Elizabeth Cazallis

Cell and Molecular Biology

Molecular Monitoring for Environment and Health – GVSU

Summer 2022 – Winter 2022

Table of Contents

Learning Objectives & Internship Objectives

The objectives of this internship were to:

1. Develop a standard operating procedure (SOP) for amplification of SARS-CoV-2 spike protein receptor binding domain (RBD).
2. Develop a method for variant classification from sequencing data.

Introduction

SARS-CoV-2 monitoring through wastewater surveillance

Wastewater is a useful tool to monitor SARS-CoV-2 community infection levels, as the virus is shed in feces. Wastewater surveillance has an outstanding epidemiologic potential that can be used as a supplementary system to clinical cases to monitor viral tracking and circulation in cities. While reported clinical cases rely on hospital, healthcare provider, and laboratory results [1], wastewater surveillance can track the spread of SARS-CoV-2 in communities and sites without the need for testing. This kind of passive surveillance is a better representation than clinical cases.

The State of Michigan SARS-CoV-2 Epidemiology – Wastewater Evaluation and Reporting (SEWER) Network is a wastewater monitoring project that uses local-coordinated projects to conduct surveillance of the virus shed in the Michigan public sewer system [2]. The SEWER Network is coordinated by MDHHS and its partnership with local health departments, Tribal communities, wastewater treatment plants (WWTPs), universities, and laboratories. MDHHS coordinates the project, data analysis, health education, and risk communication using the local data analysis from local projects that collect samples and conduct lab testing. Michigan State University (MSU) serves as the lead laboratory, standardizes testing, and provides assistance across the SEWER Network.

SEWER Network local laboratory methods for SARS-CoV-2 detection

Samples are collected from WWTPs and sentinel sites such as veteran homes, nursing homes, schools, and low income housing. These samples are solubilized and concentrated, as

viral SARS-CoV-2 RNA levels are diluted in wastewater. Viral RNA is extracted from the concentrates to use for the next step: ddPCR. RT-ddPCR of the nucleocapsid protein N determines gene copies/100 mL in samples. These results are then reported to MDHHS.

Samples containing over 9,000 gene copies/100 mL are considered for variant analysis. SEWER Network current methods of detection include qPCR and/or ddPCR. These commercial kits use probes that fluoresce upon the presence of specific amino acid mutations to identify variants. However, ddPCR and qPCR are not able to detect more than 2 mutations in a subvariant, and are unable to detect regional differences and emergent variants. Small percentages of sequences with other mutations are undetectable if there is a dominant sequence. In addition, these kits are cost prohibitive, have reliability and supply chain issues, and lag behind variant emergence. Whole genome sequencing is another option; however, this method is very expensive and needs high quality RNA. Our lab's strategy (in collaboration with Dr. Marc Johnson, University of Missouri) is to use targeted sequencing of the spike protein RBD.

Description of Work

Targeting the spike protein RBD

The choice of region to sequence (the RBD) stems from the fact that this region is under high selection pressure for new mutations and variant emergence. The RBD interacts with ACE2 receptors on human cells to gain entry. As ACE2 receptors are an exposed region, the immune system synthesizes antibodies specific to this region. Importantly, current vaccines for SARS-CoV-2 work by sending antibodies to this region.

Most mutations in the RBD are important to immune response and drive antibody escape. Therefore, mutations in this region happen with high frequency and may have greater clinical significance based on infectivity and clinical severity. Due to these selection mutations, we are targeting amino acids 319 to 541 of the spike protein RBD (Figure 1, 3).

These mutations in the spike protein RBD characterize variants. Compared to the previous Delta variants, current Omicron variants and subvariants have more mutations than previous variants (Figure 2), making these variants increasingly difficult to detect with qPCR and ddPCR.

Figure 1. Map of spike protein. Amino acids 319-541 (blue) of the RBD (shown in red) are of interest in this project. (Modified from Asif et al., J. Mol. Pathol. 2022, 3(4), 201-218; https://doi.org/10.3390/jmp3040018)

