*Modernizing

nextstrain/**lassa**

Nextstrain build for Lassa virus

- Bedford Lab Meeting -

Jennifer Chang, Ph.D.

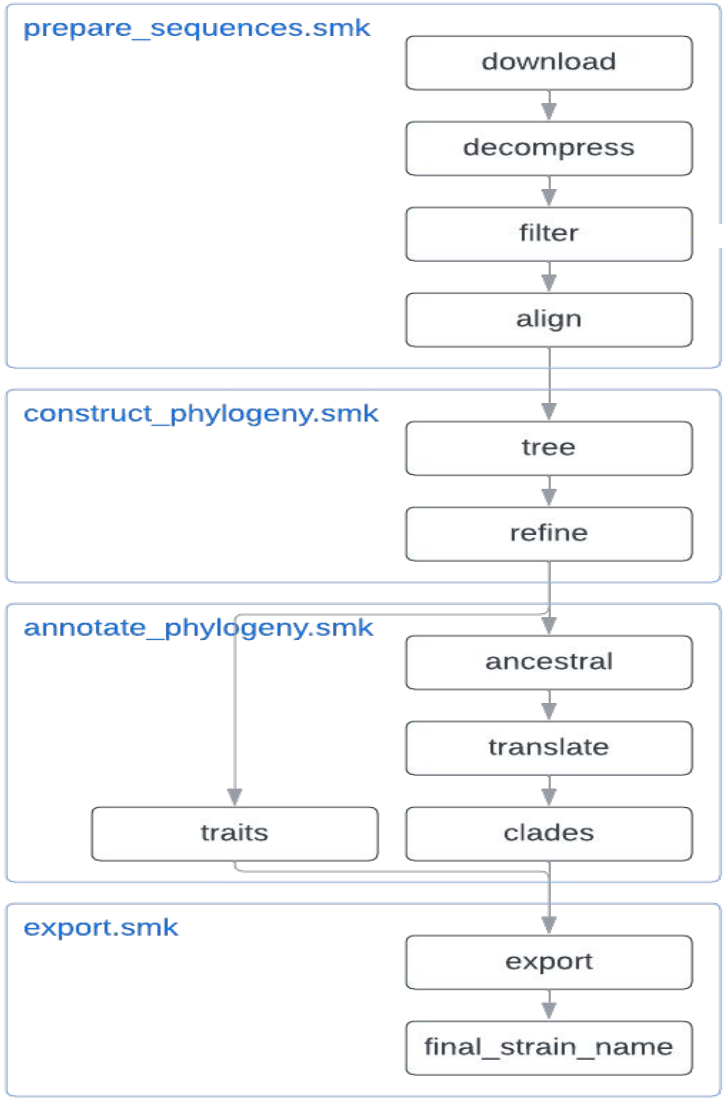
Bioinformatic Analyst III

Fred Hutchinson Cancer Center

Outline

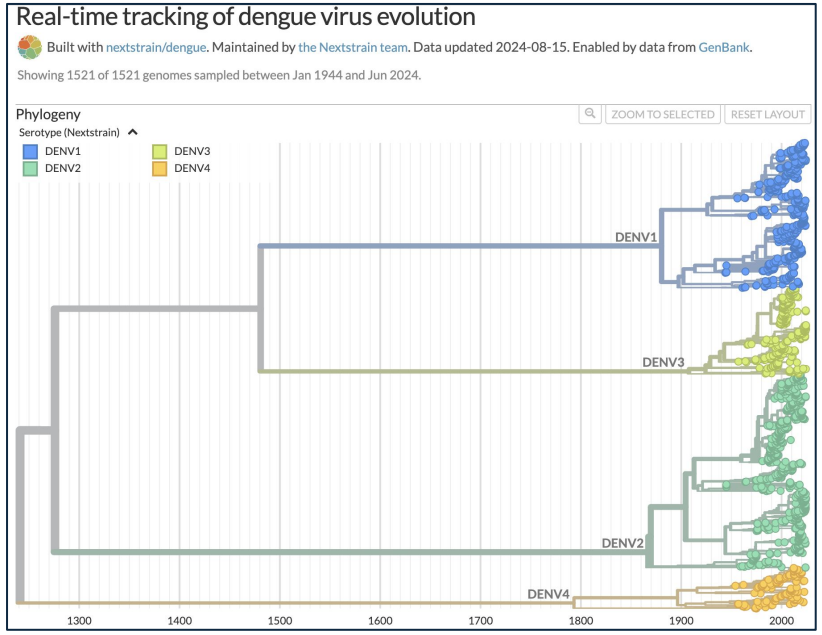
- **Lessons learned from modernizing Dengue workflows**
- About Lassa Virus
- Organizing meetings with Subject Matter Experts and feedback
- Thinking about external contributions
- Next steps

