Rigorous benchmarking of methods for SARS-CoV-2 lineage detection in wastewater

Bohdan Tyshchenko

Mentors:
Seghei Mangul (USC)
Sergey Knyazev (UCLA)
Alina Frolova (IMBG)

Mangul Lab
University of Southern California

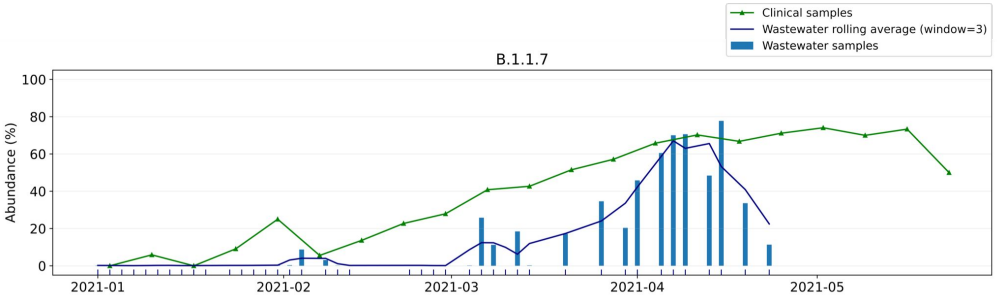
Introduction. Wastewater monitoring

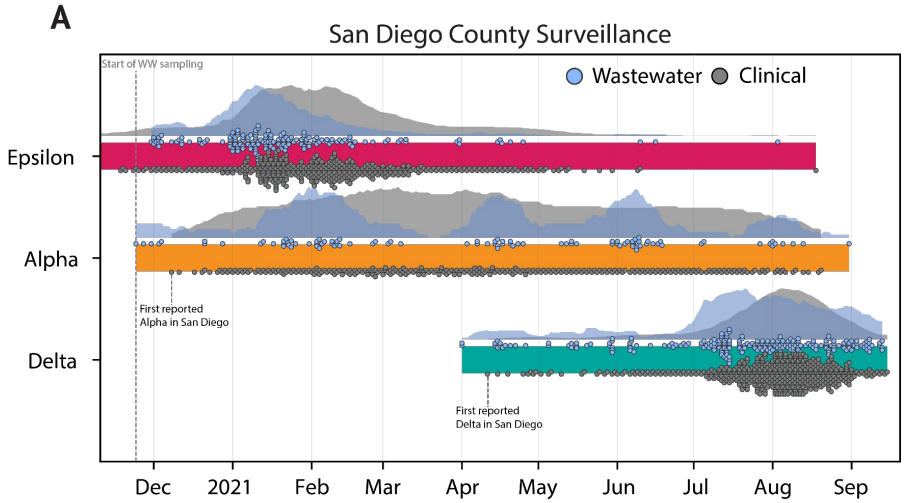
Successes:

- Identifying trends in disease spread
- Early variant detection

Issues:

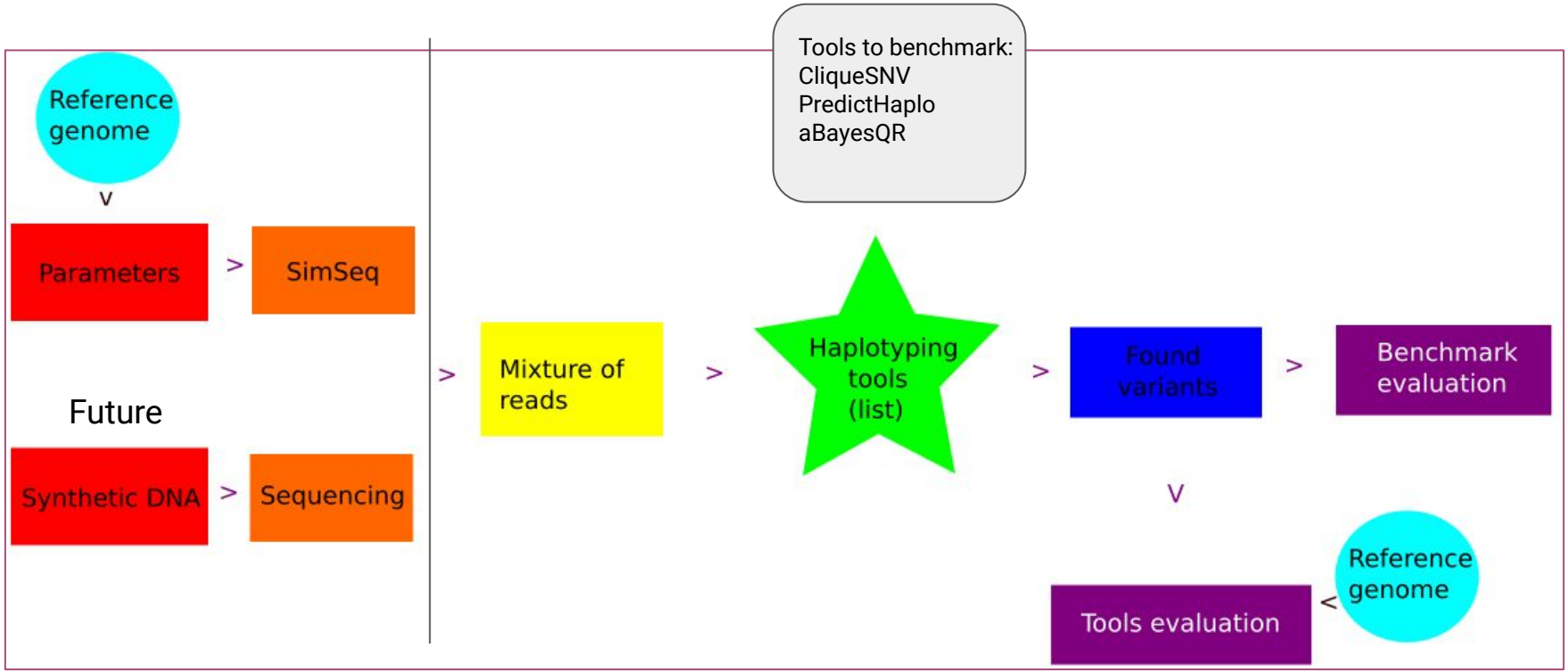
- Difficulties identifying lineages at low prevalence
- Rare lineages observed were not seen in clinical samples
- Differences in sequencing protocols lead to different detection results





