NW-PaGe
Northwest Pathogen Genomics Center of Excellence

Data Integration and Quality Assurance of Sequencing Metadata in Washington State

2024-06-07

Frank Aragona, Cory Yun, Philip Crain, Paul Lloyd, et.al
Washington State Department of Health

2024

Data Integration/Quality Assurance

Table des matières

Data Integration and Quality Assurance of Sequencing Metadata in Washington State

Frank Aragona

Cory Yun

Philip Crain

Paul Lloyd

Emily Nebergall

Cameron Ashton

Peter J Gibson

Marcela Torres

Lauren Frisbie

Allison Warren

Laura Beilsmith

Xichi Zhang

Allison Thibodeau Sarah Jinsiwale Yunpeng Yu

Topias Lemetyinen Alli Black Alex Cox

2024-03-04

Abstract Genomic surveillance is important for identifying and tracking SARS CoV-2 variants to better mitigate spread of COVID-19. Washington State Department of Health quickly increased capacity to surveil SARS CoV-2 variants by partnering with over 25 labs to collect sequencing data while developing and implementing solutions to standardize submissions and enhance data linkage and quality. High impact solutions included development of a standardized reporting template, collection of case demographics, adaptation of HL7 messages with sequencing data, and strategic utilization of external sequencing data repositories. We developed an automated pipeline that combines data science tools to ingest, clean, and link SARS CoV-2 sequencing data from multiple sources, while accounting for differing data formats and quality. This manuscript details the first version of the pipeline developed in February 2021 when processes were unstable and were being developed as they were utilized.

1 Introduction

This manuscript documents the original data integration pipeline for SARS-CoV-2 sequences for Washington State Department of Health during early 2021 to mid 2023. The Sequencing project began in February 2021 as an effort to process sequencing data into the Washington State Disease Reporting System (WDRS). In turn, the data fuels SARS-CoV-2 variant tracking and the generation of Covid-19 reports which are disseminated to the highest levels of state government. The pipeline continuously links data to our main database WDRS, where the data can be used to gain insights via surveillance reports or research [1]–[3]. Data are processed via numerous R/Python/SQL scripts and are uploaded via rosters (.csv files) into WDRS and our Molecular Epidemiology produces reports with the data. These processes ebb and flow often as changes are needed regularly in response to the data that are received. Since mid 2023 we have stopped using this particular pipeline and have built a new pipeline using a more streamlined approach. The purpose of detailing the original pipeline in this document is to give a transparent look at how data were processed under the unusual circumstances during the COVID-19 pandemic. Many aspects of this pipeline are inefficient because it was built under a rapidly changing environment, one that had never been built before, in addition to the many time constraints placed on our teams to produce data reports quickly. Some of the inefficiencies exposed in this pipeline still exist with our newer pipelines, but our teams are working to build a more sustainable way to process sequencing data of any disease type.

There are a multitude of barriers which make data processing difficult with any pipeline such as:

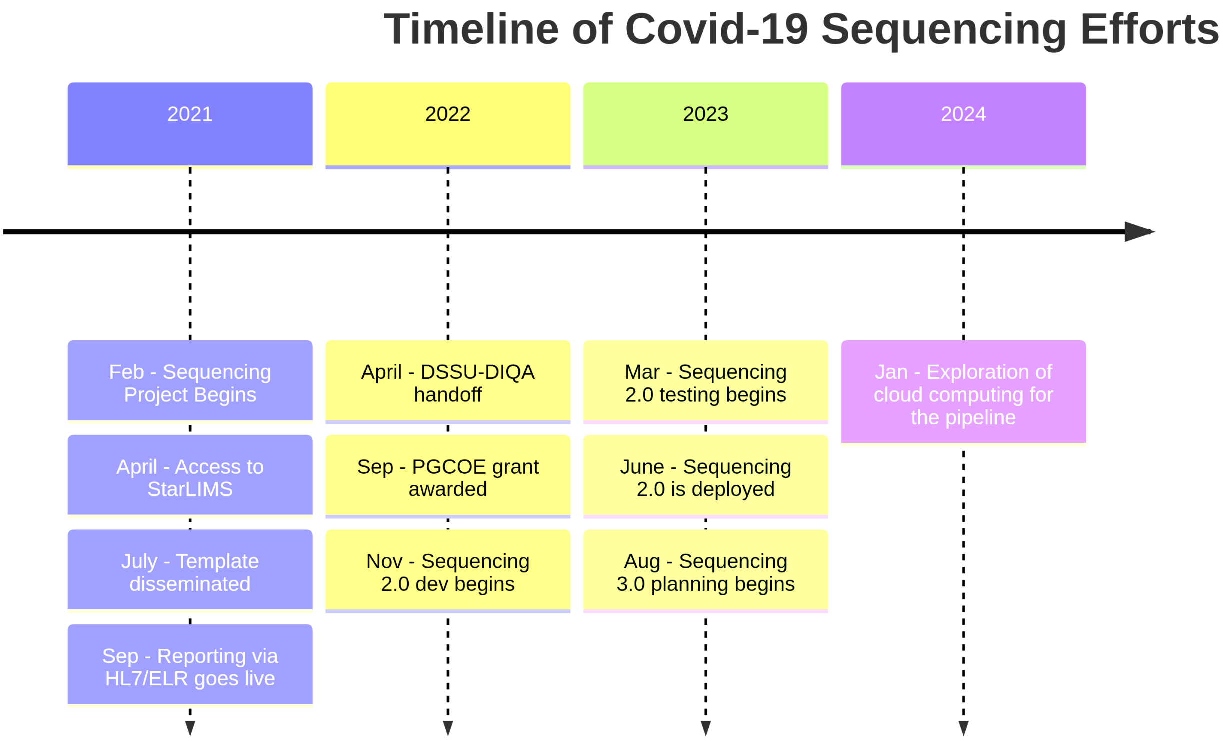
1. Data standardization; data are received multiple ways. Depending on the manner it is submitted it may not follow the standard format requested from submitters. This requires manual intervention or communication back to labs. In addition, some submitters cannot change the manner in which they submit data which makes standardization across all labs impossible. These are handled by separate processes. Occasionally, submitters may break consistency in their own manner of which they report as well.
2. Data quality; the quality of data received from labs can vary dramatically. Data needs to match between three sources: Lab Submissions, WDRS, and GISAID. When the wrong data are sent this can make matching impossible and prevents records from making it into our systems. It has been found that submitters sometimes submit incorrect ACCESSION ID's that are used to match records between systems. Without the correct data it requires a considerable amount of manual intervention to be able to roster those records. Additionally, there are considerable lag times between all three data points/repositories; when a record is submitted that are not within WDRS the records cannot be matched and rostered.
3. Technological gaps; the COVID-19 pandemic has exposed many technological gaps in our public health disease surveillance systems. Much of the technology used for processing and storing data needed to be built out during early 2020 so that we could provide disease reports in a timely manner. Therefore a lot of processes like this pipeline were built for short term needs, adding on more and more 'technological debt'. Short term so-

> lutions have consequences and our data infrastructure is not well suited for pipelines like this.

