silico analysis, cell-culture, and clinical datasets including variant-specific peptide sequences) was subjected to the PepQuery analysis of clinical MS datasets. The quality of the verified PSMs was manually interrogated using the Lorikeet visualization platform within the Multi-omics Visualization Platform (MVP) for additional evaluation. High-quality, confident peptides were confirmed for specificity to virus variants by Blast-P analysis and were annotated for associated phylogenetic lineages by Pango lineage analysis. Finally, these annotated peptides were compared to the original PeptideShaker (peptide report) and MaxQuant (peptide.txt) discovery results to confirm their initial matches to the specific protein sequences and also that they belonged to these virus variants.

The sequence database search parameters used for digestion, modifications, tolerance, and FDR estimation were consistent with the parameters mentioned in the published papers for each of these datasets (Supplementary Data 1—Table S1). Confident SARS-CoV-2 peptides from all the datasets were parsed out by eliminating the human and the common contaminant peptides. The Galaxy-based discovery workflow can be accessed here (https://usegalaxy.eu/u/galaxyp/w/coviddiscovery-workflow; last accessed 21 September 2022).

2.3. Peptide Verification Workflow

We detected 203 SARS-CoV-2 datasets (Supplementary Data 1—Figure S1). peptides from the 12 clinical MS datasets using our discovery workflow and customized protein sequence database. The 203 detected peptides were compared to the existing 803 SARS-CoV-2 peptides from previously published data [14,28], resulting in 103 unique peptides

for this current analysis (Supplementary Data 3). These peptides were then combined to create a larger SARS-CoV-2 peptide panel. The combined peptide panel of 906 peptides was then subjected to the verification workflow (Figure 2B) to confirm the veracity of the PSMs identifying these peptides, as well as the spectral quality. The peptide verification workflow performed a re-analysis of the datasets using the PepQuery tool parameters specified in (Supplementary Data 1—Table S2). PepQuery filters out putative SARS-CoV-2 PSMs that may be better matches to human or contaminant protein sequences that are present in the background reference database, as well as further confirming those PSMs that best match viral proteins. Confident viral peptides with a p-value ≤ 0.05 assigned by PepQuery were then subjected to spectral visualization and manual inspection using the Multiomics Visualization Platform (MVP) tool [29] within the Galaxy platform to ascertain the quality of verified peptides (Supplementary Data 2 and Supplementary Data 1—Table S3). As the peptide spectral quality is crucial for developing reliable targeted MS-based assays, we further validated the PepQuery-filtered peptides using criteria that included each spectrum containing three consecutive b- and/or y-ions detected, and the MS2 ion intensities considered were at least three-fold higher than the background noise (Supplementary Data 1—Figure S2). The peptides that passed the manual inspection were then subjected to BLAST-P analysis against the Wild-Type SARS-CoV-2 proteins and the non-redundant database (NCBI-nr). The wild-type SARS-CoV-2 proteome sequence was obtained from the GISAID database and represents the reference virus strain characterized at the earliest stages of the pandemic (EPI_ISL_402124—dated 30 December 2019). Additionally, the annotated viral peptides were reconfirmed as belonging to SARS-CoV-2 variants using the peptide reports from Search GUI/PeptideShaker (Peptide Report) and MaxQuant (peptides.txt) from the initial discovery analysis. This comparison along with BLAST-P analysis showed that there were 17 peptides unique to viral strains.