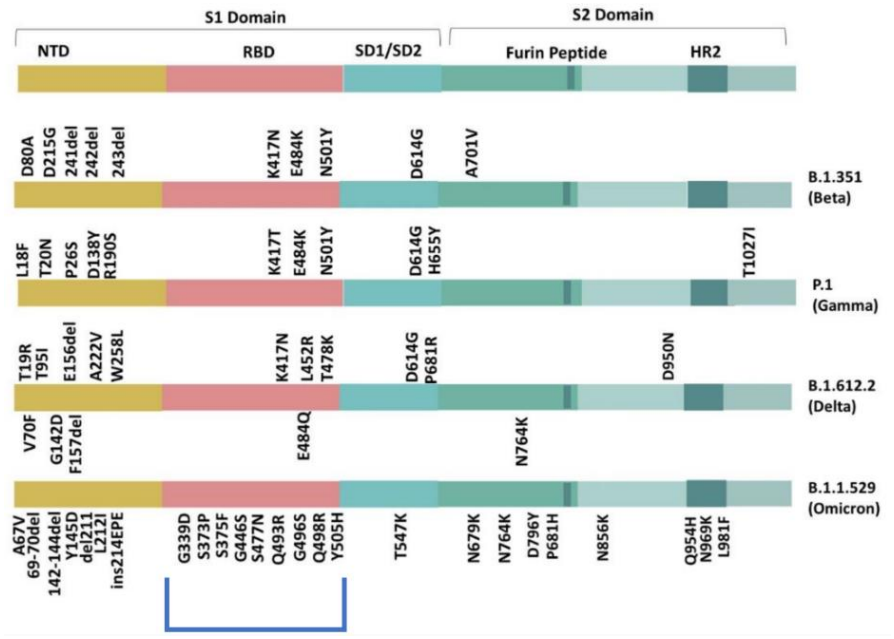

Figure 2. Amino acid mutations in the spike protein (modified from Asif et al., J. Mol. Pathol. 2022, 3(4), 201-218; https://doi.org/10.3390/jmp3040018)

Amplifying the RBD

We have based our targeted amplification methods on a paper published by [redacted] and his lab at the University of Missouri, who developed an amplification protocol and novel program sequence alignment map (SAM) refiner for amplicon sequencing [4].

Amplification by nested PCR is needed for dilute RNA, such as our viral RNA wastewater extracts, because the proportion of RNA present in samples determines the proportion of DNA after PCR. Nested PCR is a modification of PCR that is designed to improve sensitivity and specificity. This method involves two primer sets and two successive PCR reactions. In primary PCR amplification, viral RNA is reverse transcribed to generate cDNA, which serves as the template. This reaction is followed by nested PCR, which uses a high fidelity polymerase and primers further downstream and upstream from the primary PCR reactions (Figure 3). The resultant amplicon is 547 bp, covering 82% of the RBD. Due to the Johnson lab's sequencing limitations, the entire RBD cannot be sequenced and therefore we amplify only 82%. Amplicons are then verified through gel electrophoresis.

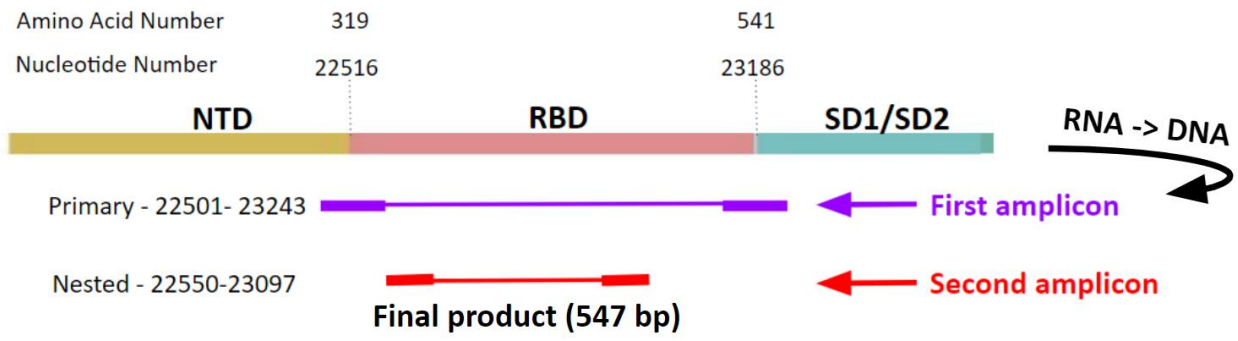

Figure 3. Nested PCR of RBD region. Primers (developed by the Johnson lab) and amplicons were mapped onto NCBI ACC# NC_045512.2

Sequencing the RBD and data analysis

Confirmed amplicons are sent to the Johnson lab to be sequenced. At the lab, deep sequencing via Illumina MiSeq is performed. These files are analyzed with a custom Python program that aligns sequences using a reference SARS-CoV-2 sequence (NC 045512.2) [4]. PCR artifacts are removed and mutations are identified. A file that is readable by Excel is returned to our lab containing nucleotide and amino acid changes (Figure 4).

