Public Health Agency of Canada | Agence de la santé publique du Canada

Canada

SARS-CoV-2 Recombinant Detection Tutorial

Katherine Eaton

2022 April 27

PROTECTING AND EMPOWERING CANADIANS
TO IMPROVE THEIR HEALTH

Overview

- **Tutorial Zip**
  - data : sequences and metadata
  - resources: metadata of pango-designation issues
  - results: output of detection tools
  - slides: presentation slides

- **Tool Demo* and Comparison (15 min)**
  - Nextclade: Identify putative recombinants.
  - sc2rf (“scarf”): Plot recombination breakpoints.
  - UShER: Refine lineage assignments.

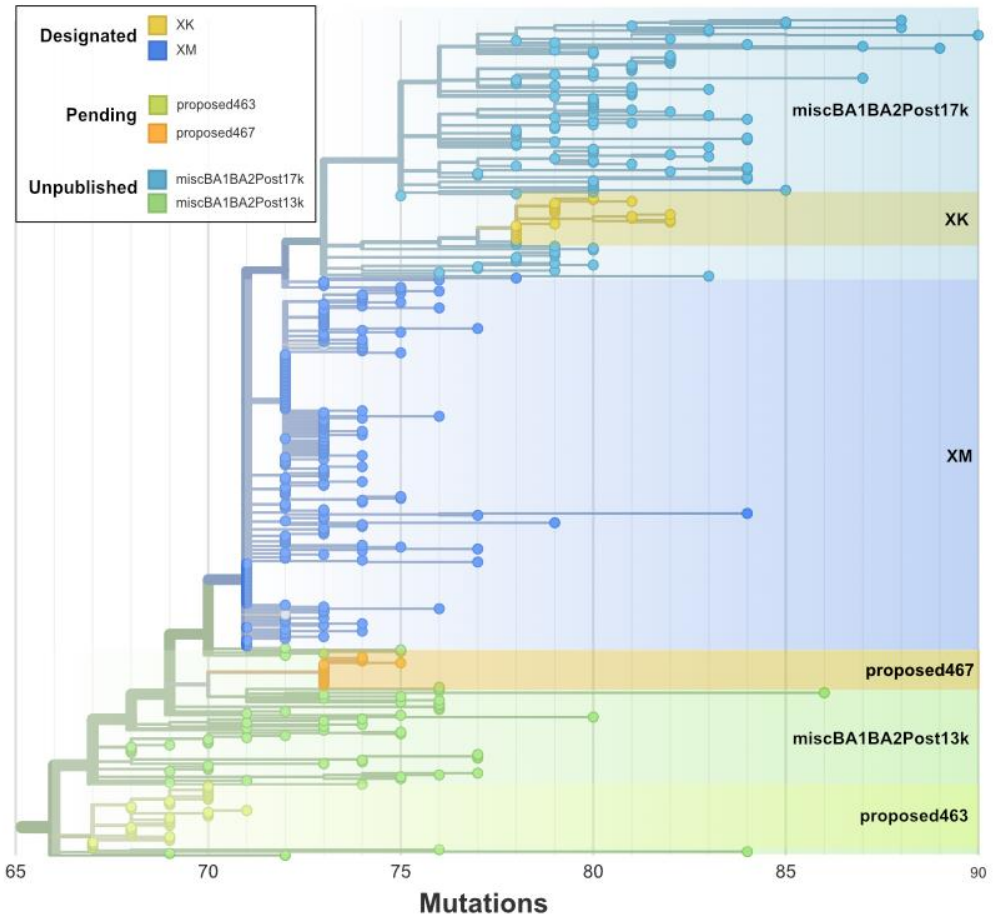
**I will demo the web browser versions. But all tools have command-line versions for high-throughput analyses.*

Why are recombinants hard to classify?

1. Algorithms assume that recombination is not occurring.

2. Rapidly evolving nomenclature that varies by program.
   - Designated (X*)
   - Pending (proposed*)
   - Unpublished (misc*)