To further annotate these sequences, a Phylogenetic Assignment of Named Global Outbreak (Pango) lineage analysis; last accessed on 3 February 2022 [30] was performed using a python script. For the Pango lineage search, a GISAID protein FASTA file (allprot0203) and an associated metadata file were downloaded from https://www.gisaid.org/; last accessed on 3 February 2022. The allprot0203.fasta file included 205,705,355 sequences for a size of 113,874,658 KB. The metadata.tsv file included 7,786,913 accession IDs for a size of 4,586,043 KB. The list of 17 peptide sequences described previously was used to subset the allprot0203.fasta file by keeping only proteins containing at least one of the peptides, with I/L residues undifferentiated. An in-house python script was used on this file for (i) the in-silico digestion of these proteins with two missed cleavages allowed; (ii) mapping of all protein sequences to each peptide from the list of 17 sequences; (iii) retrieval of all matching information (including the Pango Lineage) from the metadata file through the GISAID Accession ID; and (iv) summary of information per peptide using the "groupby" function from the python pandas package. Retained information includes (Supplementary Data 1—Table S4) the list of unique proteins, a list of Pango lineages, the first five and last five countries of appearance, the first GISAID Accession ID, and the number of associated GISAID Accession IDs. For additional verification, the peptides assigned to mutated sequences were manually aligned to the wild-type strain to verify their specificity to defined viral variants. The variants were classified according to the WHO-SARS naming system [19], and the amino acid sequence specific to these variants was confirmed by referring to the current SARS-CoV-2 lineage database [31]. The peptide verification workflow can be accessed here (https://usegalaxy.eu/u/galaxyp/w/covid-verification-workflow; last accessed 21 September 2022).

3. Results

3.1. Discovery Workflow Results

The discovery workflow performed sequence database searching using two software platforms (SearchGUI-PeptideShaker and MaxQuant). The sequence database searching

of the clinical MS datasets against the database of protein sequences specific to SARS-CoV-2 variants confidently detected 203 unique peptides (Supplementary Data 3). These peptides mainly represented structural proteins from the SARS-CoV-2 proteome, mostly from the Nucleocapsid protein. In our previous published study, a panel of 623 peptides was generated from MS data generated from cell culture and clinical samples, as well as a study using in silico-translated viral protein sequences, to generate PSMs and identify viral peptides [21,28]. We merged this list of peptides with an additional list of peptides from a study employing spectral-library searching against the wild-type SARS-CoV-2 proteome [32] to generate a list of 803 peptides. We found an overlap of 100 peptides after comparing the 203 detected peptides from the discovery workflow in our current study with 803 peptides from previous studies. All the high confidence SARS-CoV-2 peptides from our current and previous studies detected by MS/MS were merged, resulting in a total of 906 peptides.

3.2. Verification Workflow Results:

The comprehensive panel of 906 peptides was used to re-interrogate the 12 clinical MS datasets using the PepQuery tool within our verification workflow. PepQuery stringently evaluates the veracity of these putative virus-specific PSMs by the re-analysis of corresponding MS/MS spectra against the human and common contaminant protein sequences, including possible post-translational modifications [20]. Those MS/MS spectra that still best match their viral peptide sequence after this analysis are passed on for further consideration. Out of the 906 peptides, 82 high-quality peptides (Supplementary Data 4) passed the PepQuery confidence filter (p-value ≤ 0.05) and the subsequent manual spectral quality inspection using Lorikeet (Figure 3). The manual annotation and Blast-P results showed that 75.6% of these high-quality peptides were from the Nucleocapsid, 17.1% from the Spike protein, 4.9% from Membrane proteins, and 2.4% from the NS9b protein.