Lessons learned from modernizing Dengue



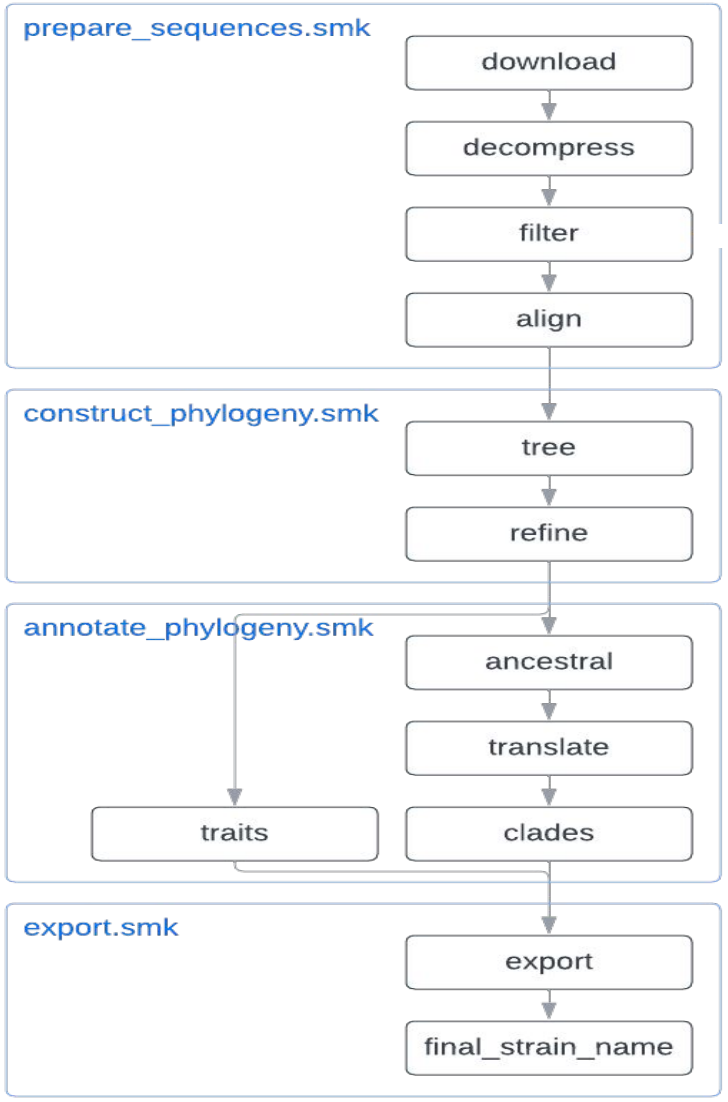
https://github.com/nextstrain/dengue

- Learned how to configure the **ingest** workflow
- Learned how to configure the **phylogenetic** workflow
- Learned how to create **gene** trees
- Learned some ways to generate a **Nextclade dataset**



Pipeline is organized according to the GitHub: nextstrain/pathogen-repo-guide

Lessons learned from modernizing Dengue



https://github.com/nextstrain/dengue

- Learned how to configure the **ingest** workflow
- Learned how to configure the **phylogenetic** workflow
- Learned how to create **gene** trees
- Learned some ways to generate a **Nextclade dataset**
- Realized too late that the dengue serotype and genotype system is not particularly clear and therefore it became very difficult to "validate the accuracy of the dataset"
- Realized that working directly with one or more Subject-Matter-Experts (SMEs) could have flagged these and other issues faster than I can read the dengue literature

