EDGE bioinformatics

COVID-19 Genome Analytics

Home Team Additional Resources

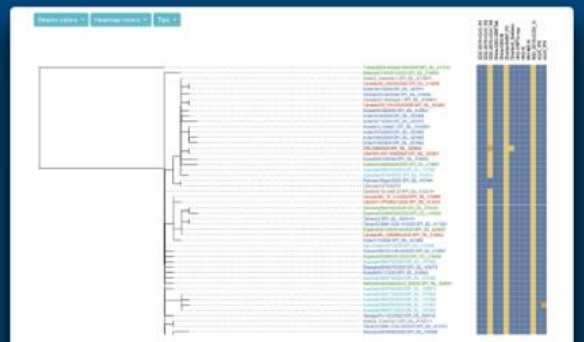
Los Alamos NATIONAL LABORATORY EST. 1943

A platform for COVID-19 analytics

EDGE bioinformatics

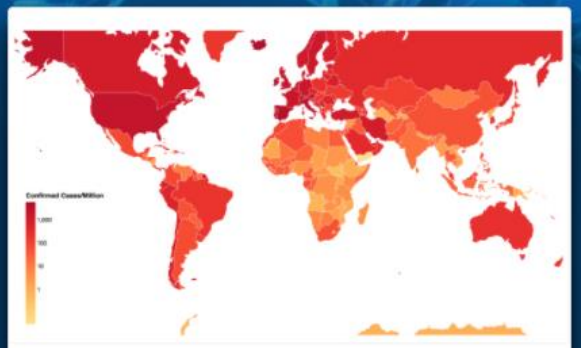
Automated workflow providing SARS-CoV-2 genomes from FASTQ files

EDGE COVID-19



COVID-19 assays screened against available SARS-CoV-2 genomes

Assay Validation



Tracking cases, deaths, and genomes

Case Counts and Genomic Data

covid19.edgebioinformatics.org

genomic STANDARDS consortium

nmdc National Microbiome Data Collaborative

Coronavirus Standards Working Group

A COVID-19 Diagnostic Standards Development Partnership

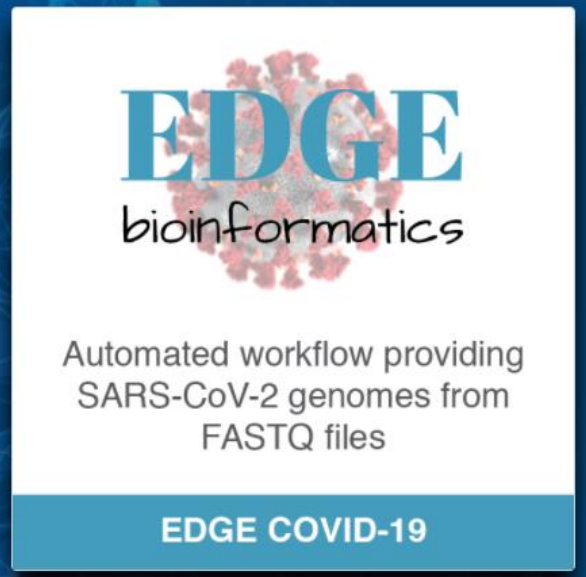
EDGE bioinformatics

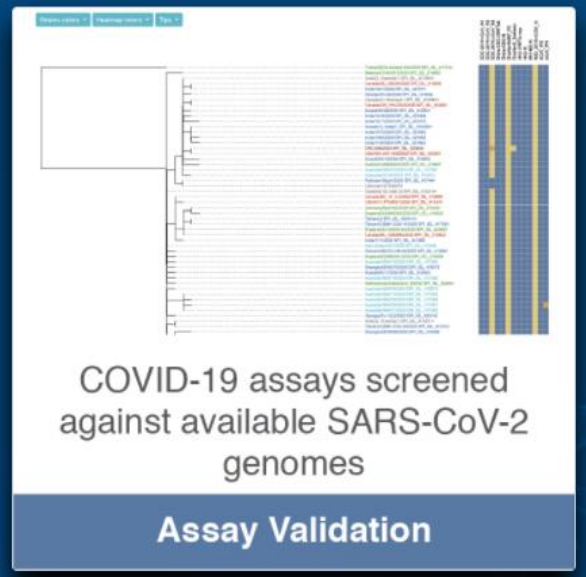
COVID-19 Genome Analytics

Home Team Additional Resources

Los Alamos NATIONAL LABORATORY

A platform for COVID-19 analytics





- As genomics is used for epi and biosurveillance of outbreak pathogens, it can help reveal where diagnostic assays/therapeutics may fail – thus we advocate for:
- 1) robust genomic data to be continually generated (even prior to outbreaks) to inform us of pathogen presence and diversity/evolution; 2) continuous tracking of mutations that may affect diagnostic assays and therapeutic targets; 3) automated re-design of assays and suggestion of alternative targets for therapeutic design

EDGE bioinformatics

COVID-19 Genome Analytics

Home Team Additional Resources

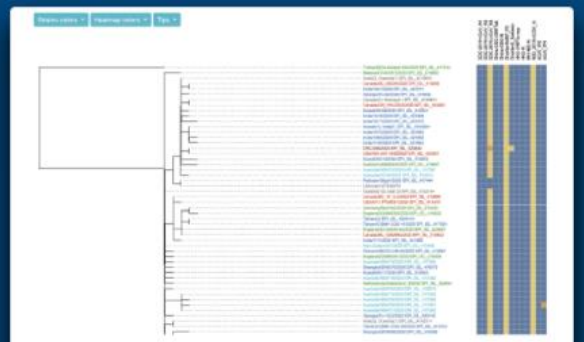
Los Alamos NATIONAL LABORATORY EST. 1943

A platform for COVID-19 analytics

EDGE bioinformatics

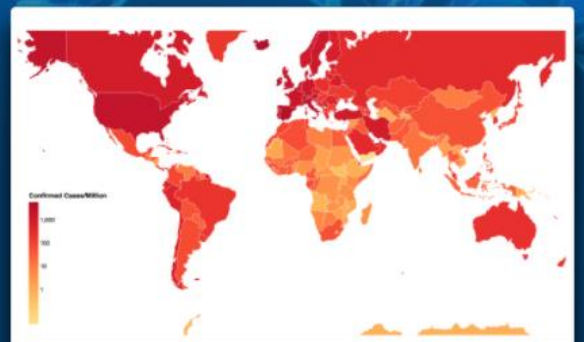
Automated workflow providing SARS-CoV-2 genomes from FASTQ files