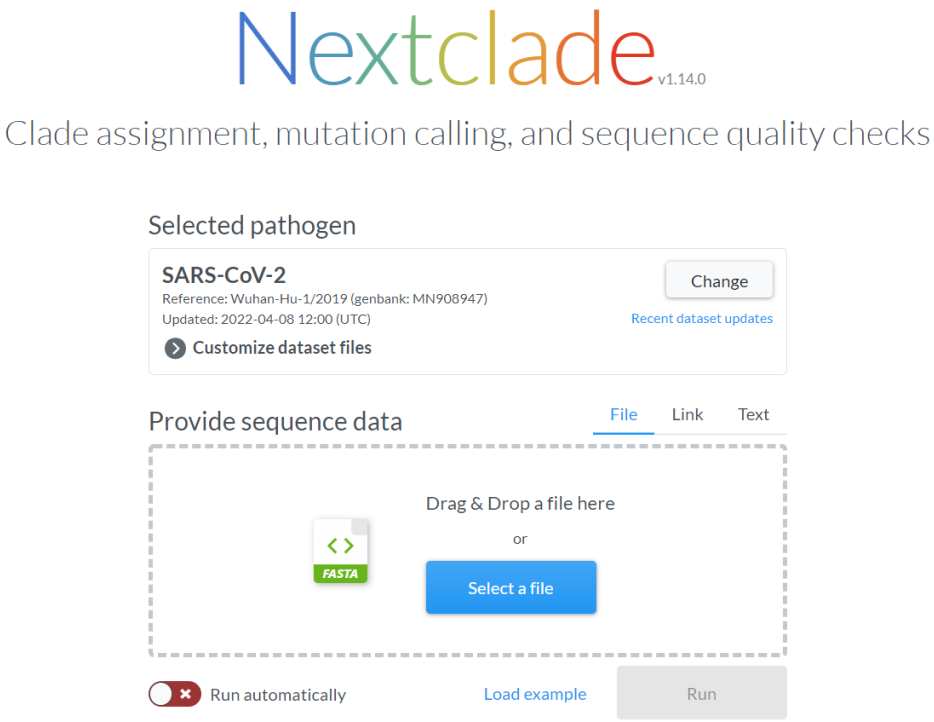
3. Not all recombinant lineages are in public databases.
   - XK, XL



Nextclade: Identify putative recombinants

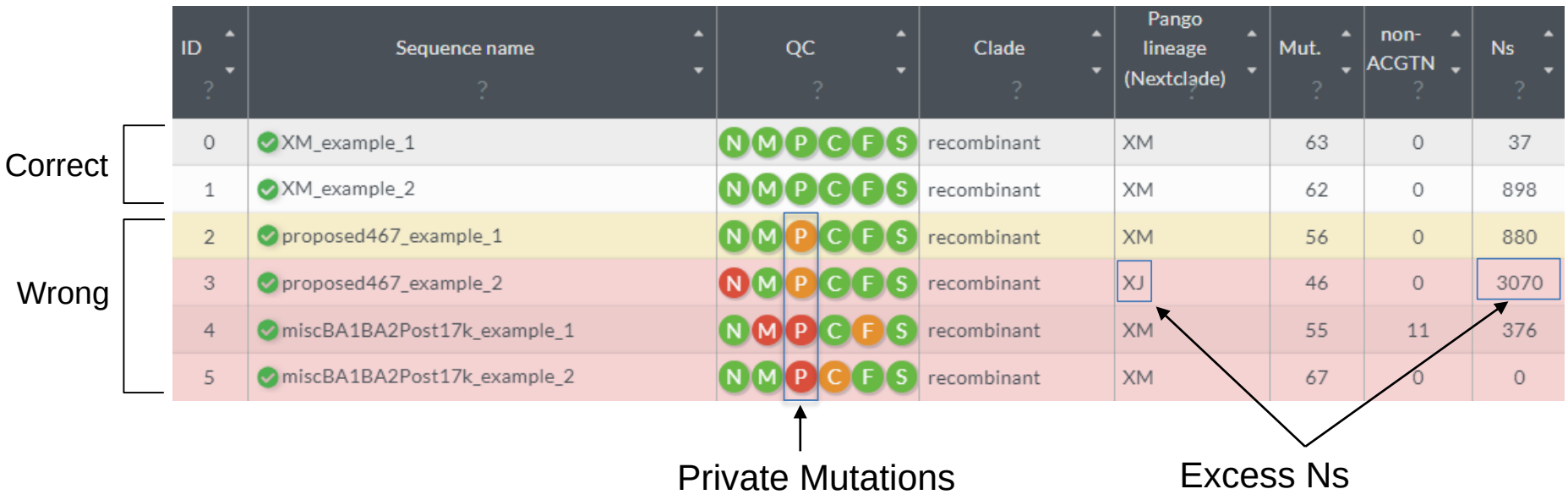
https://clades.nextstrain.org/

1. Select SARS-CoV-2
2. Upload **data/sequences.fasta.**
3. Click “Run”.



Nextclade: Identify putative recombinants

Note: Not the same method as pangolin!