I have created a user-defined dictionary that a custom Python program uses to classify variants (Figure 5). This dictionary is easily changeable to identify emergent variants. The custom Python program reads the condensed sequence files and dictionary to generate an excel file with proportional representation of all user-defined variants (Figure 6).

Finally, a weekly report of variant proportions in wastewater samples that have high viral counts. Finally, a report is created and sent to MDHHS. This data is also reported to the SEWER Network, stakeholders, and the CDC.

Sequences	Count	Abundance
A1224C(R408S) G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A14	5711	0.318
A1224C(R408S) G1251T(K417N) T1320G(N440K) T1355G(L452R) T1389C(P463P) G1430A(S477N) C14		
C1230T(I410I) G1251T(K417N) T1320G(N440K) T1355G(L452R) T1389C(P463P) G1430A(S477N) C143		
G1251T(K417N) T1320G(N440K) G1339A(G447S) T1355G(L452R) G1430A(S477N) C1433A(T478K) A14		
G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A1451C(E484A) T14	1753	0.098
G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A1451C(E484A) T14	9168	0.510
G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A1451C(E484A) T14	240	0.013
G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A1451C(E484A) T14	508	0.028
G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A1451C(E484A) T14	266	0.015
G1251T(K417N) T1320G(N440K) T1355G(L452R) G1430A(S477N) C1433A(T478K) A1451C(E484A) T14	222	0.012
G1251T(K417N) T1320G(N440K) T1355G(L452R) T1389C(P463P) G1430A(S477N) C1433A(T478K) A14		

Figure 4. Condensed nucleotide/amino acid changes and sequence proportions in a sample.

Variant name	tolerance	SNPs
BA.4/BA.5/BA.5.2.1/BF.9	0	G339D !339H !V341A !346T !K356T !368I !K444M !444T !444R !444N !446S !450D !452Q 452F
BQ.1	0	G339D !G339H !V341A !346K !346T !K356T !L368I !K444M K444T !444N !444R !V445 !G446 !N
BQ.1.1	0	G339D !G339H !V341A !346K 346T !K356T !L368I !K444M K444T !444N !444R !V445 !G446 !N
BA.4.6/BF.7.*/BF.1.*	0	G339D !G339H !V341A !346K R346T !K356T !L368I !K444M !444T !444R !444N !V445 !G446 !N
BN.1	0	!G339D 339H !V341A !346K 346T K356T !L368I !K444M !K444T !444N !444R !V445 G446S !44(
BA.2.75.2	0	!G339D 339H !V341A !346K 346T !K356T !L368I !K444M !K444T !444R !444N !V445 G446S !44
XBB*	0	!G339D 339H !V341A !346K 346T !K356T L368I !K444M !K444T !444R !444N V445 G446S !446
XBB.1.5*	0	!G339D 339H !V341A !346K 346T !K356T L368I G446S !446D !K444 V445P !L452R N460K !461
XBF*	0	!G339D 339H !V341A !346K 346T !K356T !L368I !K444 !V455 G446D !446P !N450D !L452R N4(
BR.1*	0	!G339D 339H !V341A !346K !346T !K356T !L368I K444M !K444T !V445 G446S !446D !N450D !I
CH.1.1*	0	!G339D G339H !V341A R346T !346K !K356T !L368I K444T !K444M !K444N !K444R G446S !446
BA.2.3.20	0	339D !339H !V341A !346 !356 !368 444R !444T !444M !444N !445 !446 450D 452M !452R !45
BU	0	339D !339H !341 !346 !356 !368 444M !444T !444R !444N !445 !446 !450 452R !452M !452Q
CK	0	339D !339H !341 !346 !356 !368 444N !444M !444T !444R !445 !446 !450 452R !452Q !452M
XAY.1	0	339D !339H !341 !346 !356 !368 !444 !445 446D !446S !450 452R !452M !452Q !460 477N 48(

Figure 5. User-defined dictionary to identify variants. SNPs are single amino acid mutations. The "!" symbolizes the absence of a mutation in a specific variant.