EDGE COVID-19



COVID-19 assays screened against available SARS-CoV-2 genomes

Assay Validation



Tracking cases, deaths, and genomes

Case Counts and Genomic Data

Karen Davenport

Mark Flynn

Jason Gans

Adán Myers y Gutiérrez

Bin Hu

Po-e Li

Chien-chi Lo

Elais Player

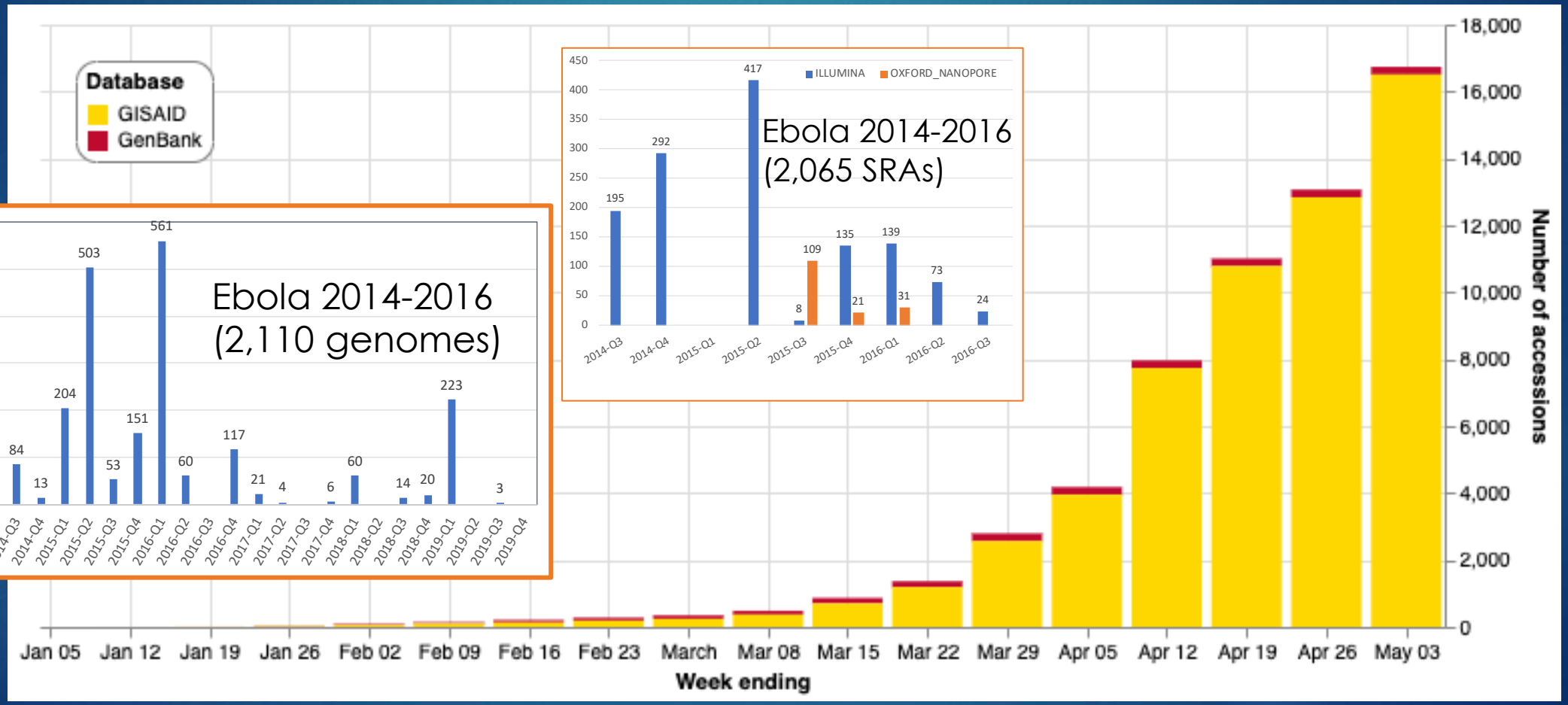
Migun Shakya

Yan Xu

Patrick Chain

Big thanks to Mike Wiley, Jason Ladner, Daryl Domman, Darrell Dinwiddie, Daesang Lee, and others for developing/using/providing feedback on initial workflows

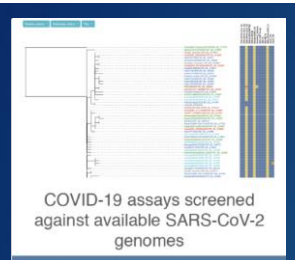
Growth of deposited SARS-CoV-2 genomes in public repos



