Coordinated efforts

- Norovirus genotypes and assignment -

Jennifer Chang, Ph.D.

Bioinformatic Analyst III

Nextstrain Team

Outline

- Collaborations and levels of commitment
- Nextclade dataset creation
- Norovirus specific topics

https://github.com/nextstrain/norovirus/issues/6

Levels of commitment - external

- Option 1: Active Code Contributor
  - Responsibilities:
    - Participate actively in the GitHub repository
    - Draft and submit GitHub issues, PRs, and reviews
  - Potential challenges:
    - Adhering to or adjusting Nextstrain GitHub contribution best-practices
  - Support:
    - I am available to guide and support people through these challenges

perhaps option 1b: Wrapping workflow in WDL

Might be faster for Nextstrain team to spin up a dataset

The reference sequences are all different lengths, so would need clarification on what region to work with

- Option 2: Reviewer
  - Responsibilities:
    - Review live [Pathogen] builds and provide feedback
    - Optionally submit GitHub issues or emails to flag any obvious errors
  - Engagement:
    - You will be emailed or pinged for reviews

Theiagen has test dataset to validate

- Option 3: SME Contributor
  - Responsibilities:
    - Regularly summarize recent [Pathogen] virus research papers in presentations to the code contributors
    - Help brainstorm and suggest new features for the public build

Needs to check with partners, to avoid getting stuck down the line

Prior Art in Nextclade dataset creation

Prior Art

- Measles: https://github.com/nextstrain/measles/pull/28/commits
- Yellow-fever: https://github.com/nextstrain/yellow-fever/pull/10/commits
- Lassa fever: https://github.com/nextstrain/lassa/pull/47/commits

Guides:

- nextstrain/nextclade_data/blob/master/docs/dataset-creation-guide.md

Blast-Based approach: -> Might be a good stop-gap until the dataset can be put together

- https://github.com/flu-crew/octoFLU/blob/ff408b0dd284cf8a127c89800c93ae2835e817eb/octoFLU.sh#L82-L87

Community Builds:

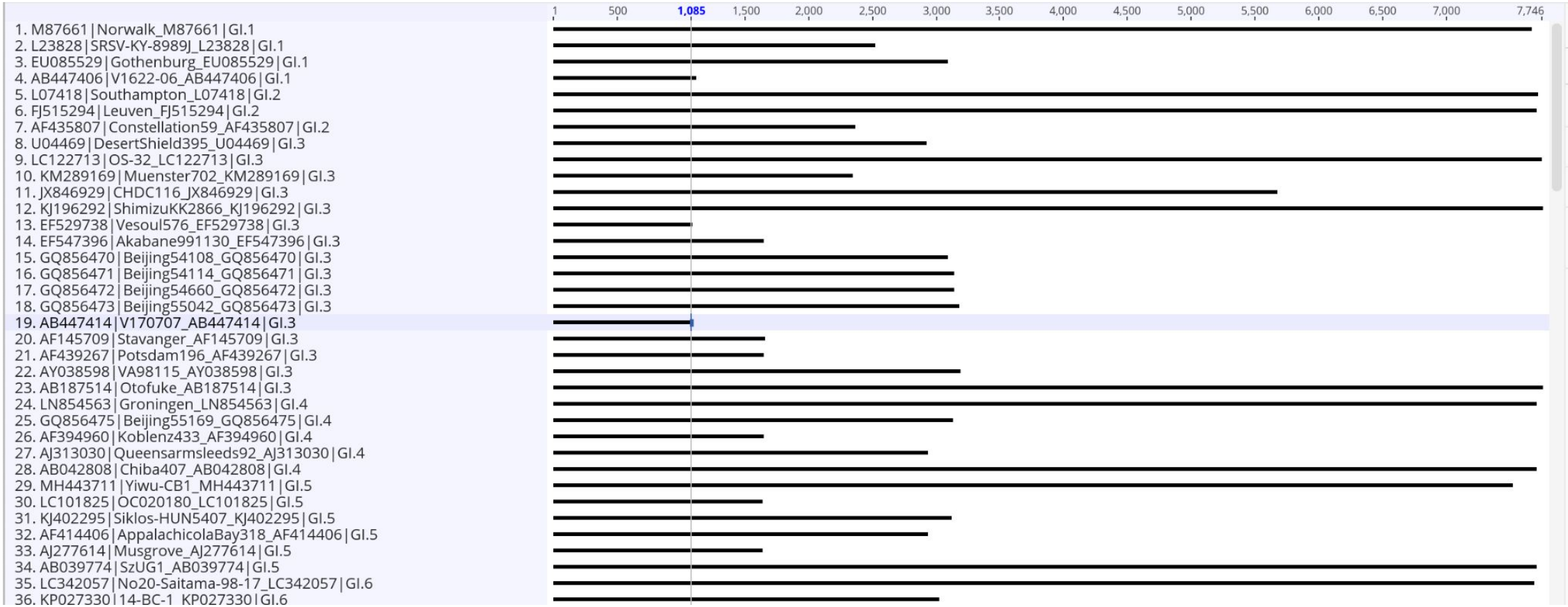
- https://github.com/mazeller/nextclade_test
- https://github.com/nextstrain/nextclade_data/tree/master/data/community
- Would Theiagen like to create a community dataset?