Code	Date	Number of Reads	BQ.1	BQ.1.1	BA.4.6/ BF.7.*/ BF.1.*	XBB*	XBB.1.5 *	BR.1*	BA.2.75. 2	BN.1	CH.1.1*	BA.4/BA.5/B A.5.2.1/BF.9	Other	Other variants	Comment
EE	221215	116272	0.79	0.011	0.198	0	0	0	0	0	0	0	0		
GO	221219	149654	0	0.991	0.004	0	0	0	0	0	0	0	0.005	'G1016A(G339D) G1037C(R346	
GR	221222	125766	0	0.67	0.12	0.195	0	0	0	0	0	0	0.014	'G1016A(G339D) G1037C(R346	
GR	221225	265980	0	0.485	0.333	0.092	0	0.079	0	0	0	0	0.011	'G1016A(G339D) G1037C(R346	
GR	221228	432658	0	0.301	0.584	0.057	0	0.048	0	0	0	0	0.01	'G1016A(G339D) G1037C(R346	
GG	221228	161160	0.986	0.005	0	0	0	0	0	0	0	0	0.007	'G1016A(G339D) C1112T(S371F	
WB	221228	118582	0	0.665	0	0	0.306	0	0	0	0	0	0.029	'GGT1015-1017CAT(G339H) G1	
GO	221229	289977	0.248	0.746	0.003	0	0	0	0	0	0	0	0.003	'G1016A(G339D) G1037C(R346	
WK	221229	159808	0.985	0.005	0	0	0	0	0	0	0	0	0.009	'G1016A(G339D) C1112T(S371F	
GR	230101	528971	0	0.311	0.478	0.106	0	0.04	0	0	0.056	0	0.01	'G1016A(G339D) G1037C(R346	

Figure 6. Excel file with proportional variant representation in samples. "Other variants" represent sequences that have not been classified. the absence of a mutation in a specific variant.

Public health applications

Using this method, we have been able to observe the change in variant composition in WWTPs over time. We have also successfully tracked the disappearance of Omicron BA.4 and BA.5 subvariants, the rise and fall of BQ.1 and its subvariants, and the emergence of the current dominant West Michigan subvariant, XBB.1.5 (Figure 7).