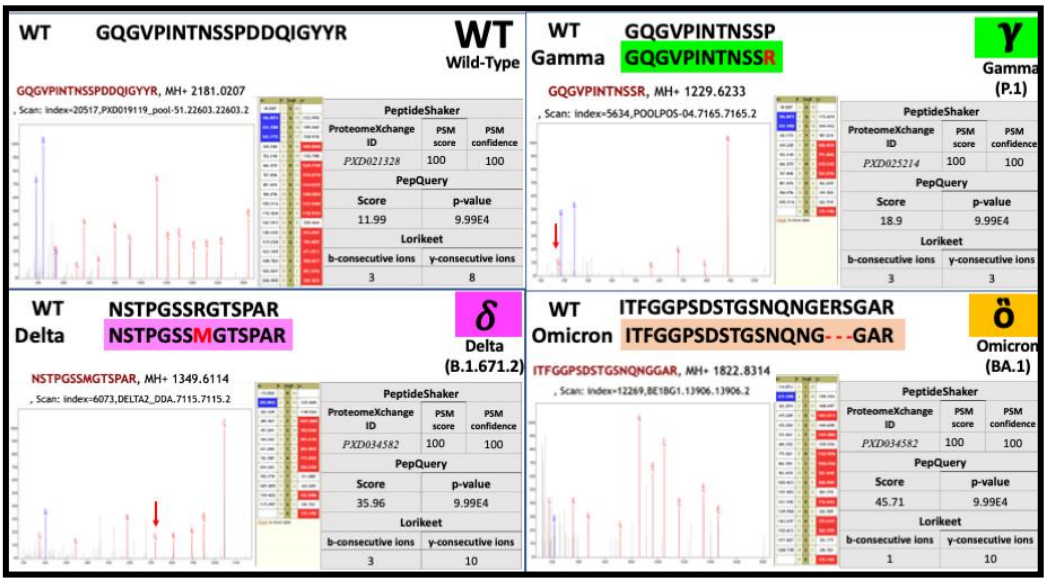


Figure 3. Representative MS/MS spectra of SARS-CoV-2 verified peptides from the clinical samples. The figure shows the manually validated and annotated MS/MS spectra resulting in these peptide sequence identifications. Representative variant-specific peptide sequences from the Nucleocapsid protein are shown, along with their alignment to the wild-type sequences and scores from the discovery, verification, and validation steps. The peptides in black font are from wild-type SARS-CoV-2, while the variant peptides are colored-coded; assignments to WHO-designated variants are shown as Gamma (green), Delta (pink), and Omicron (orange). All the amino acid sequence mutations are marked with red text, and the red arrow within the annotated MS/MS spectra designates the detection of sequence fragments at m/z values specific to fragments carrying these mutations.

Blast-P analysis along with Pango lineage analysis [30] assigned the peptides to proteins that may be specific to SARS-CoV-2 variants. The peptides in Table 1 and Supplementary Data 5 show that the verified variant peptides belonged to the nucleocapsid protein. They were further annotated using WHO nomenclature [19] and the associated Pango lineage.

Table 1. Verified peptides with variant-specific sequences to the SARS-CoV-2 nucleocapsid protein.

PEPTIDE	WT SEQUENCE	Variant (WHO name)	BLASTP IDENTITY (WT)	BLASTP IDENTITY (NR)
GQGVPINTNSSR (P80R)	GQGVPINTNSSP	P.1.(Gamma)	88.00	100.00
AYETQALPQR(D377Y)	ADETQALPQR	B.1.617.2 (Delta)	80.00	100.00
GEGVPINTNSSPDDQIGYYR (Q69E)	GQGVPINTNSSPDDQIGYYR	B.1.1.7 (Delta)	95.00	100.00
SMGTSPTRMAGNGGDAALALLLLDR (R203M & A208T)	SRGTSPARMAGNGGDAALALLLLDR	B.1.617.2 (Delta)	86.67	96.00
PGNGCDAALALLLLDR (A211P &G215C)	AGNGGDAALALLLLDR	AY.4 (Delta)	93.33	100.00
ITFGGPSDSTGSNQNGG|AR (Δ31–33)	ITFGGPSDSTGSNQNGERSGAR	BA.1 (Omicron)	86.36	100.00

The table shows different nucleocapsid peptides with sequences specific to virus variants (shown in parentheses), the wild-type sequence, and their assigned Pango lineage identifier. Mutated amino acids are shown in the red text along with the amino acid changes at their specific positions in the primary protein sequence. The Blast-P similarity is shown in both wild-type SARS-CoV-2 sequences and the Non-Redundant (NR) NCBI database. The NCBI NR-database contains many of the variant-specific sequences; thus, many of these showed 100% similarity.

To assess the specificity of these peptide sequences to viral variants, we also manually aligned them to the wild-type sequences to confirm their identities (Table 1 and Figure 3). As a result of this evaluation, we identified six peptides that belong to the variants Gamma, Delta, and Omicron. Despite the nucleocapsid being the most invariant protein sequence in SARS-CoV-2 [33], we observed sequence changes in nucleocapsid peptides specific to variants, such as P80R in Gamma, D377Y, R203M and A208T in Delta [31], and the deletion of amino acid positions 31–33 in Omicron [34]. All of the mentioned peptides were present in clinical samples obtained from COVID-19-positive patients collected at different times and geographical locations; hence, along with their rigorously verified identities from MS/MS spectral analysis, these peptides deserve consideration as optimal candidates for detection using targeted MS-proteomics (selected reaction monitoring (SRM) and parallel reaction monitoring (PRM)) [35] or potentially other detection platforms.