Pipeline is organized according to the GitHub: nextstrain/pathogen-repo-guide

Lessons learned from modernizing Dengue

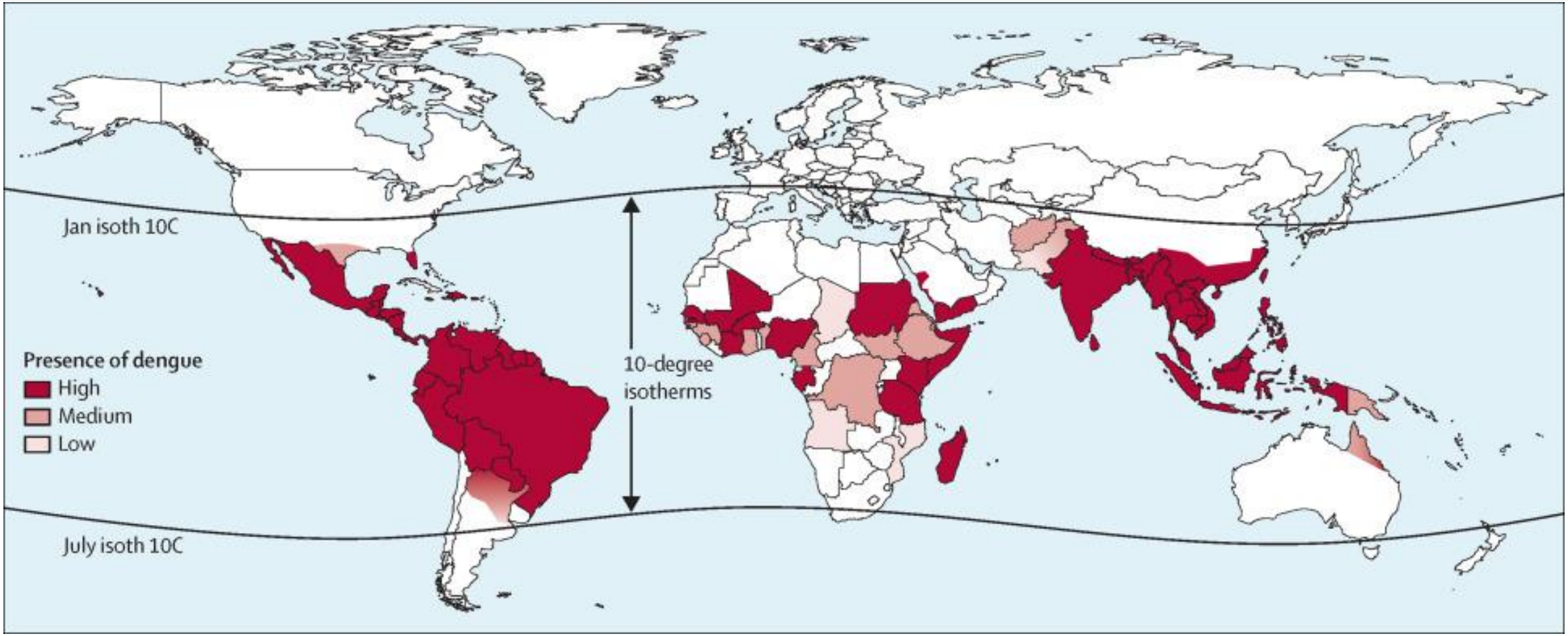
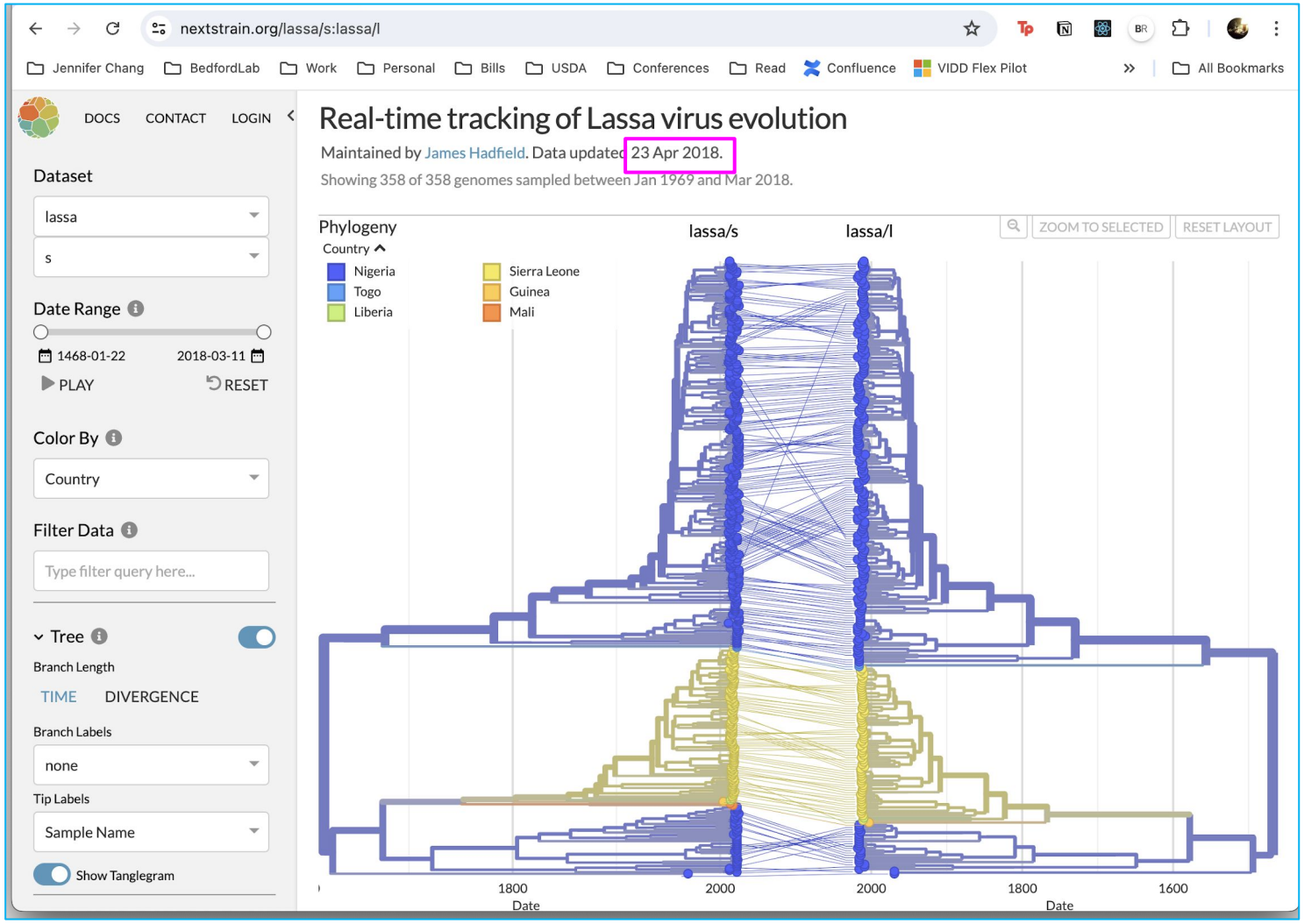


