Modernizing Mumps workflows

- Bedford Lab Meeting -

Jennifer Chang, Ph.D.

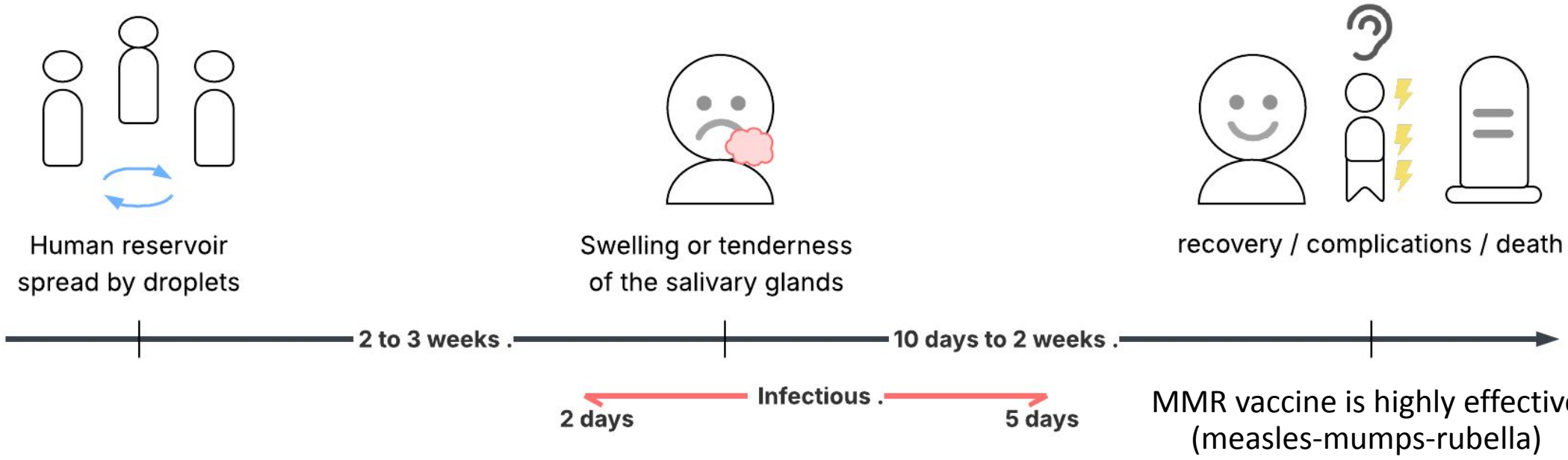
Bioinformatic Analyst III

Fred Hutchinson Cancer Center

Outline

- About Mumps
- Why align to the Pathogen Repo Guide
- Walk-through roadmap and key decisions points
  - Ingest workflow
  - Phylogenetic workflow
  - Nextclade workflow
- Next steps

About Mumps

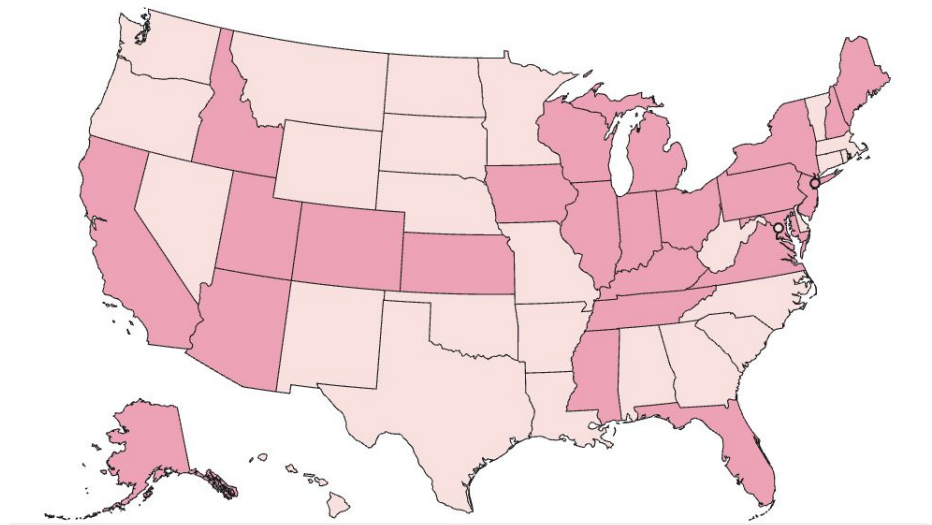


MMR vaccine is highly effective
(measles-mumps-rubella)

https://www.cdc.gov/mumps/downloads/mumps-clinical-diagnosis-fact-sheet-508.pdf

Mumps cases still occur in the USA

Reported U.S. mumps cases by jurisdiction, 2025*

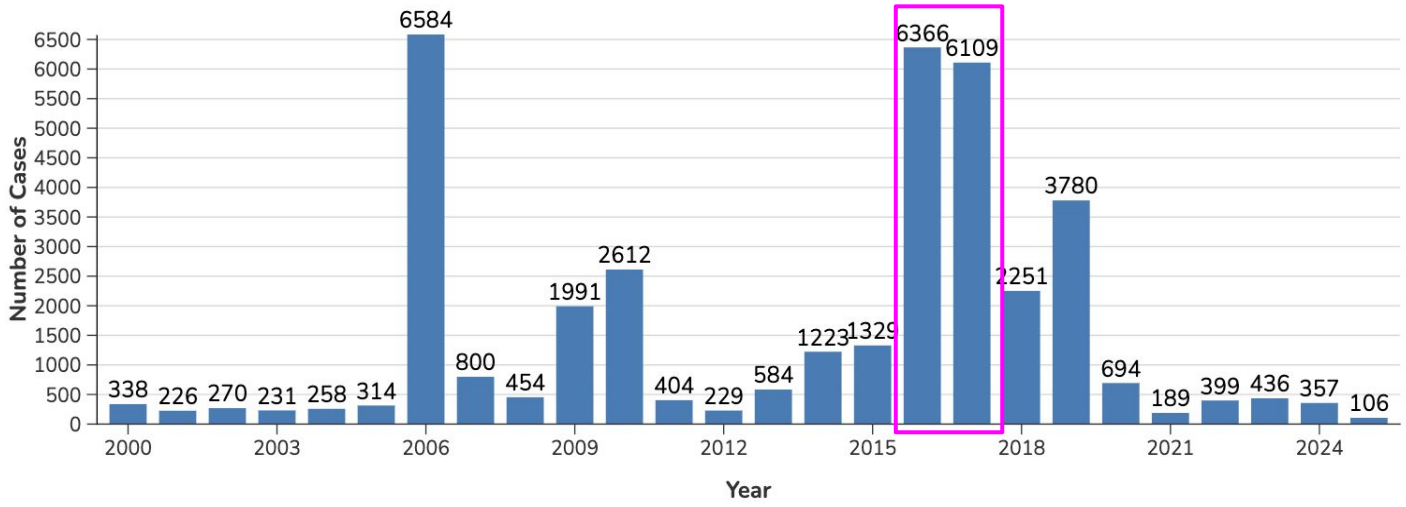


Legend

No reported cases

Reported mumps cases

Reported U.S. mumps cases by year (2000–2025)



Data Table

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
2612	404	229	584	1223	1329	6366	6109	2251	3780	694	189	399	436	357	106

*2025 Jurisdictions refer to any of the 50 states, New York City, and the District of Columbia.

**2025 map represents cases reported to CDC as of April 24, 2025; 2022–2025 case counts are preliminary and subject to change.

https://www.cdc.gov/mumps/outbreaks/index.html

Mumps cases still occur in the USA

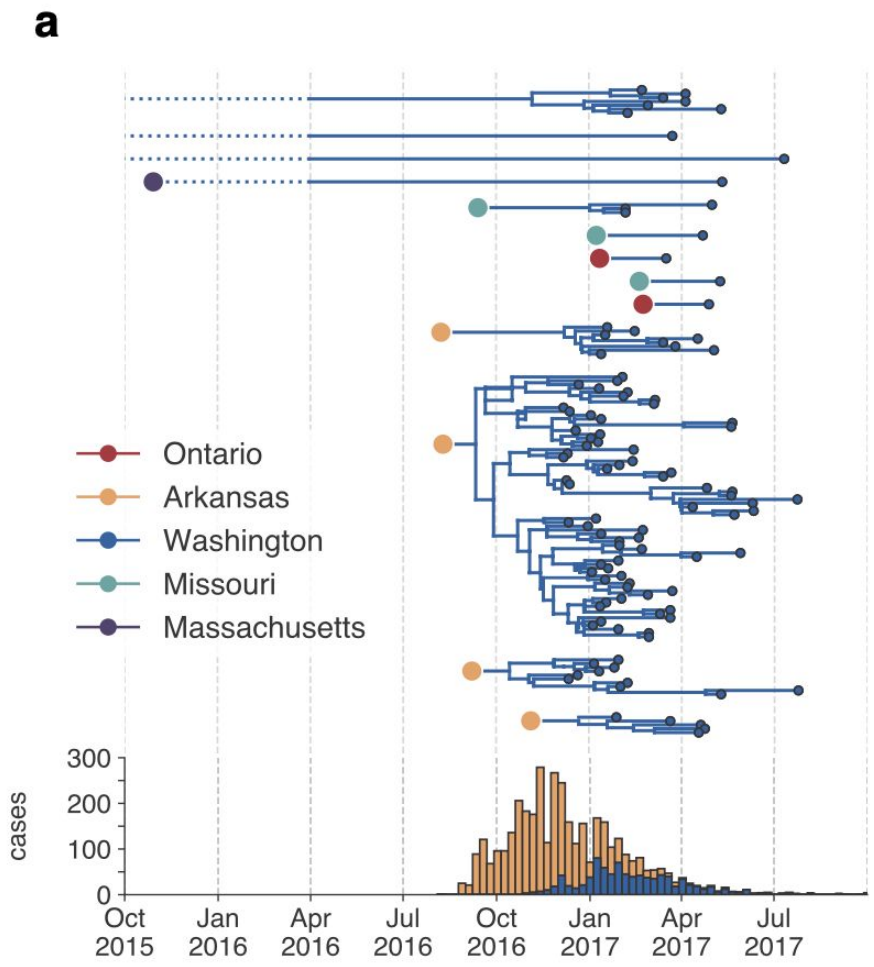
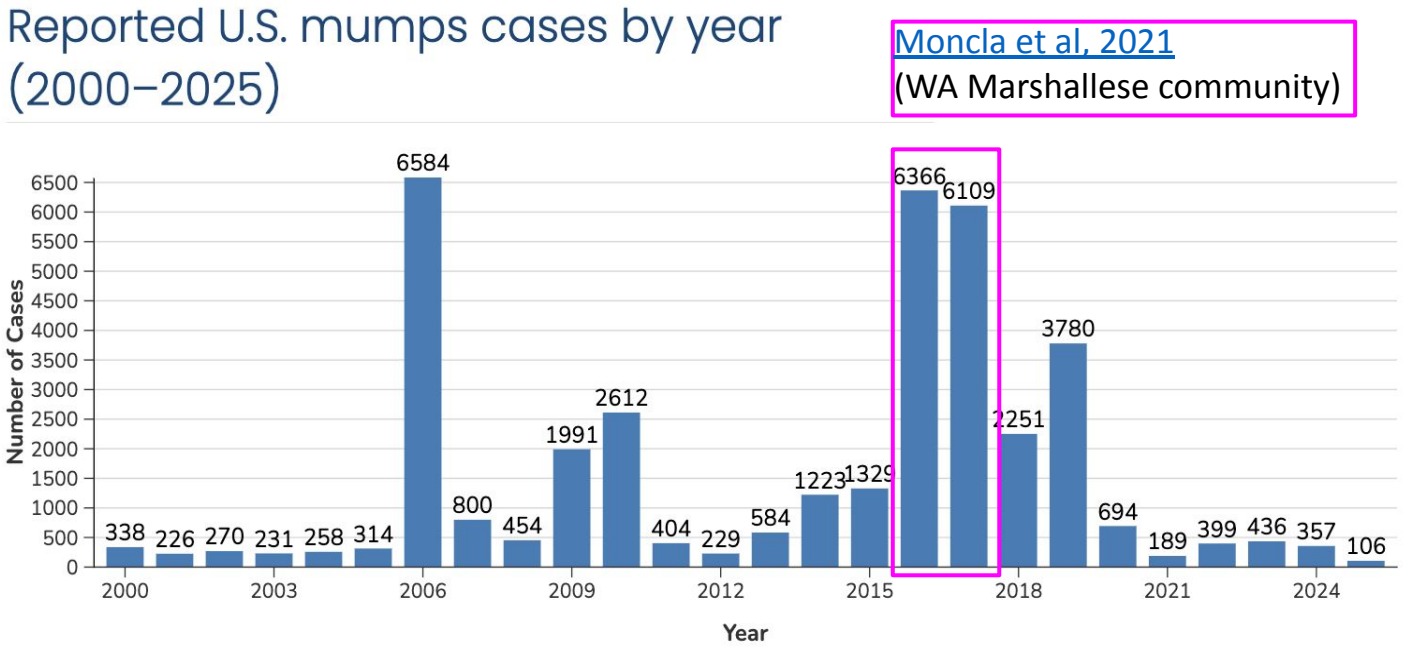


Figure 3: The mumps outbreak in Washington was seeded by approximately 13 introductions.



[Moncla et al, 2021](#)
(WA Marshallese community)

Data Table

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
2612	404	229	584	1223	1329	6366	6109	2251	3780	694	189	399	436	357	106

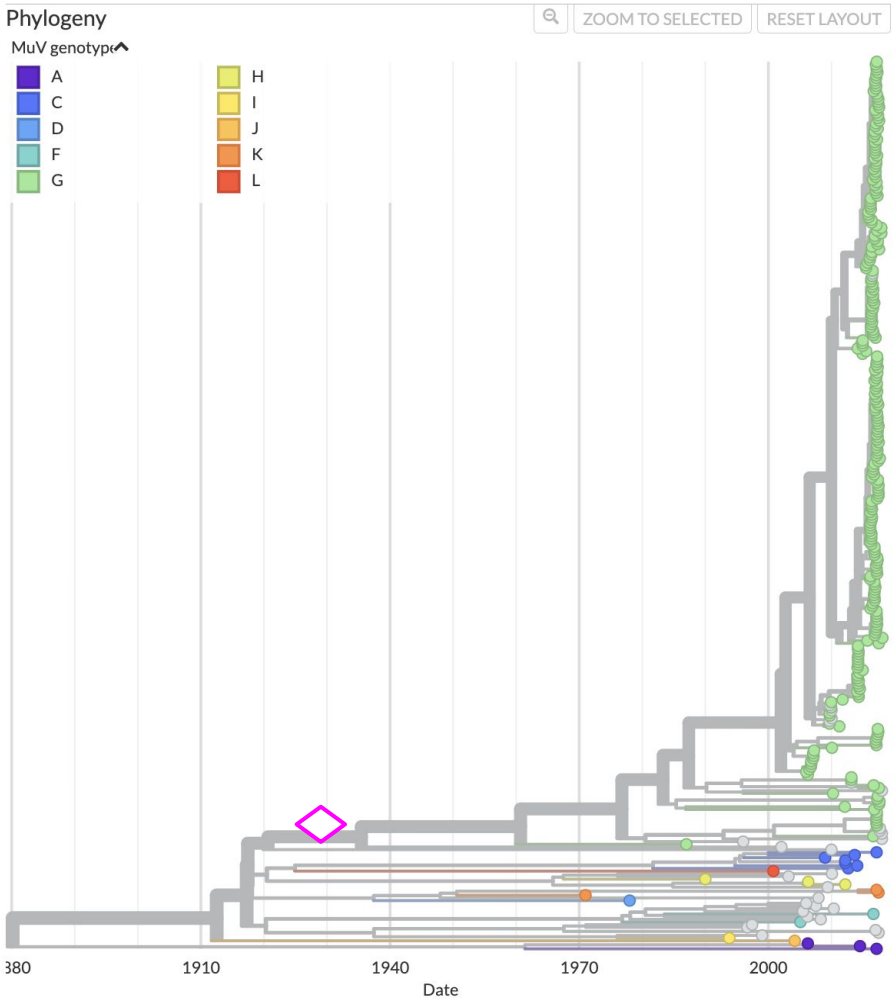
*2025 Jurisdictions refer to any of the 50 states, New York City, and the District of Columbia.

**2025 map represents cases reported to CDC as of April 24, 2025; 2022–2025 case counts are preliminary and subject to change.

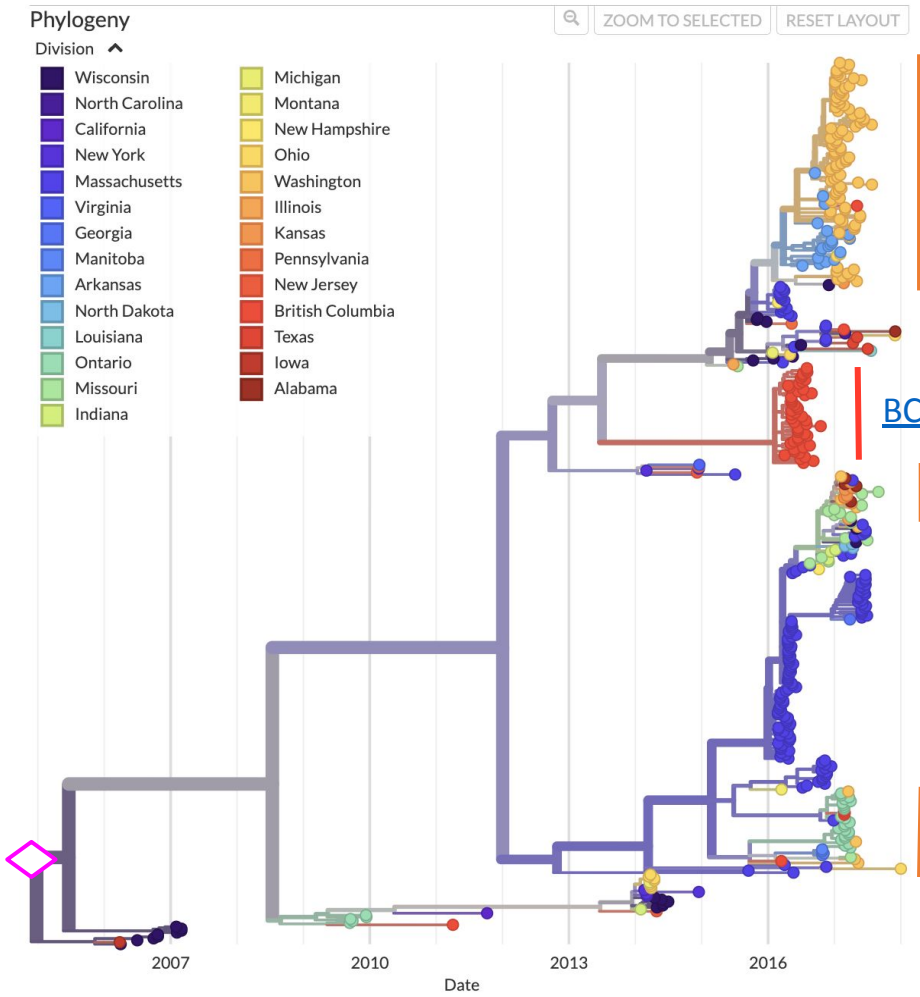
https://www.cdc.gov/mumps/outbreaks/index.html