4. Discussion and Conclusions

The different COVID-19 pandemic waves have been driven by constantly evolving virus strains that vary in their virulence, fatality, severity, and infectivity [36]. These waves have also been dynamic in the geographically affected areas and timepoints and have put strain on the healthcare system when they appear within populations. Scientists, from basic researchers to clinicians to epidemiologists, continue to monitor these emerging strains and classify them according to the WHO-SARS-CoV-2 naming system, which includes variants of concern, interest, to be monitored, or high consequence [19], depending on their contagiousness, clinical presentation, severity, and responsiveness to vaccines and/or therapies. Pango lineage nomenclature is a common way to classify distinct lineages of SARS-CoV-2 compared to the reference sequence [30]. Essential to these ongoing monitoring efforts is the ability to detect evolving sequence changes to the essential proteins of SARS-CoV-2 variants, which are the targets of rapid tests and, in some cases, vaccines and therapies. MS-based proteomics, in particular targeted methods against variant-specific peptides, offer a useful approach for such monitoring, directly characterizing these sequences from sample types commonly collected in the clinic. These methods, however, depend on verified peptide sequences specific to these variants that can serve as targets.

In this study, we presented two workflows, available within the Galaxy platform, which can identify and verify SARS-CoV-2 variant-specific peptides. These flexible workflows have the potential to detect new sequences from emerging strains on any clinical datasets analyzed using contemporary MS-based proteomics methods (e.g., data-dependent acquisition of MS/MS spectra, or even targeted parallel reaction monitoring MS/MS data [35]). These workflows are also publicly available, along with documentation, through the European-based Galaxy ecosystem [37]. We anticipate these being useful to a wide variety of researchers as COVID-19 research continues. Importantly, these workflows are composed of verification steps for peptides initially discovered by sequence database searching of MS/MS data, ensuring only verified and validated peptides are reported.

We continue to expand our panel to include peptide targets specific to emerging variants and their associated strain lineages. Our current panel contains 906 peptides shared across all strains of SARS-CoV-2, in which six variant-specific peptides useful for monitoring the infection dynamics of clinically distinct forms of the virus were characterized in this work. These peptides align to the nucleocapsid viral particles, which play critical roles in host infection mechanisms, as targets of antigen-based rapid tests, and could also help in therapeutic approaches [8,38,39]. These verified peptides are optimal for targeted MS-based proteomic assays that have been described for SARS-CoV-2 diagnostics and monitoring [13]. Given the growing emphasis on testing wastewater samples for COVID-19 surveillance [40], we also envisioned the potential use of these peptides as targets for high-sensitivity MS-based assays analyzing these emerging sample types. Such protein-based assays in environmental samples could provide an improved way forward for monitoring community infection dynamics. Although best-suited for diagnostic applications, direct monitoring of proteins specific to virus variants in communities and populations could help in the development and/or choice of the best targets for vaccines and even the development of other therapies aimed at minimizing the severe effects of infection.

In summary, we demonstrated the use of customized bioinformatic workflows to identify six confident SARS-CoV-2 variant-specific peptides suitable for MS/MS based detection in clinical samples. These peptides should be useful for developing targeted MS-based assays for the rapid and sensitive characterization of variant-specific proteins within clinical samples. Our discovery and verification workflows were developed within Galaxy and are accessible via the public and freely available European Galaxy resource (usegalaxy.eu; last accessed 21 September 2022), and as individual tools available in the Galaxy Tool Shed [41]. These workflows are flexible, with amenability to further customization for individual datasets. Although our focus has been SARS-CoV-2