Blast-based approach (might be a good stop-gap)

```
# ===== Files
export REFERENCE=norovirus_cdc_reference.fasta
export QUERY=sequences.fasta

# ===== Create your Blast Database
makeblastdb -in ${REFERENCE} -dbtype nucl

# ===== Search your Blast Database
blastn -db ${REFERENCE} -query ${QUERY} -num_alignments 1 -outfmt 6 -out results.txt

# ===== Parse results
head results.txt
```

```
OP432651    LC153121|OH16002_LC153121|GII.4 97.876  518   11    0    1    518   546   1063   0.0   896
OP432652    LC153121|OH16002_LC153121|GII.4 97.876  518   11    0    1    518   546   1063   0.0   896
OP432653    JX459908|Sydney_JX459908|GII.4  97.674  516   12    0    3    518   4847  5362   0.0   887
OP432654    LC153121|OH16002_LC153121|GII.4 97.876  518   11    0    1    518   546   1063   0.0   896
OP432655    JX459908|Sydney_JX459908|GII.4  98.062  516   10    0    3    518   4847  5362   0.0   898
```

CDC References — differing lengths!



Updated references: https://pmc.ncbi.nlm.nih.gov/articles/PMC7011714/

Active Code Contributor

- Option 1: Active Contributor
  - Responsibilities:
    - Participate actively in the GitHub repository
    - Draft and submit GitHub issues, PRs, and reviews
  - Potential challenges:
    - Adhering to or adjusting Nextstrain GitHub contribution best-practices
  - Support:
    - I am available to guide and support people through these challenges

- I can put together some slides on:
  - the pathogen repo guide
  - github commit internal practices
  - Nextclade dataset creation
- We can collaborative submit and go through the PR process

Reviewer

- Option 2: Reviewer
  - Responsibilities:
    - Review live <pathogen> builds and provide feedback
    - Optionally submit GitHub issues or emails to flag any obvious errors
  - Engagement:
    - You will be emailed or pinged for reviews

- I will email out an update with the live build
- If possible, get response if the trees look acceptable or not within a week (or set email that you are on vacation)

SME Contributor

- Option 3: SME Contributor
  - Responsibilities:
    - Regularly summarize recent <pathogen> virus research papers in presentations to the code contributors
    - Help brainstorm and suggest new features for the public build

- If you'd be willing to compile and share a powerpoint summarizing the literature
- Schedule a later meeting to go through the slides

Nextstrain GitHub Standards

Why adhere 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

Nextstrain GitHub Practices

Consistency and Reproducibility

To ensure consistency, try to make sure all pathogen repos adhere to the pathogen-repo-guide

- https://github.com/nextstrain/pathogen-repo-guide
- Mostly this means following the file structure

Continuous Improvement and Adaptability

A high degree of comments on PR

- Submit Github Issues or PRs to best practices to incorporate new tools and methodologies
- Flexibly explore new features on dev-branches to be reviewed by wider team
- Surface new methods and analysis discussions in slack and on github

Steps for adding features

- Create a GitHub Issue
  - this is where we discuss potential solutions
- Create a Fork or PR linked to the Issue
  - this is where we implement potential solutions
  - this often turns into a long dialogue on various aspects of the solution
  - can trigger the creation of other github issues/PRs
  - is not guaranteed to be merged in
- If approved, use GitHub rebase to clean up commits on PR
  - this cleans up github commit history
  - this helps incorporate changes that have already been in merged into the repository
- Pick merge or squash merge
  - use a "merge" commit if there are multiple changes that we want to preserve history
  - use squash merge if there is one small or minor change