Existing Mumps Nextstrain site

global tree



na (north-america) tree



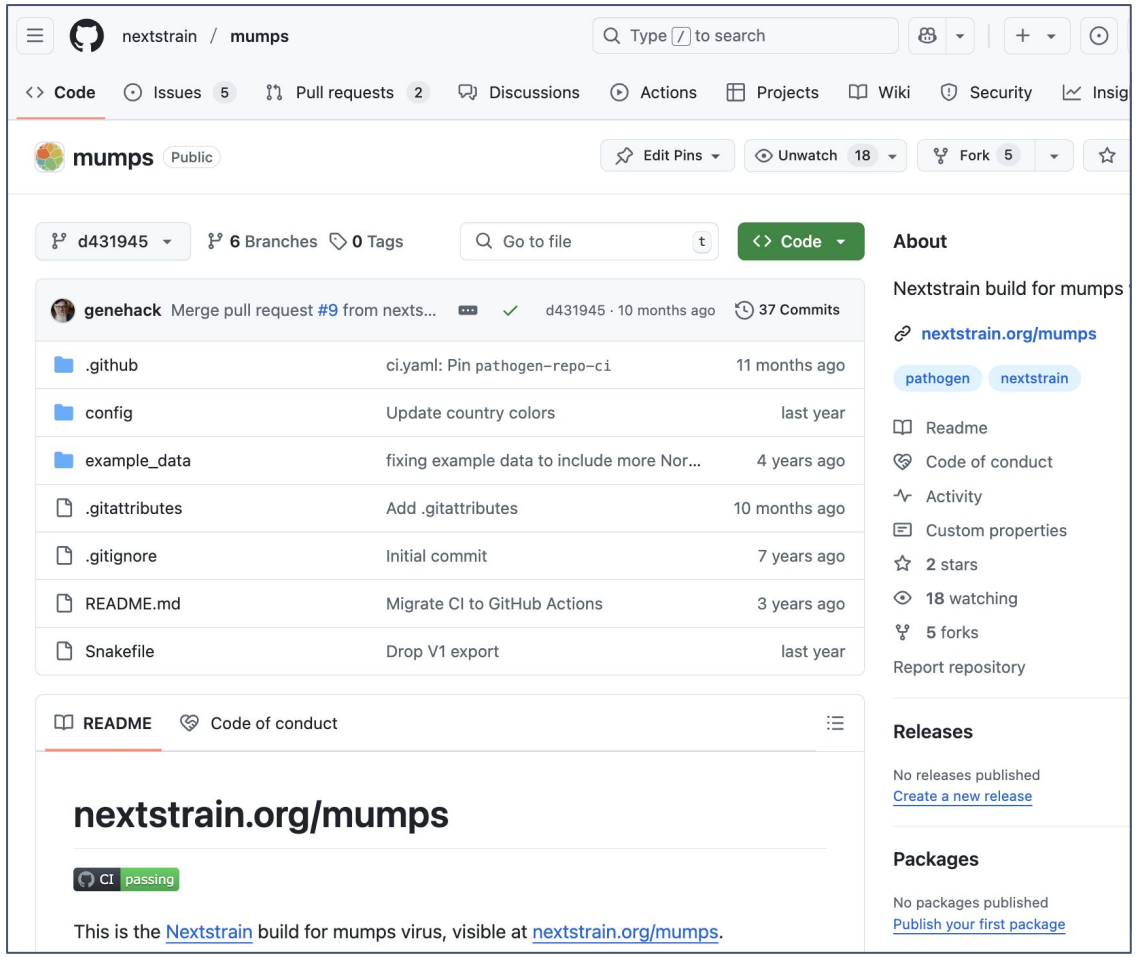
Moncla et al, 2021
WA <-> AR dynamics

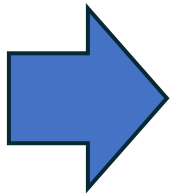
BC CDC sequences

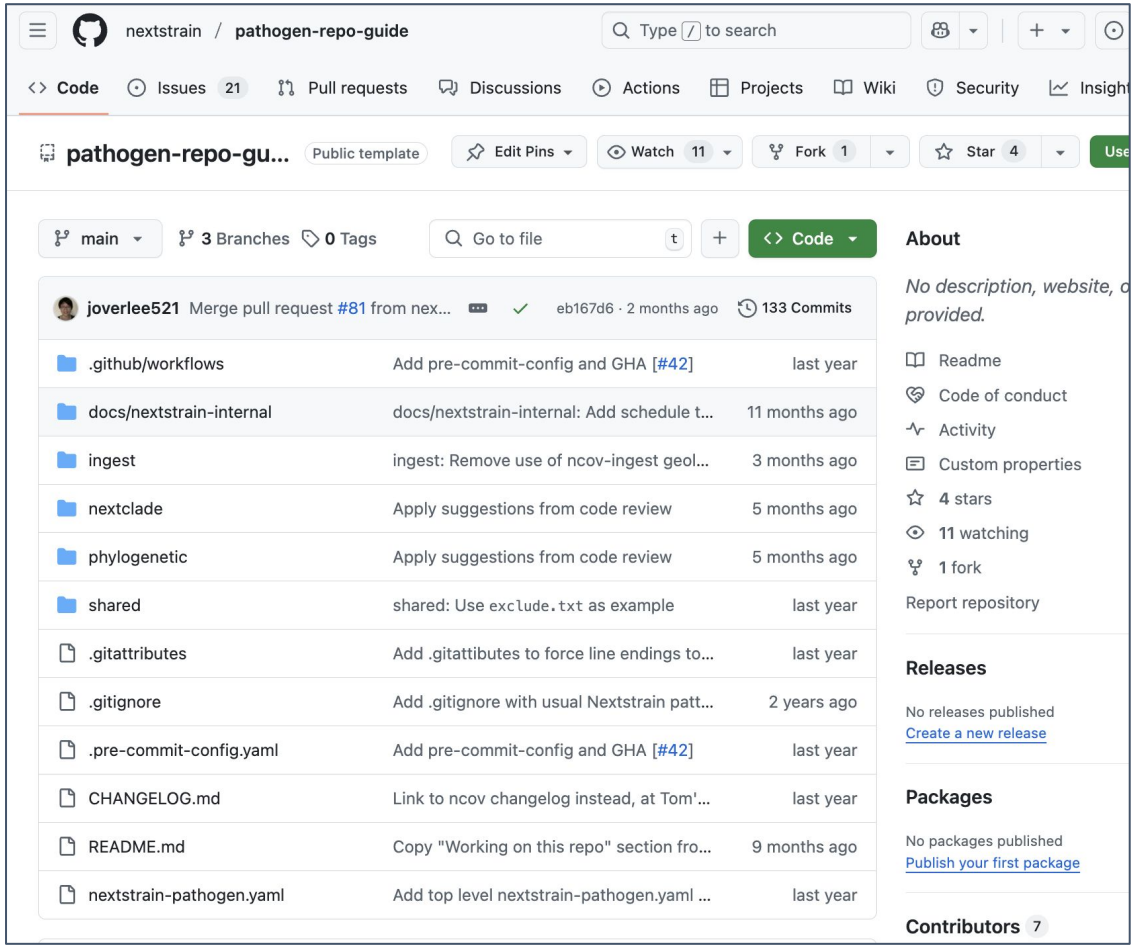
WA <-> other state

WA <-> Canada

Update Mumps Nextstrain workflow







- Refactor code into subdirectories (workflows)
- Connect GitHub Action automation

Pathogen-repo-guide

Why align to a pathogen repo guide?

Nextstrain GitHub Practices

Consistency and Reproducibility

Nextstrain's focus on pathogen genomics requires a high degree of consistency in data analysis workflows. By implementing best practices, particularly in Snakemake workflows, we ensure:

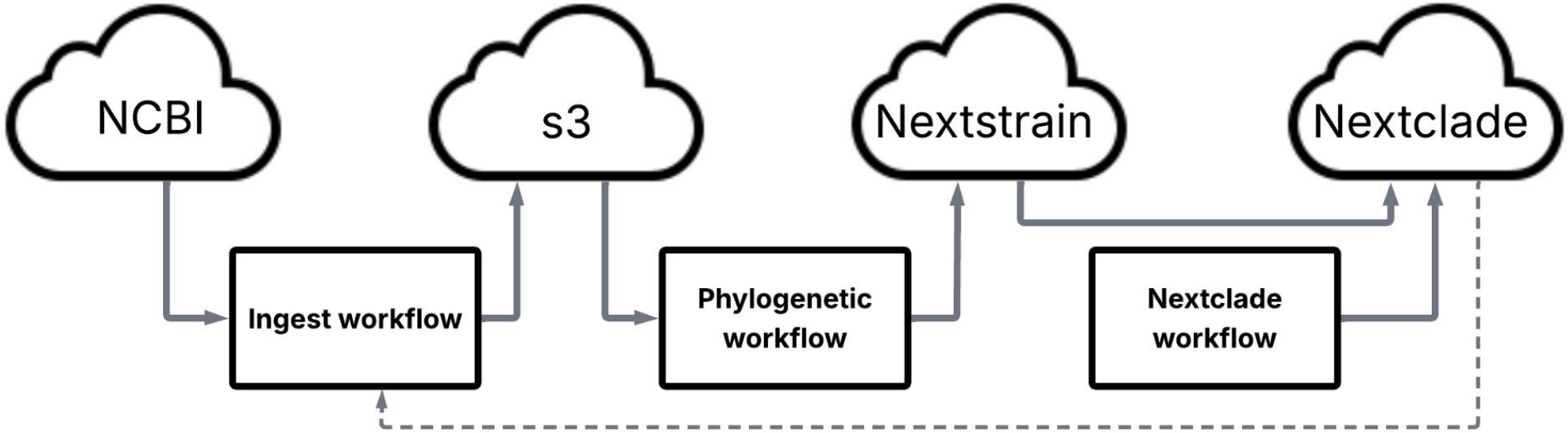
- Reproducible analysis across different datasets and pathogens
- Uniform coding standards that facilitate easier code review and maintenance
- Consistent file structures and naming conventions

Continuous Improvement and Adaptability

The field of pathogen genomics is rapidly evolving, and Nextstrain's best practice aims to collaboratively adapt and maintain high quality by:

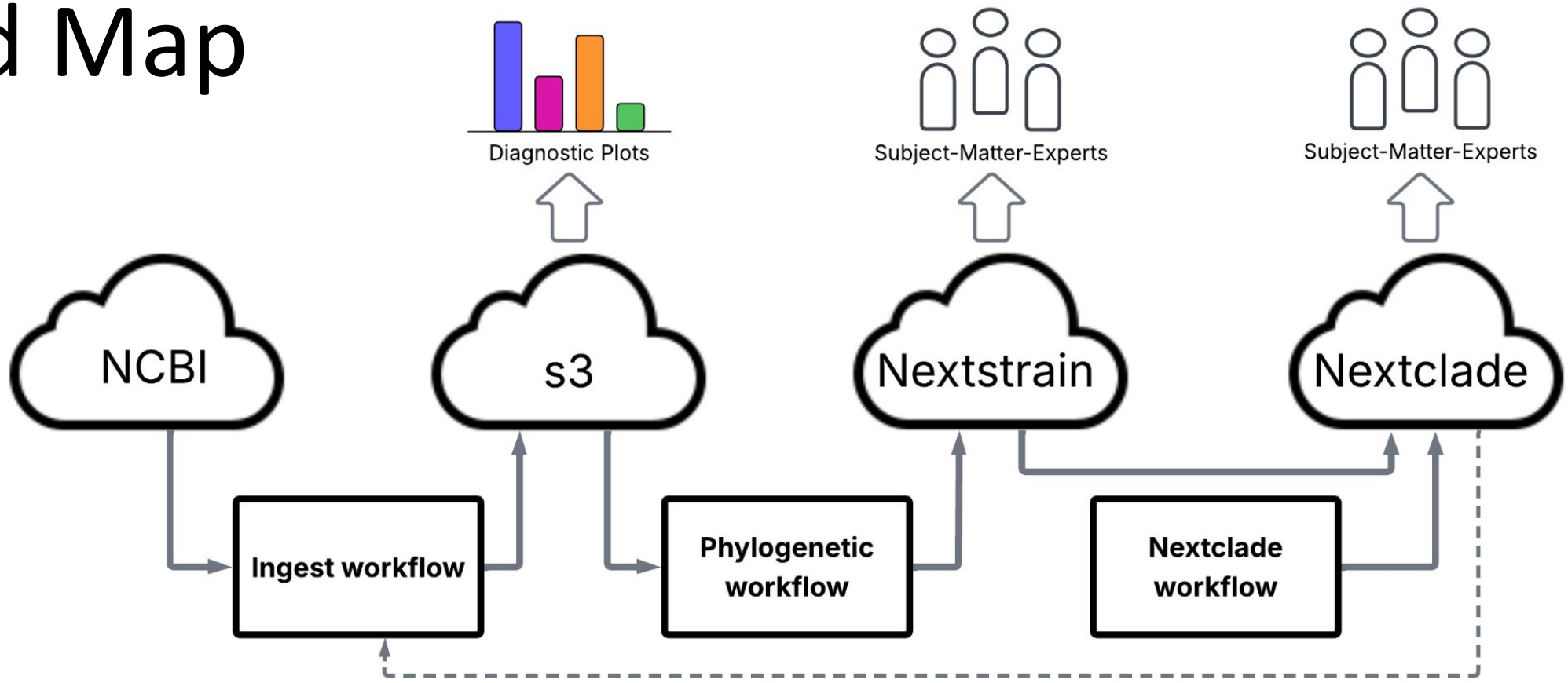
- Regular review and updates to best practices to incorporate new tools and methodologies
- Some flexibility to adapt workflows for different pathogens and analysis requirements

Road Map



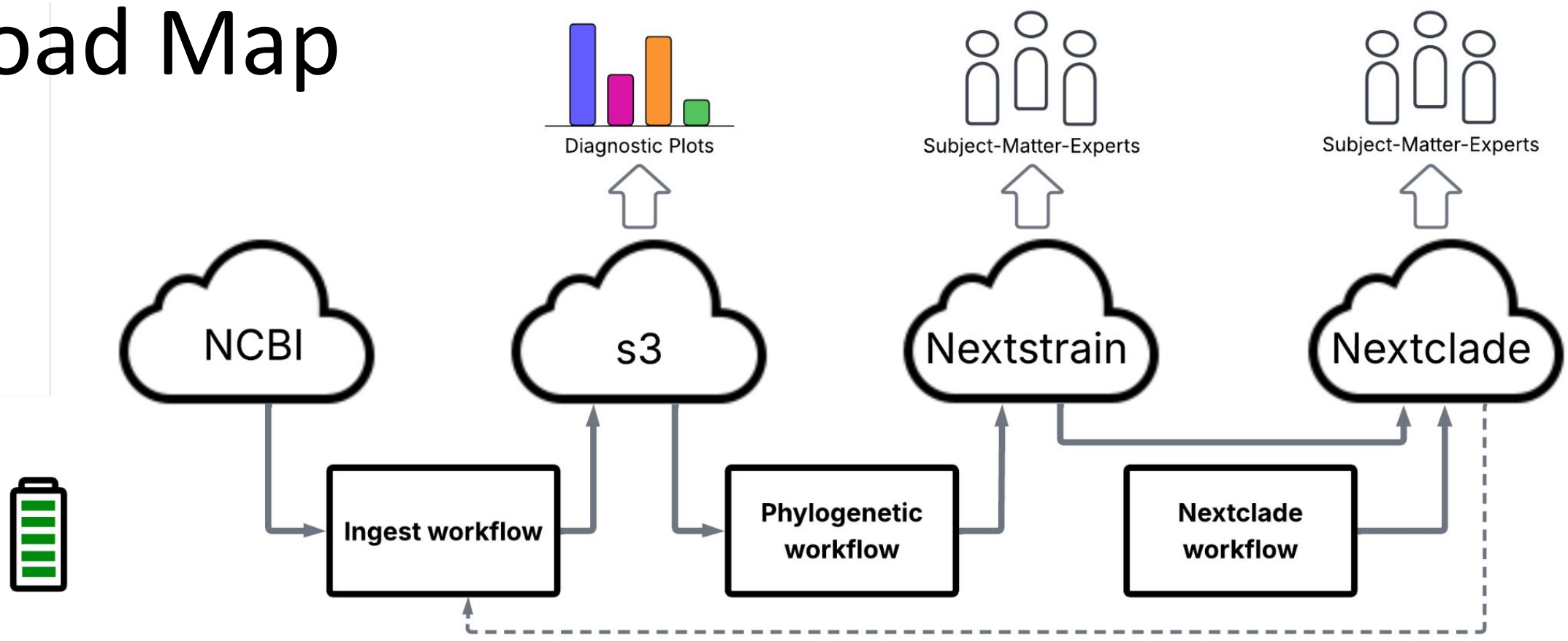
Overview of implementing the pathogen-repo-guide

Road Map



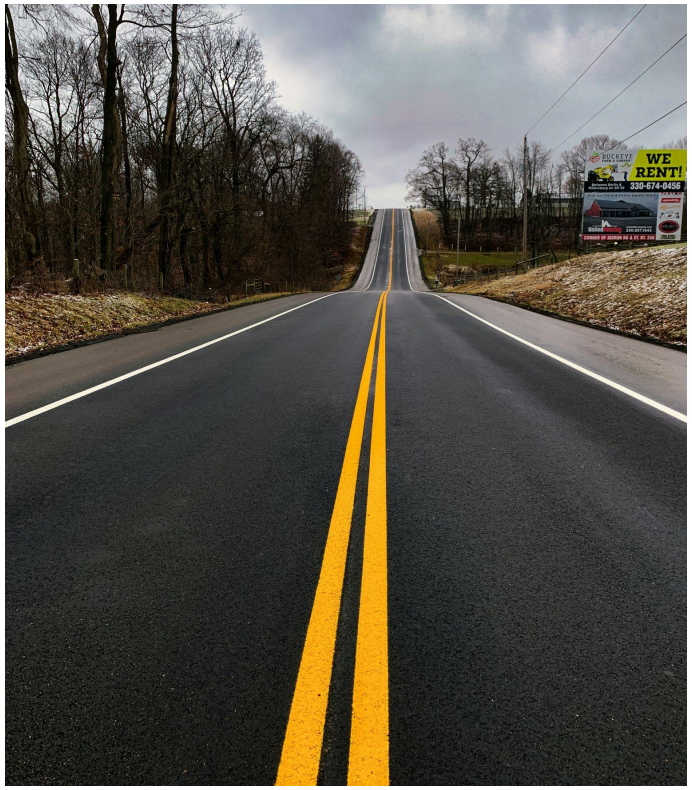
Add some checkpoints to collect feedback

Road Map



Pace yourself so you don't burn out

GitHub PRs - levels of effort for review



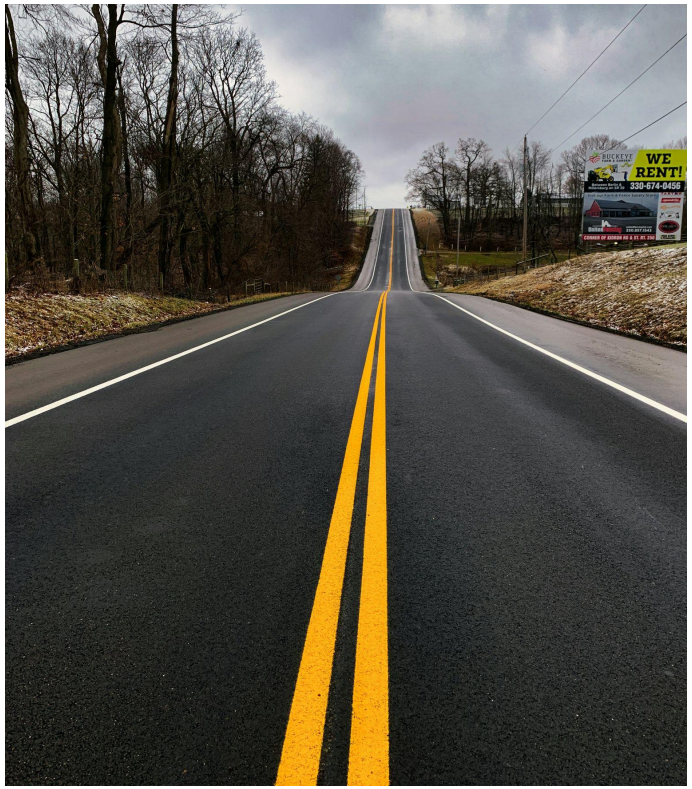
pathogen-repo-guide

in another repo

add a new standard

Photos by Leslie Saunders, Jens Lelie, and Joshua Earle on Unsplash

GitHub PRs - levels of effort for review



pathogen-repo-guide

(e.g. ingest workflow)

in another repo

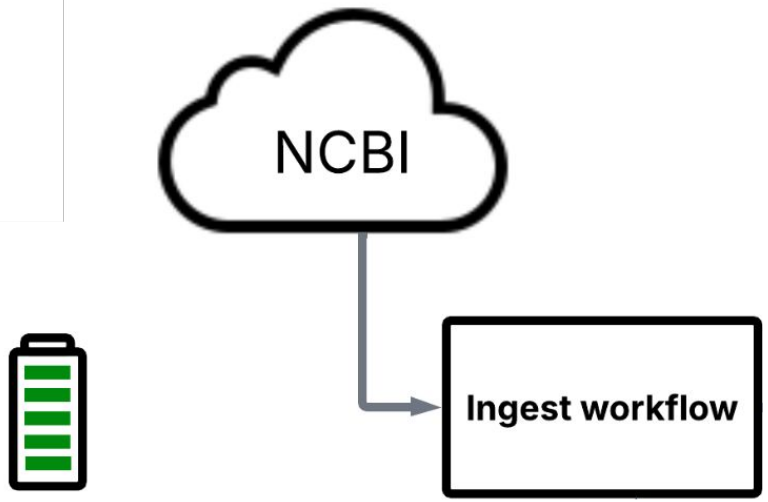
(e.g. copy an existing script)

add a new standard

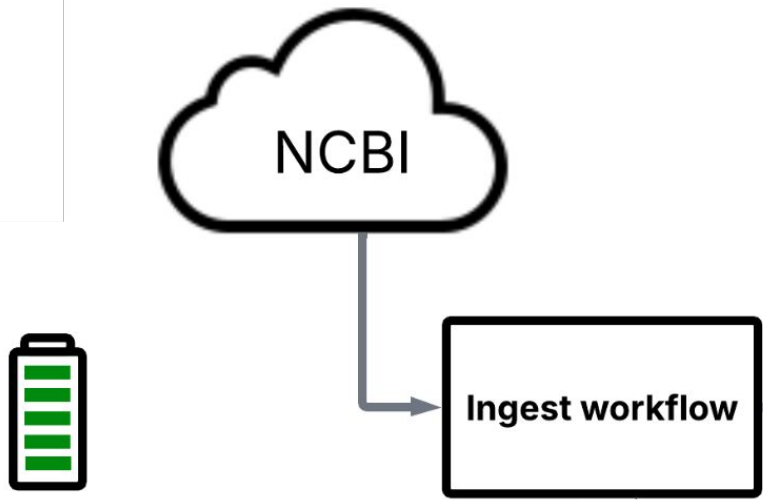
(e.g. add a new tool/feature)

Photos by Leslie Saunders, Jens Lelie, and Joshua Earle on Unsplash

Road Map



Road Map



maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602

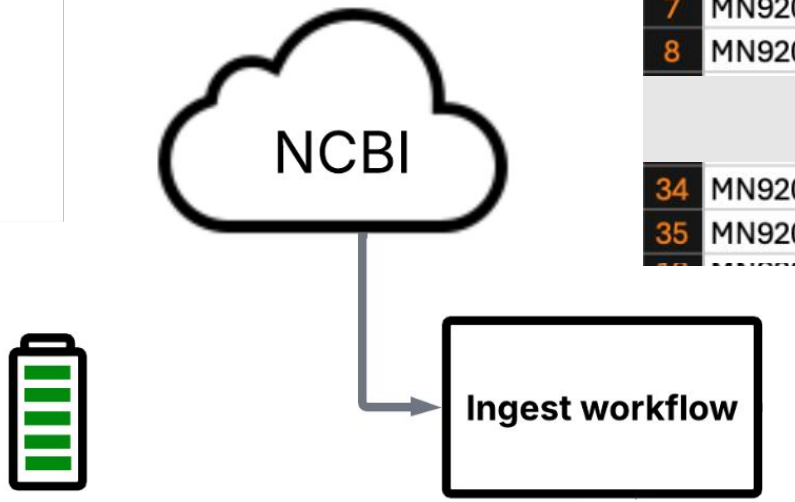
Guide for ingesting public data from NCBI. The workflow outputs a pair of curated metadata and sequences that can be used as input for the phylogenetic workflow.