characterization, our workflows could be adapted to MS-based proteomics data from other pathogenic organisms, providing value to the broader infectious disease research community.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v14102205/s1, Supplemental Data S1*:* The supplementary consists of the following tables and figures. Supplementary Table S1: Search parameters for SearchGUI to analyze clinical datasets in Galaxy. Parameters were based on instrumentation, data generation, and data analysis protocols detailed in the original publications. Supplementary Table S2: Parameters for the PepQuery search engine for the verification of clinical datasets in Galaxy. Supplementary Figure S1: Comparison of the 203 peptides obtained from the Discovery workflow to the 803-peptide panel from published data provides us with 100 unique peptides from the present study. Supplementary Figure S2: Manual inspection of the spectra to validate peptide quality. The criteria for an accepted spectrum were that the spectra containing product ions should be at least a three-fold higher intensity than the noise level and the spectra should have at least three consecutive b- and/or y-ions in their series. Supplementary Table S3: Number of peptide-spectral matches (PSMs) and peptide intensities of the detected variant peptides from the 12 datasets. The PSM numbers reported here are PSMs that qualify after PepQuery analysis (PSM rank output) with a *p*-value of 0.05. The peptide intensity values reported in parentheses have been extracted from MaxQuant or FlashLFQ software outputs. Supplementary Table S4: Pango lineage associated with the peptide sequences. Supplemental Data S2: MS/MS spectra of SARS-CoV-2 verified 82 peptides using the Lorikeet Spectral viewer within the Multiomics visualization platform (MVP). Supplemental Data S3: Unique peptides identified from each of the ProteomeXchange datasets along with all identified peptides. Supplemental Data S4: Tabular file containing the following: column (1) Peptides detected from SearchGUI-PeptideShaker (SGPS) and MaxQuant; column (2) the list of 803 peptides; column (3) 906 peptide panel for PepQuery; column (4) 82 peptides that passed spectral visualization after PepQuery validation. Supplemental Data S5: Table shows 82 PepQuery and spectral quality-verified peptides with protein assignment, Pango lineage identifier, first observed date, and their Blast-P identity with the Wild-Type (Wuh-Cor-1) sequence.

Author Contributions: Conceptualization, P.D.J., S.M., J.A. and T.J.G.; methodology, S.M.; software, B.A.G., J.E.J., and R.W.; validation, S.M., and O.P.; formal analysis, S.M., O.P. and A.T.R. ; investigation, S.M. and P.D.J. resources, B.A.G., J.E.J.; data curation, V.M.C. and O.P.; writing—original draft preparation, S.M.; writing—review and editing, P.D.J. and T.J.G.; visualization, S.M.; supervision, P.D.J. and T.J.G.; project administration P.D.J. and T.J.G.; funding acquisition, T.J.G. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge funding for this work from the grant National Cancer Institute—Informatics Technology for Cancer Research (NCI-ITCR) grant 1U24CA199347 to T.J.G. The European Galaxy server that was used for data analysis is partly funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and the German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)).

Institutional Review Board Statement: Ethical review and approval were waived for this study as all the data used for this manuscript were from publicly available resources.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data have been made available via the Zenodo platform-https://doi.org/10.5281/zenodo.7153337. The datasets were downloaded according to their PXD identifier from ProteomeXchange using url: http://proteomecentral.proteomexchange.org/cgi/GetDataset; last accessed 4 October 2022. The Galaxy workflows along with a sample input data and results are available via https://covid19.galaxyproject.org/proteomics/; last accessed 21 September 2022 and the Galaxy HUB— https://galaxyproject.org/community-projects/covid-proteomics/;last accessed 4 October 2022.