The original pipeline aggregates sequencing data submitted via secure file transfer, electronic lab reporting, lab management software, and open access data repositories. The process incorporates robust solutions to data challenges, including lagged data availability, variable data formatting, record duplication, and missing data. After aggregation, the pipeline connects records to COVID-19 case data via sequence to diagnostic test identifiers. We used patient demographics and string matching to link sequences to cases that are missing the unique identifiers. Internal databases are compared to external repositories to identify new results, track missing data, onboard new sequencing partners, and validate data. This pipeline was able to mostly automate aggregation, cleaning, and linkage of SARS CoV-2 genomic surveillance data, minimizing manual work and hastening availability of data for analysis and reporting.

Challenges remain despite these improvements in data standardization and management. Barriers include differences in HL7 message reporting capabilities among submitters, inconsistencies in virus naming conventions, challenges in pulling data from public repositories, and limitations within our current internal database infrastructure. These factors increase likelihood of errors, require data processing logic unique to individual submitters and require manual intervention. Continued development of national standards can address these issues. Data accessibility can be improved by encouraging open-source sharing, especially to repositories that enable full programmatic data sourcing for all users. A community of practice across departments of health to discuss storage methods, processing pipelines, and matching approaches would enhance current practices, and support greater consistency and interoperability across public health.

1.1 Timeline


Timeline of Covid-19 Sequencing Efforts
2021
2022
2023
2024
Feb - Sequencing Project Begins
April - Access to StarLIMS
July - Template disseminated
Sep - Reporting via HL7/ELR goes live
April - DSSU-DIQA handoff
Sep - PGCOE grant awarded
Nov - Sequencing 2.0 dev begins
Mar - Sequencing 2.0 testing begins
June - Sequencing 2.0 is deployed
Aug - Sequencing 3.0 planning begins
Jan - Exploration of cloud computing for the pipeline

The sequencing metadata linkage project began in February 2021 by the Data Science Support Unit (DSSU). A pipeline was needed to process and upload sequencing data to Washington Disease Reporting System (WDRS) for variant tracking/generation of reports to the governor's office. The pipeline links sequencing metadata to case data in WDRS, links sequences to GISAID (Global Initiative on Sharing All Influenza Data, a public sequencing repository), and it cleans and transforms non-standardized data.

The project originally started as a group of individual contributors writing R scripts to handle needs for pulling data and cleaning, matching, and transforming non-standardized sequencing data. During this period, urgency was prioritized at the cost of technical debt: it was a "build as we go" mentality given the time restraints during the pandemic. This scenario made for segmented processes and no true workflow. During height of the 2021-2022 period this pipeline would typically process 1000+ records per week.

2 General Overview

Sequencing data gets sent through our pipeline through multiple processes and the pipeline works in the following steps:

1. A submitter sends us the sequencing data three different ways;
    - as tabular data via secure file transfer (SFT)
    - as tabular data that is 'scraped' web-scraping of their dashboard as is the case with our PHL
    - via ELR or electronic lab reporting that is automatically connected to our database
2. Our pipeline will extract, transform, and link that data to a case in WDRS
3. The pipeline then performs quality checks to make sure errors or data leaks did not occur

The three main routes that a submitter can send us data through are detailed in Figure 1 under Template Submitters (secure file transfer of tabular data), PHL (webscraping of tabular data), and HL7 messages (secure data transfer for ELR).

First, the Template Submitters script processes the majority of data received by external submitters. Sequencing data from external submitters are received via `.csv` files in a template format which they uploaded to our secure file transfer (SFT) portal.

Second, the The PHL Roster script processes the data received from PHL (Public Health Laboratory), our internal laboratory. Sequencing data from PHL is pulled from an internal dashboard, 'StarLIMS'. The logic for both processes are similar. There is an attempt to link the sequencing data to the patient-level data using a `SEQUENCE_CLINICAL_ACCESSION` ; an accession ID that should match between the patient-level data and the specimen sequenced by laboratories (note: while the terms 'Laboratory' and 'Submitter' may be used interchangeably at times they are not the same). In many cases, this accession ID is unable to match. This may be due to multiple reasons ranging from lag times to data quality issues. In this case, if demographics have been provided by the submitter an attempt will be made to match based on these demographic variables (name, date of birth, etc.).

Lastly, the ELR Roster script processes the data from external submitters that have been received via HL7 messages and populated in the table in WDRS named `[dbo].[DD_ELR_DD_ENTIRE]` . In this case, the ELR (electronic laboratory reporting) process performs the matching of sequencing data to the patient-level data. Since the sequencing data and patient-level data are already tied for these records received via ELR, the logic for the ELR Roster script is predominately transformation to a format acceptable for import into the `[dbo].[DD_GCD_COVID_19_FLATTENED]` table. All records, regardless of the process through which they are received/processed are uploaded to the `[dbo].[DD_GCD_COVID_19_FLATTENED]` table via .csv files in a roster format. These rosters are the end product and output of all three processes.