1. select virus
2. glance at all NCBI fields pulled
3. run ingest and curation
4. push curated files to s3

https://docs.nextstrain.org/en/latest/tutorials/creating-a-pathogen-repo/creating-an-ingest-workflow.html

Road Map

	Accession	Isolate Lineage	Virus Name	Virus Taxon	Isolate Collection date	Length
2	MN920136.1	MMR17-1436	Mumps virus genotype G	1384672	2017-04-20	316
3	MN920137.1	MMR17-1437	Mumps virus genotype G	1384672	2017-04-13	316
4	MN920138.1	MMR17-1438	Mumps virus genotype G	1384672	2017-04-18	316
5	MN920139.1	MMR17-1439	Mumps virus genotype G	1384672	2017-04-15	316
6	MN920140.1	MMR17-1440	Mumps virus genotype G	1384672	2017-04-13	316
7	MN920141.1	MMR17-1441	Mumps virus genotype G	1384672	2017-04-18	316
8	MN920142.1	MMR17-1442	Mumps virus genotype G	1384672	2017-04-18	316
			....			
34	MN920168.1	MMR19-0001	Mumps virus genotype C	1428458	2018-12-24	316
35	MN920169.1	MMR19-0036	Mumps virus genotype C	1428458	2018-04-17	316

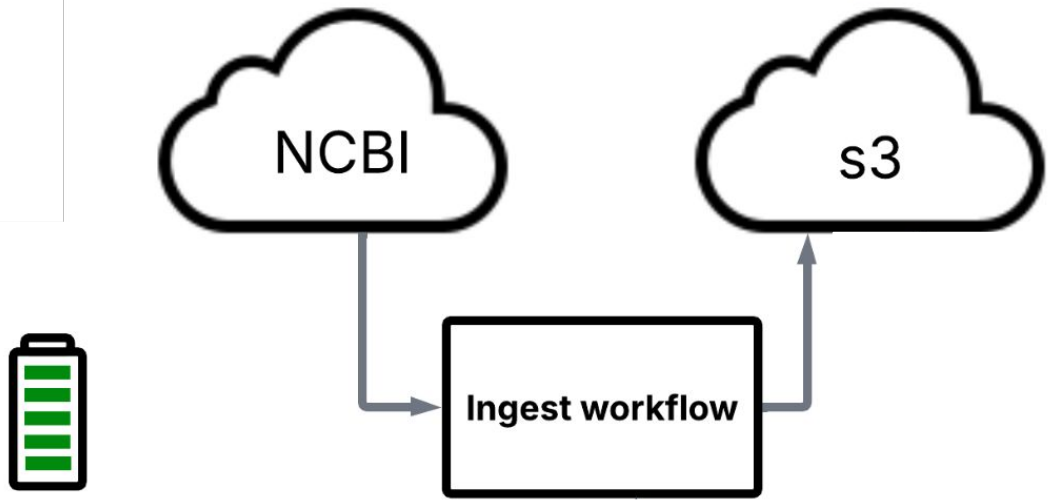


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

1. select virus
2. glance at all NCBI fields pulled
3. run ingest and curation
4. push curated files to s3

https://docs.nextstrain.org/en/latest/tutorials/creating-a-pathogen-repo/creating-an-ingest-workflow.html

Road Map

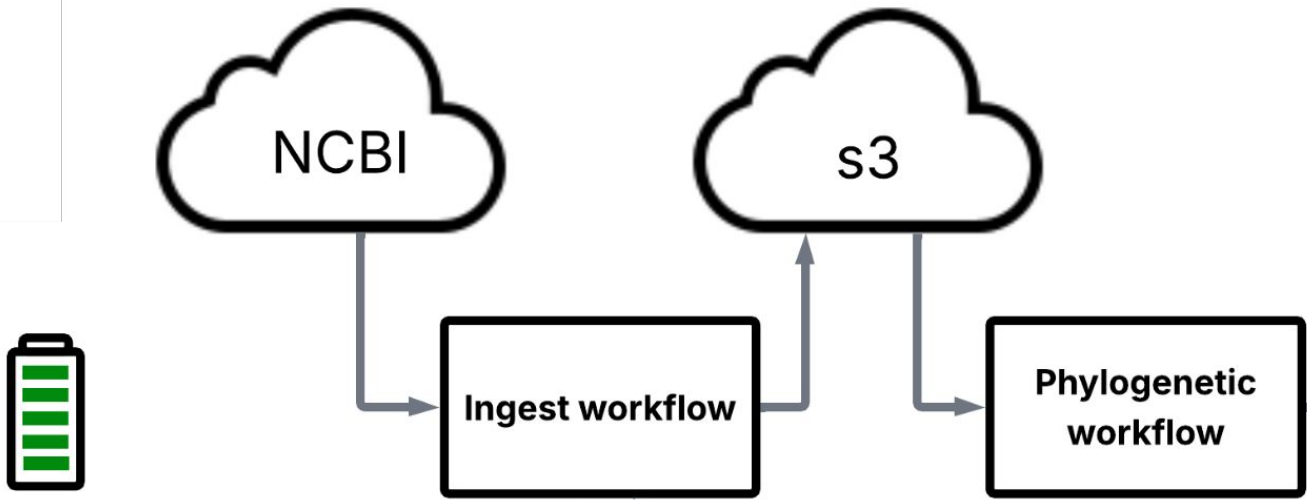


- Open an Ingest PR
- Push metadata and sequences to s3

maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

https://docs.nextstrain.org/en/latest/tutorials/creating-a-pathogen-repo/creating-an-ingest-workflow.html

Road Map



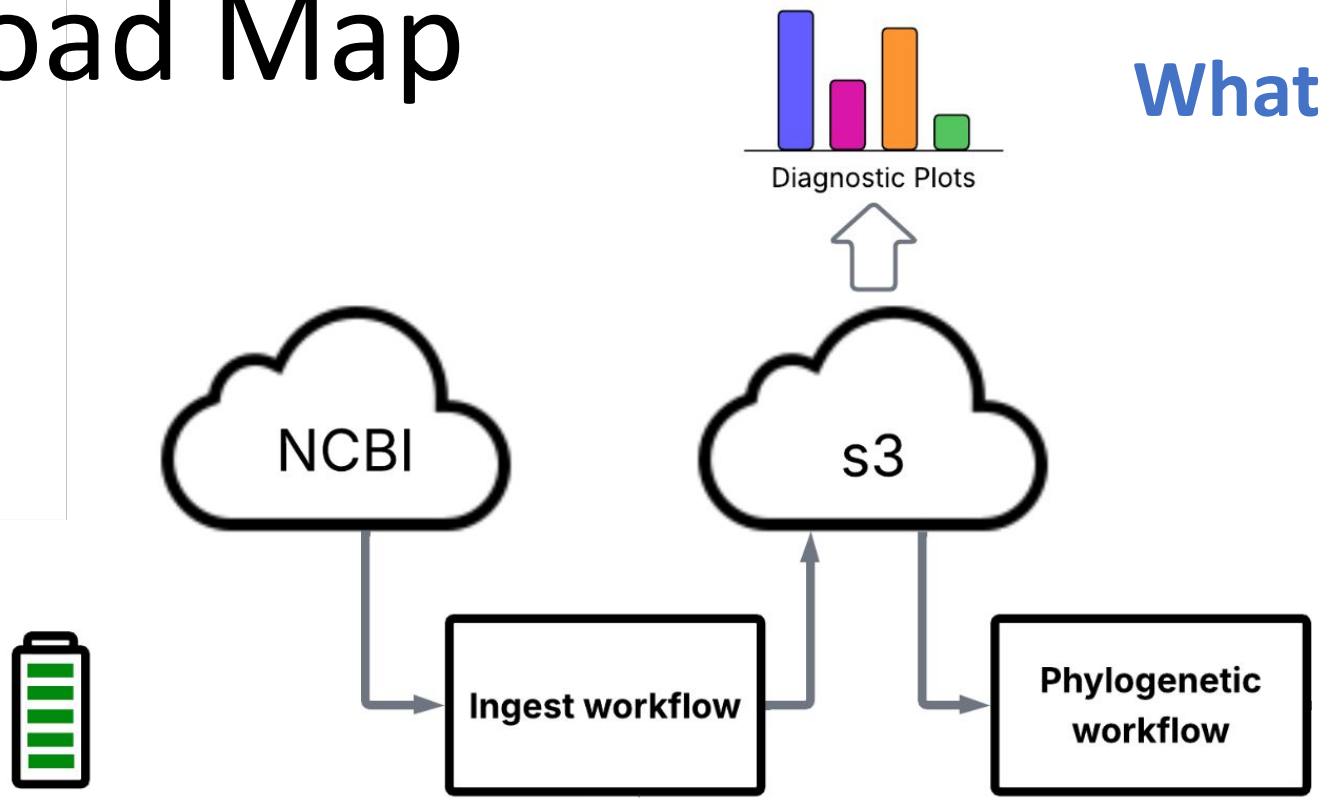
- Open an Ingest PR
- Push metadata and sequences to s3
- Start phylogenetic workflow on a new branch

maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

global tree
north-america tree

https://docs.nextstrain.org/en/latest/tutorials/creating-a-phylogenetic-workflow.html

Road Map



What diagnostic plots?

- Open an Ingest PR
- Push metadata and sequences to s3
- Start phylogenetic workflow on a new branch
- Create diagnostic plots from the metadata.tsv

maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

global tree
north-america tree

https://docs.nextstrain.org/en/latest/tutorials/creating-a-phylogenetic-workflow.html

Explore the Mumps data - Diagnostic plots

Pull metadata.tsv from s3 bucket

```
pathogen <- "mumps"

# Download and decompress file
data_file=paste0(pathogen, "/metadata.tsv")
if (!file.exists(paste0(pathogen, "/metadata.tsv"))) {
  system(paste("mkdir -p ", pathogen, sep=""))
  file_name <- "metadata.tsv"
  zst_file <- paste0(file_name, ".zst")
  url <- paste0("https://data.nextstrain.org/files/workflows/", pathogen,
                "/metadata.tsv.zst" )
  system(paste0("curl -L ", url, " -o ", pathogen,"/", zst_file))
  system(paste0("zstd -d ", pathogen, "/", zst_file," -o ", pathogen, "/",
                file_name))
}

data <- readr::read_delim(paste0(pathogen, "/metadata.tsv"), delim="\t")
```

Generalize and import diagnostic plots

```
source("shared_functions.R")

diagnostic_length_plot(
  data = data,
  approx_length = 15384, # nt length of Mumps genome
  binwidth = 200,
  percentage_y = 9000    # y-coordinate for min-length labels
)
```

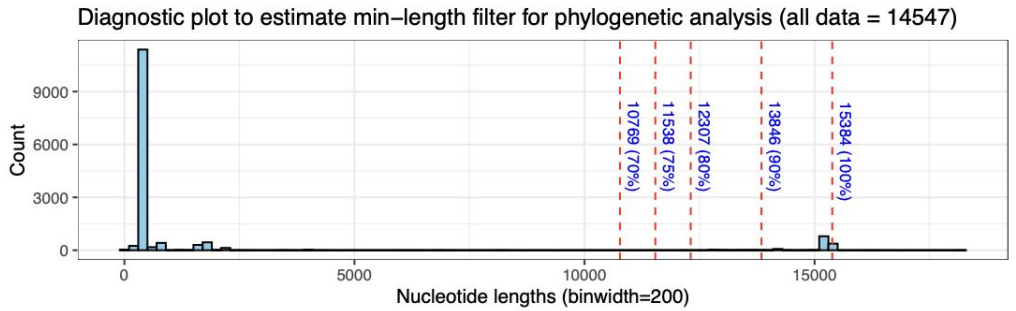
GitHub: j23414/generated-reports/reports/mumps.pdf

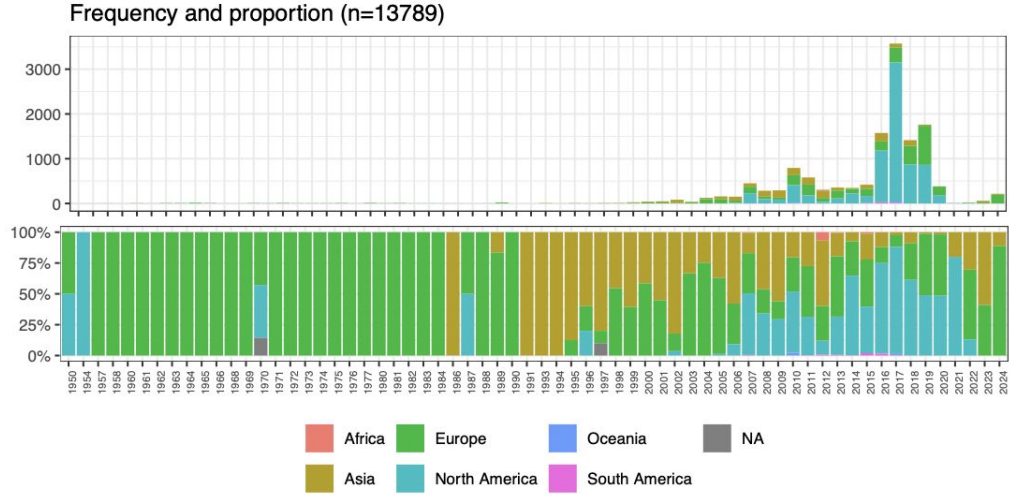
Mumps notes last update: April 12, 2025

Background

Exploratory Graphics

NCBI GenBank records





Explore the Mumps data - tables

Table 1: Top 25 most frequent sequence submitters with their region and countries

authors	n	regions	countries
hiebert et al.	3174	North America	Canada
mcnall et al.	1502	North America	USA
castellanos et al.	1061	Europe	Spain
wharton et al.	945	North America, Oceania	USA, Micronesia, Marshall Islands
cui et al.	514	Europe, Asia, North America	United Kingdom, China, Dominican Republic
bodewes et al.	488	Europe	Netherlands
gouma et al.	485	Europe, North America	Netherlands, USA
rota et al.	425	North America, NA	USA, NA
kidokoro et al.	392	Asia, NA	Japan, Mongolia, NA
frost et al.	319	Europe, North America	United Kingdom, Canada
bryant et al.	315	North America	USA
kim et al.	285	North America, Asia, NA	USA, South Korea, NA, Korea
hickman et al.	254	North America	USA
peran-ramos et al.	253	Europe	Spain
aoki et al.	249	Asia	Japan
gavilan et al.	227	Europe	Spain
rubalskaia et al.	218	Europe	Russia
ma et al.	178	Asia	China
catellanos et al.	172	Europe	Spain
byrne et al.	169	North America	USA
moncla et al.	166	North America	USA
anton et al.	155	Europe	Spain
shah et al.	154	Europe, North America	Netherlands, Canada
rivailler et al.	144	Oceania, North America	Guam, USA
jin et al.	142	NA, Europe, Asia, North America	NA, Sweden, Japan, United Kingdom, USA, Germany, Malaysia,

Earliest and latest submitters

Table 2: Top 20 earliest records

date_adjusted	accession	strain	country	authors
1950-01-01	JQ946042	MuVi/Taylor.GBR/0.50s	United Kingdom	Jin et al.
1950-01-01	KF876715	MuVi/Kilham.USA/0.50[A]	USA	Jin et al.
1954-01-01	KX136900	MuVi/Albany.USA/0.54[A]	USA	Gouma et al.
1954-01-01	KX136946	MuVi/Albany.USA/0.54[A]	USA	Gouma et al.
1954-01-01	KX136993	MuVi/Albany.USA/0.54[A]	USA	Gouma et al.
1957-01-01	KX136901	MuVi/NLD/0.57[L]	Netherlands	Gouma et al.
1957-01-01	KX136947	MuVi/NLD/0.57[L]	Netherlands	Gouma et al.
1957-01-01	KX136994	MuVi/NLD/0.57[L]	Netherlands	Gouma et al.
1958-01-01	JQ034458	Enders-58-2	United Kingdom	Cui et al.
1958-01-01	JQ034507	ENDERS-58-2	United Kingdom	Cui et al.
1960-01-01	KX136902	MuVi/NLD/0.60/1[L]	Netherlands	Gouma et al.
1960-01-01	KX136903	MuVi/NLD/0.60/2[L]	Netherlands	Gouma et al.
1960-01-01	KX136948	MuVi/NLD/0.60/1[L]	Netherlands	Gouma et al.
1960-01-01	KX136949	MuVi/NLD/0.60/2[L]	Netherlands	Gouma et al.
1960-01-01	KX136995	MuVi/NLD/0.60/1[L]	Netherlands	Gouma et al.
1960-01-01	KX136996	MuVi/NLD/0.60/2[L]	Netherlands	Gouma et al.
1961-01-01	KX136904	MuVi/NLD/0.61[D]	Netherlands	Gouma et al.
1961-01-01	KX136950	MuVi/NLD/0.61[D]	Netherlands	Gouma et al.
1961-01-01	KX136997	MuVi/NLD/0.61[D]	Netherlands	Gouma et al.
1962-01-01	KX136905	MuVi/NLD/0.62/1	Netherlands	Gouma et al.

Table 3: Top 20 latest records

date_adjusted	accession	strain	country	authors
2024-03-22	PQ311690	MuVs/KIPMR/Chennai.IND/4.24/3[C]	India	Kaveri et al.
2024-03-23	PQ311691	MuVs/KIPMR/Chennai.IND/4.24/4[C]	India	Kaveri et al.
2024-03-23	PQ311692	MuVs/KIPMR/Chennai.IND/4.24/5[C]	India	Kaveri et al.
2024-03-25	PQ311693	MuVs/KIPMR/Chennai.IND/5.24[C]	India	Kaveri et al.
2024-05-14	PQ311694	MuVs/KIPMR/Chennai.IND/3.24[C]	India	Kaveri et al.
2024-05-14	PQ311695	MuVs/KIPMR/Chennai.IND/3.24/2[C]	India	Kaveri et al.
2024-05-14	PQ311696	MuVs/KIPMR/Chennai.IND/3.24/3[C]	India	Kaveri et al.
2024-05-14	PQ311697	MuVs/KIPMR/Chennai.IND/3.24/4[C]	India	Kaveri et al.
2024-05-17	PQ451425	MuVs/Odisha.INDIA/20.24[C]	India	Mamidi et al.
2024-05-17	PQ451426	MuVs/Odisha.INDIA/20.24/1[C]	India	Mamidi et al.
2024-05-17	PQ451427	MuVs/Odisha.INDIA/20.24/2[C]	India	Mamidi et al.
2024-05-17	PV072739	MuVs/Odisha.INDIA/20.24/8[C]	India	Mishra et al.
2024-05-27	PV072740	MuVs/Odisha.INDIA/20.24/9[C]	India	Mishra et al.
2024-06-07	PQ001008	MuVs/Makhchkala.RUS/23.24[C]	Russia	Zamotaeva et al.
2024-06-07	PQ001009	MuVs/Makhchkala.RUS/23.24[G]	Russia	Zamotaeva et al.
2024-06-19	PV072734	MuVs/Odisha.INDIA/20.24/3[C]	India	Mishra et al.
2024-07-05	PV072735	MuVs/Odisha.INDIA/20.24/4[C]	India	Mishra et al.
2024-07-31	PV072738	MuVs/Odisha.INDIA/20.24/7[C]	India	Mishra et al.
2024-08-09	PV072737	MuVs/Odisha.INDIA/20.24/6[C]	India	Mishra et al.
2024-08-29	PV072736	MuVs/Odisha.INDIA/20.24/5[C]	India	Mishra et al.

Literature review - collect papers

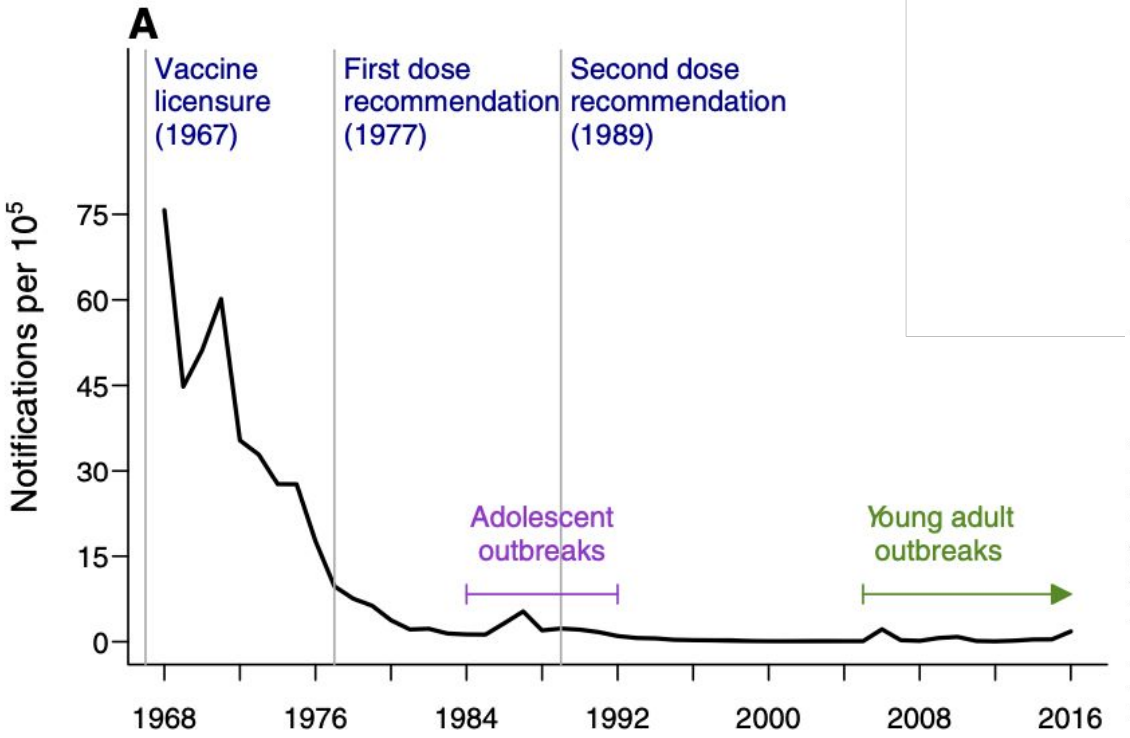
- What is the **earliest study** identifying Mumps as a viral disease?
- When was the Mumps virus first **sequenced**, and which genome region was published?
- When was the first Mumps **vaccine** developed, and how effective was it?
- Have there been major updates to the vaccine? What prompted them?
- What **genotype** system has been used for Mumps virus strains?
- What are some of the **key open questions** in recent Mumps virus research?