Acknowledgments: We would like to acknowledge the European Galaxy team for providing us with infrastructure for our research. We would like to thank Monica E. Kruk for her assistance in generating online materials.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Home—Johns Hopkins Coronavirus Resource Center Available online: https://coronavirus.jhu.edu (accessed on 24 June 2022).
2. Dyer, O. Covid-19: Variants Are Spreading in Countries with Low Vaccination Rates. *BMJ* **2021**, *373*, n1359. https://doi.org/10.1136/BMJ.N1359.
3. Wu, Y.; Kang, L.; Guo, Z.; Liu, J.; Liu, M.; Liang, W. Incubation Period of COVID-19 Caused by Unique SARS-CoV-2 Strains: A Systematic Review and Meta-analysis. *JAMA Netw. Open.* **2022**, *5*, e2228008. https://doi.org/10.1001/jamanetworkopen.2022.28008.
4. Mantas, J. The Importance of Health Informatics in Public Health During the COVID-19 Pandemic. *Stud. Health Technol. Inform.* **2020**, *272*, 487–488. https://doi.org/10.3233/SHTI200602.
5. Jackson, L.A.; Anderson, E.J.; Rouphael, N.G.; Roberts, P.C.; Makhene, M.; Coler, R.N.; McCullough, M.P.; Chappell, J.D.; Denison, M.R.; Stevens, L.J.; et al. An MRNA Vaccine against SARS-CoV-2—Preliminary Report. *N. Engl. J. Med.* **2020**, *383*, 1920–1931. https://doi.org/10.1056/NEJMOA2022483.
6. Lazarus, J.V.; Ratzan, S.C.; Palayew, A.; Gostin, L.O.; Larson, H.J.; Rabin, K.; Kimball, S.; El-Mohandes, A. A Global Survey of Potential Acceptance of a COVID-19 Vaccine. *Nat. Med.* **2021**, *27*, 225–228. https://doi.org/10.1038/S41591-020-1124-9.
7. Lukas, H.; Xu, C.; Yu, Y.; Gao, W. Emerging Telemedicine Tools for Remote COVID-19 Diagnosis, Monitoring, and Management. *ACS Nano* **2020**, *14*, 16180–16193. https://doi.org/10.1021/ACSNANO.0C08494.
8. Yadav, R.; Chaudhary, J.K.; Jain, N.; Chaudhary, P.K.; Khanra, S.; Dhamija, P.; Sharma, A.; Kumar, A.; Handu, S. Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19. *Cells* **2021**, *10*, 821. https://doi.org/10.3390/CELLS10040821.
9. Nagura-Ikeda, M.; Imai, K.; Tabata, S.; Miyoshi, K.; Murahara, N.; Mizuno, T.; Horiuchi, M.; Kato, K.; Imoto, Y.; Iwata, M.; et al. Clinical Evaluation of Self-Collected Saliva by Quantitative Reverse Transcription-PCR (RT-QPCR), Direct RT-QPCR, Reverse Transcription–Loop-Mediated Isothermal Amplification, and a Rapid Antigen Test to Diagnose COVID-19. *J. Clin. Microbiol.* **2020**, *58*, e01438–e1520. https://doi.org/10.1128/jcm.01438-20.
10. Kriegova, E.; Fillerova, R.; Kvapil, P. Direct-RT-QPCR Detection of SARS-CoV-2 without RNA Extraction as Part of a COVID-19 Testing Strategy: From Sample to Result in One Hour. *Diagnostics* **2020**, *10*, 605. https://doi.org/10.3390/diagnostics10080605.
11. Corman, V.M.; Landt, O.; Kaiser, M.; Molenkamp, R.; Meijer, A.; Chu, D.K.W.; Bleicker, T.; Brünink, S.; Schneider, J.; Schmidt, M.L.; et al. Detection of 2019 Novel Coronavirus (2019-NCoV) by Real-Time RT-PCR. *Eurosurveillance* **2020**, *25*, 2000045. https://doi.org/10.2807/1560-7917.es.2020.25.3.2000045.