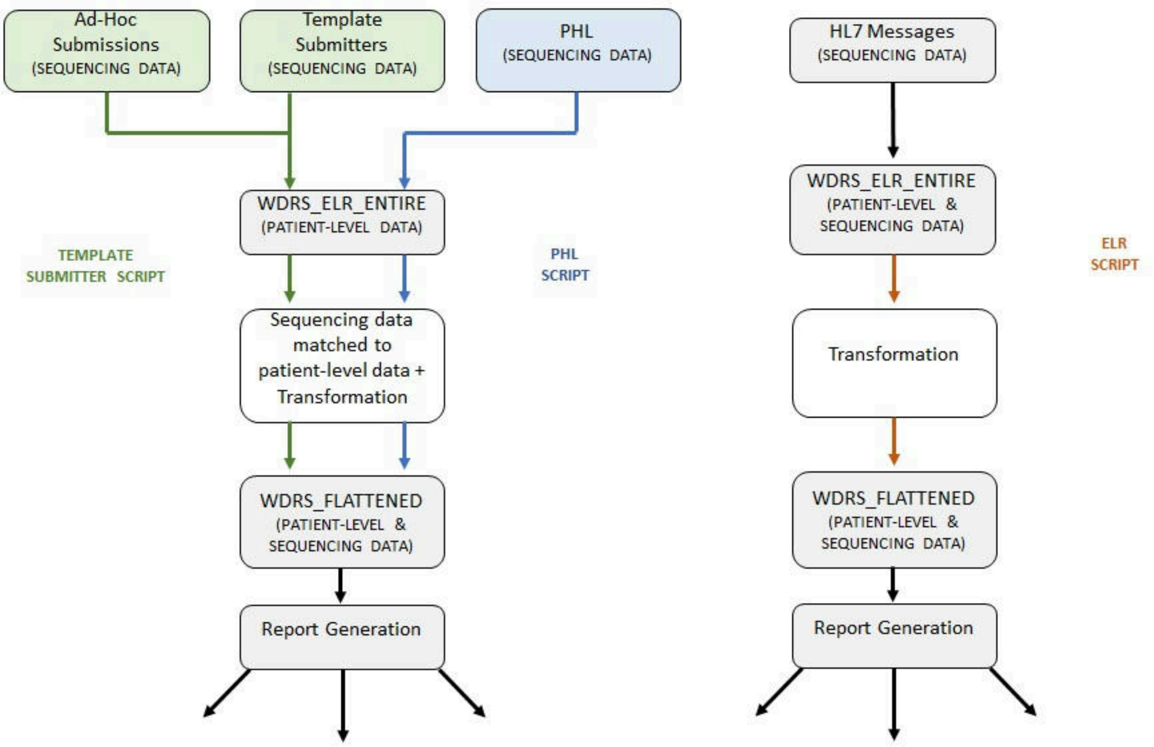

Ad-Hoc Submissions (SEQUENCING DATA)
Template Submitters (SEQUENCING DATA)
PHL (SEQUENCING DATA)
HL7 Messages (SEQUENCING DATA)
WDRS_ELR_ENTIRE (PATIENT-LEVEL DATA)
WDRS_ELR_ENTIRE (PATIENT-LEVEL & SEQUENCING DATA)
TEMPLATE SUBMITTER SCRIPT
PHL SCRIPT
ELR SCRIPT
Sequencing data matched to patient-level data + Transformation
Transformation
WDRS_FLATTENED (PATIENT-LEVEL & SEQUENCING DATA)
WDRS_FLATTENED (PATIENT-LEVEL & SEQUENCING DATA)
Report Generation
Report Generation

Figure 1 — Overview of sequencing pipeline

In addition to sequencing data submissions and WDRS, there is a third component; GISAID (Global Initiative on Sharing Avian Influenza Data). GISAID is an initiative that provides open-access to genomic data of influenza viruses and more importantly for the purposes of this project, SARS-CoV-2; coronavirus responsible for the COVID-19 pandemic). This database is available online to the public and holds genomic data of sequenced specimens from across the globe. In theory, submitters should be sending their sequencing data to the DOH and GISAID. When this happens there is consistency between both databases and the data received from submitters Figure 2

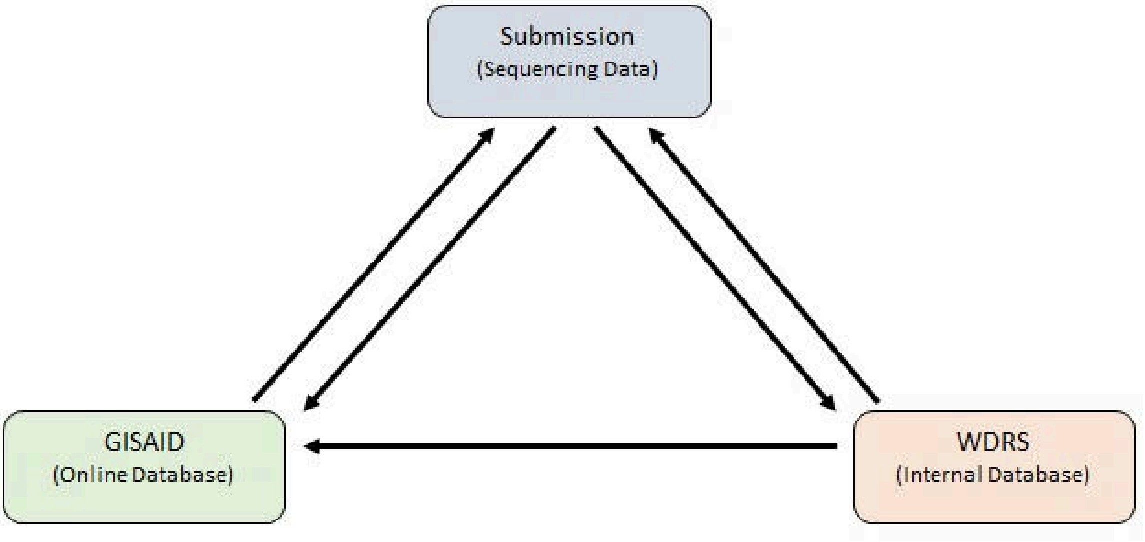


Figure 2 — Sequence submissions should match WDRS and GISAID

It should be noted that in some instances sequencing data is manually entered by creating events within WDRS. However, this practice is not common and generally should be avoided if at all possible in order to prevent non-standard entries and potential data quality issues.

2.1 Laboratories and Submitters

Through various contracts and collaboration with external agencies the DOH receives sequencing data from numerous submitters. Laboratories are the entities which perform the sequencing. Submitters are the entities that relay/send sequencing data to the DOH. Many of our submitters perform both the sequencing and submission of data. There are some submitters that send sequencing data to the DOH on behalf of multiple laboratories as well. Therefore, some laboratories that perform the sequencing do not directly submit the sequencing data themselves. Data quality is a significant issue across all submission routes listed below. This is mainly due to the fact that there are no national standards as how sequencing data should be transmitted.

2.2 External Dependencies and Data Pulls

Prior to any processing of sequencing data received from submitters there are numerous data pulls and external dependencies that must be completed. Below is a brief description of each process Figure 3.

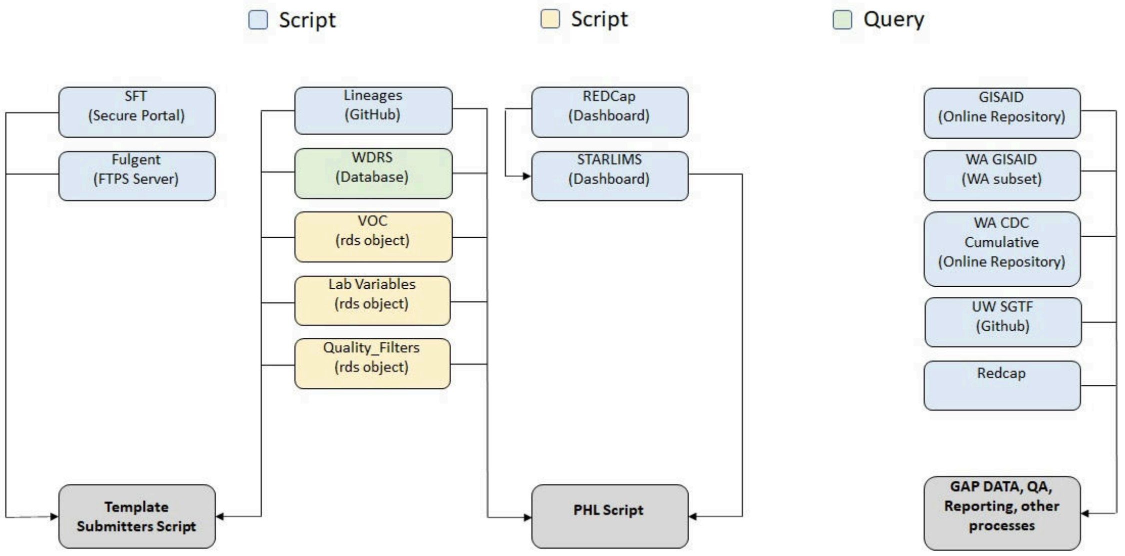


Figure 3 — External Data Pulls

The scripts below are responsible for pulling submissions from their corresponding locations and dropping a file into the submissions folder in the Network Drive so that it can be picked up by the roster scripts and processed into WDRS.