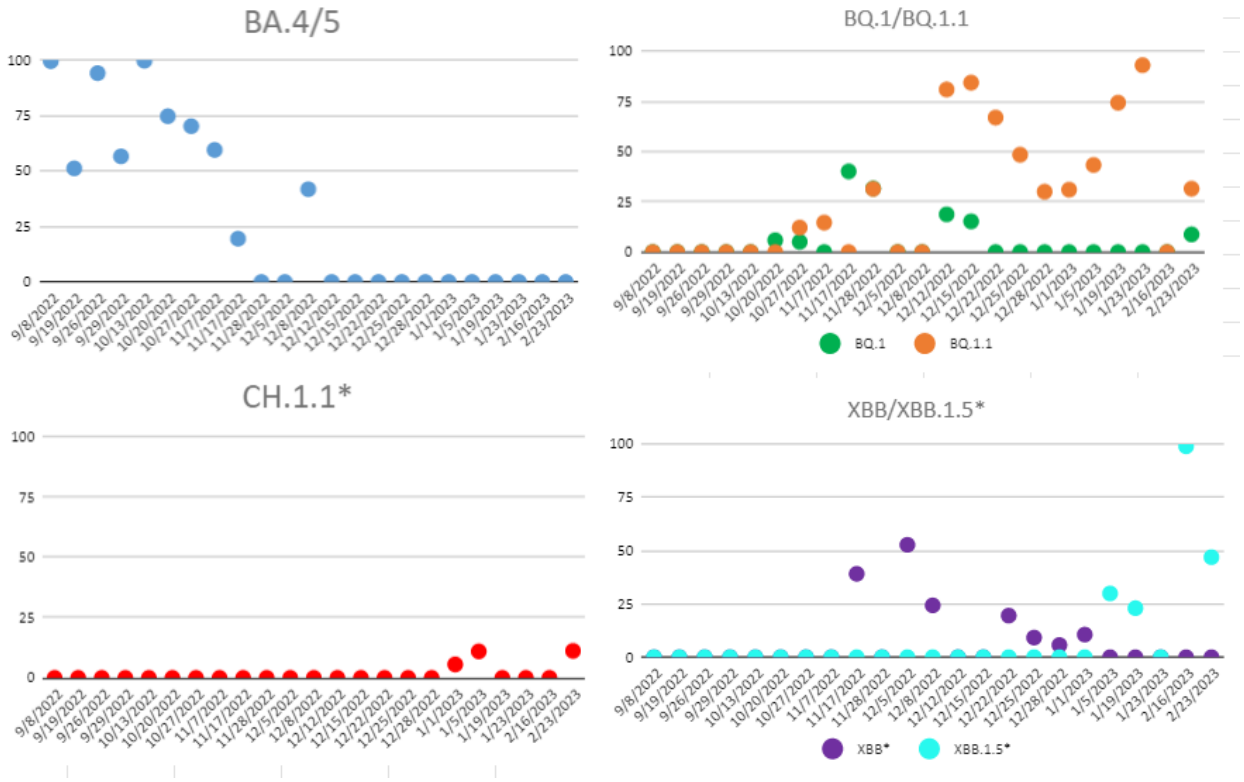

Figure 7. Variant composition in a local WWTP over time. BA.4/5 disappeared in late November, BQ.1/1 emerged in late October and is declining, CH.1.1* rose and fell in January, XBB emerged in November and declined early January, and XBB.1.5 emerged in January and continues to rise.

Discussion of Internship

I was able to achieve the objectives, develop a protocol for the RBD amplification and sequence analysis, in this internship. I collaborated with Dr. [redacted] and my PIs to develop an SOP for nested PCR amplification of the RBD that has been approved by MDHHS and can be used by any lab in the SEWER Network. Additionally, I am working on processing samples from other labs in the network that do not have appropriate equipment to generate amplicons. I have created a dictionary for a python program to read, which is able to identify

circulating and expected emerging variants. I have used this dictionary and program to determine proportional variant representations in the surrounding county.

In this internship, I've become very comfortable with running nested PCR reactions and the attention to detail required when handling RNA. I've learned how SOPs need to be constructed before they are disseminated to other labs. I have also become familiar with running Python as well as how Python is able to read dictionaries. I am comfortable speaking with and collaborating with our partners and I feel confident when I present my internship work; my PIs urged me to present at the Michigan Academy of Science, Arts, and Literature (MASAL). Here, I presented my internship work to other scientists; the abstract is published in the proceedings. Presenting at MASAL gave me experience presenting research in a professional environment.

This internship required a significant amount of understanding of how PCR works, how to perform sterile work, handling RNA, and compiling and communicating data. I feel that CMB 626, CMB 505, and CMB 506 gave me the tools and understanding I needed in order to succeed in my laboratory work. Presentation skills I learned in CMB 501 helped me create and present successful presentations. I have been applying the skills I learned in the PSM-CMB program and continue to do so by teaching upcoming lab members on technique, training health department members, and writing communication papers.

Some challenges I experienced were amplification failure and mixed variant identification. I spoke with my PIs about failed PCRs after several re-runs. To remedy this, I ran the PCR with a dilution, which resulted in an amplicon. Inhibitors in wastewater extracts may

hinder PCR reactions, and now any failed amplifications are re-run with a dilution. Creating the variant dictionary was also quite challenging. Sometimes sequences were classified as “mixed”, because the program was not able to classify the sequence as a specific variant. To fix this, I added more “not” expressions in my dictionary for any amino acid mutation that was not present in a variant. After adding stricter classifications in my dictionary, the program has not classified any variants as “mixed”.

This internship experience was extremely valuable. I became even more comfortable with wet lab work and was able to add more lab skills to my resume. Working as a laboratory team helped me develop my teamwork skills as well as simulated a laboratory team in the industry. I could not have asked for better mentors; they supported me through every aspect of this project. I thoroughly enjoyed collaborating with them to improve our methods as well as preparing our work to present at conferences. My experience here, and my team and mentors, have helped me become a desirable employee, as I was hired by my dream company.

References

1. CDC. (2022, October 5). *Cases, Data, and Surveillance*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html

2. *Wastewater Surveillance for COVID-19*. (n.d.). Retrieved March 30, 2023, from https://www.michigan.gov/coronavirus/stats/wastewater-surveillance/wastewater-surveillance-for-covid-19

3. Steffen, T. L et al., (2020). The receptor binding domain of SARS-CoV-2 spike is the key target of neutralizing antibody in human polyclonal sera. *bioRxiv* (p. 2020.08.21.261727). https://doi.org/10.1101/2020.08.21.261727

4. Gregory, D. et al., (2021). Monitoring SARS-CoV-2 populations in wastewater by amplicon sequencing and using the novel program SAM refiner. Viruses, 13(8): 1647. doi: 10.3390/v13081647.