Figure 1: The global dengue burden, 2014 (Guzman and Harris, The Lancet, 2015)

Outline

- Lessons learned from modernizing Dengue workflows
- **About Lassa Virus**
- Organizing meetings with Subject Matter Experts and feedback
- Thinking about external contributions
- Next steps

Modernizing the Lassa build



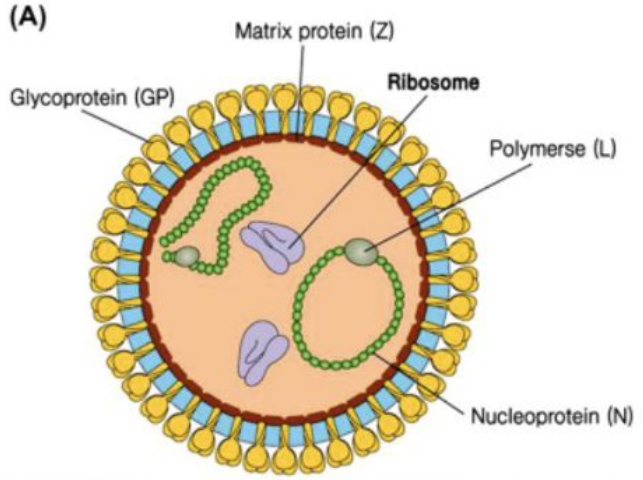
- Update the dataset
- Adhere to the pathogen-repo-guide
- Move and refactor ‘phylogenetic’ workflows into subdirectories
- Connect GitHub action automation
- Have a team of SME (subject-matter-experts) to check the validity of the site and make suggestions
- Do not “scoop” people’s work, provide a general analysis