2.2.a Data Pulls

sel_Dashboard_All.Rmd performs the task of pulling data from across three separate dashboard within starLIMS. This data is aggregated then placed into the submission folder for PHL in the network drive for processing.

sft_main.py performs the task of pulling all data from the individual submitter folders within the SFT, routing the downloaded files to the correct submitter folders in the network drive, deleting out the old files, and keeping a log. Additionally, an email is sent out to the correct stakeholders each day on what submission were uploaded to the SFT (if any) and notifies of any new labs that have uploaded a submission for the first time.

2.2.b External Processes

wa_cdc_pull.Rmd performs the task of pulling data from the CDC for specimens sequenced by laboratories under the CDC for the state of Washington so that it can be picked up by multiple QA scripts and utilized by other other stakeholders.

lineages_main.py performs the task of pulling data from a .txt in a GitHub repository containing the latest Covid Lineages and dropping them in the lineages folder in the network drive so that it can be picked up by multiple scripts and utilized by other stakeholders. The .txt file in the repository is the same file used update the Cov-Lineages site (https://cov-lineages.org/lineage_list.html).

2.2.b.a SGTF

To assist with the monitoring of Omicron, five labs are submitting S-gene target failures (SGTF) data to WADOH via Redcap or GitHub (UW only). These submissions are standardized and compiled each day to calculate % SGTF by epidemiological week.

sgtf_compile_daily.Rmd performs the task of compiling all the templates and performing the necessary calculations. Submissions are downloaded by Molecular Epi each day and placed into the network drive to be picked up by sgtf_compile_daily.Rmd

uw_sgtf.Rmd performs the task of pulling the latest SGTF file from UW's GitHub repository, routing the downloaded files to the correct folder in the Network Drive, and keeping a log.

2.3 Example Datasets

In 2021, data was sent from sequencing labs to us via tabular files. There were no standards between submitting labs, and for a given submitter the format of the tabular files would often change between each submission as well. It was impossible to process these data without editing scripts each time to account for a varying format. Initially, all data was received via non-standardized tabular files, including data sent from our public health lab (PHL). We did not have access to their starLIMS database at the time. Figure 1 below is an example of the data sent to us in tabular format. Figure 2 shows data sent in tabular form from our public health laboratory (PHL) and Figure 3 is an example of data sent from the University of Washington Virology Lab (UW Virology).

Figure 1 — example of tabular datasets sent to the Department of Health from sequencing labs in 2021

Variable	Description
Accession	identifier that links a sequence to a test
COLLECTION_SAMPLE_ID	the identifier that linked a sequence to a test
ORIG_ACCESSION_NUMBER	the identifier that linked a sequence to a test
PAT_FIRST_NAME	patient first name
PAT_LAST_NAME	patient last name
DATE_OF_BIRTH	date of birth
PAT_ADDRESS_1	address
PAT_CITY	city
PAT_STATE	state
PAT_ZIP	zip code
Phone	phone number
Original Physician	doctor name

Figure 2 — example of tabular datasets sent to the Department of Health from a Public Health Lab (PHL) during 2021

Variable	Description
LIMS	the laboratory information management system (LIMS)
Project	the reason for sequencing
Investigator_sample_id	an identifier to the sample
collection_date	the specimen collection date
age	patient age
county	patient county
sex	patient sex
specimen_id	the identifier linking to the original PCR covid test
submitting_lab	the lab submitting the sequence
note	free text note field

Figure 3 — example of tabluar datasets sent to the Department of Health by UW Virology during 2021

Variable	Description
uwnum	the identifier of the sequence that links to GISAID
acc_num	the accession number linking to the positive covid PCR test
collection date	specimen collection date
loc_state	patient state
fullname	the full GISAID identifier name, such as HCoV19-USA/WA-#####/2021 (not the patient's name)
shortname	a partial GISAID identifier name, such as WA-#######

As you can see from the tables above, data sent via tabular format early in the project was not standardized and could not be processed automatically due to it constantly changing field names and descriptions.

3 Roster Workflows

There are three main workflows for rostering sequencing data into WDRS. ELR, PHL, and Template scripts as detailed in Section 2. All three will output a roster, and the Roster Compile script will combine them and send the data to be rostered into WDRS. See Figure 4 for a high level overview. The elr.Rmd script pulls sequencing metadata from WDRS, transforms it, and runs QA checks on it before putting it into a usable roster. template_submitters.Rmd performs the task of processing template submission. PHL.Rmd performs the task of processing all phl records. Both the template and phl processes operate based on similar logic from a high level view, but there are significant differences between each script.

The `SEQUENCE_CLINICAL_ACCESSION` variable is used to find a matching event within the `[dbo].[DD_ELR_DD_ENTIRE]` table. The `CASE_ID` (WDRS variable that is an identifier for a disease case, not a person) for that event is pulled then assigned to the corresponding record. At this point, if an event using the accession ID cannot be found for a record it will be routed to two different processes depending on if the sequencing data received has patient demographics attached to it; FIRST_NAME, LAST_NAME, MIDDLE_NAME, DOB. If the record received contains patient demographic it is routed to the fuzzy matching process, an attempt to match to the correct event will be made using the demographic information. If the record received contains no patient demographic it is routed to the keep na process (Section 3.4), the record will be retained and an attempt to match via the accession ID will regularly be made in case the corresponding event populates in WDRS later on.

Once a record has been matched to an event it will undergo transformation to clean and standardize the matched data into a roster format. Some submitters do not provide the full `GISAID_ID` in the submission. In this case, the `SEQUENCE_ACCESSION` can be constructed from their internal accession is inputted into the `GISAID_ID` column. This happens during the transformation process and the resulting `SEQUENCE_ACCESSION` should match to what is in GISAID.

Records are then put through a series of quality filters to check for QA issues. All records that pass this series of QA checks will then populated into a final roster then outputted to be picked up by the roster compile script Section 3.1.d and sent to data support for upload to WDRS.