	ID	Sequence name	QC	Clade	Pango lineage (Nextclade)	Mut.	non-ACGTN	Ns
Correct	0	XM_example_1	N M P C F S	recombinant	XM	63	0	37
Correct	1	XM_example_2	N M P C F S	recombinant	XM	62	0	898
Wrong	2	proposed467_example_1	N M P C F S	recombinant	XM	56	0	880
Wrong	3	proposed467_example_2	N M P C F S	recombinant	XJ	46	0	3070
Wrong	4	miscBA1BA2Post17k_example_1	N M P C F S	recombinant	XM	55	11	376
Wrong	5	miscBA1BA2Post17k_example_2	N M P C F S	recombinant	XM	67	0	0

Private Mutations

Excess Ns

Use the download button in the top right to download the alignment (nextclade.aligned.fasta) for a later step.

sc2rf: Plot recombination breakpoints

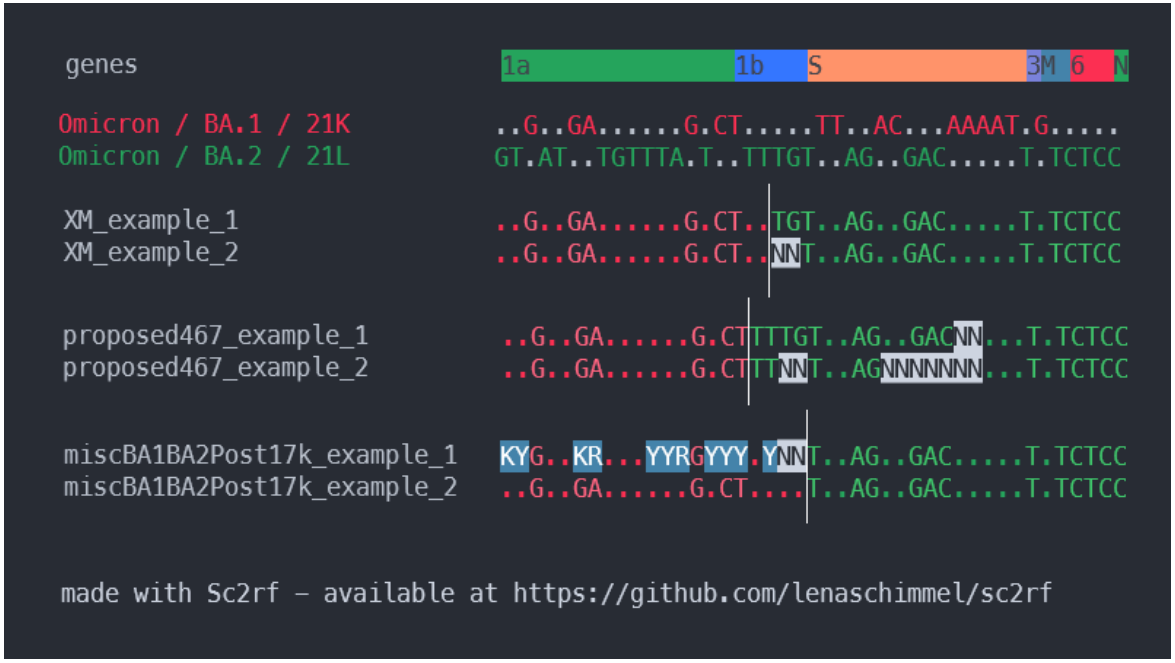
INSTALL

```
git clone https://github.com/lenaschimmel/sc2rf
cd sc2rf
pip install -r requirements.txt
```