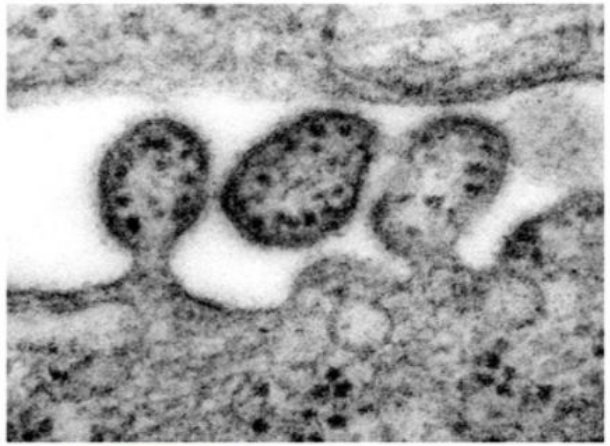
About Lassa Virus

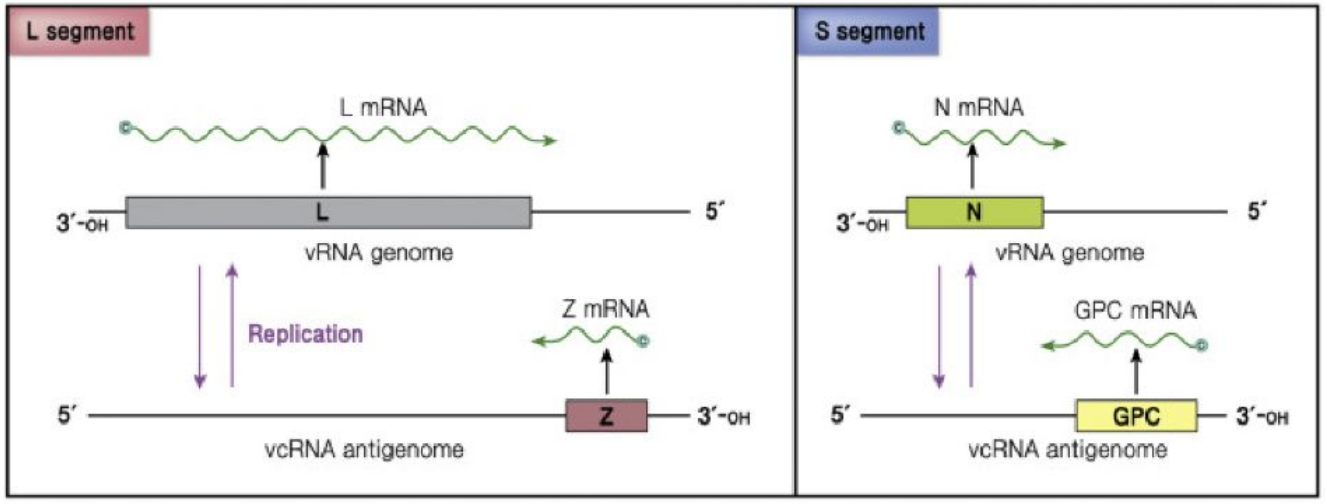
- Biosafety Level 4 (BSL-4)
- Causes severe disease with high mortality rates (15-20% in hospitalized patients)
- Spread by rats found in parts of West Africa although person-to-person transmission can occur
- No licensed vaccine
- Treatment options limited
- Stable as an aerosol, remains infections for several days outside host

Background

Lassa has two segments "L" and "S" from Chapter 16 of "Molecular Virology of Human Pathogenic Viruses" by Wang-Shick Ryu



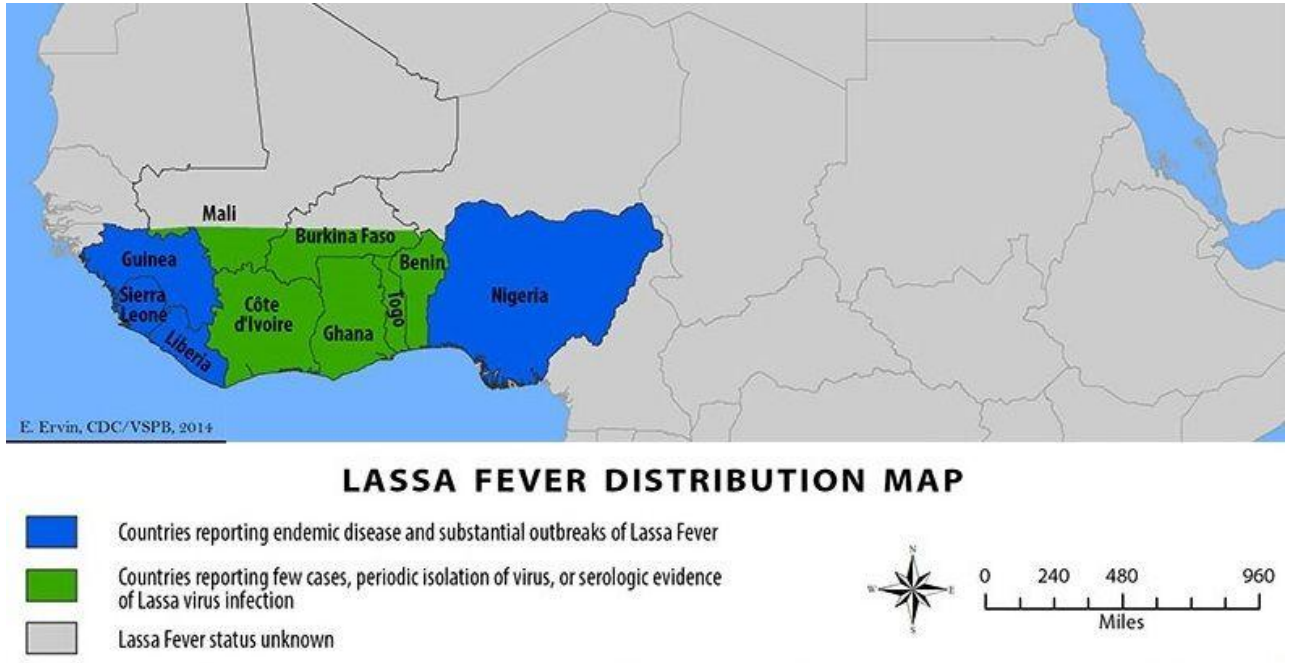




List of potential SMEs

[redacted]

- Other suggested contacts?