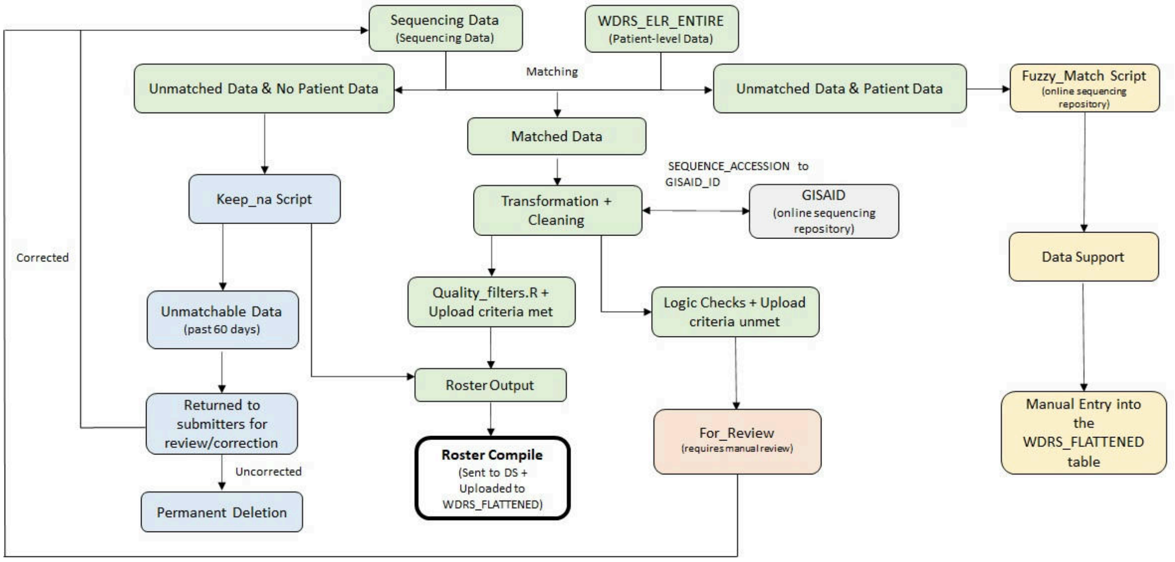

Sequencing Data
(Sequencing Data)
WDRS_ELR_ENTIRE
(Patient-level Data)
Matching
Unmatched Data & No Patient Data
Unmatched Data & Patient Data
Fuzzy_Match Script
(online sequencing repository)
Matched Data
SEQUENCE_ACCESSION to GISAID_ID
Keep_na Script
Transformation + Cleaning
GISAID
(online sequencing repository)
Corrected
Data Support
Unmatchable Data
(past 60 days)
Quality_filters.R + Upload criteria met
Logic Checks + Upload criteria unmet
Roster Output
Returned to submitters for review/correction
For_Review
(requires manual review)
Manual Entry into the WDRS_FLATTENED table
Roster Compile
(Sent to DS + Uploaded to WDRS_FLATTENED)
Uncorrected
Permanent Deletion

Figure 4 — Roster Compile

3.1 Roster Scripts

This section gives more details about each roster script and a high level diagram following the process.

Legend:

External Data Owner

Internal Data Owner

Data processing (software)
Data processing (scripts)
Database

3.1.a ELR

Electronic Lab Reporting for sequencing went live in September 2021. These are records with Covid PCR tests processed by WELRS/DRIVE and sent to WDRS. See Figure 5 below. For more details on the script, see the ELR notebook. From a high-level overview, the script will:

- WELRS/DRIVE process, match, and fill the entire/lab tables but not sequencing table
- No QA processing
- Sequencing table is built as an after thought
- WELRS/DRIVE is somewhat of a black box to us (changes without knowing, don't have oversight on mismatches)
- Our ELR script will extract from the entire table, transform, QA, and send via roster

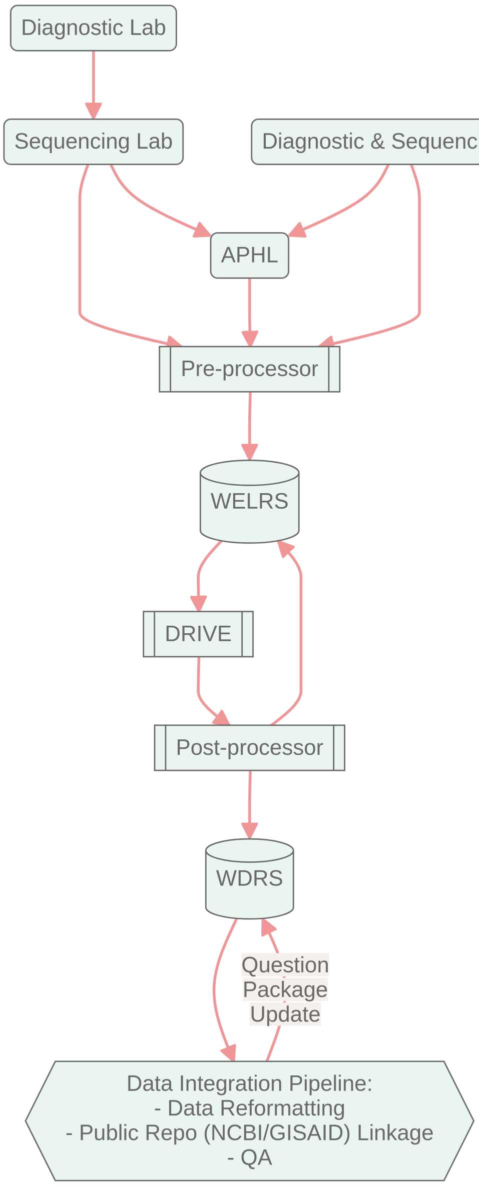


Figure 5 — ELR submission to WDRS workflow

3.1.b PHL

Access to starLIMS, our Public Health Laboratory (PHL) system, was granted in April 2021. However, there was no API or underlying database access so the pipeline needed to scrape data from the starLIMS dashboard. It would download the `.xslx` files from starLIMS and then use identifiers to match sequences to a case in our database, WDRS. See Figure 6 below for a high level summary. This process can get complicated for a multitude of reasons mainly due to challenges with our underlying data infrastructure. For more details on the workflow and to view those challenges, see more details here and for all script details see the PHL notebook. From a high-level overview, the script will:

- Scrape from StarLIMS

- Match to a WDRS case
- If no match based on `FILLER_ORDER_NUM` then match on demographics
- Uses a processed file to eliminate feedback loops (prevent failed records from being processed every run)
- Fields may change in starLIMS without our knowledge