12. Foster, M.W.; Gerhardt, G.; Robitaille, L.; Plante, P.L.; Boivin, G.; Corbeil, J.; Moseley, M.A. Targeted Proteomics of Human Metapneumovirus in Clinical Samples and Viral Cultures. *Anal. Chem.* **2015**, *87*, 10247–10254. https://doi.org/10.1021/acs.analchem.5b01544.
13. Cardozo, K.H.M.; Lebkuchen, A.; Okai, G.G.; Schuch, R.A.; Viana, L.G.; Olive, A.N.; dos Lazari, C.S.; Fraga, A.M.; Granato, C.F.H.; Pintão, M.C.T.; et al. Establishing a Mass Spectrometry-Based System for Rapid Detection of SARS-CoV-2 in Large Clinical Sample Cohorts. *Nat. Commun.* **2020**, *11*, 6201. https://doi.org/10.1038/s41467-020-19925-0.
14. Rajczewski, A.T.; Mehta, S.; Nguyen, D.D.A.; Grüning, B.; Johnson, J.E.; McGowan, T.; Griffin, T.J.; Jagtap, P.D. A Rigorous Evaluation of Optimal Peptide Targets for MS-Based Clinical Diagnostics of Coronavirus Disease 2019 (COVID-19). *Clin. Proteomics* **2021**, *18*, 15. https://doi.org/10.1186/S12014-021-09321-1/FIGURES/7.
15. Thuy-Boun, P.S.; Mehta, S.; Gruening, B.; Mcgowan, T.; Nguyen, A.; Rajczewski, A.T.; Johnson, J.E.; Griffin, T.J.; Wolan, D.W.; Jagtap, P.D. Metaproteomics Analysis of SARS-CoV-2-Infected Patient Samples Reveals Presence of Potential Coinfecting Microorganisms. *J. Proteome Res.* **2021**, *20*, 1451–1454. https://doi.org/10.1021/ACS.JPROTEOME.0C00822/SUPPL_FILE/PR0C00822_SI_002.PDF.
16. Afgan, E.; Nekrutenko, A.; Grüning, B.A.; Blankenberg, D.; Goecks, J.; Schatz, M.C.; Ostrovsky, A.E.; Mahmoud, A.; Lonie, A.J.; Syme, A.; et al. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2022 Update. *Nucleic Acids Res.* **2022**, *50*, W34. https://doi.org/10.1093/NAR/GKAC247.
17. Hiltemann, S.; Rasche, H.; Gladman, S.; Hotz, H.-R.; Larivière, D.; Blankenberg, D.; Jagtap, P.D.; Wollmann, T.; Bretaudeau, A.; Goué, N.; et al. Galaxy Training: A Powerful Framework for Teaching! *bioRxiv* **2022**. https://doi.org/10.1101/2022.06.02.494505.
18. COVID-19 Analysis on Usegalaxy ★. Available online: https://covid19.galaxyproject.org (accessed on 3 July 2022).
19. Tracking SARS-CoV-2 Variants. Available online: https://www.who.int/activities/tracking-SARS-CoV-2-variants (accessed on 3 July 2022).
20. Wen, B.; Wang, X.; Zhang, B. PepQuery Enables Fast, Accurate, and Convenient Proteomic Validation of Novel Genomic Alterations. *Genome Res.* **2019**, *29*, 485–493. https://doi.org/10.1101/GR.235028.118/-/DC1.
21. Gouveia, D.; Miotello, G.; Gallais, F.; Gaillard, J.C.; Debroas, S.; Bellanger, L.; Lavigne, J.P.; Sotto, A.; Grenga, L.; Pible, O.; et al. Proteotyping SARS-CoV-2 Virus from Nasopharyngeal Swabs: A Proof-of-Concept Focused on a 3 Min Mass Spectrometry Window. *J. Proteome Res.* **2020**, *19*, 4407–4416. https://doi.org/10.1021/ACS.JPROTEOME.0C00535/SUPPL_FILE/PR0C00535_SI_008.XLSX.
22. Bankar, R.; Suvarna, K.; Ghantasala, S.; Banerjee, A.; Biswas, D.; Choudhury, M.; Palanivel, V.; Salkar, A.; Verma, A.; Singh, A.; et al. Proteomic Investigation Reveals Dominant Alterations of Neutrophil Degranulation and MRNA Translation Pathways in Patients with COVID-19. *iScience* **2021**, *24*, 102135. https://doi.org/10.1016/J.ISCI.2021.102135.