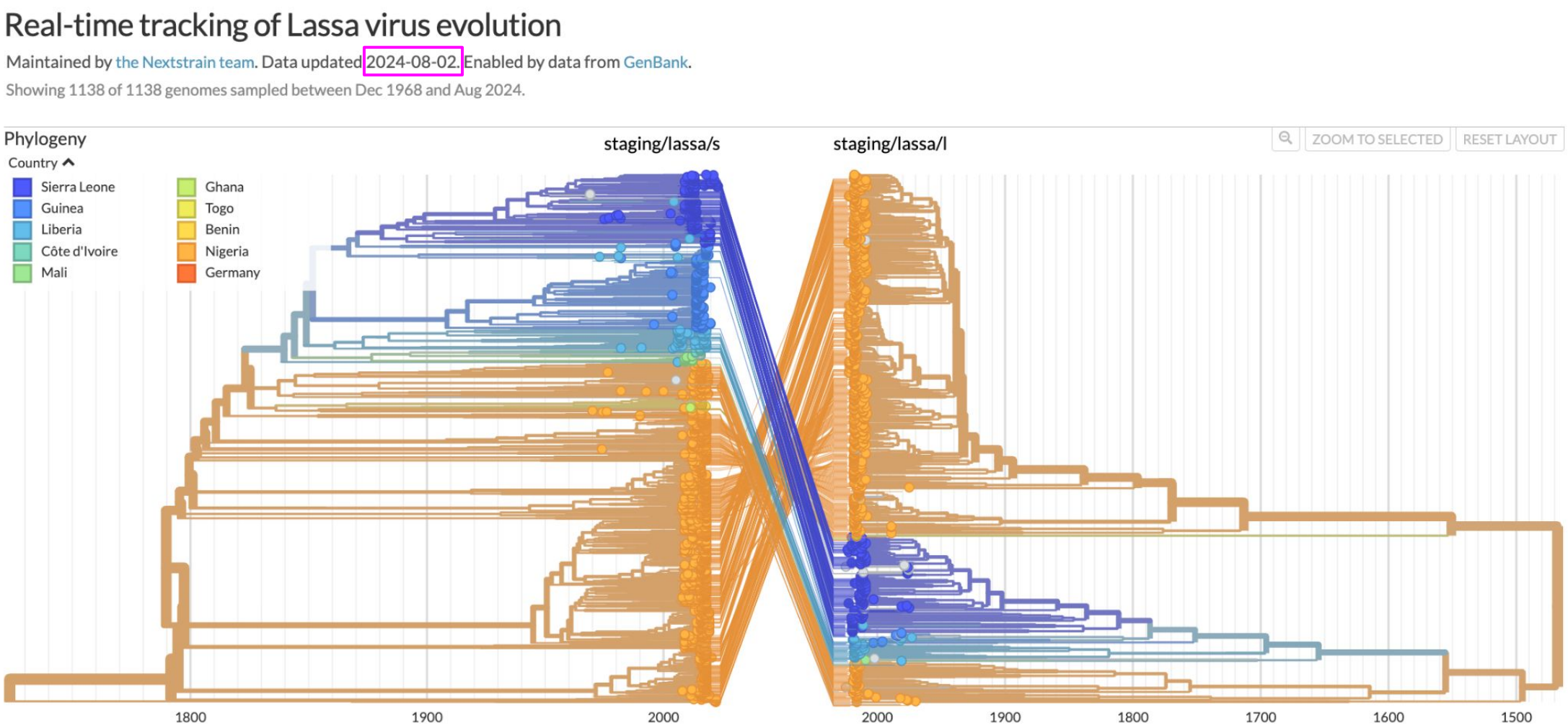


https://www.cdc.gov/lassa-fever/about/index.html

Outline

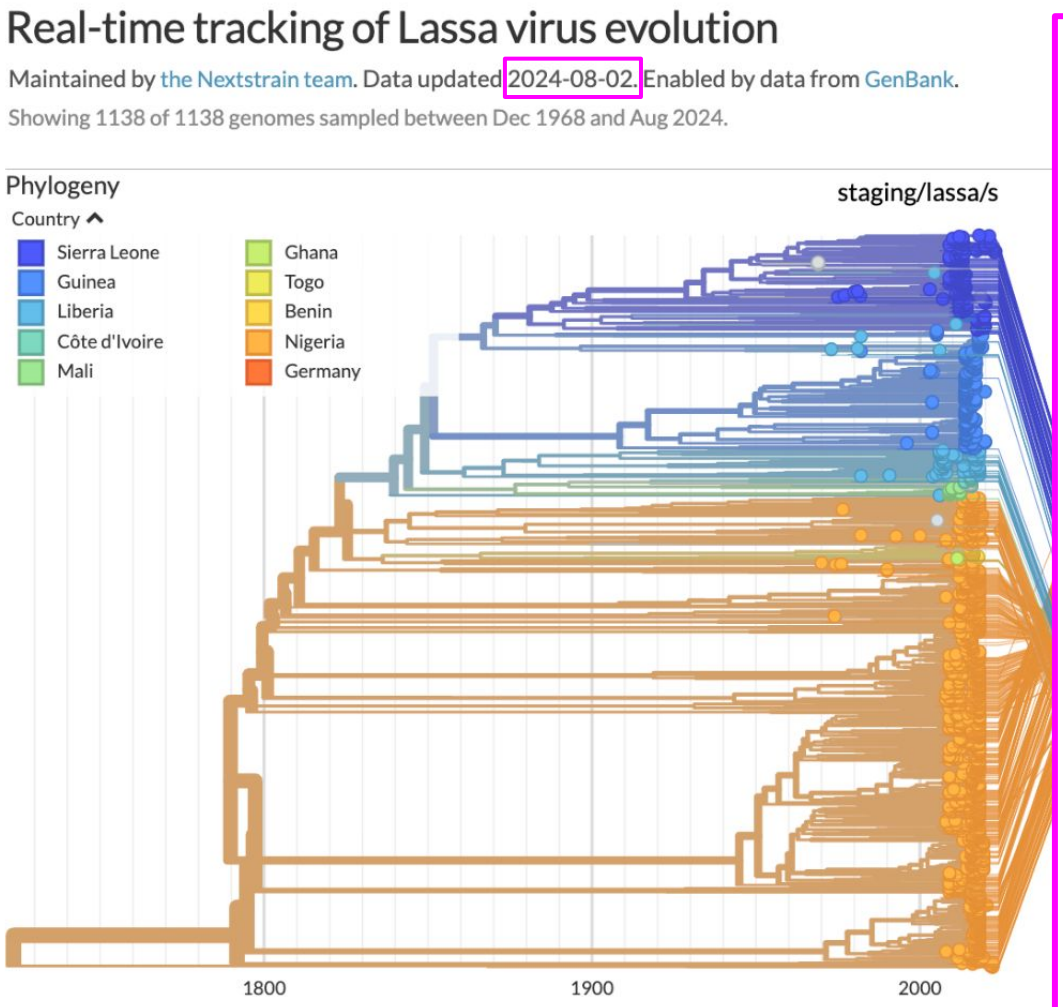
- Lessons learned from modernizing Dengue workflows
- About Lassa Virus
- **Organizing meetings with Subject Matter Experts and feedback**
  - [redacted]
- Thinking about external contributions
- Next steps