Images: Karthikeyan, Smruthi & Levy, Joshua & Hoff. *et al.* Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. (2021). https://doi.org/10.1101/2021.12.21.21268143.

Proposed plan



Pipeline

Setup environments for tools

Generate sequencing reads (SimSeq, Python)

Align reads (bwa tool)

Run tools, saving outputs (Bash)

Align predicted variants with reference (ClustalW)

Picard

SAMtools

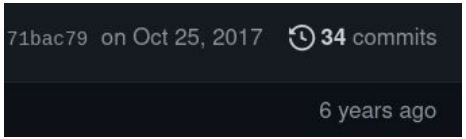
bwa - Burrows-Wheeler Alignment Tool

CLUSTAL

ANACONDA.

Challenges

- Poorly supported tools



- Remote

Results

deepwebhoax / **benchmarking-lineage-detection** Public

Languages

Python 53.4% Shell 46.6%

fullen_5_default_50000_ReadGenerator.sam	Added datasets
fullen_5_ratios_50000_ReadGenerator.sam	Added datasets
spikes_5_default_50000_ReadGenerator.sam	Added datasets
spikes_5_ratios_10000_ReadGenerator.sam	Added datasets
spikes_5_ratios_20000_ReadGenerator.sam	Added datasets
spikes_5_ratios_40000_ReadGenerator.sam	Added datasets
spikes_5_ratios_50000_ReadGenerator.sam	Added datasets
spikes_5_ratios_50000_ReadGenerator_s.sam	Added datasets

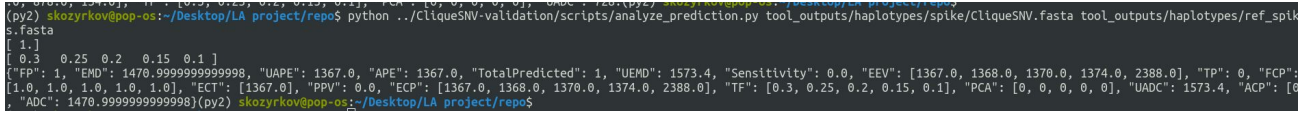
align.sh	MOD: better output filenames
config.yaml	ADD: snakemake test version
gen_spike.py	minor improvements
ham.py	hamming distance calculator
mix6.fasta	ADD: CliequeSNV results, envi setup and WIV04 reference
ref_trim.py	minor improvements
session_setup.sh	ADD: CliequeSNV results, envi setup and WIV04 reference
sim.py	minor improvements
simulate_variants.sh	ADD: simulation script
snakefile	ADD: snakemake test version

Sergey-Knyazev / **CliqueSNV-validation** Public

vtsyvina / **CliqueSNV** Public

Results

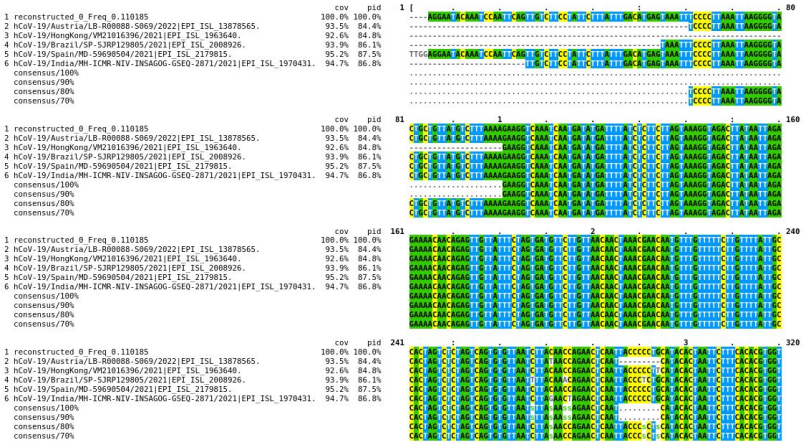
Getting scores



	CliqueSNV				PredictHaplo				aBayesQR			
	Precision	Recall	Variants detected	EMD	Precision	Recall	Variants detected	EMD	Precision	Recall	Variants detected	EMD
Spike Mixture	0	0	1	1470	0	0	7	2047	0	0	2	802
Full genome Mixtures	0	0	1	14977	0	0	5	21542	0	0	3	1204

Future work

- Use multiple-sequence alignment to improve scores
- Benchmark other tools
- Pipeline tool reproducible and expandable



Thank you!