RUN

```
./sc2rf.py results/nextclade.aligned.fasta --csvfile sc2rf.csv
```

OUTPUT

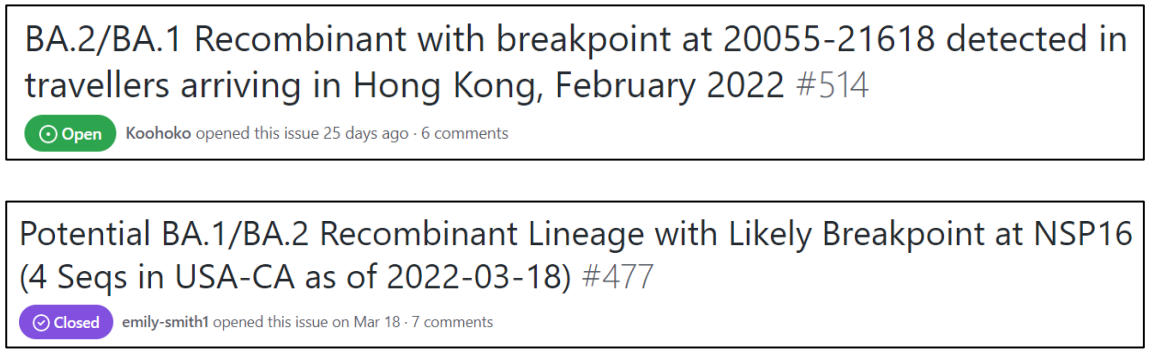


sc2rf: Plot recombination breakpoints

1. Compare the breakpoints to those in pango-designation repository issues.

2. Explore the breakpoints from sc2rf with the “breakpoints” column in:
   - **resources/pango-designation_issues_update.tsv**

3. Example
   - miscBA1BA2Post17k has a breakpoint between 20055 and 21618.
   - Matches breakpoints in Issue#477 and Issue#514.

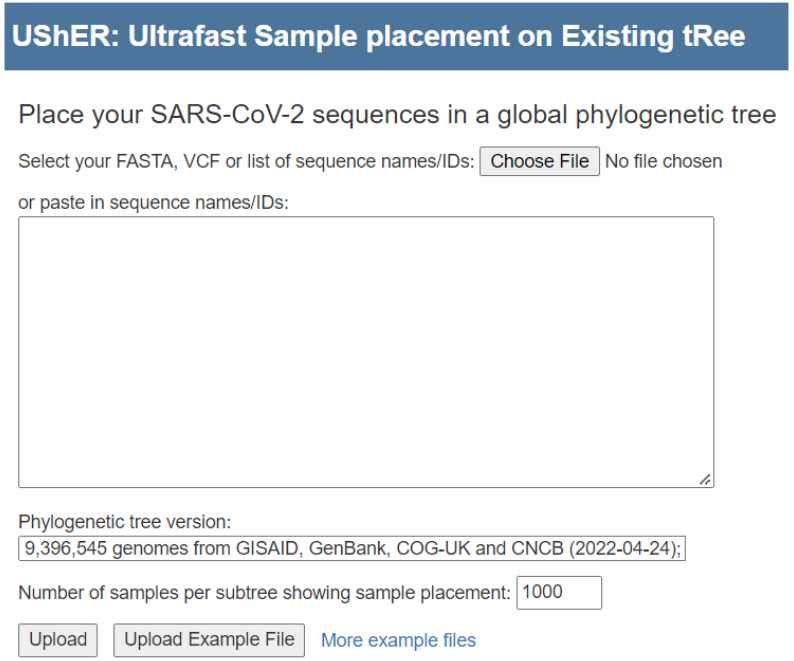


UShER: Refine lineage assignments

1a. Run on UCSC server:
https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

1b. Run on local machine:
https://shusher.gi.ucsc.edu/

2. Upload **data/sequences.fasta**



UShER: Refine lineage assignments

1. Lineages assignments by different algorithms currently differ more than they agree.
2. For pending or unpublished recombinants, UShER is the most accurate.
3. For designated recombinants, UShER and Nextclade are comparable.
4. Pangolin (PLEARN and PUSHER mode) is the most inaccurate.

UShER Assignment

pangolin Assignment

Nextclade Assignment

Fasta Sequence	Pango lineage (?)	Lineage of neighbor (?)	Subtree number (?)	#Maximally parsimonious placements (?)	Nextclade Lineage
XM_example_1	XM	XM	1 (view in Nextstrain)	1	XM
XM_example_2	XM	Unassigned	1 (view in Nextstrain)	1	XM
proposed467_example_1	proposed467	BA.1	1 (view in Nextstrain)	1	XM
proposed467_example_2	proposed467	BA.1	1 (view in Nextstrain)	1	XJ
miscBA1BA2Post17k_example_1	miscBA1BA2Post17k	Unassigned	1 (view in Nextstrain)	1	XM
miscBA1BA2Post17k_example_2	miscBA1BA2Post17k	Unassigned	1 (view in Nextstrain)	1	XM

BONUS Slides

- **Updating pango-designations metadata:**

```
pip install click pandas numpy
resources/pango-designation_issues.py > issues.tsv
```

- **Running the tutorial with [ncov-recombinant](): **

```
git clone --recursive https://github.com/ktmeaton/ncov-recombinant.git
cd ncov-recombinant

mamba env create -f workflow/envs/environment.yaml
conda activate ncov-recombinant

snakemake --profile profiles/tutorial
```