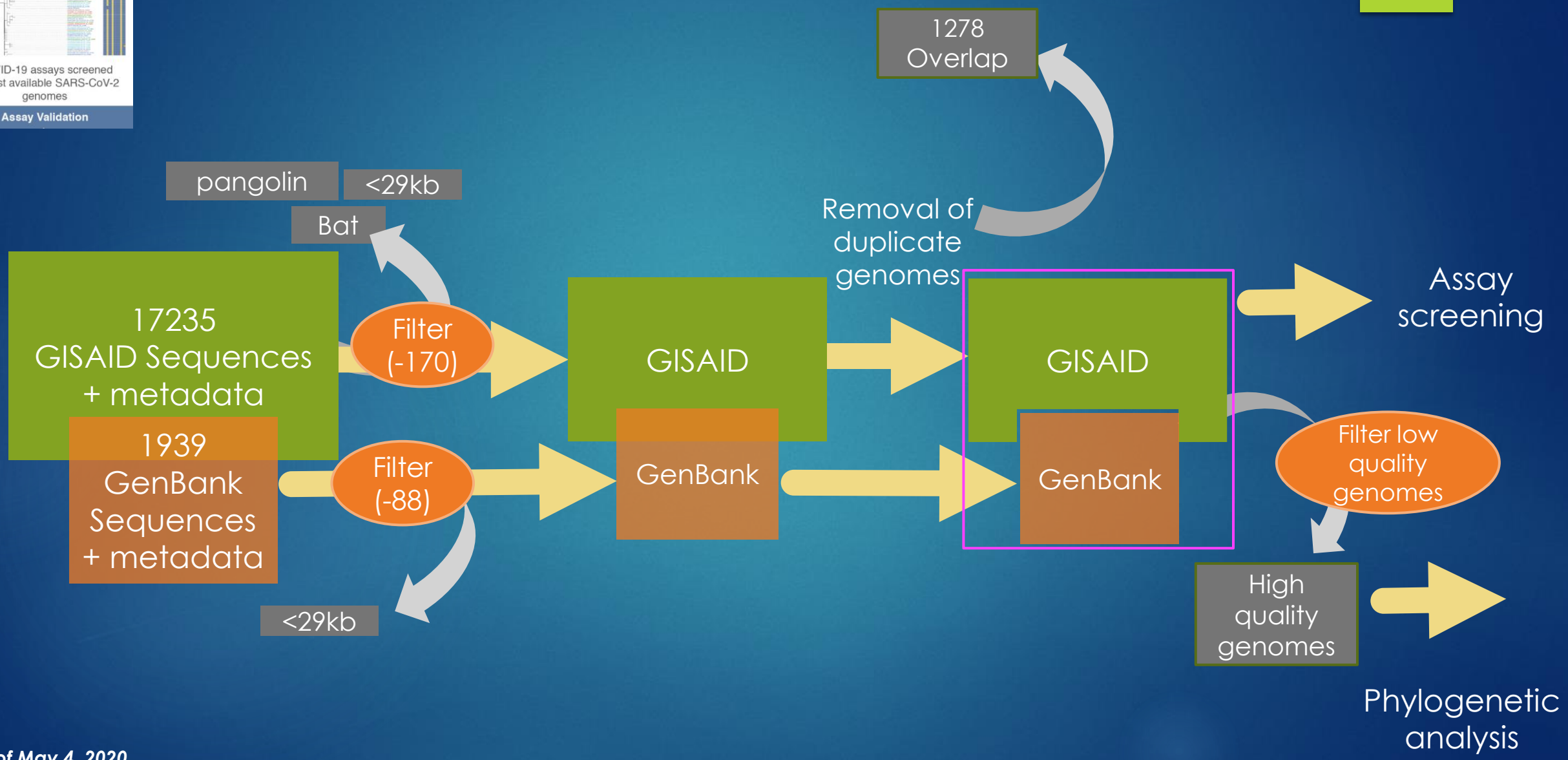
EDGE bioinformatics

COVID-19 Genome Analytics

Los Alamos NATIONAL LABORATORY EST. 1943

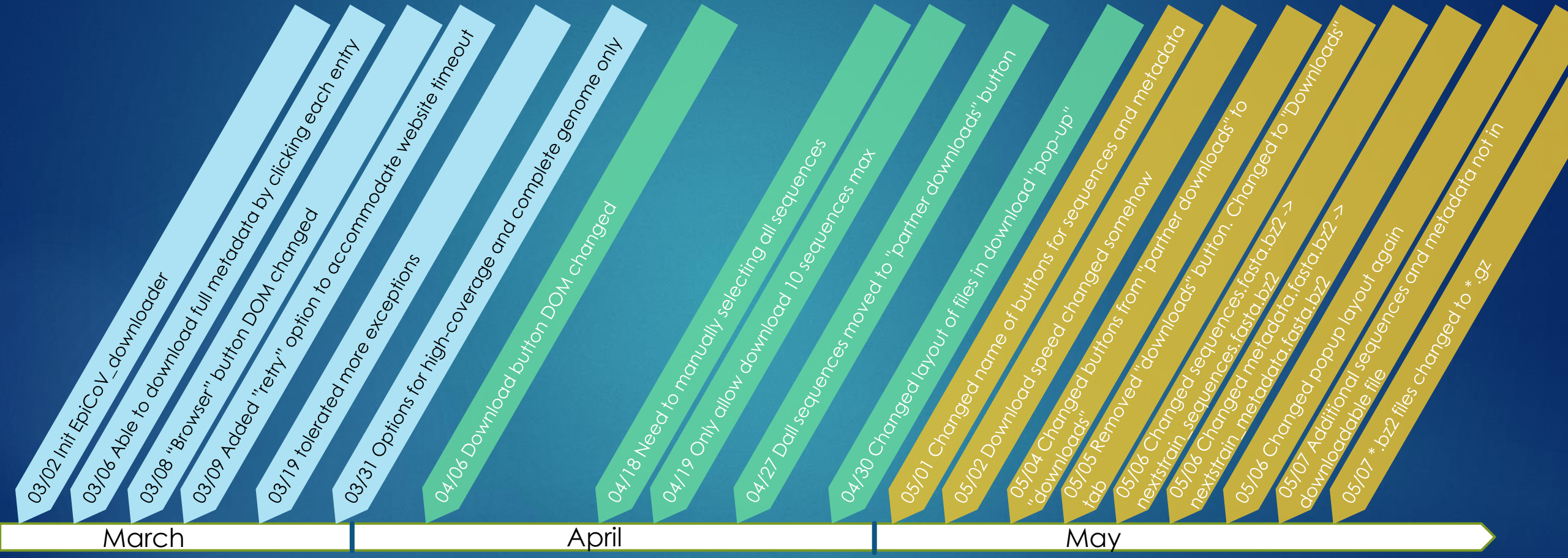


Assay Validation



As of May 4, 2020

Modifications of scraper to accommodate changes on GISAID site

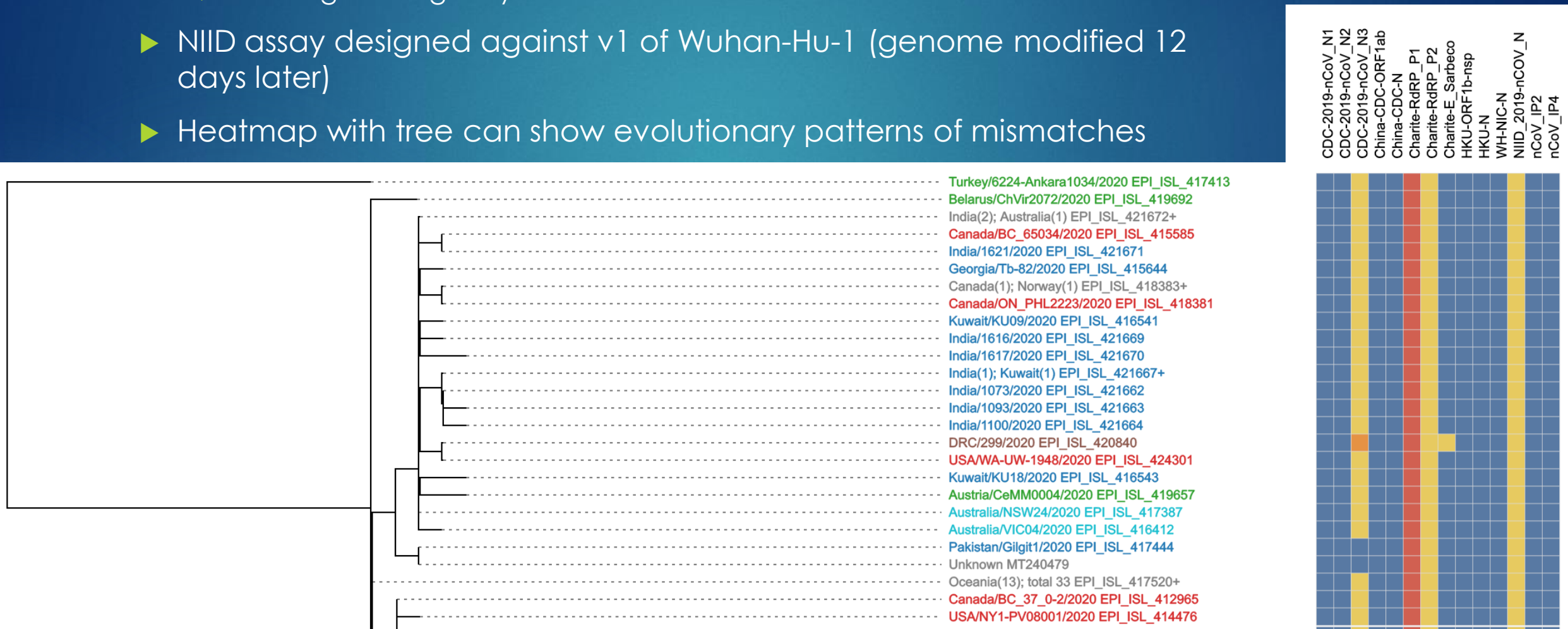


March 2nd – May 7th, 2020

https://github.com/poeli/EpiCoV_downloader

Tree/heatmap view of assay results

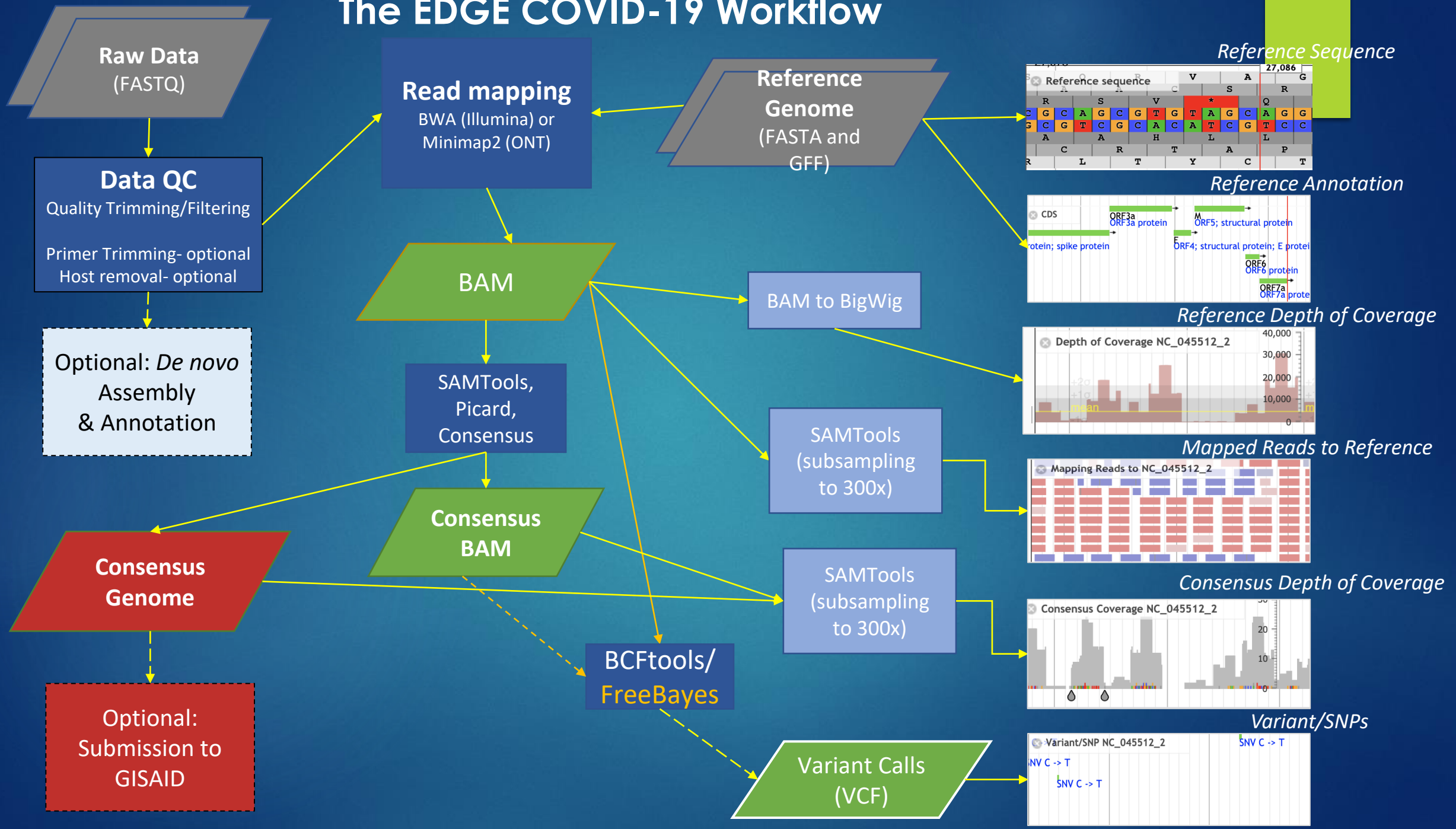
- Charité: probe with two mms (P1), reverse primer with one mm (P2)
  - P2 designed originally for SARS and SARS-like bat coronaviruses
- NIID assay designed against v1 of Wuhan-Hu-1 (genome modified 12 days later)
- Heatmap with tree can show evolutionary patterns of mismatches



Toward needed standards for outbreak genomics

- Poor metadata makes even high quality genomes less relevant
- Issues with genomes can mislead – including evolution/diversity, diagnostics, etc.
- Many sequencing approaches/platforms – and for any one:
  - many different pipelines, tools, parameters/cutoffs, decision trees – results in less accurate interpretation of results
- Difficult for a number of labs to run any local analyses
- Difficult to obtain a high quality genome and to submit it to a public repo
  - What to do with gaps, low coverage, quasispecies variants, etc.