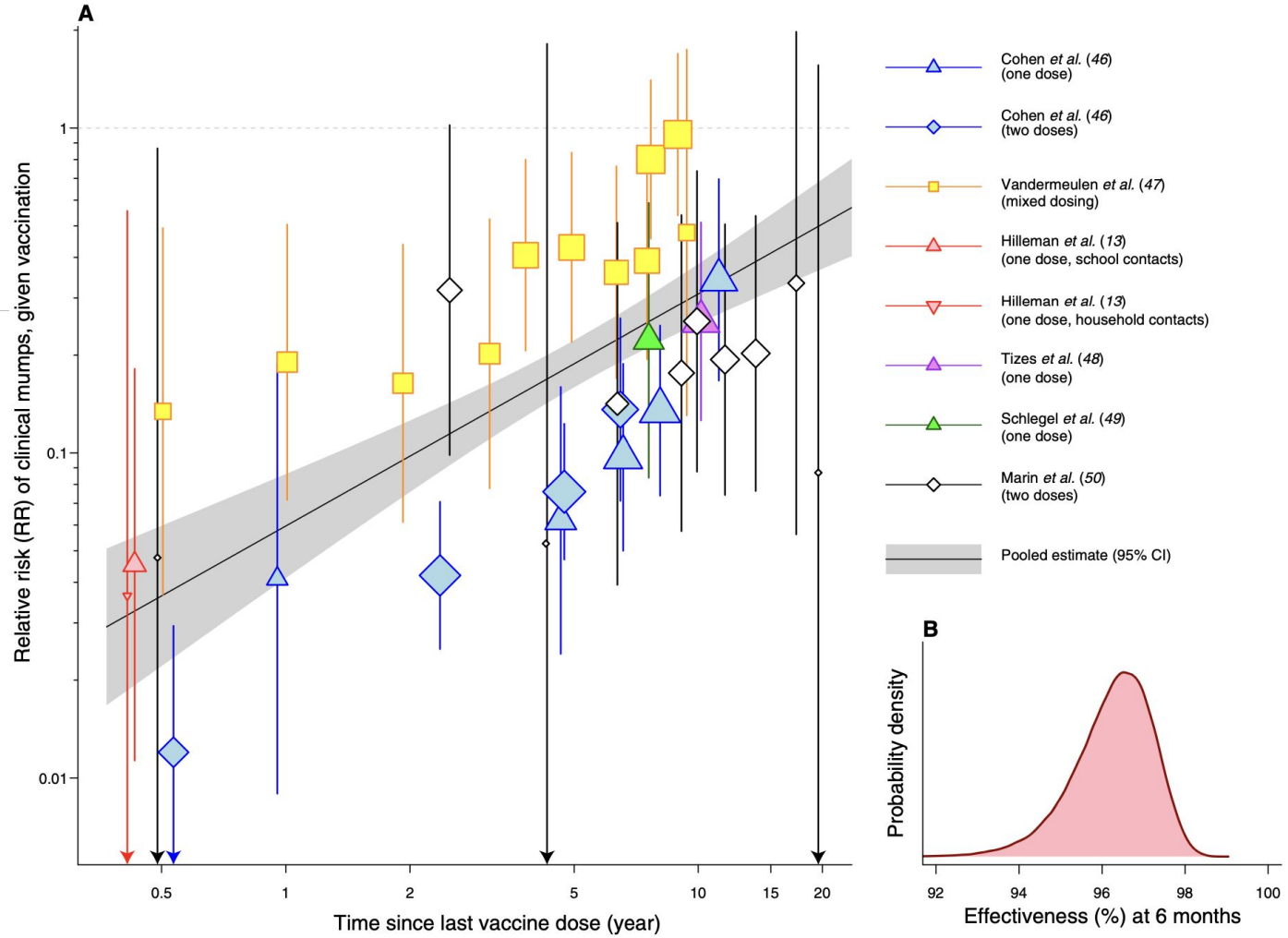
Literature review - collect papers

- What is the **earliest study** identifying Mumps as a viral disease?
  - (Johnson and Goodpasture, 1933) - "An investigation of the etiology of Mumps"
- When was the Mumps virus first **sequenced**, and which genome region was published?
  - 1989 - gene segments (F, HN, **SH**, …) perhaps Takeuchi et al; full genome perhaps Mori
- When was the first Mumps **vaccine** developed, and how effective was it?
  - (Hillman et al, 1967) - "Live Attenuated Mumps Virus Vaccine" Jeryl-Lynn Strain
- Have there been major updates to the vaccine? What prompted them?
  - Yes, side effects or regional genotype (Urabe, Leningrad-3, Rubini, Miyahara)
- What **genotype** system has been used for Mumps virus strains?
  - (Jin et al, 2005; Jin et al, 2015; WHO 2012) tables of SH, HN, and full genome
- What are some of the **key open questions** in recent Mumps virus research?
  references

  - CDC Mumps website; Moncla et al, 2021; Lewnard & Grad, 2018

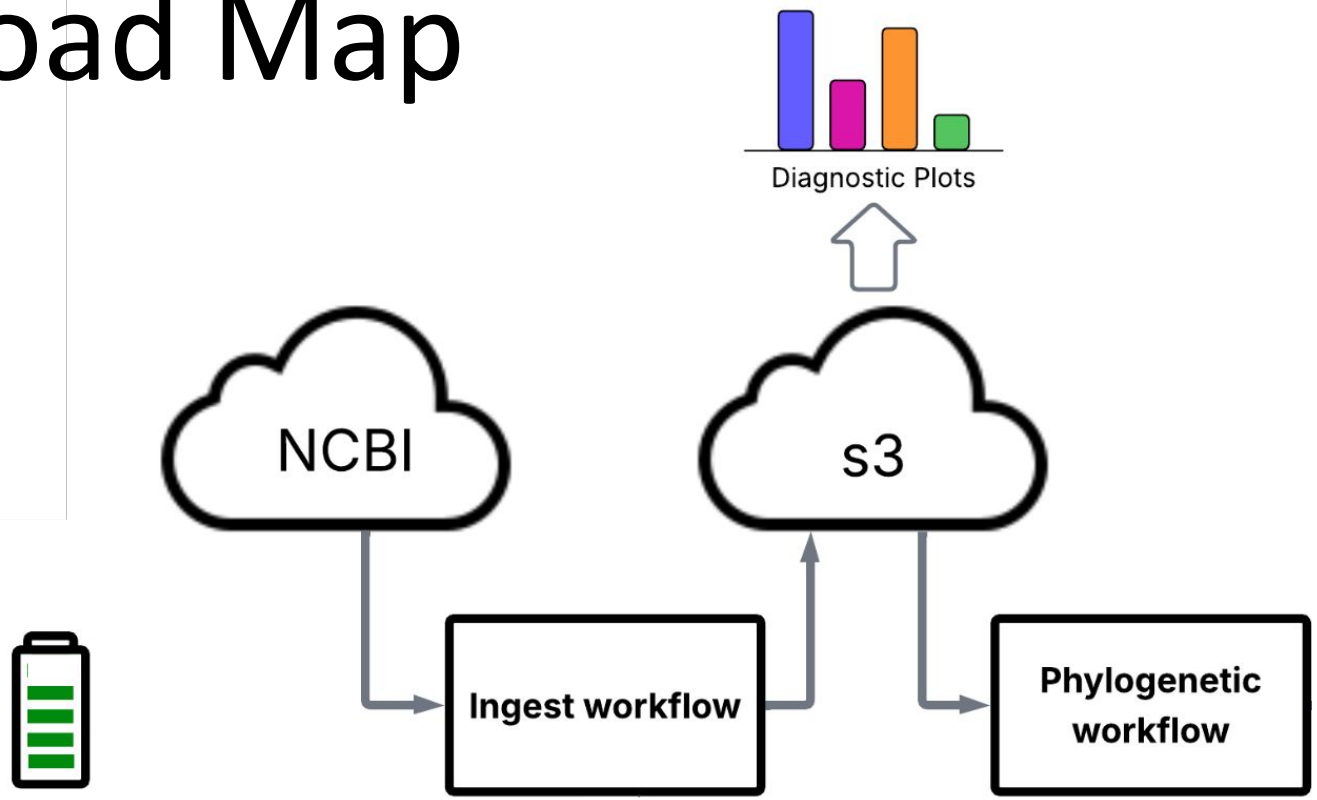
Immune waning > antigenic advance





Lewnard & Grad, 2018; Rubin et al, 2011

Road Map



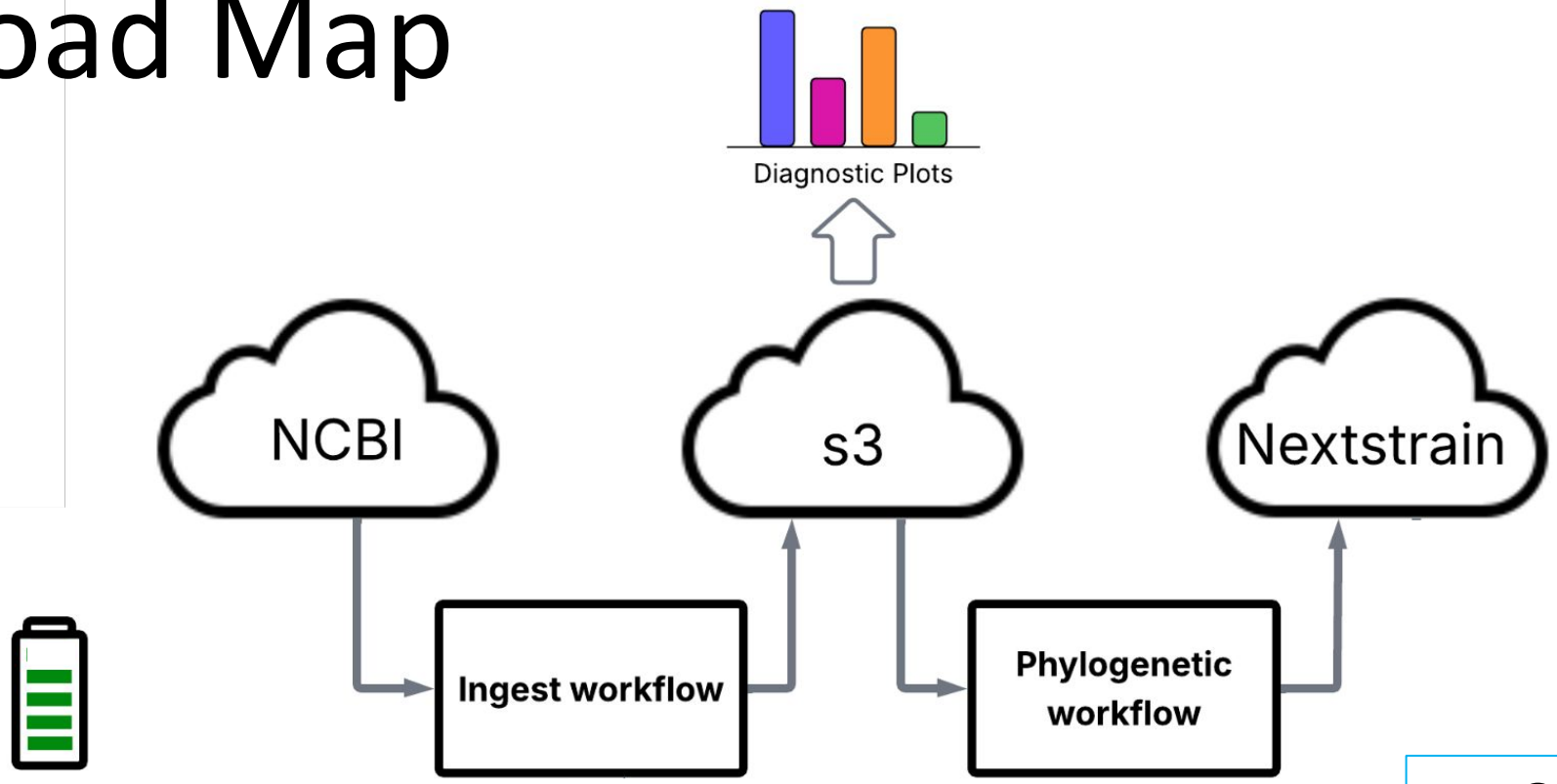
- Create diagnostic plots from the metadata.tsv
  - pick a min-length
  - sampling
  - ID potential SMEs

maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

global tree
north-america tree
add frequencies
maybe SH tree

https://docs.nextstrain.org/en/latest/tutorials/creating-a-phylogenetic-workflow.html

Road Map



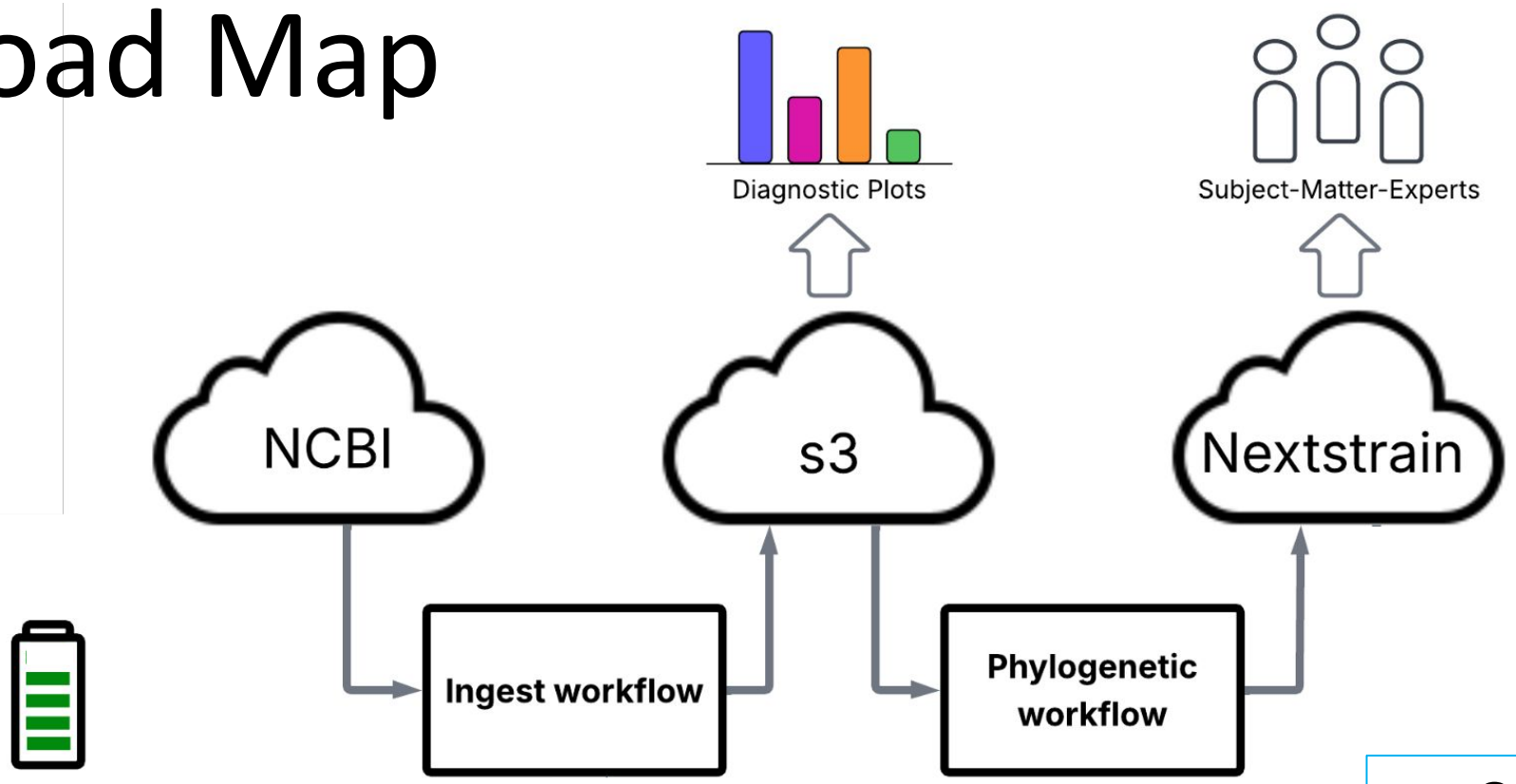
maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

global tree
north-america tree
add frequencies
maybe SH tree

- Open a Phylogenetic PR
- Push to staging

https://docs.nextstrain.org/en/latest/tutorials/creating-a-phylogenetic-workflow.html

Road Map

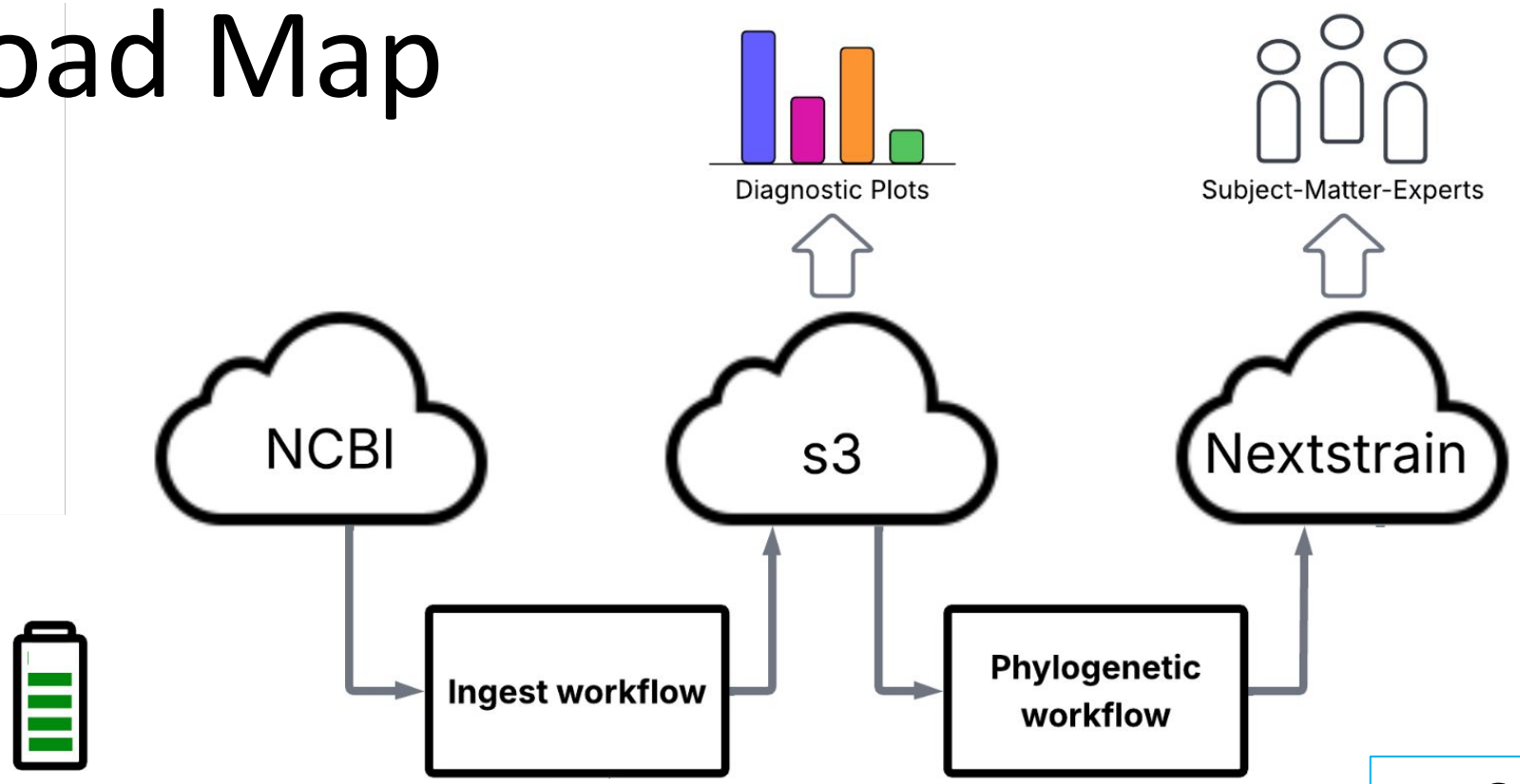


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

global tree
north-america tree
add frequencies
maybe SH tree

- Open a Phylogenetic PR
- Push to staging
- Contact SME's

Road Map

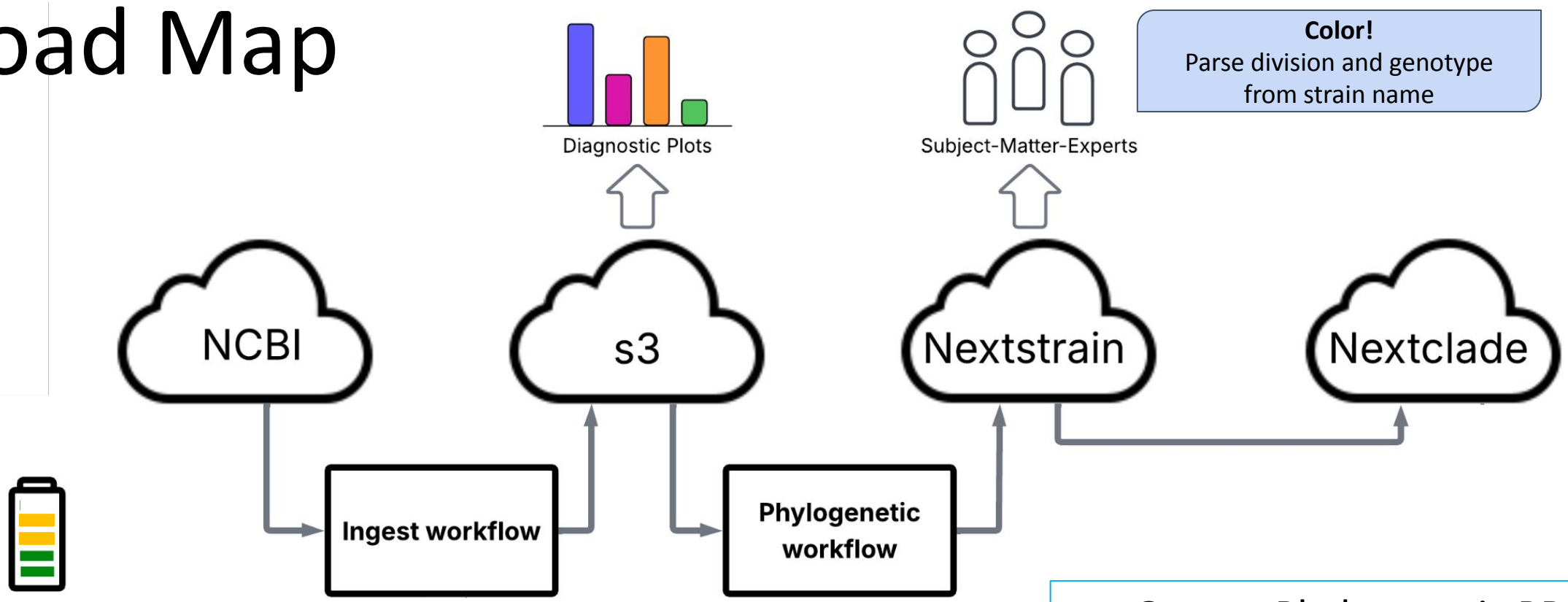


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon

global tree
north-america tree
add frequencies
maybe SH tree

- Open a Phylogenetic PR
- Push to staging
- Contact SME's
- Ingest is not merged yet, so can still adjust ingest