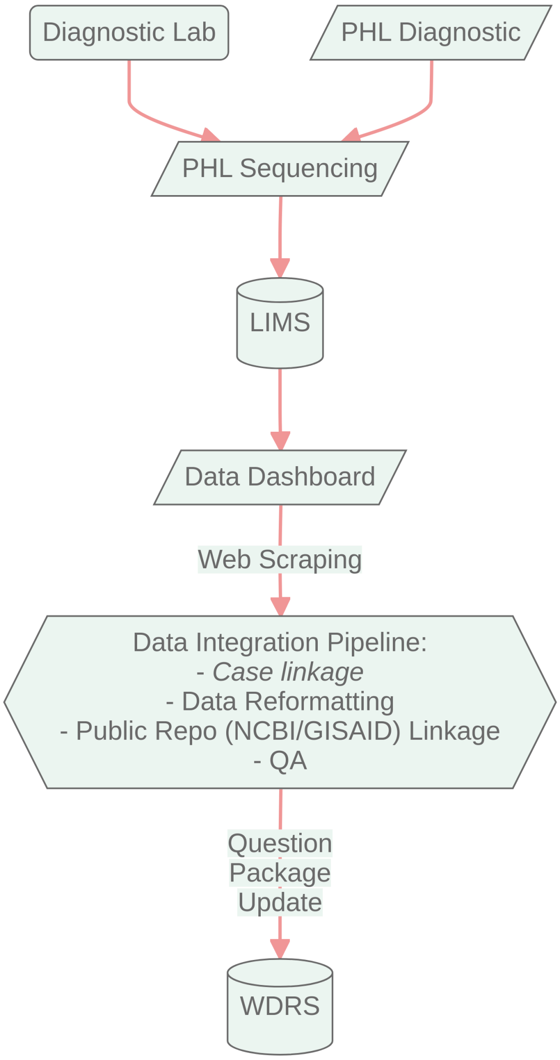


Figure 6 — PHL submission to WDRS workflow

3.1.c Template

There are still labs that cannot send us data via ELR or PHL and could only send us tabular files. The Department of Health has a secure file transfer (SFT) site that the labs could send data to and that we could pull from. We did not have a way to pull from our own SFT site so we needed to scrape this data as well. In July 2021, Cory Yun developed a template (see Figure 4) for labs to fill the sequencing data instead of labs sending us data with no standards. See Figure 7 below. For script details see the template submitters notebook. From a high-level overview, the script will:

- Labs send us a `.csv` file into our MFT site
- All data follows a specific template created by Cory
- Scrape the site and download the `.csv` files for each lab
- Format, find a match based on `FILLER_ORDER_NUM` or demographics

Figure 4 — Template Data Variables

Variable	Description	Example
LAB_ACCESSION_ID	id matching a sequence to a PCR test	alphanumeric string
GISAID_ID	identifier of sequence in GISAID	USA/WA-X/2020
SPECIMEN_COLLECTION_-DATE	collection date of specimen	mm-dd-yyyy
SUBMITTING_LAB	lab name	UW Virology
SEQUENCE_REASON	reason for sequencing	SENTINEL SURVEILLANCE
SEQUENCE_STATUS	complete or failed sequence	COMPLETE
PANGO_LINEAGE	lineage call in GISAID	BA.1
FIRST_NAME	patient first name	
LAST_NAME	patient last name	
MIDDLE_NAME	patient middle name	
DOB	date of birth	
ALTERNATIVE_ID	alternative identifier	alphanumeric string

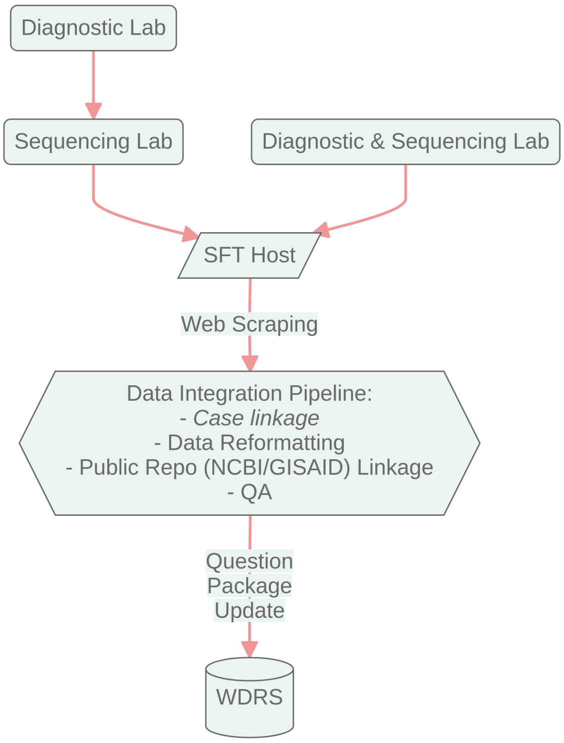
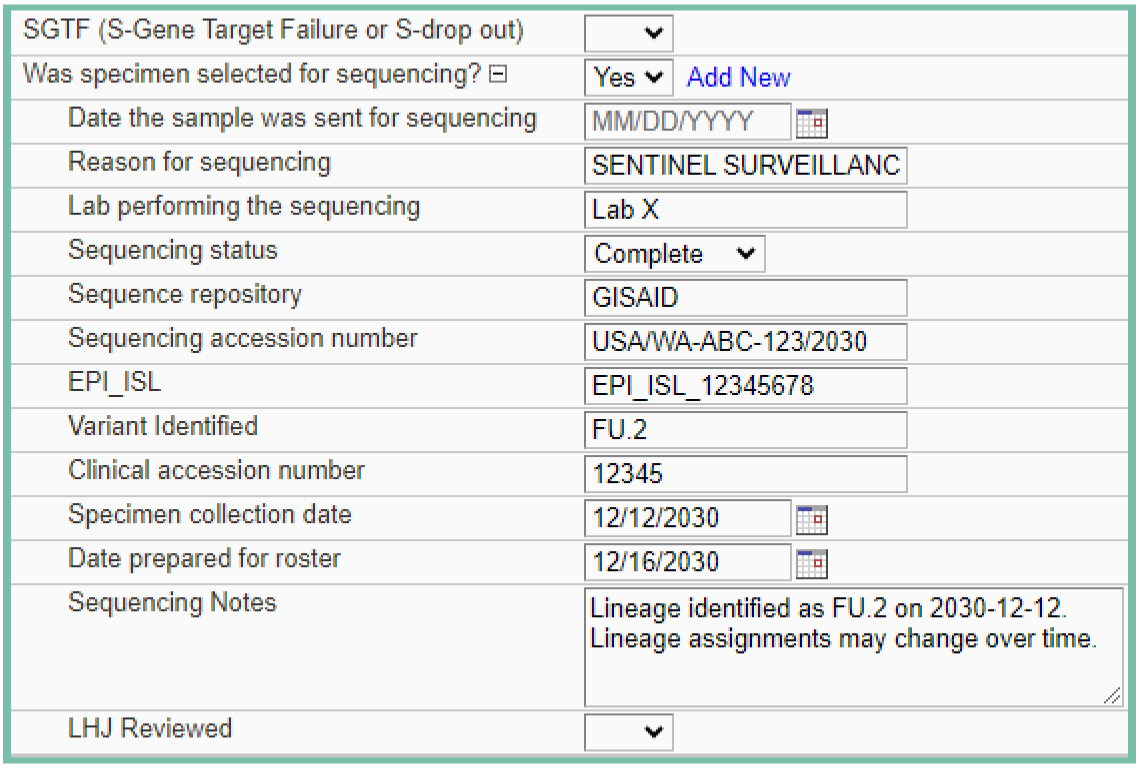


Figure 7 — Template submission to WDRS workflow

3.1.d Roster Compile

After the Template, PHL, and ELR scripts run, they all output a `.csv` file into a folder called `write_roster_here` in our network drive. The Roster Compile script reads in all of these files, combines them, and runs additional QA checks on them before outputting the results into one file for output to the WDRS database. Then our Data Support team will upload the file into WDRS where it will update the results into the flattened table. Each row will match a `CASE_ID` in WDRS and the sequencing event is added to the cases external data as seen below in Figure 8.