The EDGE COVID-19 Workflow



The EDGE COVID-19 Workflow

EDGE bioinformatics

Automated workflow providing SARS-CoV-2 genomes from FASTQ files

EDGE COVID-19

Raw Data (FASTQ)

Reference Genome

QC

FaQCs
Porechop + Nanoplot

Options:

- Primer Trimming (eg. ARTIC)
- Human host removal

Optional: *De novo* Assembly & Annotation

Reads Mapping

BWA mem
minimap2

BAM

Consensus Genome Generation

Consensus Genome (FASTA)

Optional: Submission to GISAID

Variant Calling

BCFtools/VCFutils

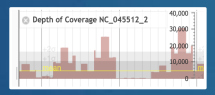
- Mapping quality [60];
- max_depth [300x];
- minimum read depth [5];
- minimum number of alternate bases [3];
- minimum ratio of alternate bases [0.3];
- SNP within INT bp around a gap to be filtered [3];
- window size for filtering adjacent gaps [10];
- min P-value for end distance bias [0.0001];
- maximum fraction of reads supporting an indel [0.5];
- min P-value for strand bias [1e-10];

JBrowse

Reference genome

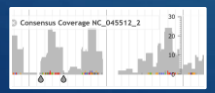
Reference annotation

Coverage track



Mapped reads

Consensus coverage



The EDGE COVID-19 Workflow

EDGE bioinformatics

Automated workflow providing SARS-CoV-2 genomes from FASTQ files

EDGE COVID-19

Raw Data (FASTQ)

Reference Genome

QC

FaQCs
Porechop + Nanoplot
Options:
- Primer Trimming (eg. ARTIC)
- Human host removal

Optional: *De novo* Assembly & Annotation

Reads Mapping

BWA mem
minimap2

BAM

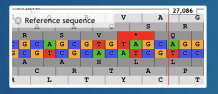
Consensus Genome Generation

Consensus Genome (FASTA)

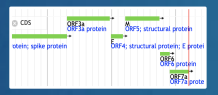
Optional: Submission to GISAID

JBrowse

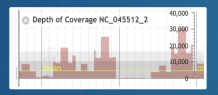
Reference genome



Reference annotation



Coverage track



Mapped reads

Consensus coverage

Variant Calling

BCFtools/VCFutils
- Mapping quality [60];
- max_depth [300x];
- minimum read depth [5];
- minimum number of alternate bases [3];
- minimum ratio of alternate bases [0.3];
- SNP within INT bp around a gap to be filtered [3];
- window size for filtering adjacent gaps [10];
- min P-value for end distance bias [0.0001];
- maximum fraction of reads supporting an indel [0.5];
- min P-value for strand bias [1e-10];

Consensus Genome Generation

Filter reads

SAMtools/Picard/Consensus
1. PCR duplicate removal
2. Mapping Quality (< 60)
3. Base Quality (<5 ONT; <20 Illumina)
4. BAQ (Illumina)