Road Map

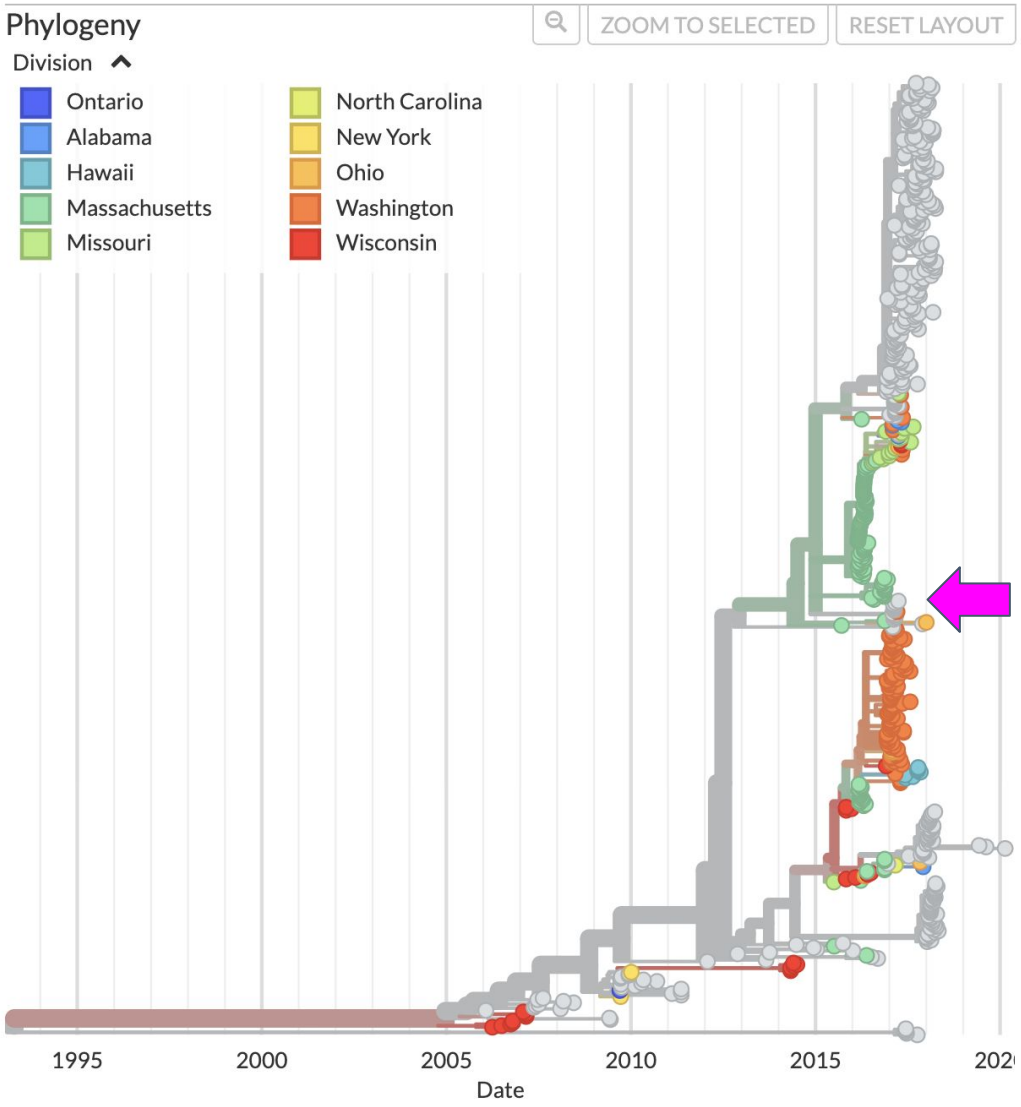


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

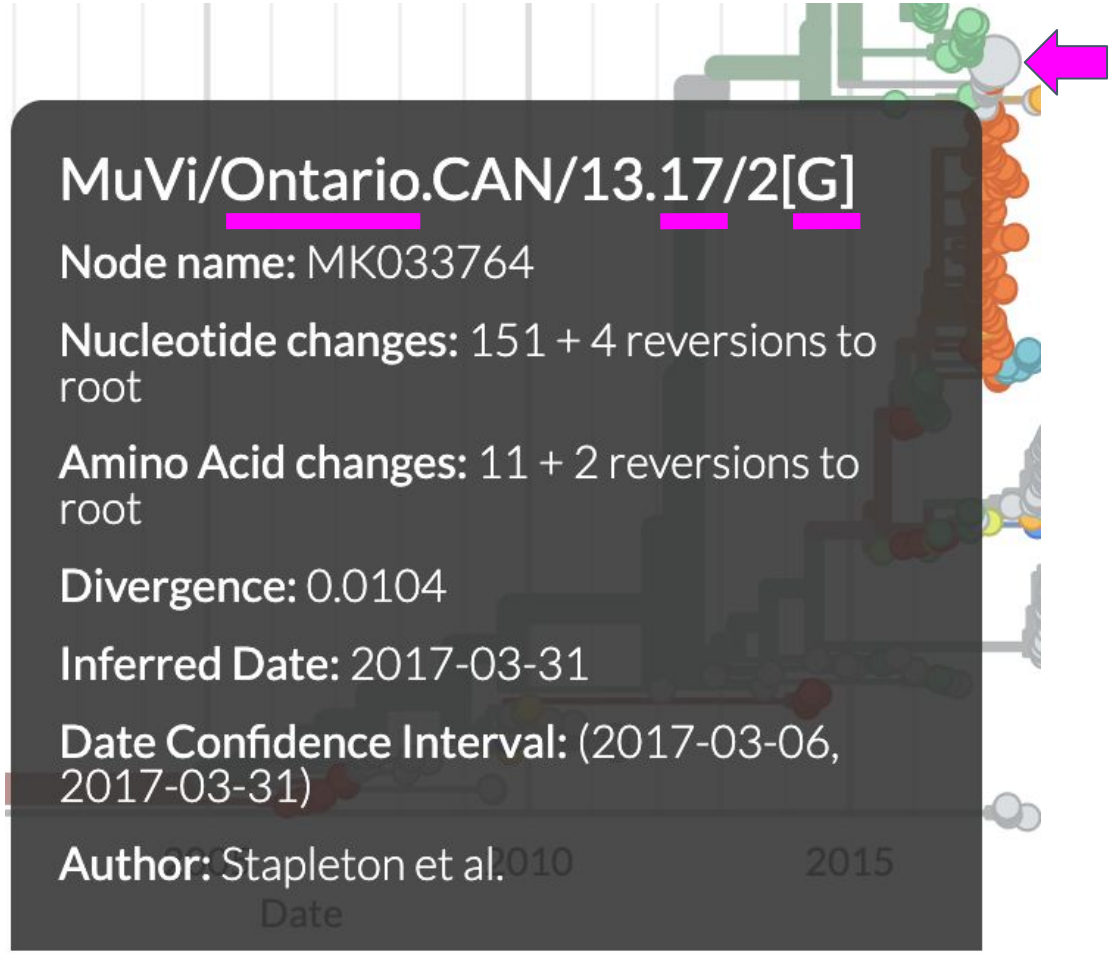
global tree
north-america tree
add frequencies
maybe SH tree

- Open a Phylogenetic PR
- Push to staging
- Contact SME's
- Ingest is not merged yet, so can still adjust ingest

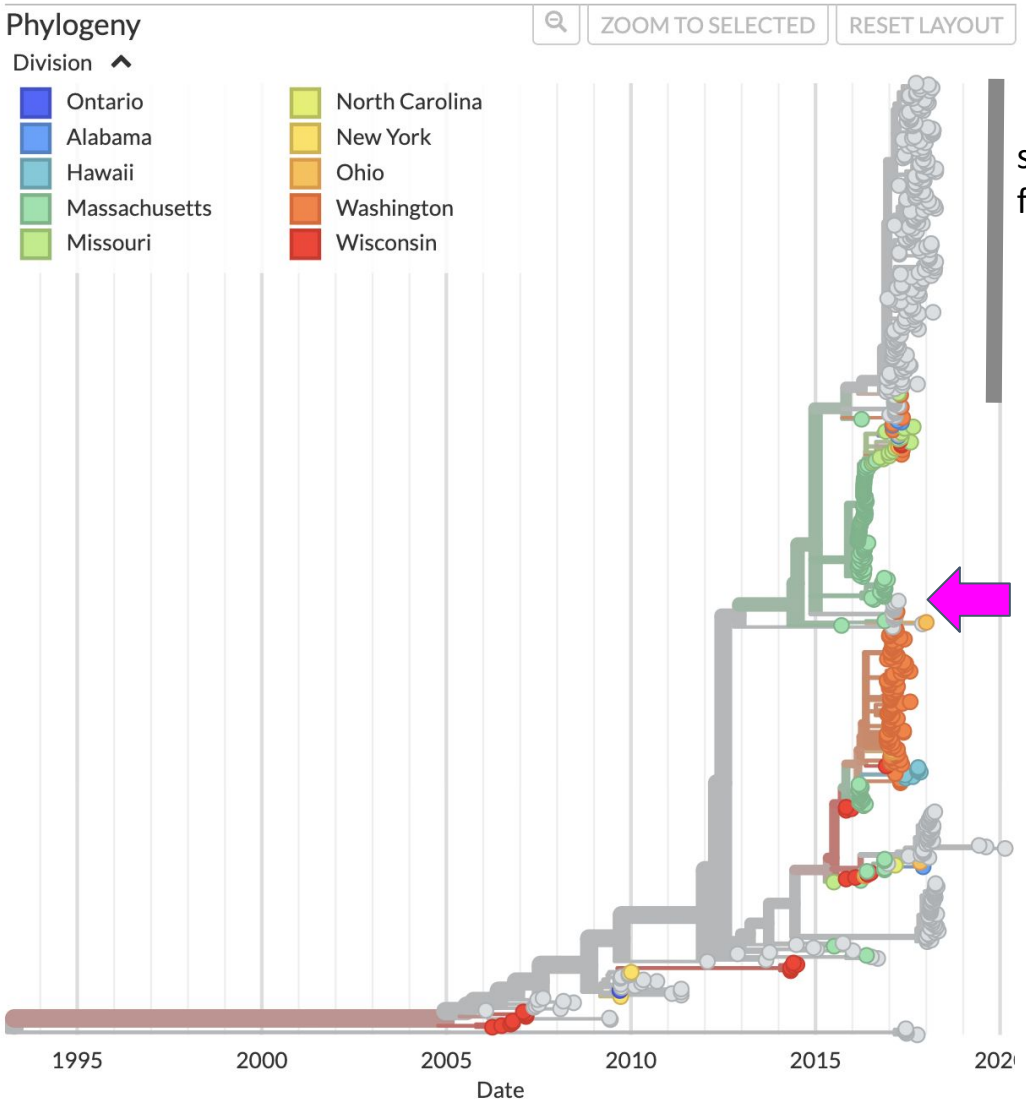
Infer division from strain name



[north-america tree](#) (color by division)

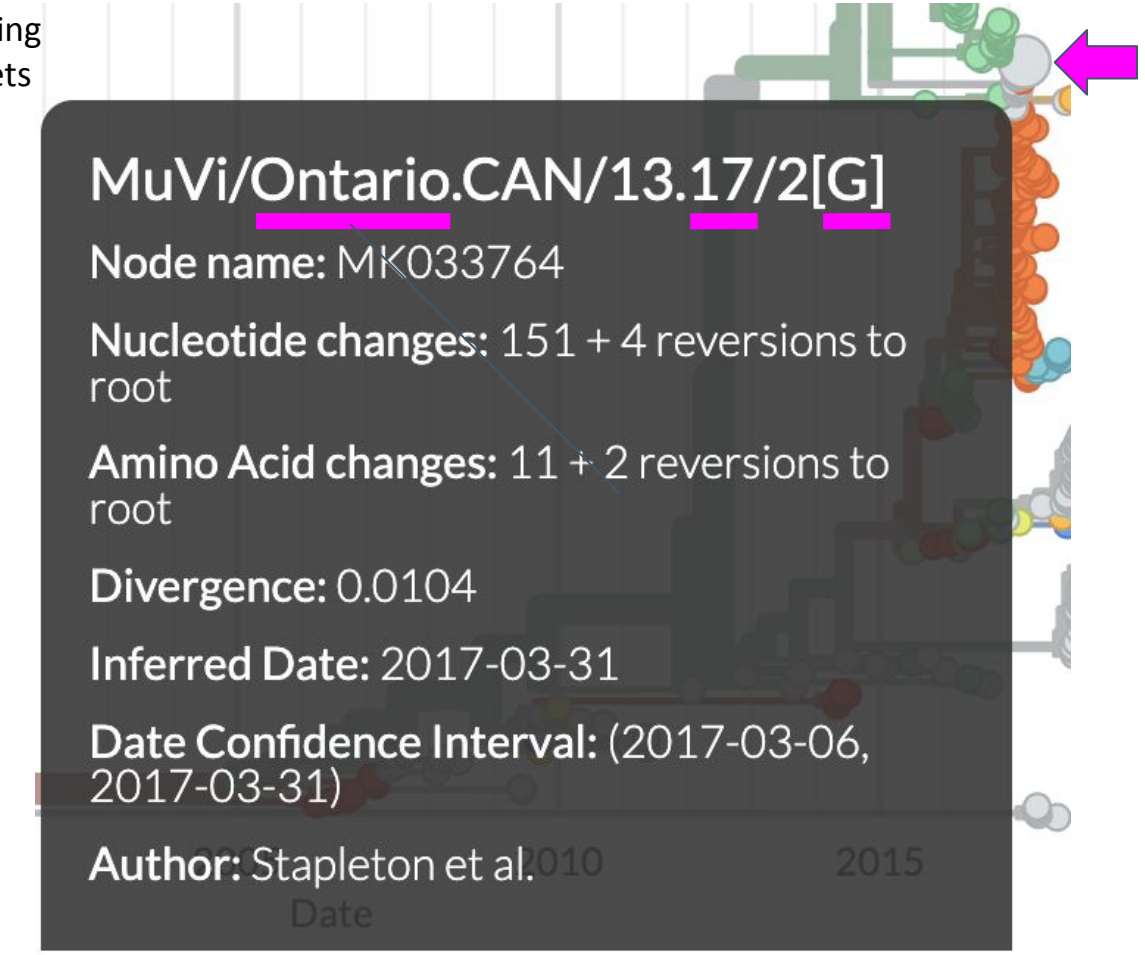


Infer division from strain name



[north-america tree](#) (color by division)

strain name missing from NCBI datasets



Missing strain names

	accession	accession_version	strain	date	region	country	division	location	length	H
2	MN920136	MN920136.1	MMR17-1436	4/20/17	North Americ	Canada			316	H
3	MN920137	MN920137.1	MMR17-1437	4/13/17	North Americ	Canada			316	H
4	MN920138	MN920138.1	MMR17-1438	4/18/17	North Americ	Canada			316	H
5	MN920139	MN920139.1	MMR17-1439	4/15/17	North Americ	Canada			316	H
6	MN920140	MN920140.1	MMR17-1440	4/13/17	North Americ	Canada			316	H
7	MN920141	MN920141.1	MMR17-1441	4/18/17	North Americ	Canada			316	H
8	MN920142	MN920142.1	MMR17-1442	4/18/17	North Americ	Canada			316	H
9	MN920143	MN920143.1	MMR17-1443	4/16/17	North Americ	Canada			316	H
10	MN920144	MN920144.1	MMR17-1444	4/18/17	North Americ	Canada			316	H
11	MN920145	MN920145.1	MMR17-1445	4/18/17	North Americ	Canada			316	H
12	MN920146	MN920146.1	MMR17-1446	4/18/17	North Americ	Canada			316	H
13	MN920147	MN920147.1	MMR17-1447	4/20/17	North Americ	Canada			316	H
14	MN920148	MN920148.1	MMR17-1448	4/20/17	North Americ	Canada			316	H
15	MN920149	MN920149.1	MMR17-1461	4/18/17	North Americ	Canada			316	H
16	MN920150	MN920150.1	MMR17-1463	4/19/17	North Americ	Canada			316	H

measles -> submitted a request for NCBI datasets to support "strain" (and "isolate") fields

Go to:

```
LOCUS       MN920136                316 bp    cRNA    linear   VRL 19-JAN-2020
DEFINITION  Mumps virus genotype G strain MuVs/Manitoba.CAN/16.17/8[G] small
            hydrophobic protein (SH) gene, complete cds.
ACCESSION   MN920136
VERSION     MN920136.1
KEYWORDS    .
SOURCE      Mumps virus genotype G
  ORGANISM  Mumps virus genotype G
            Viruses; Riboviria; Orthornavirae; Negarnaviricota;
            Haploviricotina; Monjiviricetes; Mononegavirales; Paramyxoviridae;
            Rubulavirinae; Orthorubulavirus; Orthorubulavirus parotitidis.
REFERENCE   1  (bases 1 to 316)
  AUTHORS   Hiebert,J. and Severini,A.
  TITLE     Mumps surveillance in Canada
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 316)
  AUTHORS   Hiebert,J. and Severini,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-JAN-2020) Viral Exanthemata & STDs, National
            Microbiology Laboratory, Public Health Agency of Canada, 745 Logan
            Avenue, Winnipeg, Manitoba R3E 3L5, Canada
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..316
                     /organism="Mumps virus genotype G"
                     /mol_type="viral cRNA"
                     /strain="MuVs/Manitoba.CAN/16.17/8[G]"
                     /isolate="MMR17-1436"
                     /isolation_source="buccal swab"
                     /host="Homo sapiens"
                     /db_xref="taxon:1384672"
                     /geo_loc_name="Canada"
                     /collection_date="2017-04-20"
                     /note="genotype: G"
     gene            51..224
                     /gene="SH"
     CDS             51..224
                     /gene="SH"
                     /codon_start=1
                     /product="small hydrophobic protein"
                     /protein_id="QHG14343.1"
                     /translation="MPAIQPPLYLTFLLLILLYLIITLYVWIILTVTYKTAVRHAALY
                     QRSFFHWSFDHSL"
ORIGIN
        1 aagaatgaat ctcatggggt cgtaacgtct cgtgaccctg ccgttgcact atgccggcga
       61 tccaaccccc attatacctc acatttctat tgctaattct tctttatctg atcataactt
      121 tgtatgtctg gattatatta actgttactt ataagactgc ggtgcgacat gcagcactgt
      181 accagagatc cttctttcac tggagtttcg atcactcact ctaagaagat ccccagttag
      241 gacaagttcc gatccatcat gcaagaacaa tctgcatttg aataatgccg ttcaatcatg
```

Different coders, different solutions

oropouche - custom script

mumps - custom script

```
149     rule parse_strain_from_genbank:
150         input:
151             genbank="data/genbank.gb",
152         output:
153             strain_names="data/strain_names.tsv",
154         benchmark:
155             "benchmarks/parse_strain_from_genbank.txt",
156         params:
157             annotation="strain",
158         shell:
159             r"""
160             ./scripts/parse-genbank-annotations.py \
161               --annotation {params.annotation} \
162               --silent-no-match \
163               {input.genbank:q} \
164             > {output.strain_names:q}
165             """
```

ebola - new tool: bio

```
150     rule parse_genbank_to_ndjson:
151         input:
152             genbank="data/genbank.gb",
153         output:
154             ndjson="data/ncbi_entrez.ndjson",
155         benchmark:
156             "benchmarks/parse_genbank_to_ndjson.txt"
157         shell:
158             r"""
159             bio json {input.genbank:q} \
160             | jq -c '.[] | {{accession: .record.accessions[0], strain: .record.strain[0]}}' \
161             > {output.ndjson:q}
162             """
```

rubella

-> also tested out bio + custom script to get genotype

-> tested parsing all fields from the GB or the XML files (SeqIO)