SGTF (S-Gene Target Failure or S-drop out)	
Was specimen selected for sequencing? ⊟	Yes Add New
Date the sample was sent for sequencing	MM/DD/YYYY
Reason for sequencing	SENTINEL SURVEILLANC
Lab performing the sequencing	Lab X
Sequencing status	Complete
Sequence repository	GISAID
Sequencing accession number	USA/WA-ABC-123/2030
EPI_ISL	EPI_ISL_12345678
Variant Identified	FU.2
Clinical accession number	12345
Specimen collection date	12/12/2030
Date prepared for roster	12/16/2030
Sequencing Notes	Lineage identified as FU.2 on 2030-12-12. Lineage assignments may change over time.
LHJ Reviewed	

Figure 8 — WDRS Front End

3.2 For Review

The pipeline attempts to link sequencing data to cases in WDRS. Some records have quality issues and cannot be processed in our system. These data are tagged and saved in a separate folder where our team reviews them and attempts to fix and re-process them. Figure 5 is an example of the sort of issues that get tagged in our pipeline:

Figure 5 — For review quality issue tags

Variable	Description
QA_CASE_ID	Missing CASE_ID from WDRS
QA_SCA_NA	Clinical Accession identifier is missing
QA_SCA_INT_DUPE	Clinical Accession duplicate in file
QA_SCA_WDRS_DUPE	Clinical Accession duplicate found in WDRS
QA_SA_INT_DUPE	Accession duplicate in file
QA_SA_WDRS_DUPE	Accession duplicate found in WDRS
QA_SEQ_VARIANT	Variant not in list of VOC
QA_SEQ_STAT	Status error (labeled complete sequence when it was failed)
QA_SEQ_REASON	Unknown sequence reason
QA_SEQ_NOTES	Sequence note not formatted
QA_COLLECT_DATE	Match found but collection dates >14 days
QA_OTHER	Other formatting issues
sum	Total number of errors found

3.3 Fuzzy Matching Review

When records cannot be linked via accession identifier the pipeline attempts to match a sequence to a PCR test in WDRS via demographics (first name, last name, date of birth, and collection date). The fuzzy matching script uses string distances to match names from a submitter to names in WDRS and determine the highest likelihood of a correct link.

There may be quality issues with the demographics and the fuzzy matching script tags issues and outputs them into a fuzzy matching review folder where our team will manually review the errors and re-process the files. Figure 6 is an example of the files the fuzzy matching process outputs:

Figure 6 — Fuzzy matching review quality issue tags

File	Description
Fuzzy bad rows	error in a column other than demographics columns
Fuzz 1	best match was a name levenshtein distance of 1
Fuzz 2	best match was a name levenshtein distance of 2
Fuzz 3	best match was a name levenshtein distance of 3
Did_not_match	no match was found
Fuzzy perfect	perfect match found

3.4 Keep NA

A sequenced specimen may not initially match to our database (WDRS) for many reasons. A case may not have been updated at the time our pipeline tried to match the sequence to the PCR test, or a sequence may simply not match to a case in WDRS. Our Keep NA script reads in all the data that could not be matched in previous pipeline runs and attempts to match them again in the case that new and updated case data is in WDRS. If an unmatched record is in our archive for more than 60 days the Keep NA script will remove it from the list and keep it in an archived file. We made this decision because the vast majority of records that are in Keep NA for over 60 days have never matched to any case in WDRS.

4 QA Processes

4.1 Gap data

gap_data.Rmd performs the task of identifying and tracking the number of sequencing records for the state of WA that have been submitted to GISAID but are missing from WDRS. As previously mentioned, submitters should be sending records to both the DOH and GISAID. This process uses the `SEQUENCE_ACCESION` (`GISAID_ID`) to identify any records from GISAID that are not in WDRS for the state of WA. An excel file containing two pivot tables and metadata is output. Each pivot table contains either the number or proportion of records missing in WDRS, this information is the submission date (month-year) and the submitting lab. The metadata contains each record and accompanying information pulled from GISAID.

This output is utilized by Data Support and other stakeholders for two main reasons. First, to reach out to submitters to regarding missing records. Second, to identify any new that are submitting to GISAID regularly and should potentially be onboarded.

4.2 WDRS Logic Checks

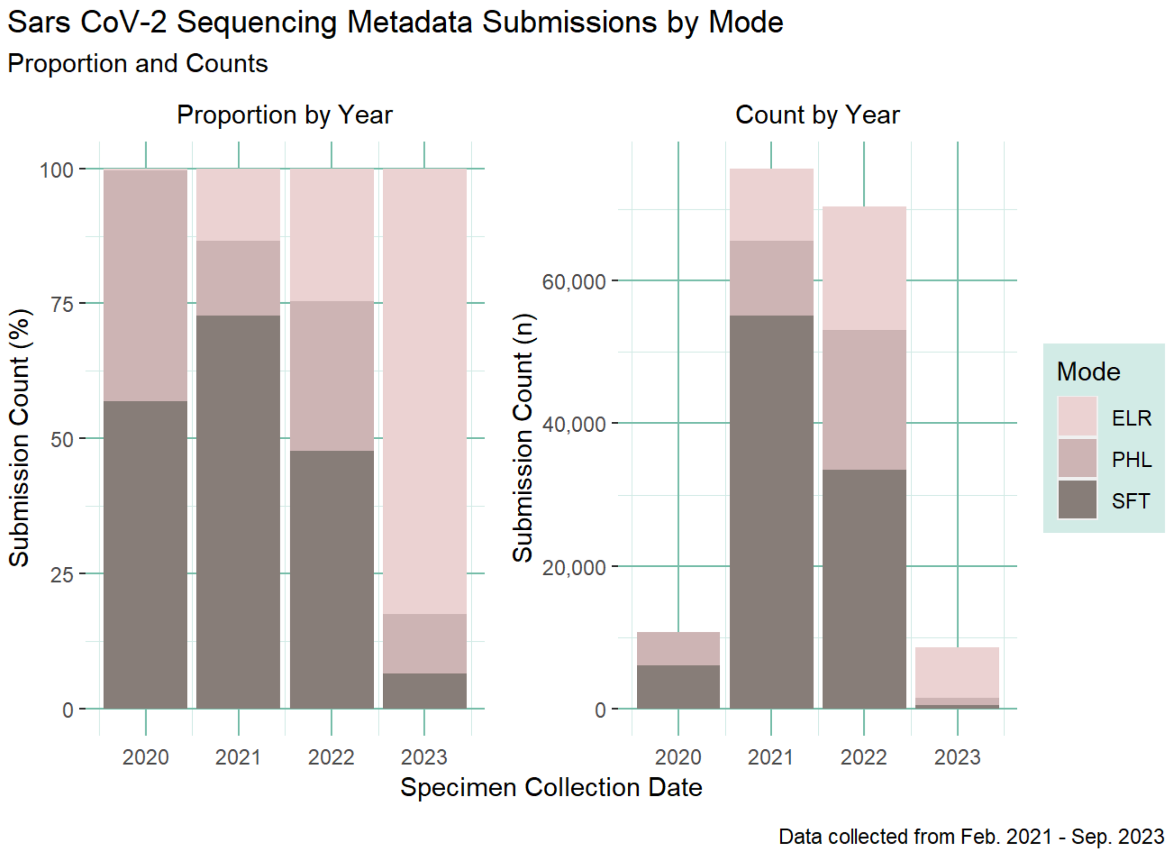
`wdrs_logic_checks.R` pulls the sequencing data from our database and runs checks on them to confirm that there are no issues with the data uploaded. See Figure 7 below for more info.