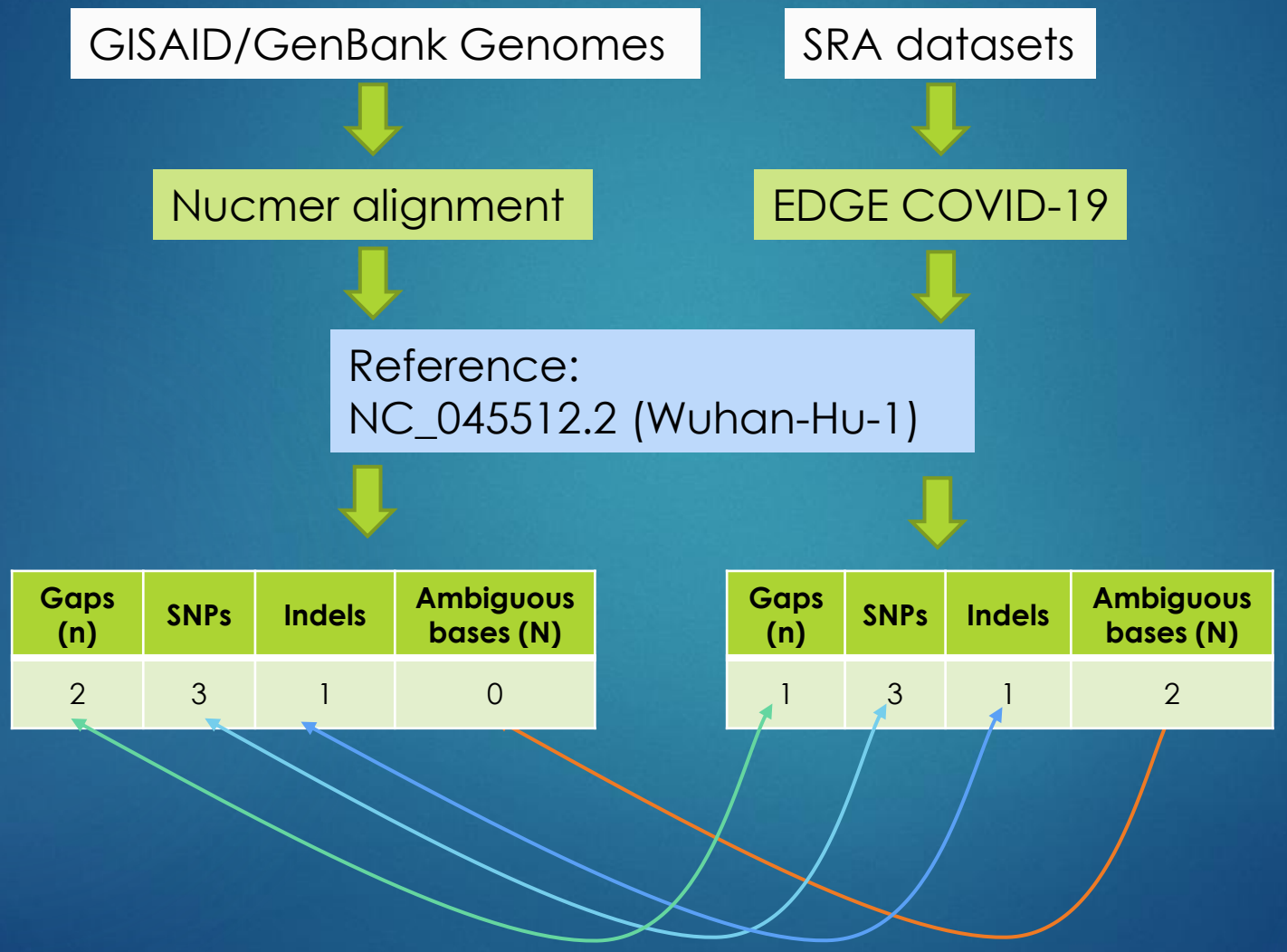
--> consensus BAM

0 reads mapped → Gap → n

Min. coverage > 5
Alt. base/deletion ratio > 0.5 → SNPs, indels → A T G C

Min. coverage < 5
Max alt. base ratio < 0.5 → Ambiguous → N

Comparing EDGE COVID-19 results to deposited genomes



Gaps (n)	SNPs	Indels	Ambiguous bases (N)
2	3	1	0

Gaps (n)	SNPs	Indels	Ambiguous bases (N)
1	3	1	2

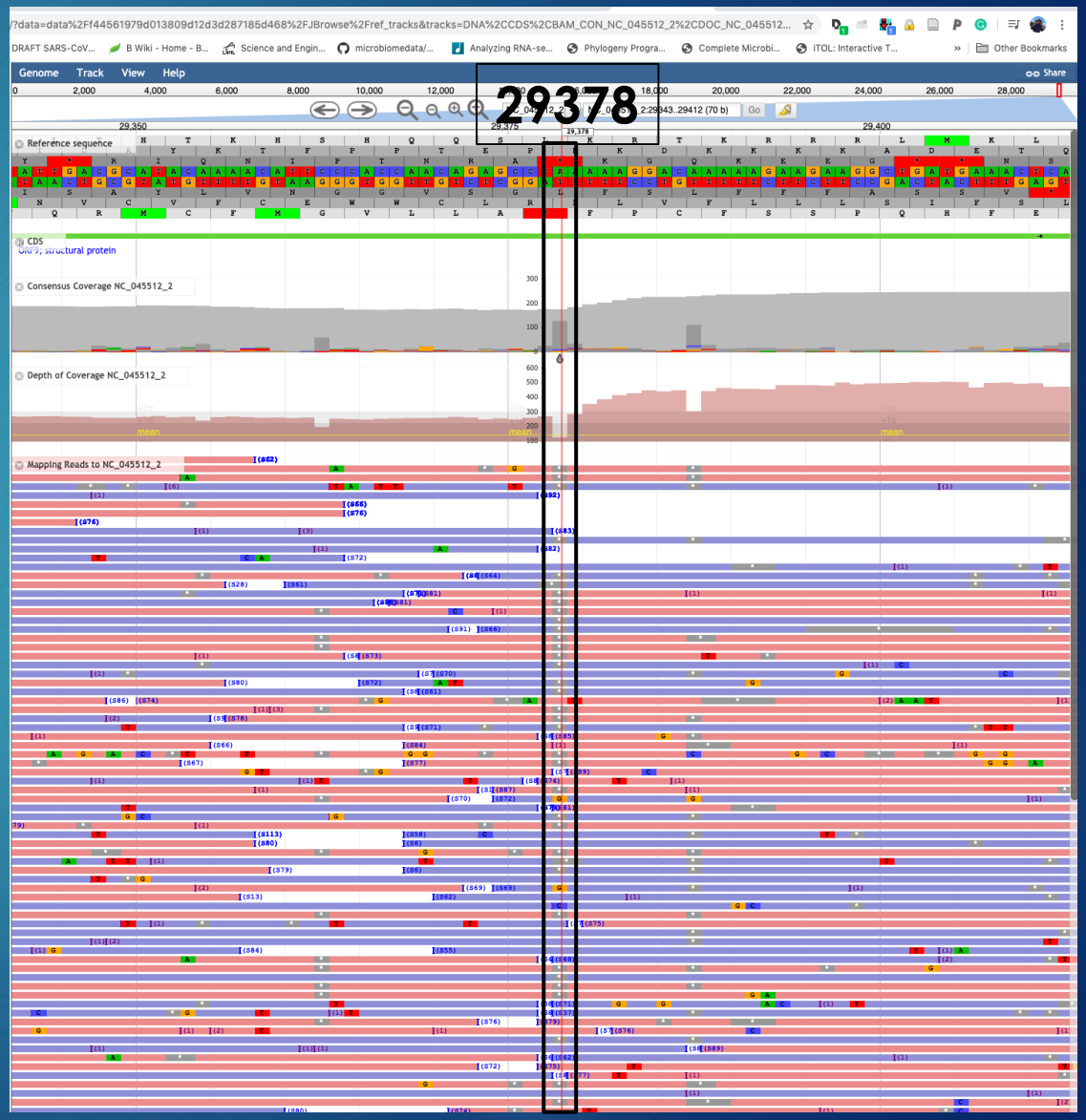
Identical SNPs in EDGE COVID-19 and gisaid/genbank genomes

EDGE Project Name	SRR	tech	SNPs (gisaid/genbank)	SNPs (EDGE COVID-19)	SNPs match?
FDA_ILLUMINA_SHOTGUN	SRR11393704	Illumina	3	3	✓
FDA_ILLUMINA_CAPTURE	SRR11409417	Illumina	3	3	✓
NEPAL-61_ILLUMINA_PCR	SRR11177792	Illumina	1	1	✓
TIGER_NY_ILLUMINA_ARTIC	SRR11587600	Illumina	6	6	✓
USA_WI1_ONT_vero76	SRR11140749	ONT	1	1	✓
HKU-902a_ONT_SHOTGUN	SRR11178057	ONT	1	1	✓
WA-0711_ONT_PCR	SRR11637325	ONT	5	5	✓
VIC07_ONT_ARTICv1	SRR11397722	ONT	5	5	✓

EDGE Project Name	SRR	platform	indels (gisaid/genbank)	indels (EDGE COVID-19)	indels match?
FDA_ILLUMINA_SHOTGUN	SRR11393704	Illumina	0	0	✓
FDA_ILLUMINA_CAPTURE	SRR11409417	Illumina	0	0	✓
NEPAL-61_ILLUMINA_PCR	SRR11177792	Illumina	0	0	✓
TIGER_NY_ILLUMINA_ARTIC	SRR11587600	Illumina	0	0	✓
USA_WI1_ONT_vero76	SRR11140749	ONT	1	1	✓
HKU-902a_ONT_SHOTGUN	SRR11178057	ONT	0	1	✗
WA-0711_ONT_PCR	SRR11637325	ONT	0	0	✓
VIC07_ONT_ARTICv1	SRR11397722	ONT	0	0	✓