Lassa -> also tested out bio to get strain names and L/S names

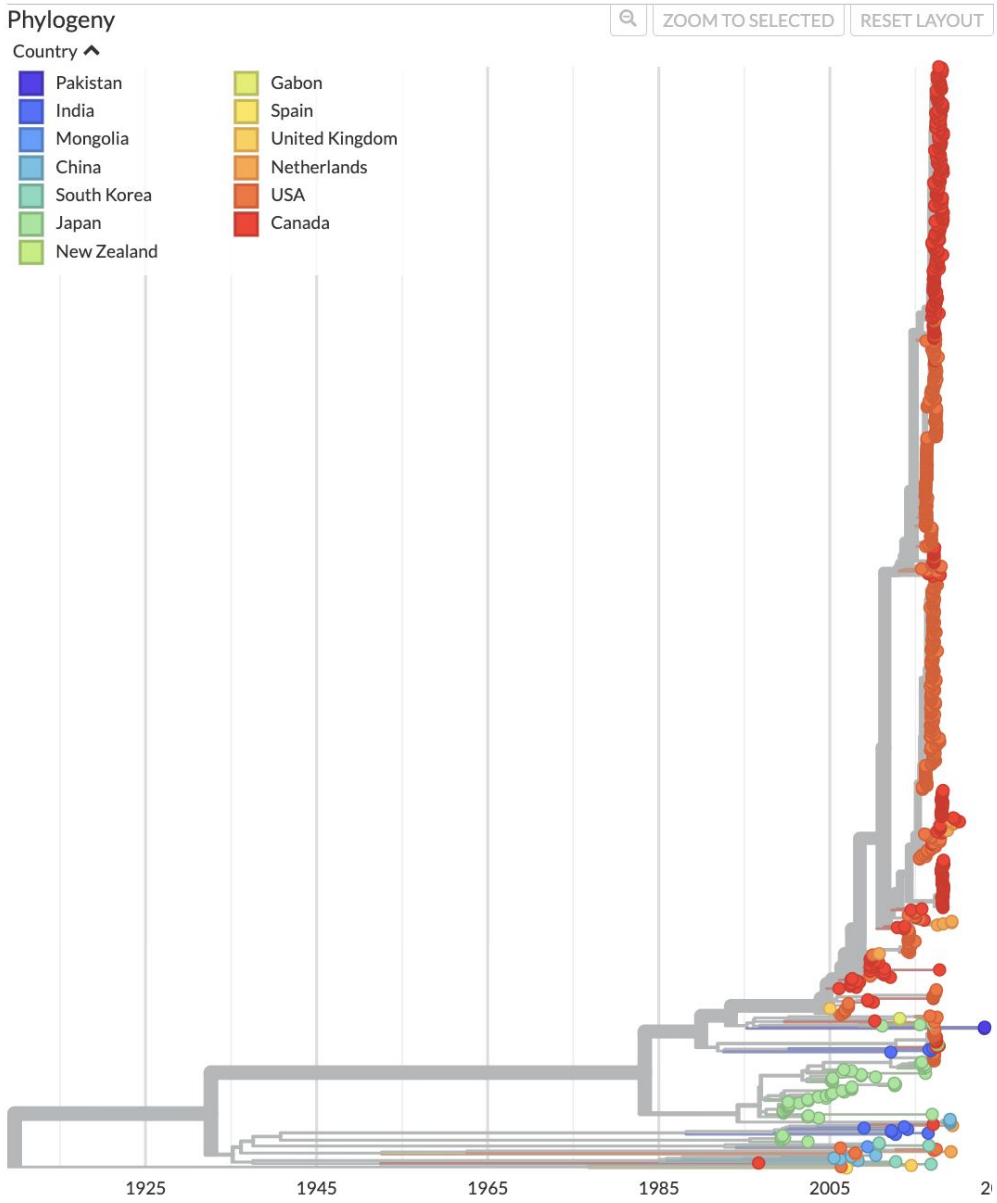
github issue discussion

Pathogen-repo-guide: annotations.tsv

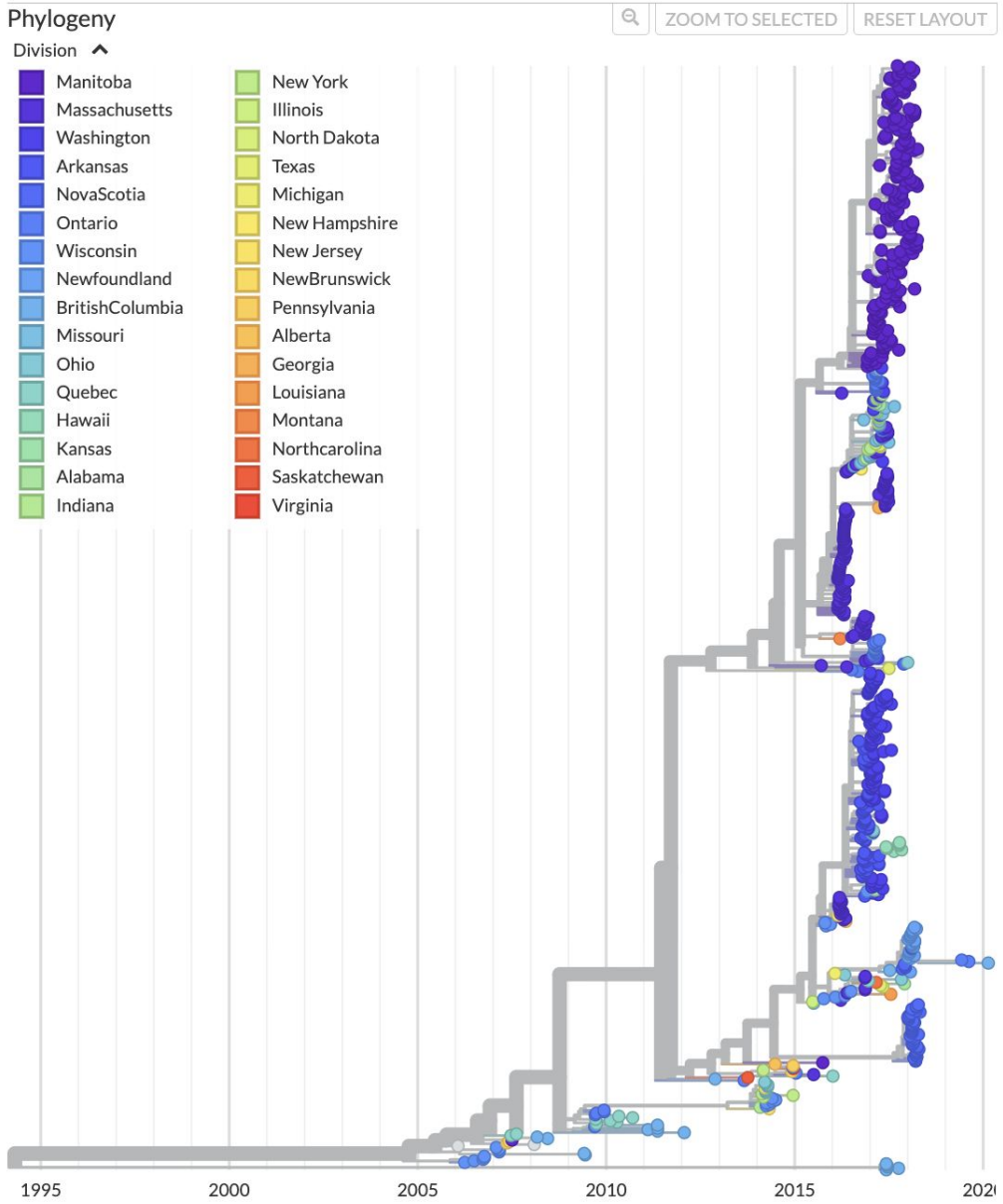
```
# Manually curated annotations TSV file
# The TSV should not have a header and should have exactly three columns:
# id to match existing metadata, field name, and field value
# If there are multiple annotations for the same id and field, then the last value is used
# Lines starting with '#' are treated as comments
# Any '#' after the field value are treated as comments.
# This is a workaround since NCBI datasets pulls "isolate" instead of "strain" annotations
# This step can be dropped if we come up with a solution to issue in https://github.com/nextstrain/mumps/issues/15
MN920136    strain  MuVs/Manitoba.CAN/16.17/8[G]
MN920137    strain  MuVs/Manitoba.CAN/15.17/8[G]
MN920138    strain  MuVs/Manitoba.CAN/16.17/9[G]
MN920139    strain  MuVs/Manitoba.CAN/15.17/9[G]
MN920140    strain  MuVs/Manitoba.CAN/15.17/10[G]
MN920141    strain  MuVs/Manitoba.CAN/16.17/10[G]
MN920142    strain  MuVs/Manitoba.CAN/16.17/11[G]
MN920143    strain  MuVs/Manitoba.CAN/15.17/11[G]
MN920144    strain  MuVs/Manitoba.CAN/16.17/12[G]
MN920145    strain  MuVs/Manitoba.CAN/16.17/13[G]
MN920146    strain  MuVs/Manitoba.CAN/16.17/14[G]
MN920147    strain  MuVs/Manitoba.CAN/16.17/15[G]
MN920148    strain  MuVs/Manitoba.CAN/17.17/8[G]
MN920149    strain  MuVs/Saskatchewan.CAN/16.17[G]
MN920150    strain  MuVs/Saskatchewan.CAN/16.17/2[G]
MN920151    strain  MuVs/Saskatchewan.CAN/16.17/3[G]
MN920152    strain  MuVs/Saskatchewan.CAN/16.17/4[G]
MN920153    strain  MuVs/Saskatchewan.CAN/16.17/5[G]
MN920154    strain  MuVs/Saskatchewan.CAN/16.17/6[G]
MN920155    strain  MuVs/Saskatchewan.CAN/16.17/7[G]
MN920156    strain  MuVs/British Columbia.CAN/16.17[G]
MN920157    strain  MuVs/Manitoba.CAN/16.17[G]
```

- Gives the team enough time to discuss code implementations
- Speeds up getting the site updated for next round of SME review
- Can be easily replaced by the final code implementation

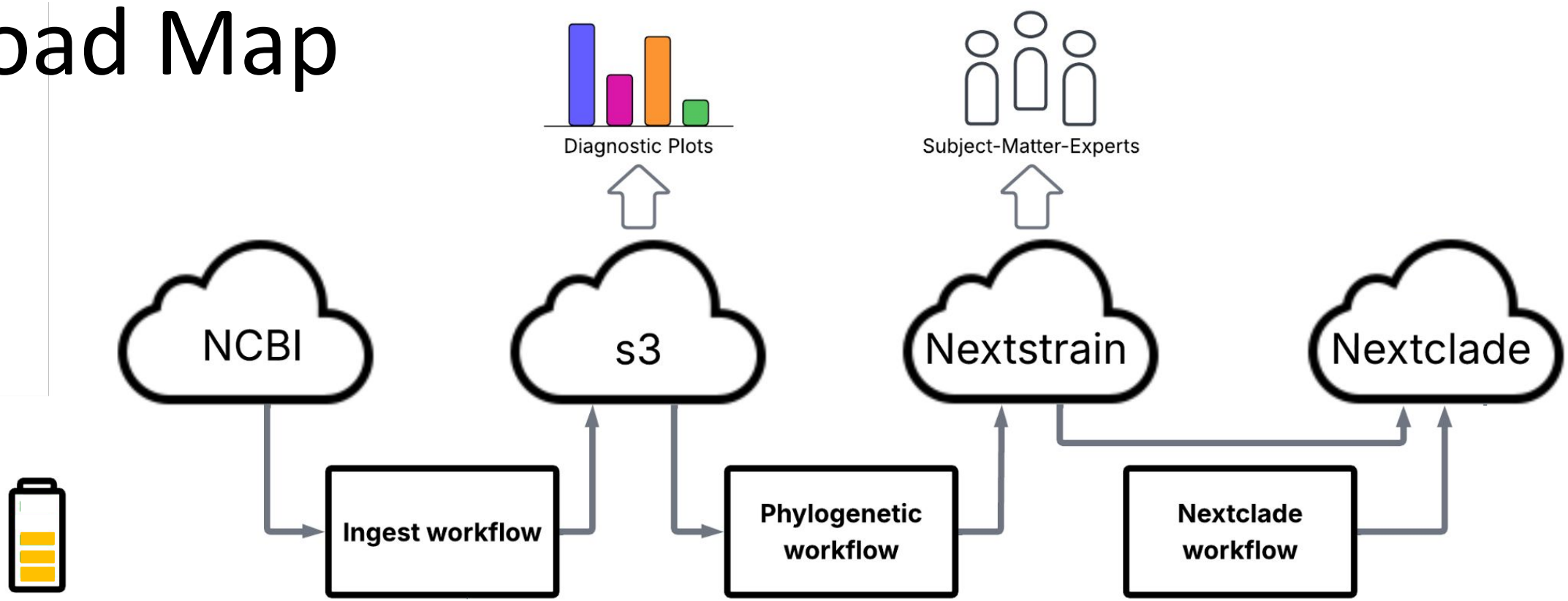
global



north-america



Road Map



maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

global tree
north-america tree
add frequencies
maybe SH tree

- SME review again
- Push staging to live site
- It's time to start the Nextclade workflow

Literature review - collect papers

- What is the **earliest study** identifying Mumps as a viral disease?
  - (Johnson and Goodpasture, 1933) - "An investigation of the etiology of Mumps"
- When was the Mumps virus first **sequenced**, and which genome region was published?
  - 1989 - gene segments (F, HN, **SH**, …) perhaps Takeuchi et al; full genome perhaps Mori
- When was the first Mumps **vaccine** developed, and how effective was it?
  - (Hillman et al, 1967) - "Live Attenuated Mumps Virus Vaccine" Jeyrl-Lynn Strain
- Have there been major updates to the vaccine? What prompted them?
  - Yes, side effects or regional genotype (Urabe, Leningrad-3, Rubini, Miyahara)
- What **genotype** system has been used for Mumps virus strains?
  - (Jin et al, 2005; Jin et al, 2015; WHO 2012) tables of SH, HN, and full genome
- What are some of the **key open questions** in recent Mumps virus research?
  references

  - CDC Mumps website; Moncla et al, 2021;

Nextclade Dataset

Table 1. The proposed reference strains and global distribution of mumps genotypes

Genotype	Reference strain	GB accession no	Country (IS03) and year identified
A	**End/USA45***	D90231	USA45, 50, 63; SWE69, 93; CHE74;
	SBL-1/SWE69*	D00663	DEU87, 92; CAN88
	JL/US63 (vaccine)	D90232	
	Rubini (vaccine)	X72944	
B	**Urb/Jap67**	D90236	JPN67-95: GBR89, 90
	Mat/Jap84	D90233	
	Miya (vaccine)*	D90234	
	Hoshino (vaccine)	AB003414	
C	**Bf/UK75***	X63709	GBR75, 80s, 90, 98-2000; SWE80s, 92;
	Bm1/UK90	DS26771.DAT	DEU87, 92, 93; CHE95; PRT96; LTU98-00
D	**Ge9/Gem77**	DS26771.DAT	DEU77; PRT96; GBR96, 97, 99;
	Islip1/UK97	AF142766	LTU99; DEN80s, 90s, 01; JPN93
E (C)	**Ed2/UK88**	X63711	GBR88
F	**WLZ1/CNA95**	Z77158	CHN95; GBR 99; SWE71, 72, 84;
	WSH1/CNA96	Z77160	
G	**Glouc1/UK96***	AF142764	GBR91, 96-05; JPN99-05
	UK01-22	AY380075	
H	**Be1/UK88**	DS26771DAT	GBR88, 95, 96, 98-01;
	ManchS1/UK95*	AF142771	CHE 95, 98-00; KOR99; JPN97
I	**Odate-1**	D86174	JPN93; KOR97-01
	AA12/Korea97*	AF180374	
J	**MP94H/JNP94**	AB03417	JPN94, GBR97
	Loug1/UK97	AF142770	
K	**DK81/01 (DMK81)**	AF365891	DNK81-88
L	**Fukuoka49/JPN00***	AB105483	JPN00-01
	Tokyo S-III-10/JPN01*	AB105480	
**	Leningrad 3 (vaccine)	AY493374	RUS53
	L-Zagreb (vaccine)	AJ272363	
**	Tay/UK50s*	AF142774	GBR50s
**	UNK02-19*	AY380077	GBR02

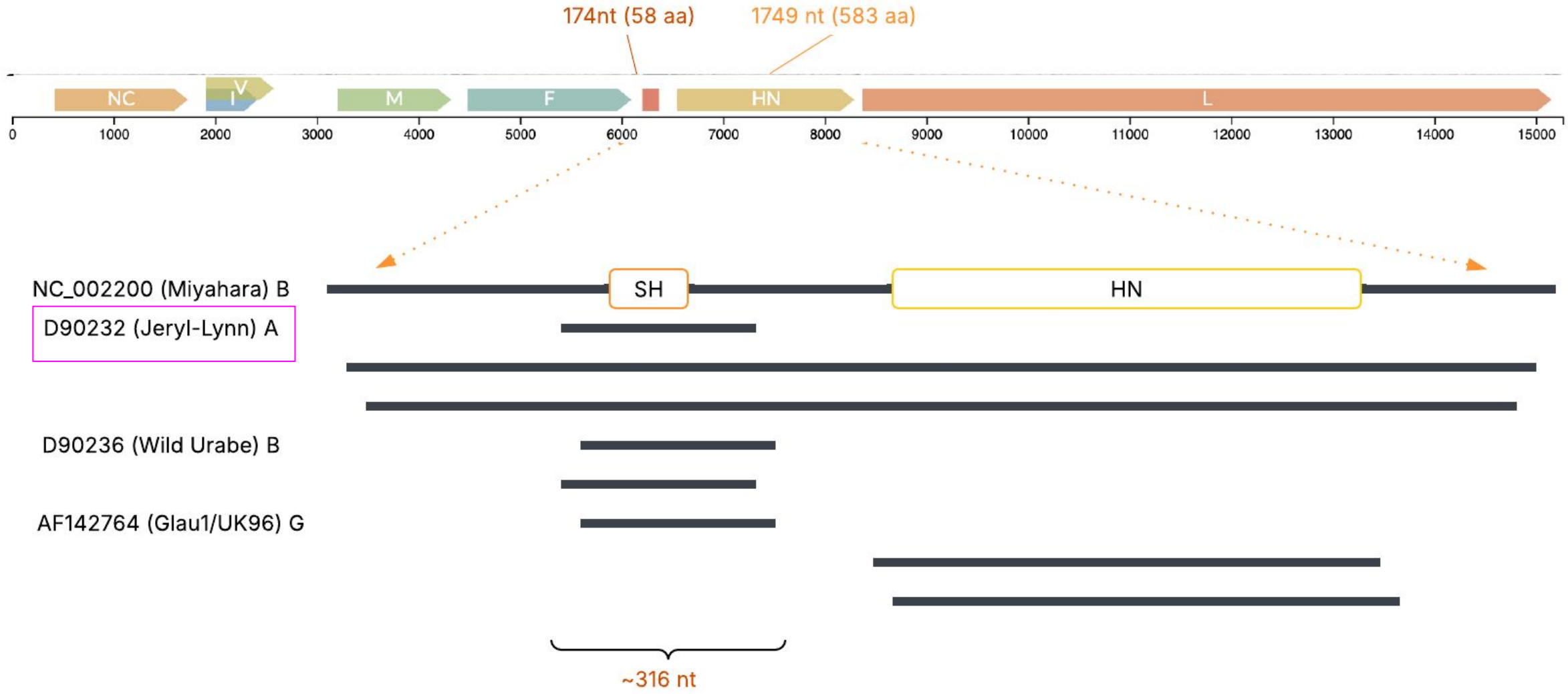
*Isolate available; **Reference strains for potential new genotypes

Table 1. Mumps genotype reference strains and those with full genome sequence for analysis (73 sequences)

Genotype (no. seqs)	*Reference strain × no. of identical sequences	GenBank accession number: Full genome (identical sequences)	GenBank accession number: SH/HN of ref strain
A (22)	*MuVi/Boston.USA/0.45[A]	GU980052	
	*MuVi/Pennsylvannia.USA/13.63[A] (VAC)x13	AF338106 (BD293023, EA500331–2, AX081133, DI021804, DI064912, AX081123, FJ211584–6, BD293022, AF201473)	
	MuVi/JL2.USA/0.63(VAC)-Ax5	AF345290 (EA500333,BD293024, DI035997, HQ416907)	
	MuVi/JL2.USA/0.63-Ax2	FN431985 (AX081134)	
	MuVi//JL5.S79/CHN-A	HQ416906	
B (14)	*MuVi/Urabe.JPN/0.67[B]x10	AB000388 (AB000386–7, AF314558–62, FJ375177–8)	JQ945269/JQ946041
	*MuVi/Himeji.JPN/24.00[B]	–	
	MuVi/Y213.JPN/0.0[B]	AB576764	
	MuVi/Miyahara.JPN/vac[B] x2	AB040874 (NC002200)	
	Muvi/Hoshino.JPN/vac[B]	Ab470486	
C (2)	*MuVi/Zagreb.HRV/39.98[C]	EU370206	JQ945268/JQ999999
	*MuVi/Stockholm.SWE/46.84[C]	–	
	MuVi/Drag94.RUS/0.94[C]	AY669145	
D (1)	*MuVi/Ge9.DEU/0.77[D]	*KF878076*	JQ945275/JQ946039
	*MuVi/Nottingham.GBR/19.04[D]	–	JQ034452/JQ034464
F (7)	*MuVi/Shandong.CHN/4.05[F]	KF042304	
	*MuVi/Zhejiang.CHN/11.06/1[F]	KF170917	
	MuVi/SP-A.Yunnan.CHN/0.05-Fx3	FJ556896 (EU884413, DQ649478)	
	MuVi/Zhejiang.CHN/16.08/2-F	KF170918	
	MuVi/Zhejiang.CHN/26.05-F	KF17091	
G (9)	*MuVi/Gloucester.GBR/32.96[G]	AF280799	EU597478/JQ946046
	*MuVi/Sheffield.GBR/1.05[G]	–	
	MuVi/Split.CRO/05.11[G]x5	JN635498 (JX287387, JX287389–91)	
	MuVi/Iowa.USA/0.06-Gx2	JX287385 (JN012242)	
	MuVi/Du.CRO/0.05-G	EU370207	
H (5)	*MuVi/Bedford.GBR/0.89[H]	*KF878077*	JQ945273/JQ946035
	*MuVi/Ulaanbaatar.MNG/22.09[H]	AB600843	
	MuVi//1961.USA/0.88[H]	AF467767	
	MuVi/Mass.USA/4.10[H]	JX287388	
	MuVi/Novosibirsk.RUS/10.03[H]	AY681495	
I (3)	*MuVi/Akita.JPN/42.93[I]x2	*KF878078* (AB600942)	JQ945274/JQ946037
	*MuVi/Dg1062.KOR/46.98[I]	AY309060	
J (1)	*MuVi/Leeds.GBR/9.04[J]	*KF878079*	JQ945271/JQ946033
	*MuVi/Sapporo.JPN/12.00[J]	–	AB105475/JQ946044
K (2)	*MuVi/RW154.USA/0.70s[K]	*KF878080*	JQ945276/JQ946040
	*MuVi/Stockholm.SWE/26.83[K]	–	JQ945270/JQ946045
	MuVi/California.USA/50.07/1-K	JX287386	
L (1)	*MuVi/Fukuoka.JPN/41.00[L]	*KF878081*	AB105483/JQ946036
	*MuVi/Tokyo.JPN/6.01[L]	–	AB105480/JQ946043
N (5)	*MuVi/Vector.RUS/0.53[N] (VAC)x3	AY508995 (JF727651–2)	
	*MuVi/L-Zagreb.HRV/0.71[N] (VAC)x2	AY685920 (AY685921)	
Unclassified	MuVi/Taylor.GBR/0.50s	–	AF142774/JQ946042
	MuVi/Tokyo.JPN/0.93	–	AB003415/AB003415
	MuVi/London.GBR/3.02	*KF878082*	AY380077/JQ946038

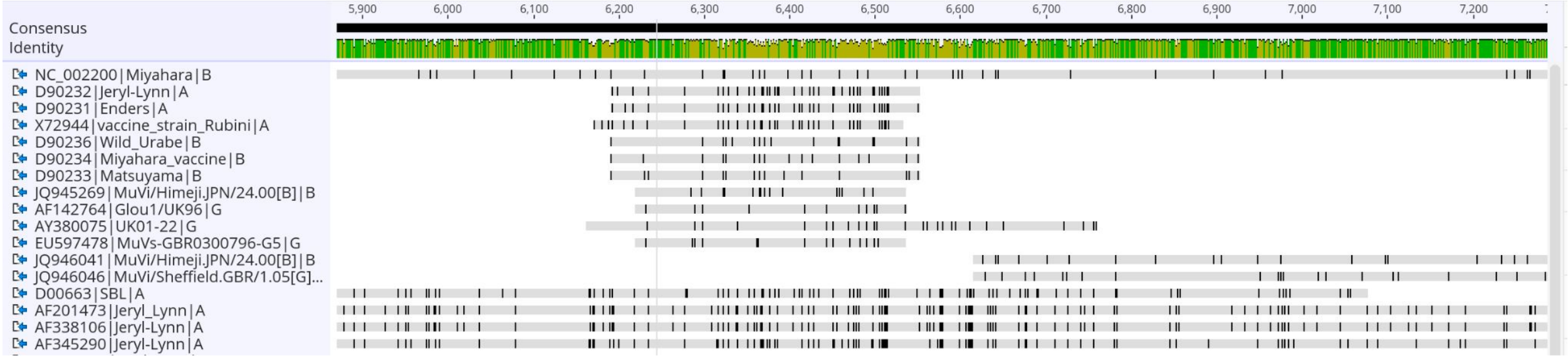
Notes: Italic: sequenced in this study;
*Reference strains.