23. Ihling, C.; Tänzler, D.; Hagemann, S.; Kehlen, A.; Hüttelmaier, S.; Arlt, C.; Sinz, A. Mass Spectrometric Identification of SARS-CoV-2 Proteins from Gargle Solution Samples of COVID-19 Patients. *J. Proteome Res.* **2020**, *19*, 4389–4392. https://doi.org/10.1021/ACS.JPROTEOME.0C00280.
24. Chavan, S.; Mangalaparthi, K.K.; Singh, S.; Renuse, S.; Vanderboom, P.M.; Madugundu, A.K.; Budhraja, R.; McAulay, K.; Grys, T.E.; Rule, A.D.; et al. Mass Spectrometric Analysis of Urine from COVID-19 Patients for Detection of SARS-CoV-2 Viral Antigen and to Study Host Response. *J. Proteome Res.* **2021**, *20*, 3404–3413. https://doi.org/10.1021/ACS.JPROTEOME.1C00391.
25. Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**, *17*, 2552–2555. https://doi.org/10.1021/acs.jproteome.8b00175.
26. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R.A.; Olsen, J.V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011**, *10*, 1794–1805. https://doi.org/10.1021/pr101065j.
27. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11*, 2301–2319. https://doi.org/10.1038/nprot.2016.136.
28. St-Germain, J.R.; Astori, A.; Raught, B. A SARS-CoV-2 Peptide Spectral Library Enables Rapid, Sensitive Identification of Virus Peptides in Complex Biological Samples. *J. Proteome Res.* **2021**, *20*, 2187–2194. https://doi.org/10.1021/ACS.JPROTEOME.1C00048.
29. Mcgowan, T.; Johnson, J.E.; Kumar, P.; Sajulga, R.; Mehta, S.; Jagtap, P.D.; Griffin, T.J. Multi-Omics Visualization Platform: An Extensible Galaxy Plug-in for Multi-Omics Data Visualization and Exploration. *Gigascience* **2020**, *9*, giaa025. https://doi.org/10.1093/gigascience/giaa025.
30. O'Toole, Á.; Pybus, O.G.; Abram, M.E.; Kelly, E.J.; Rambaut, A. Pango Lineage Designation and Assignment Using SARS-CoV-2 Spike Gene Nucleotide Sequences. *BMC Genomics* **2022**, *23*, S12864–S022. https://doi.org/10.1186/S12864-022-08358-2.
31. Cov-Lineages. Available online: https://cov-lineages.org/constellations.html (accessed on 3 July 2022).
32. Zeng, H.L.; Chen, D.; Yan, J.; Yang, Q.; Han, Q.Q.; Li, S.S.; Cheng, L. Proteomic Characteristics of Bronchoalveolar Lavage Fluid in Critical COVID-19 Patients. *FEBS J.* **2021**, *288*, 5190–5200. https://doi.org/10.1111/FEBS.15609.
33. Wang, H.; Li, X.; Li, T.; Zhang, S.; Wang, L.; Wu, X.; Liu, J. The Genetic Sequence, Origin, and Diagnosis of SARS-CoV-2. *Eur. J. Clin. Microbiol. Infect. Dis.* **2020**, *39*, 1629–1635. https://doi.org/10.1007/s10096-020-03899-4.
34. World Health Organization. *Enhancing Response to Omicron SARS-CoV-2 Variant: Technical Brief and Priority Actions for Member States*; View Most Current Version A. Context; WHO: Geneva, Switzerland, 2022.
35. Rauniyar, N. Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. *Int. J. Mol. Sci.* **2015**, *16*, 28566. https://doi.org/10.3390/IJMS161226120.
36. Callaway, E. Beyond Omicron: What's next for COVID's Viral Evolution. *Nature* **2021**, *600*, 204–207. https://doi.org/10.1038/D41586-021-03619-8.
37. Proteomics | COVID-19 Analysis on Usegalaxy ★. Available online: https://covid19.galaxyproject.org/proteomics (accessed on 3 July 2022).
38. Huang, Y.; Yang, C.; Xu, X.F.; Xu, W.; Liu, S.W. Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19. *Acta Pharmacol. Sin.* **2020**, *41*, 1141–1149. https://doi.org/10.1038/S41401-020-0485-4.
39. Rahman, M.S.; Islam, M.R.; Alam, A.S.M.R.U.; Islam, I.; Hoque, M.N.; Akter, S.; Rahaman, M.M.; Sultana, M.; Hossain, M.A. Evolutionary Dynamics of SARS-CoV-2 Nucleocapsid Protein and Its Consequences. *J. Med. Virol.* **2021**, *93*, 2177–2195. https://doi.org/10.1002/JMV.26626.
40. Brumfield, K.D.; Leddy, M.; Usmani, M.; Cotruvo, J.A.; Tien, C.-T.; Dorsey, S.; Graubics, K.; Fanelli, B.; Zhou, I.; Registe, N.; et al. Microbiome Analysis for Wastewater Surveillance during COVID-19. *MBio* **2022**, *13*, e00591-22. https://doi.org/10.1128/MBIO.00591-22.
41. Galaxy | Tool Shed. Available online: https://toolshed.g2.bx.psu.edu (accessed on 3 July 2022).