Differences in indel between EDGE COVID-19 and GISAID/GENBANK

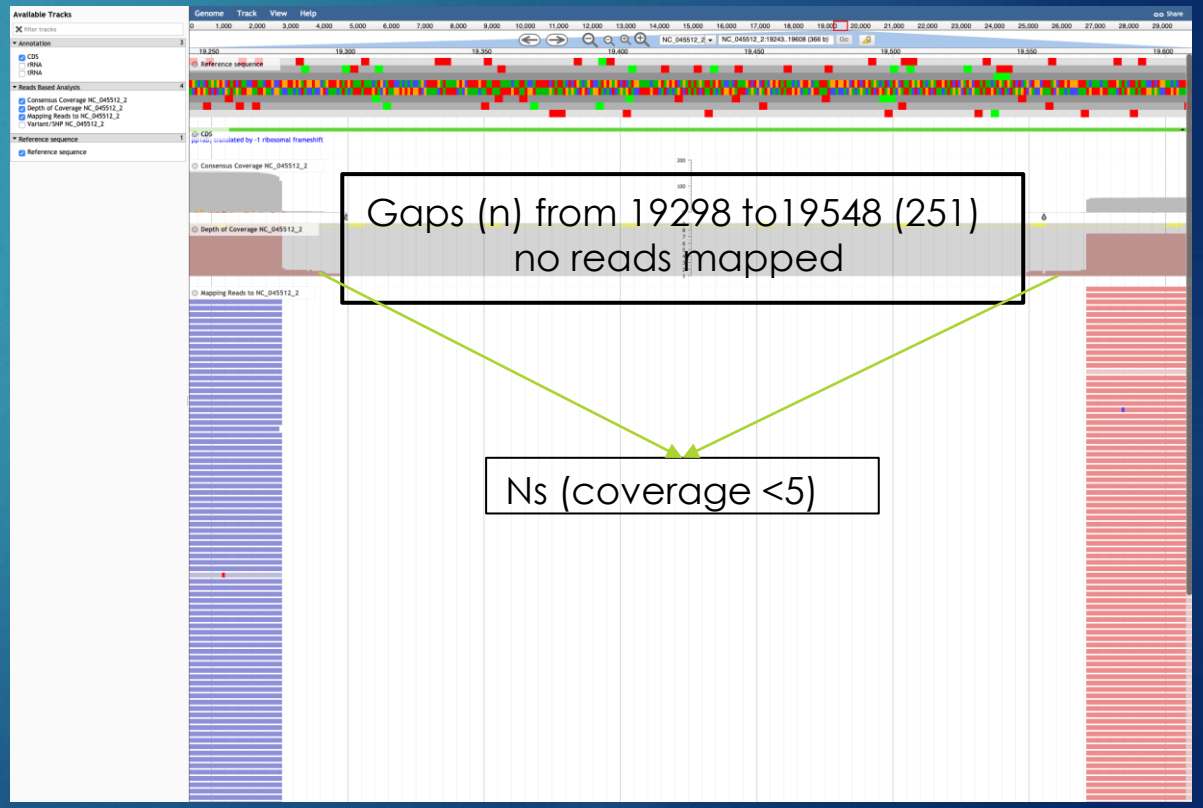
- In sample HKU-902a
  - EDGE COVID-19 added one indel
  - GISAID genome (EPI_ISL_434563) did not have any indels
- Deletion at position 29378
- 54% of 164X coverage show deletion events



Differences in Ambiguous bases (Ns) and gaps (ns)

- In WGS of SARS-CoV-2 from tiger
  - EDGE COVID-19 genome has **174nt Ns** and **268nt gaps** within the genome.
    - 174nt positions with coverage < 5
    - 251nt in 1 gap had 0 reads mapped
  - GISAID genome (EPI_ISL_420293) has **0 Ns** and **0 gaps**

- Why is the difference?
  - Threshold coverage less than five?
  - Gaps filled with reference genome?
    - 251 gap region nucleotide in GISIAD genome matches reference (NC_045512_2) 100%