Pick a reference for SH region

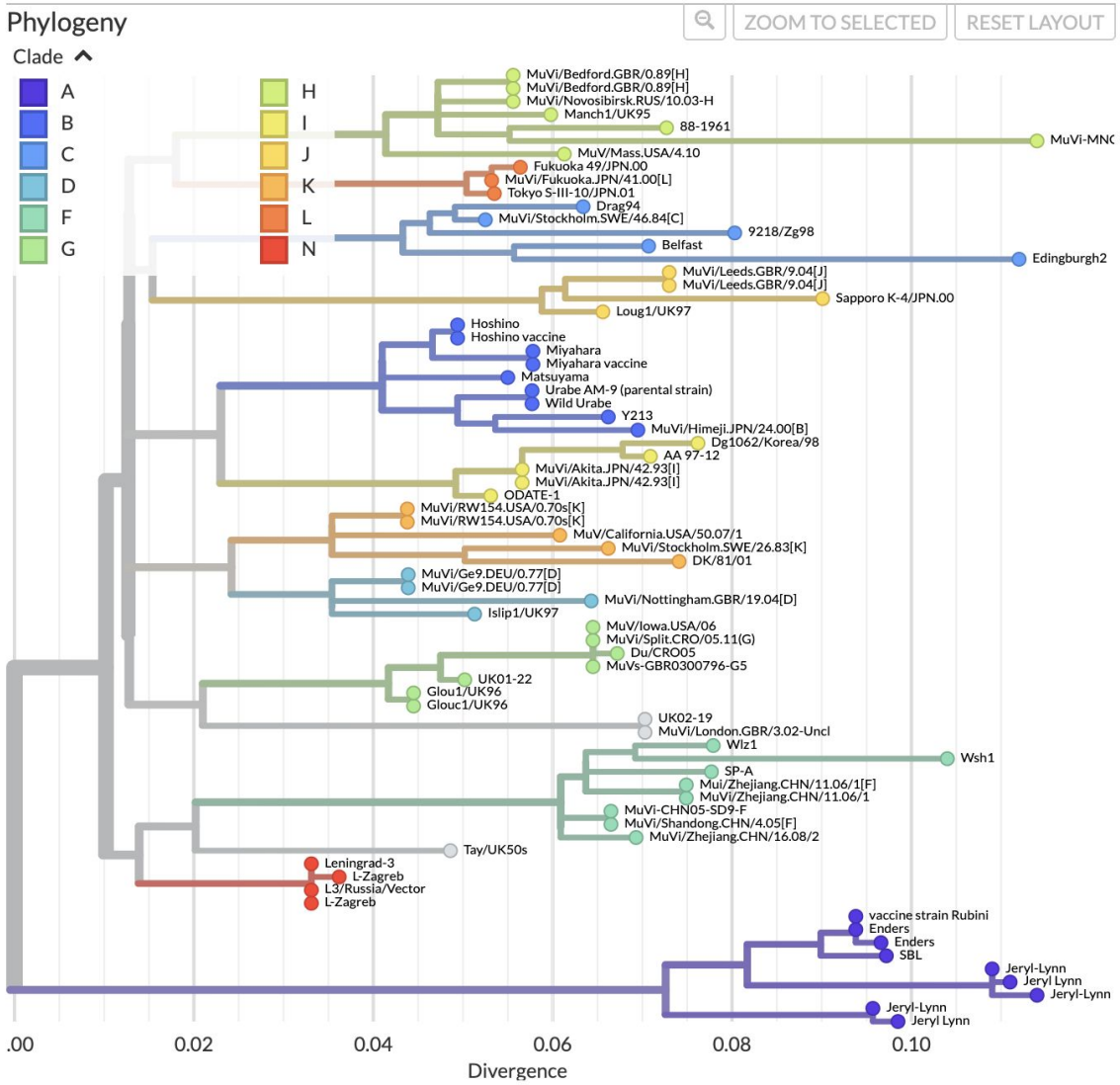


*

Align the scaffold sequences



Draft Nextclade Dataset



Differences to the global tree

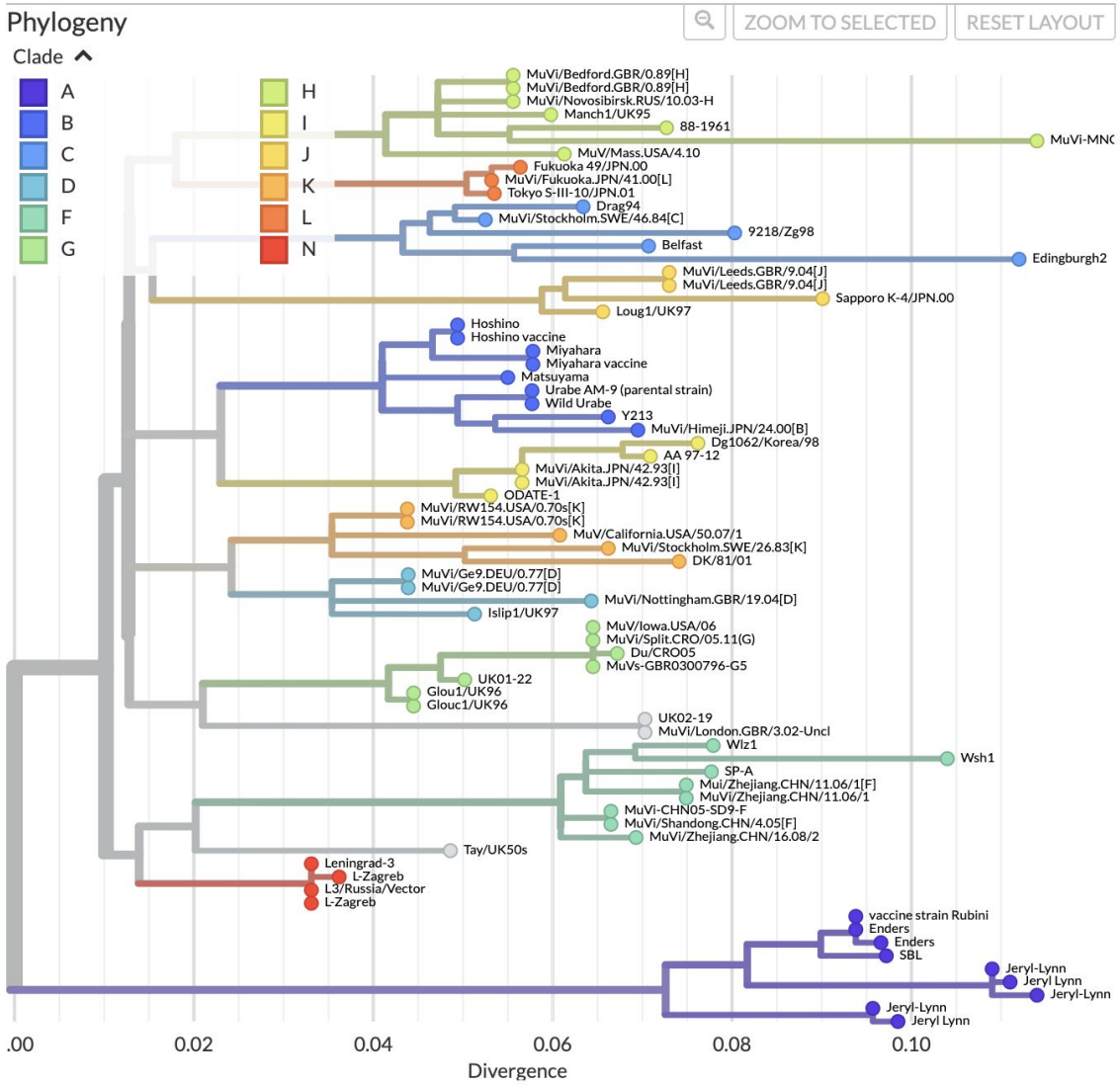
- Only SH **region**
- Mid point root
- Includes B and N genotype strains
- There are reference unknown/unclassified genotype strains
- Jeryl Lynn is a bit over represented

Nextclade Alignment parameters

- Nextclade preset alignments (short sequences)
- Adjusting alignments so that a reference of 316 nt doesn't cause queries of 15,000 nt to fail (300/15000 nt =0.02)


```
{
  "alignmentParams": {
    "minLength": 80,
    "penaltyGapExtend": 1,
    "penaltyGapOpen": 4,
    "penaltyGapOpenInFrame": 4,
    "penaltyGapOpenOutOfFrame": 6,
    "penaltyMismatch": 1,
    "scoreMatch": 4,
    "noTranslatePastStop": false,
    "excessBandwidth": 9,
    "terminalBandwidth": 80,
    "allowedMismatches": 12,
    "minMatchLength": 30,
    "maxAlignmentAttempts": 5,
    "includeReference": true,
    "includeNearestNodeInfo": true,
    "retryReverseComplement": true,
    "minSeedCover": 0.01
  },
```

Draft Nextclade Dataset

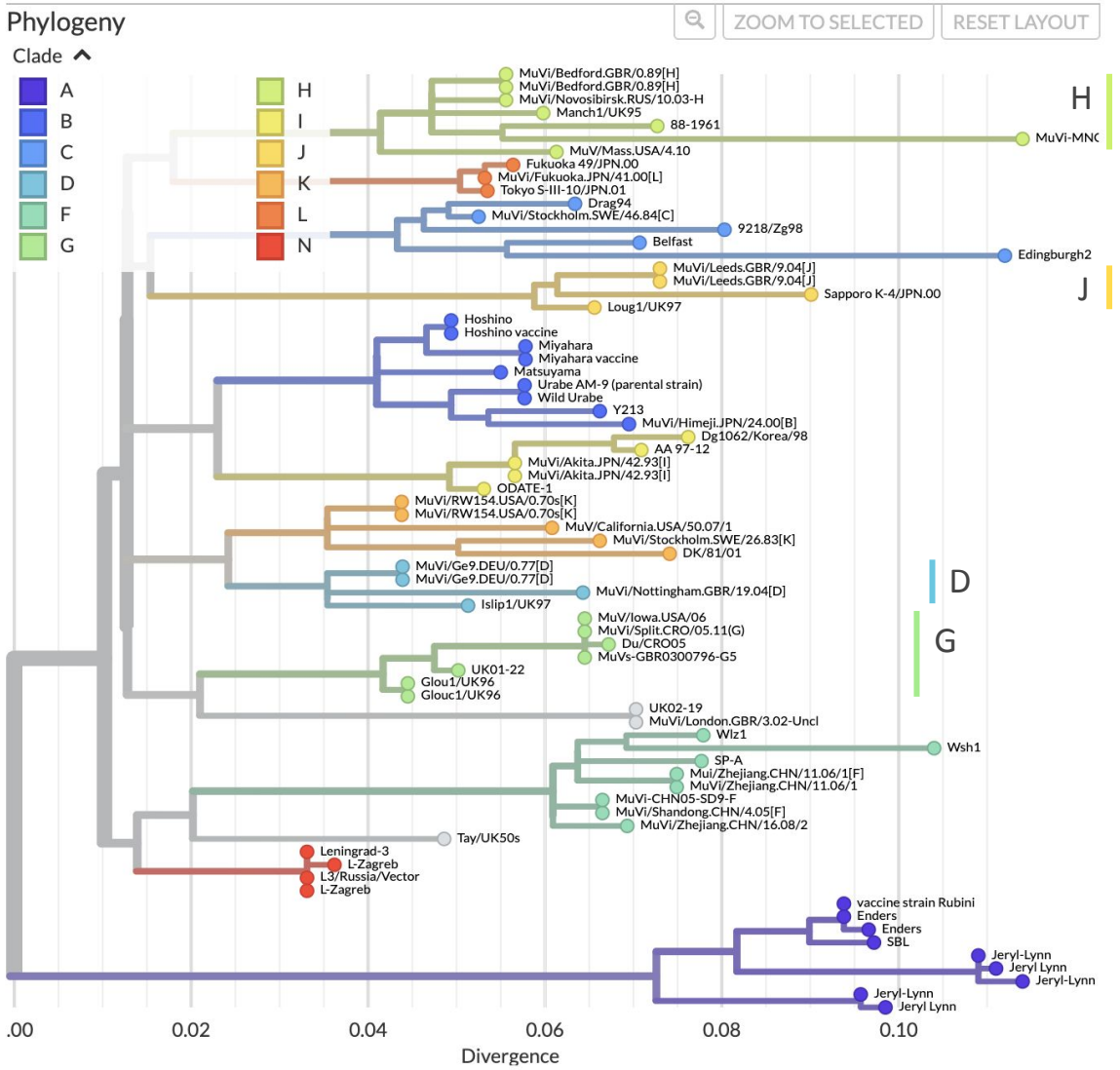


Draft Mumps Nextclade dataset results

GenBank MuV_genotype

Count of accession / Row Labels	Column Labels: A	B	C	D	F	G	H	I	J	K	L	N	(blank)	Grand Total
A	9												44	53
B		35											16	51
C			213										29	242
D				101		3							86	190
D1				4										4
F					273								157	430
G				3		8791	1		1				2426	11222
G1						56								56
G2						179								179
H							155	1					33	189
H1							9							9
H2							2							2
I								95					8	103
J									27				9	36
K										77			30	107
K/M										1				1
L											6		21	27
M										1			1	2
N												14	6	20
(blank)	16	64	76	42	198	545	102	33	69	53	2	6	419	1625
Grand Total	**25**	**99**	**289**	**150**	**471**	**9574**	**269**	**129**	**97**	**132**	**8**	**20**	**3285**	**14548**

Draft Nextclade Dataset

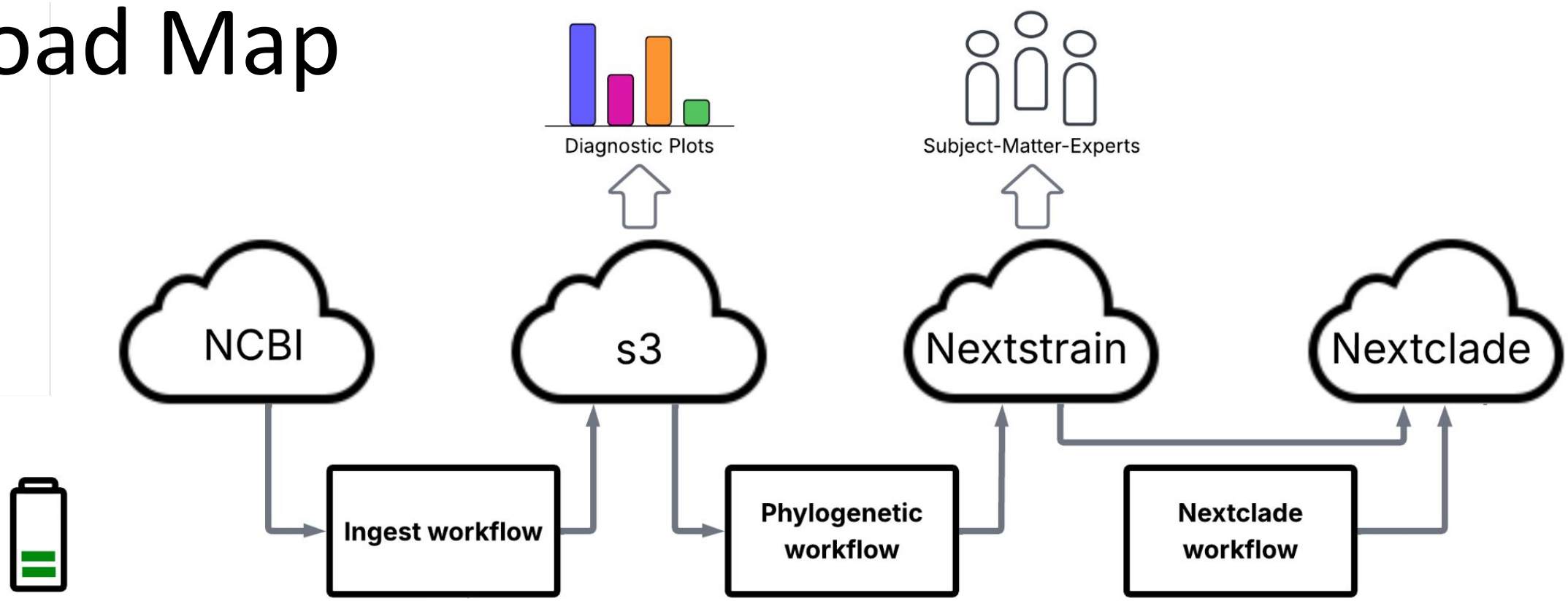


Draft Mumps Nextclade dataset results

GenBank MuV_genotype

Count of accession	Column Labels													
Row Labels	A	B	C	D	F	G	H	I	J	K	L	N	(blank)	Grand Total
A	9												44	53
B		35											16	51
C			213										29	242
D				101		3							86	190
D1				4										4
F					273								157	430
G				3		8791	1		1				2426	11222
G1						56								56
G2						179								179
H							155	1					33	189
H1							9							9
H2							2							2
I								95					8	103
J									27				9	36
K										77			30	107
K/M										1				1
L											6		21	27
M										1			1	2
N												14	6	20
(blank)	16	64	76	42	198	545	102	33	69	53	2	6	419	1625
Grand Total	**25**	**99**	**289**	**150**	**471**	**9574**	**269**	**129**	**97**	**132**	**8**	**20**	**3285**	**14548**