Figure 7 — WDRS QA Checks

QA Check	Description
SEQUENCE_REASON is NULL	The reason for sequencing is used to determine sentinel surveillance counts and cannot be null
SEQUENCE_REASON not standardized	If the reason has a typo or unexpected value
SEQUENCE_VARIANT_OPEN_TEXT filled but SEQUENCE_STATUS is not COMPLETE	Status must be complete when the variant is filled
SEQUENCE_ACCESSION number NULL but status not FAILED/LOW QUALITY	A sequence identifier must be provided for complete sequences
SEQUENCE_VARIANT_OPEN_TEXT exists but SEQUENCE_ACCESION number is null	A sequence identifier must be provided for sequences with lineage calls
SEQUENCE_VARIANT not of concern/ interest - check or update list	Lineage has a typo or not a variant of concern
SEQUENCE_LAB not standardized - check or update list	The lab name is not standardized to our database standards
SEQUENCE_SPECIMEN_COLLECTION_DATE out of range. Before 1/05/2020 or after today's date	The sequence collection date is in the future or before 2020 (invalid)
SEQUENCE_SPECIMEN = 'No' but sequencing data attatched	Database error
SEQUENCE_ACCESSION number and SEQUENCE_CLINICAL_ACCESSION numbers missing	A sequence needs the identifiers attached
Unexpected characters in a column	Database or submitter error (usually typos or wrong value in a column)
Lineage found in SEQUENCE_NOTES but SEQUENCE_VARIANT_OPEN_TEXT is NULL	Database error
SEQUENCE_STATUS = 'Complete' and SEQUENCE_VARIANT_OPEN_TEXT is NULL	Sequence needs a lineage call if status is complete
Duplicate - SCA, SA and Variant duplicated	Duplicate identifier values found in database

5 Results

During the February 2021 to September 2023 period we processed a total of 172,050 sequences. These data were most commonly processed via SFT (secure file transfer) of tabular datasets (see Figure 9 below)



Sars CoV-2 Sequencing Metadata Submissions by Mode

Proportion and Counts

Data collected from Feb. 2021 - Sep. 2023

Figure 9 — Count and proportion of sequencing metadata submissions by mode

96% of those sequences were successfully matched to a case in WDRS, while 3% had no match. Less than 1% of the records had quality issues that could not be resolved and are still archived in our for review process. See Figure 8 for more details.

Figure 8 — Count of sequences matching to WDRS cases

Covid Sequencing Pipeline Counts

Location	Count
For Review	220 (0.13%)
Fuzzy Review	569 (0.33%)
Keep NA	5,710 (3.32%)
WDRS	165,551 (96.22%)
Total	172,050 (-)

When stratifying by lab/submitter in Figure 9, we can see that most of the sequences were submitted by 4 labs. Over 40% of the sequences were submitted by University of Washington Virology Lab, followed by our own PHL with 19%, Labcorp with 15% and Northwest Genomics with 11%.

Figure 9 — Count of sequences by lab and status during the sequencing pipeline 1.0 phase

Number of Sequences by Lab

Before 2023-06-01 switch to 2.0 pipeline

Sequencing Lab	Count	Percent of Total Sequences
UW Virology	69,799	42.2%
PHL	32,845	19.8%
Labcorp	26,597	16.1%
NW Genomics	18,532	11.2%
Quest	4,121	2.5%
Altius	3,696	2.2%
Fulgent Genetics	2,859	1.7%
PHL/Bedford	2,857	1.7%
SCAN/Bedford	1,004	0.6%
Aegis	943	0.6%
Curative Labs	649	0.4%
KP WA Research Inst	281	0.2%
USAFSAM	275	0.2%
CDC	211	0.1%
Providence Swedish	173	0.1%
Helix	151	0.1%
Lauring Lab	118	0.1%
Atlas Genomics	89	0.1%
Boise VA	67	0%
OHSU	61	0%
SFS/Bedford	53	0%
IDBOL	40	0%
Gravity Diagnostics	36	0%
ASU	33	0%
NA	18	0%
OSPHL	15	0%
USAMRIID	9	0%
Infinity Biologix	7	0%
Grubaugh Lab	2	0%
Montana Public Health Lab	2	0%
Flow Diagnostics	1	0%
Grittman Medical Center	1	0%
NW GENOMICS	1	0%
Naval Health Research Center	1	0%
Providence_Swedish	1	0%
SCAB/Bedford	1	0%
SFS/Bedford	1	0%
The Jackson Laboratory	1	0%

source notebook.

Bibliography

[1] H. N. Oltean *et al.*, “Sentinel Surveillance System Implementation and Evaluation for SARS-CoV-2 Genomic Data, Washington, USA, 2020–2021 - Volume 29, Number 2—February 2023 - Emerging Infectious Diseases Journal - CDC,” doi: 10.3201/eid2902.221482.

[2] H. N. Oltean *et al.*, “Changing Genomic Epidemiology of COVID-19 in Long-Term Care Facilities during the 2020–2022 Pandemic, Washington State,” *BMC Public Health*, vol. 24, p. 182–183, Jan. 2024, doi: 10.1186/s12889-023-17461-2.

[3] C. Wagner *et al.*, “Positive Selection Underlies Repeated Knockout of ORF8 in SARS-CoV-2 Evolution.” Accessed: Mar. 18, 2024. [Online]. Available: https://www.medrxiv.org/content/10.1101/2023.09.21.23295927v1