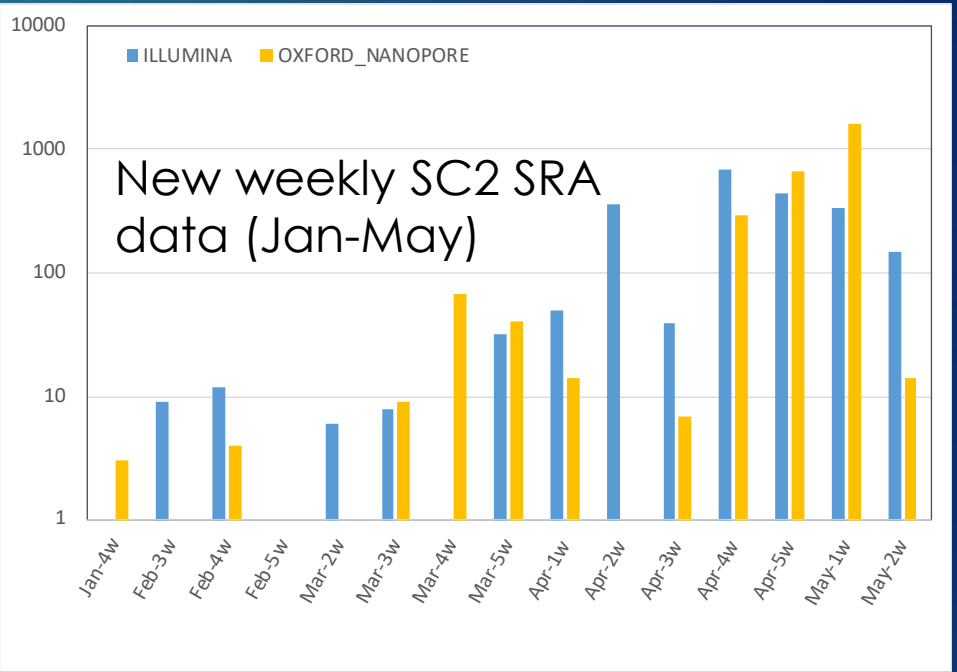
EDGE bioinformatics

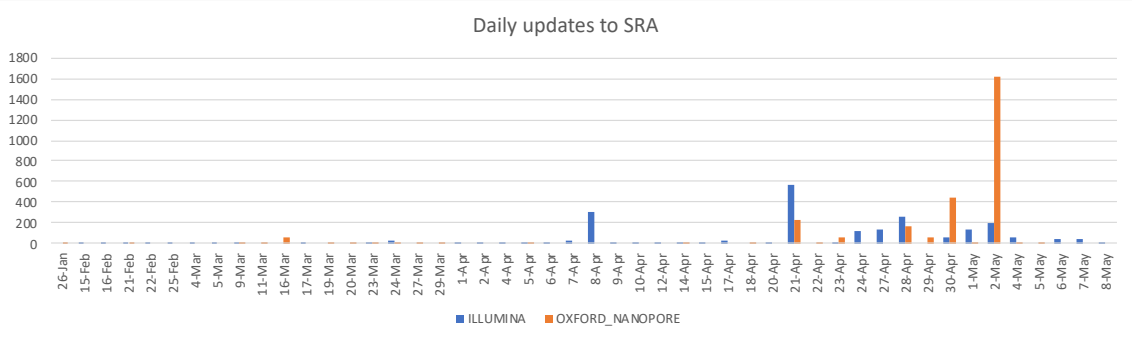
COVID-19 Genome Analytics

Los Alamos NATIONAL LABORATORY EST. 1943

ReleaseDate 16	Sample Name 17	SRA Study 18	Library Name 19	Collection_Date 20
2020-04-18	SARS-CoV-2/Valencia003/human/2020/ES	ERP120836		
2020-04-18	SARS-CoV-2/Valencia004/human/2020/ES	ERP120836		
2020-04-18	SARS-CoV-2/Valencia005/human/2020/ES	ERP120836		
2020-04-18	SARS-CoV-2/Valencia008/human/2020/ES	ERP120836		
2020-04-18	SARS-CoV-2/Valencia006/human/2020/ES	ERP120836		
2020-04-18	SARS-CoV-2/Valencia007/human/2020/ES	ERP120836		
2020-04-30	DK/ALAB-HH-08/2020	ERP121327	DK/ALAB-HH-08/2020	2020-03-17
2020-04-30	DK/ALAB-HH-11/2020	ERP121327	DK/ALAB-HH-11/2020	2020-03-18
2020-04-30	DK/ALAB-HH-13/2020	ERP121327	DK/ALAB-HH-13/2020	2020-03-19
2020-04-30	DK/ALAB-HH-20/2020	ERP121327	DK/ALAB-HH-20/2020	2020-03-26
2020-04-30	DK/ALAB-HH-66/2020	ERP121327	DK/ALAB-HH-66/2020	2020-04-29
2020-04-30	DK/ALAB-HH-84/2020	ERP121327	DK/ALAB-HH-84/2020	2020-05-12
2020-04-30	DK/ALAB-HH-86/2020	ERP121327	DK/ALAB-HH-86/2020	2020-05-13
2020-04-30	DK/ALAB-SSI-108/2020	ERP121327	DK/ALAB-SSI-108/2020	2020-05-27
2020-04-30	DK/ALAB-SSI-109/2020	ERP121327	DK/ALAB-SSI-109/2020	2020-05-28
2020-04-30	DK/ALAB-SSI-129/2020	ERP121327	DK/ALAB-SSI-129/2020	2020-06-12
2020-04-30	DK/ALAB-SSI-132/2020	ERP121327	DK/ALAB-SSI-132/2020	2020-06-13
2020-04-30	DK/ALAB-SSI-158/2020	ERP121327	DK/ALAB-SSI-158/2020	2020-06-30
2020-04-30	DK/ALAB-SSI-163/2020	ERP121327	DK/ALAB-SSI-163/2020	2020-07-04
2020-04-30	DK/ALAB-SSI-164/2020	ERP121327	DK/ALAB-SSI-164/2020	2020-07-05
2020-04-30	DK/ALAB-SSI-166/2020	ERP121327	DK/ALAB-SSI-166/2020	2020-07-06
2020-04-30	DK/ALAB-SSI-167/2020	ERP121327	DK/ALAB-SSI-167/2020	2020-07-07
2020-04-30	DK/ALAB-SSI-395/2020	ERP121327	DK/ALAB-SSI-395/2020	2020-10-26

Matching genomes to SRA





True disease forecasting??

COVID-19 Genome Analytics

- Genomic data as of today
  - # of genomes in GISAID = 23,381
  - # of genomes* in GenBank = 2,435
  - # of SRA experiments = 5,385
- Most of these data are connected somehow, but can we connect them?
  - No specific feature in SRA experiment records that indicate if the genome has been deposited to either genbank or gisaid.
  - Best way is matching the *Library Name* with GISAID metadata, but not always consistent.

SRX8255490: Severe acute respiratory syndrome coronavirus 2
1 ILLUMINA (NextSeq 550) run: 362,448 spots, 106.7M bases, 39.5Mb downloads

Design: ARTIC V3 amplicons, Nextera XT library, minimap2 v2.17, ivar v1.2.1, samtools v1.10. Using minimap2, short reads mapped to SARS-CoV-2 NCBI accession MN908947.3. Using samtools, proper_pairs (samflag 2) mapping to MN908947.3 retained, unmapped reads (samflag 4) discarded (to filter out non-SARS-CoV-2 cDNA). Filtered reads submitted to NCBI

Submitted by: The Peter Doherty Institute for Infection and Immunity

Study: Severe acute respiratory syndrome coronavirus 2 Genome sequencing
PRJNA613958 • SRP253798 • All experiments • All runs
show Abstract

Sample: SARS-Cov-2 VIC1273
SAMN14839002 • SRS6598939 • All experiments • All runs
Organism: Severe acute respiratory syndrome coronavirus 2

Library:
Name: VIC1273_illumina
Instrument: NextSeq 550
Strategy: AMPLICON
Source: VIRAL RNA
Selection: PCR
Layout: PAIRED

Runs: 1 run, 362,448 spots, 106.7M bases, 39.5Mb

Run	# of Spots	# of Bases	Size	Published
SRR11695894	362,448	106.7M	39.5Mb	2020-05-06

Full

SRX8264257: Sample 32
1 ILLUMINA (Illumina MiSeq) run: 555,303 spots, 164.6M bases, 83Mb downloads

Design: mutation detection

Submitted by: Paragon Genomics

Study: High sensitivity detection of SARS-CoV-2 using multiplex PCR and a multiplex-PCR-based metagenomic method
PRJNA614546 • SRP253783 • All experiments • All runs
show Abstract

Sample: CHLA9
SAMN14829711 • SRS6602226 • All experiments • All runs
Organism: Severe acute respiratory syndrome coronavirus 2

Library:
Name: Sample 32
Instrument: Illumina MiSeq
Strategy: AMPLICON
Source: VIRAL RNA
Selection: PCR
Layout: PAIRED

Runs: 1 run, 555,303 spots, 164.6M bases, 83Mb

Run	# of Spots	# of Bases	Size	Published
SRR11704822	555,303	164.6M	83Mb	2020-05-06

ID: 10766238