Road Map

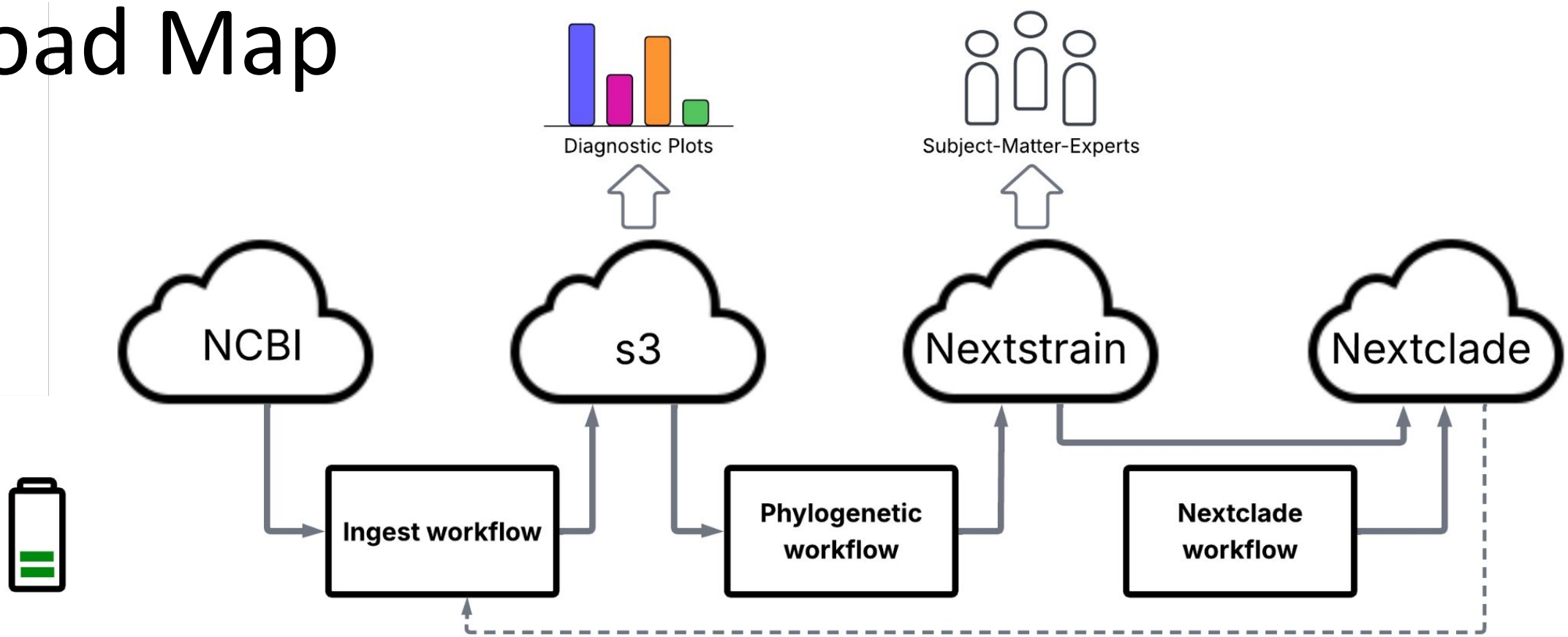


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

global tree
north-america tree
add frequencies
maybe SH tree

- Open Nextclade PR
- Connect to Ingest
- More diagnostic plots

Road Map

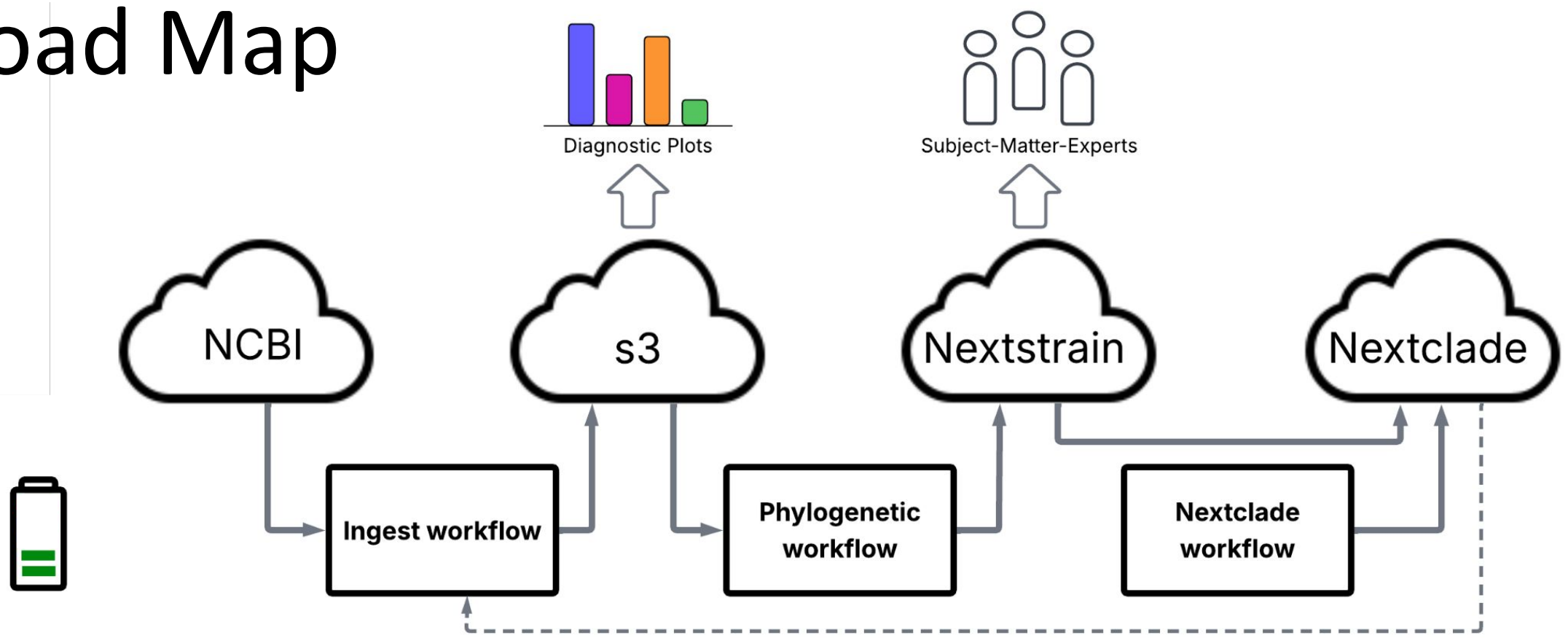


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

global tree
north-america tree
add frequencies
maybe SH tree

- Open Nextclade PR
- Connect to Ingest
- More diagnostic plots

Road Map

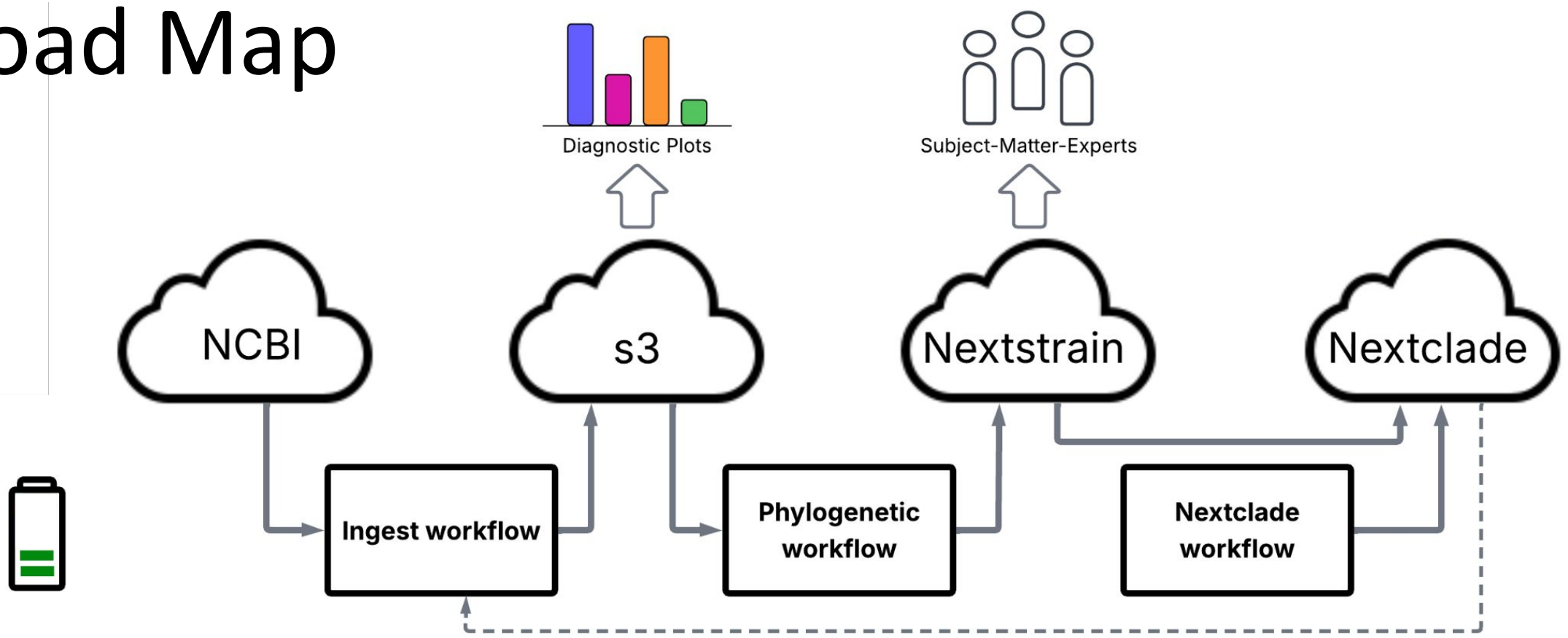


maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

global tree
north-america tree
add frequencies
maybe SH tree

Jin and WHO scaffold strains
Jeryl Lynn SH region reference
adjust nextclade parameters
maybe PR to Nextclade data

Road Map

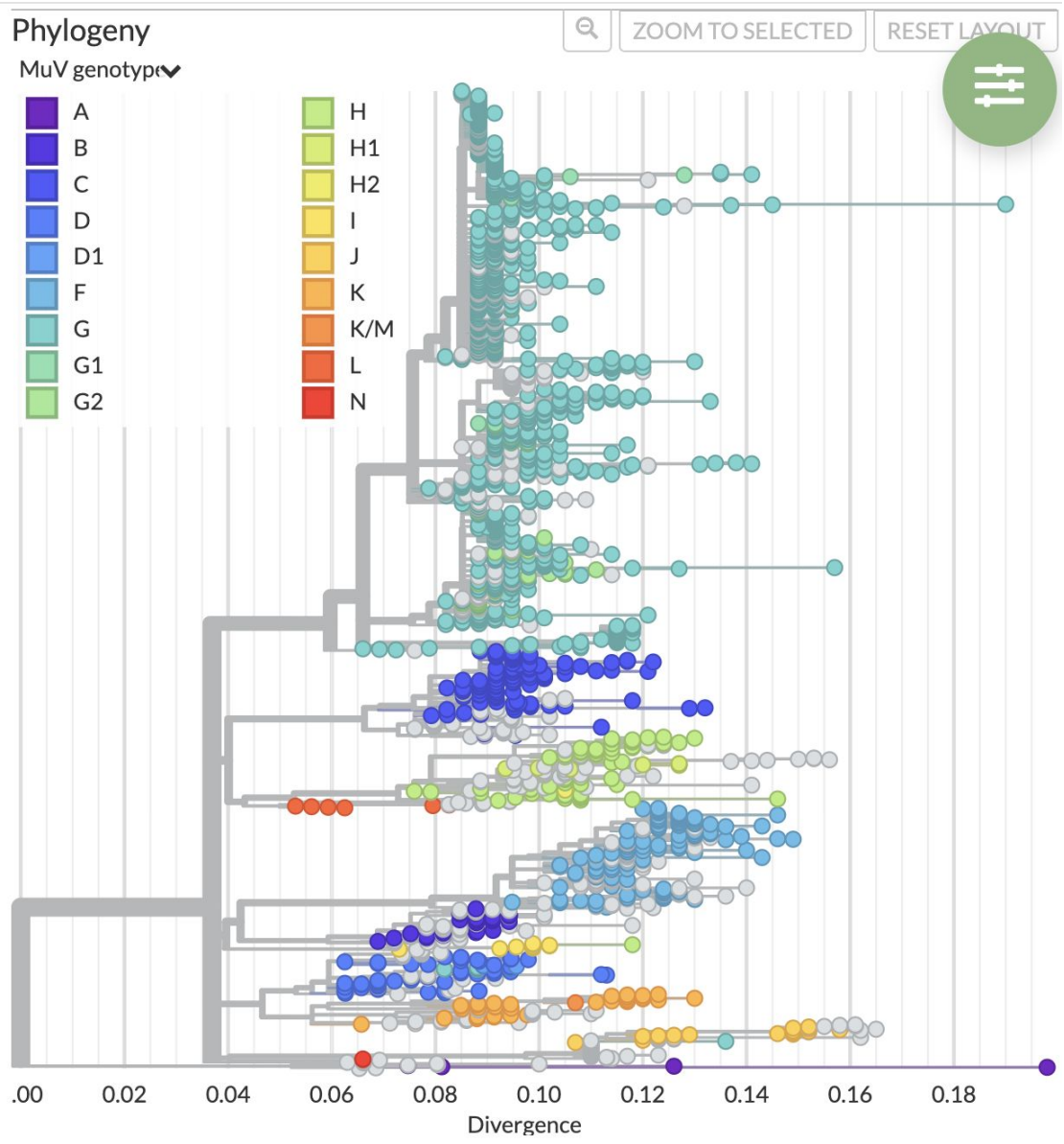


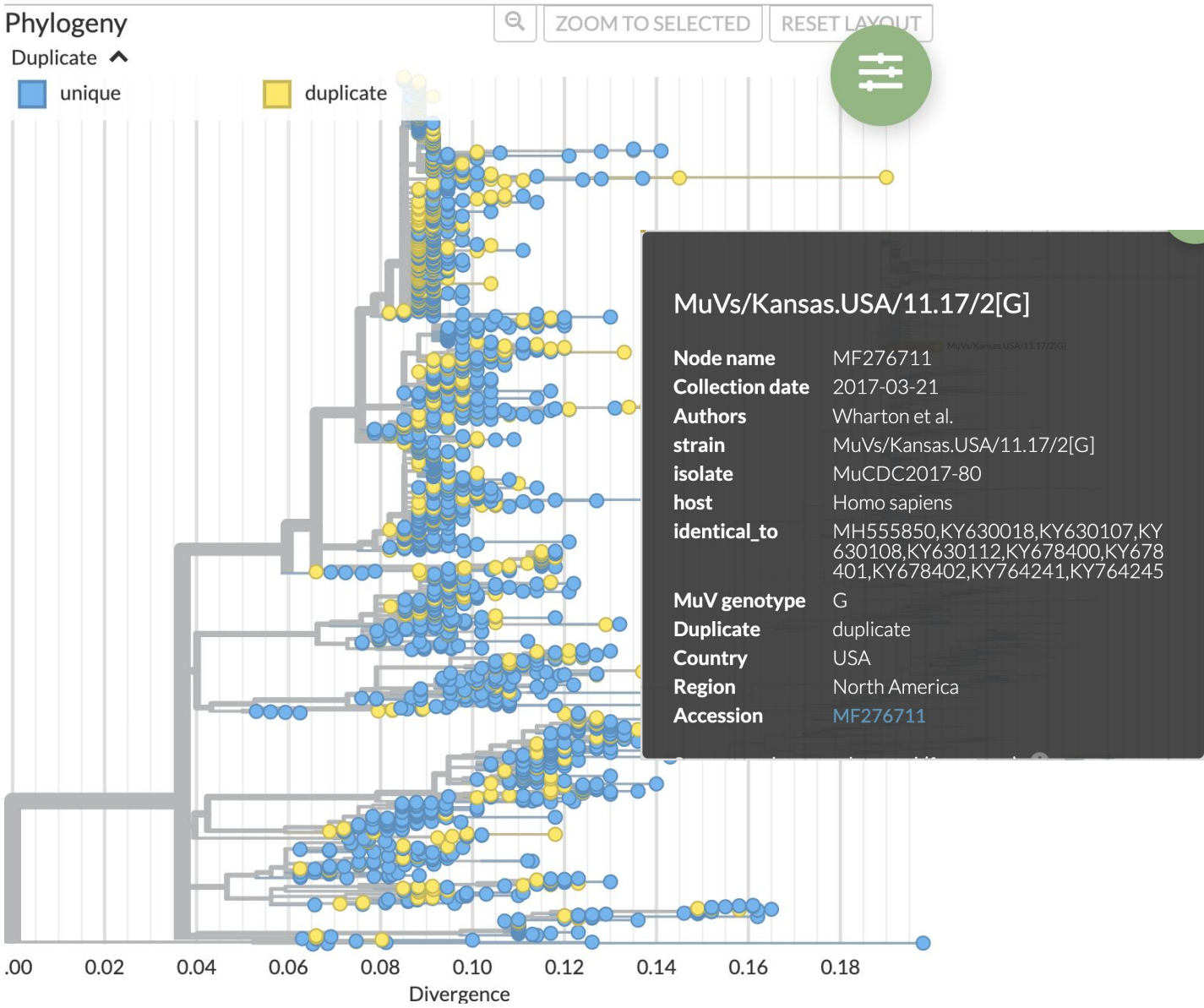
maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

global tree
north-america tree
add frequencies
maybe SH tree

Jin and WHO scaffold strains
Jeryl Lynn SH region reference
adjust nextclade parameters
maybe PR to Nextclade data

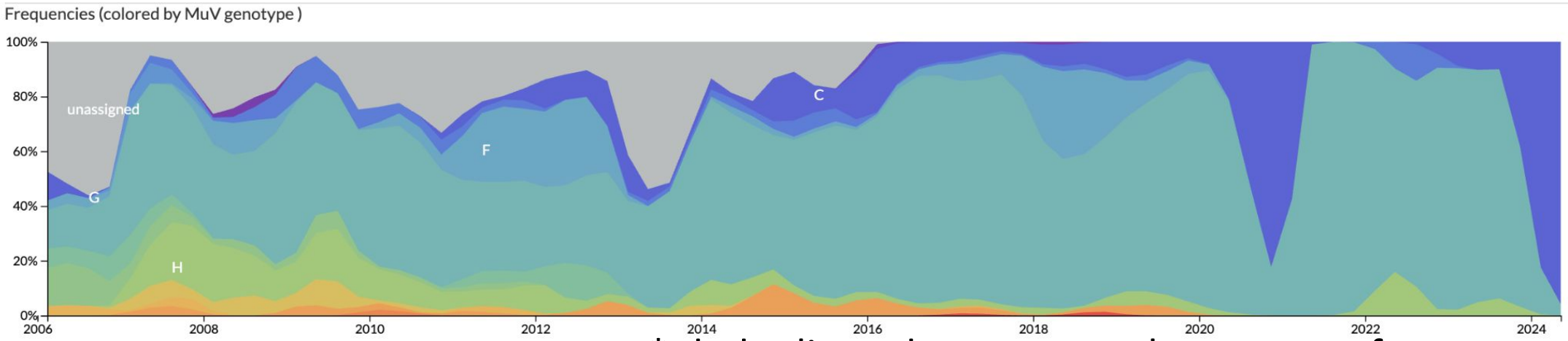
SH Tree - merge identical samples





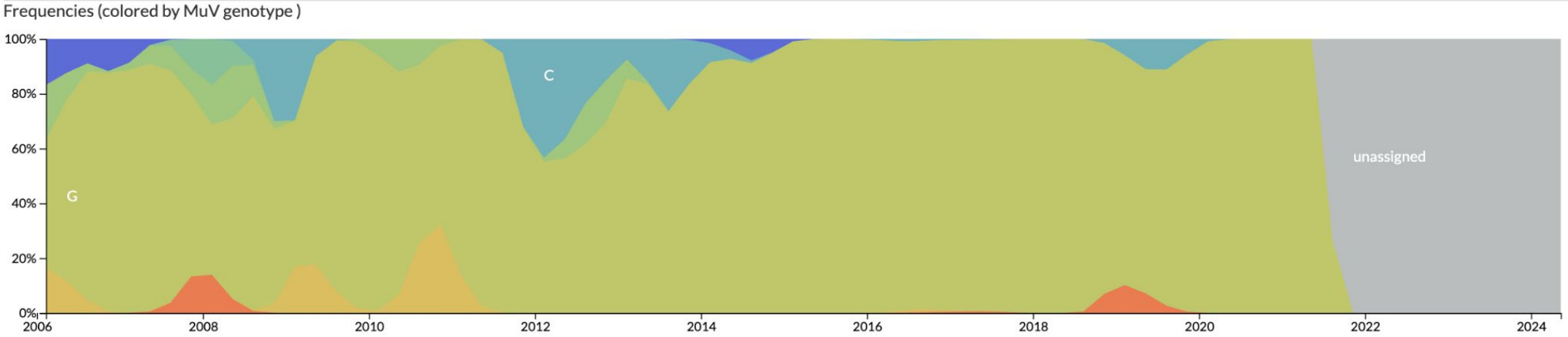
SH Tree - more recent data

SH

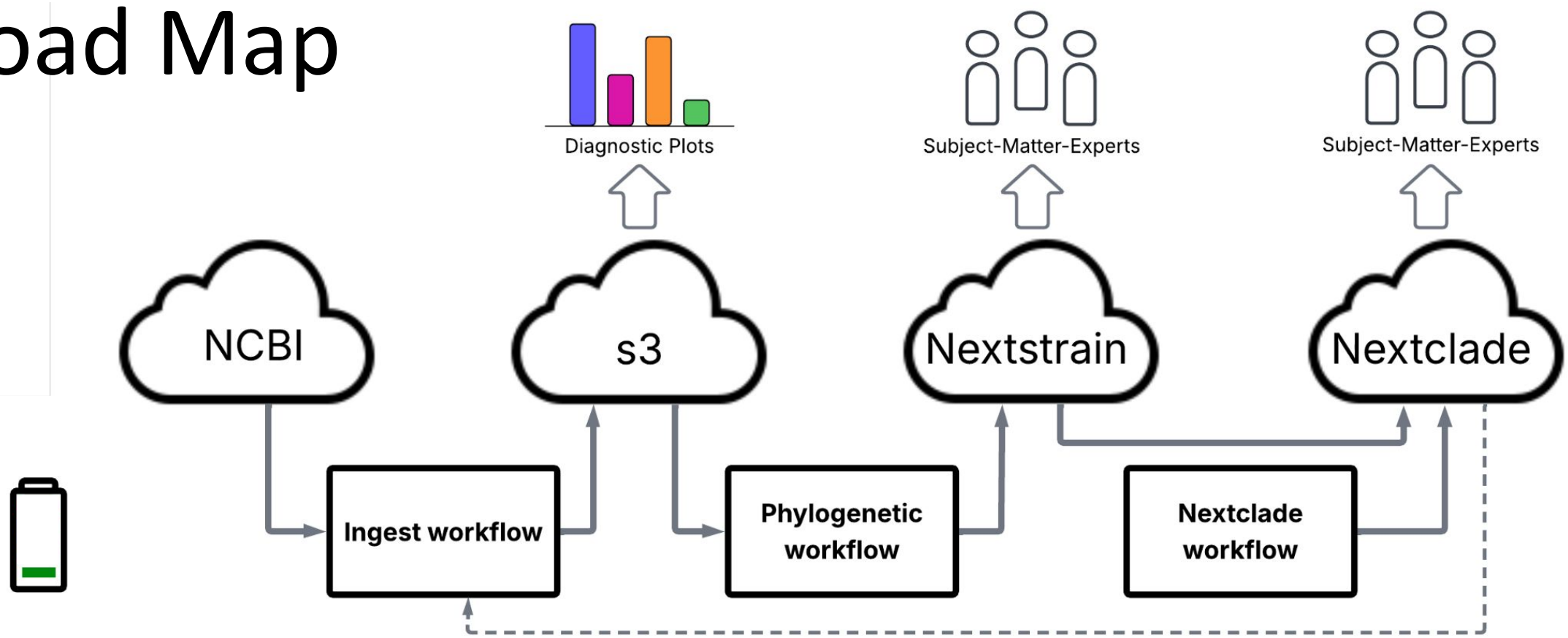


* deduplicated, so may not be accurate frequency

global



Road Map



maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

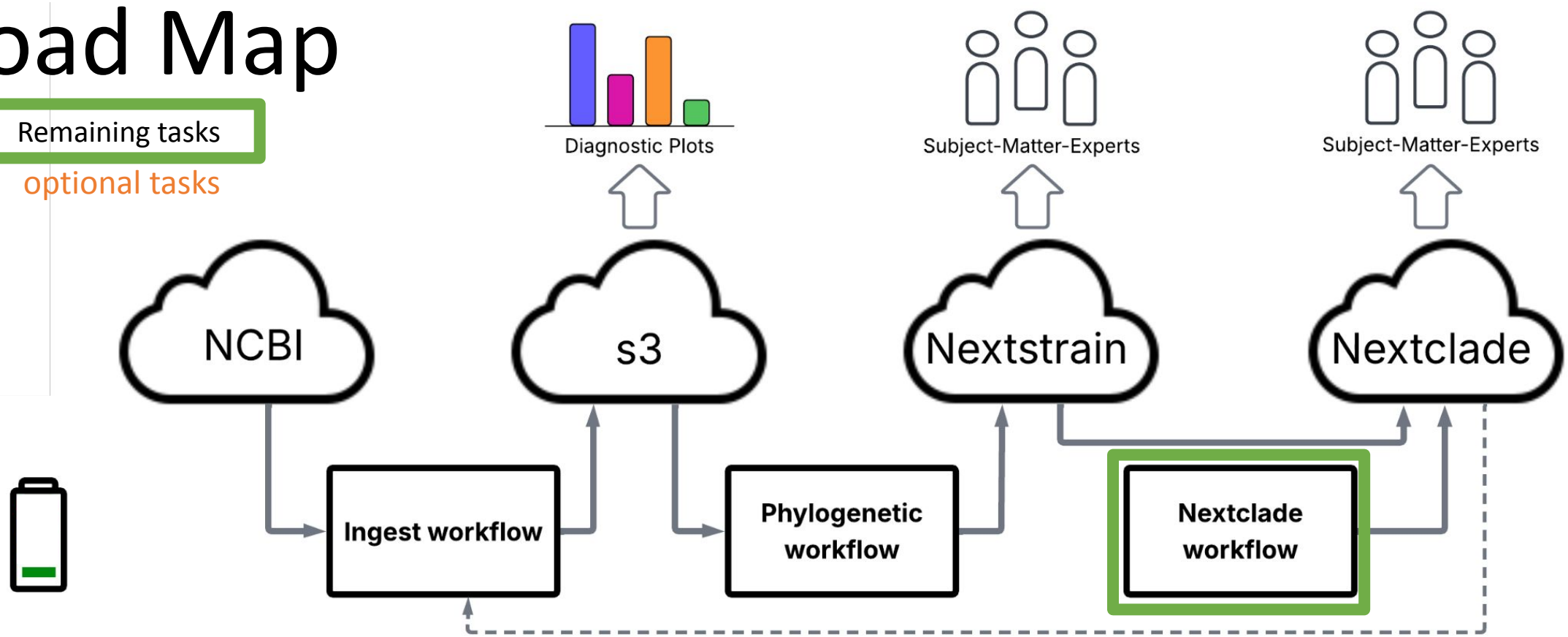
global tree
north-america tree
add frequencies
maybe SH tree

Jin and WHO scaffold strains
Jeryl Lynn SH region reference
adjust nextclade parameters
maybe PR to Nextclade data
maybe HN and full genome datasets

Road Map

Remaining tasks

optional tasks



maybe spike in BC CDC sequences
mumps NCBI Taxon ID: 2560602
parse genotype from Virus Taxon
parse division and genotype from strain
pull strain name -> parse genbank

global tree
north-america tree
add frequencies
maybe SH tree

Jin and WHO scaffold strains
Jeryl Lynn SH region reference
adjust nextclade parameters
maybe PR to Nextclade data
write mumps blog post
HN or full genome Nextclade dataset