Update the Lassa trees



- https://next.nextstrain.org/staging/lassa/l

Update the Lassa trees

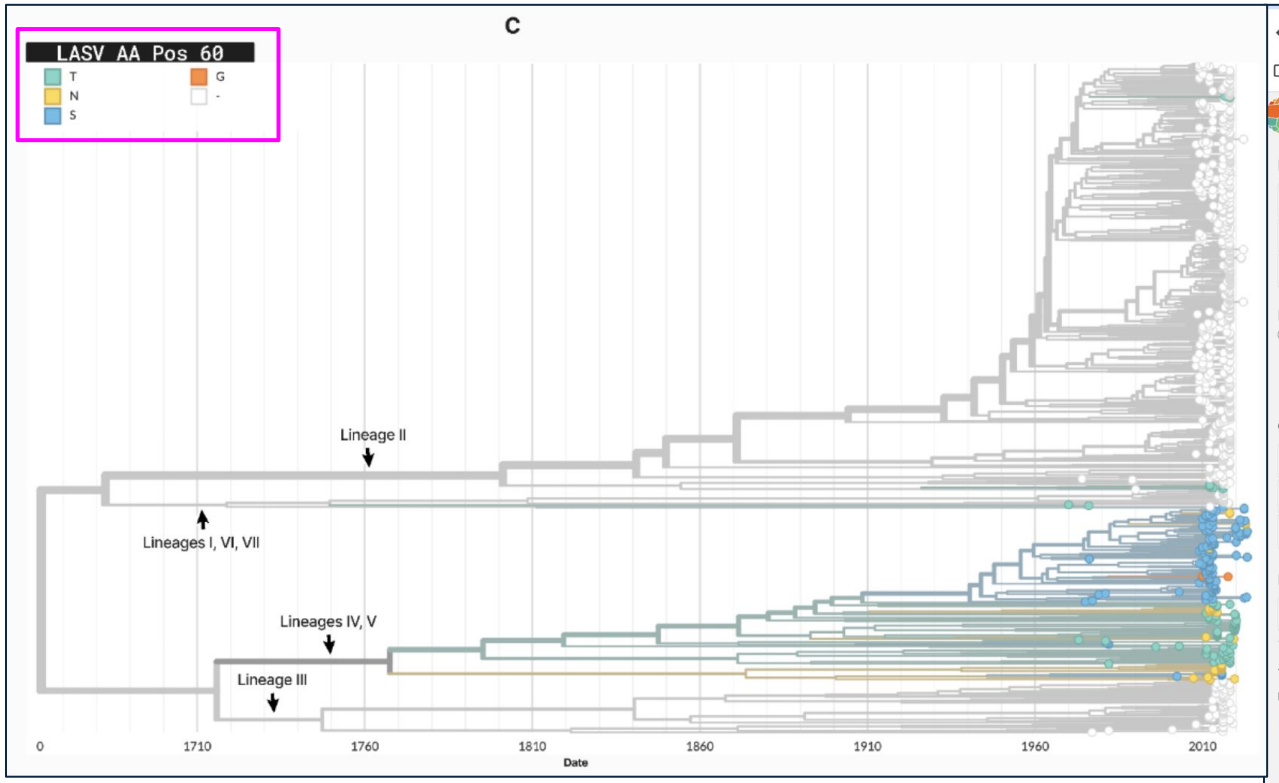


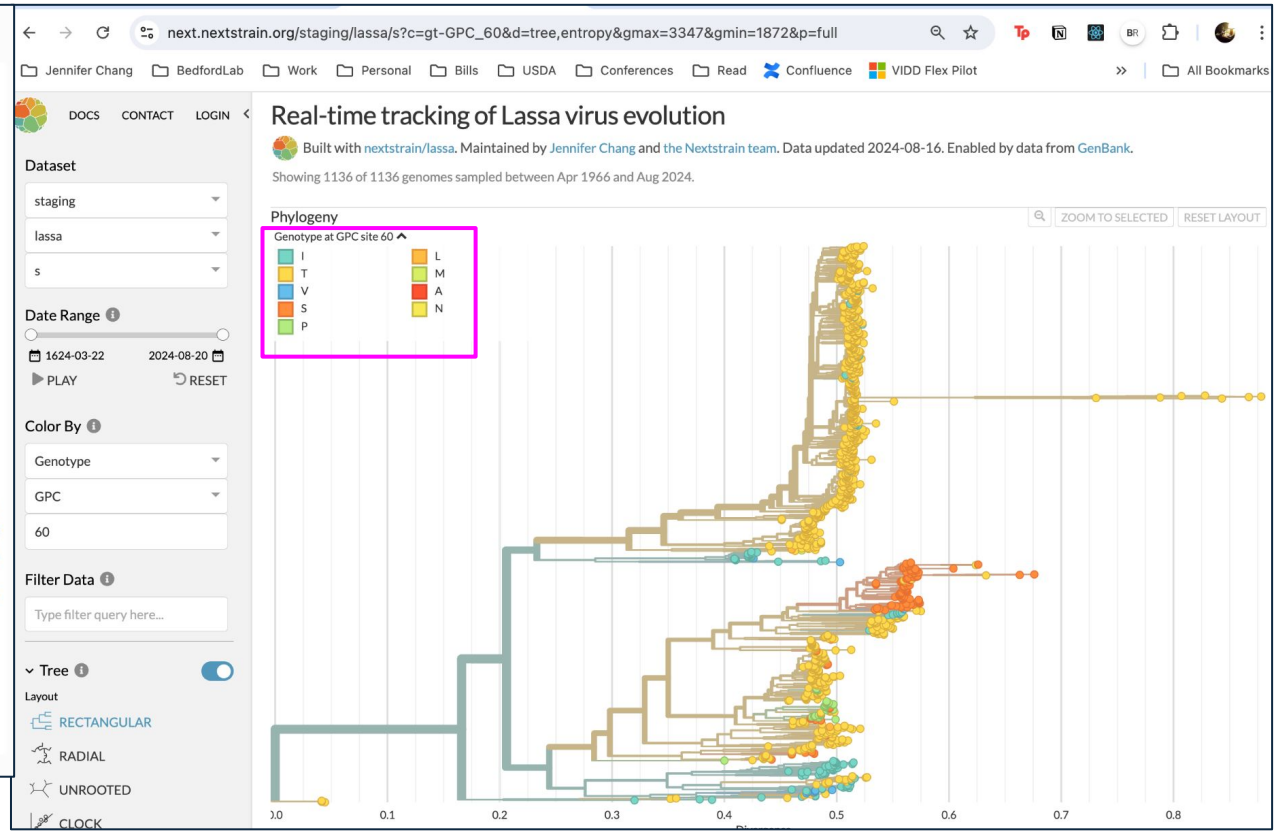
Questions for the Subject Matter Experts (SMEs)

1. Are there any obvious issues with the trees?
2. Should we extract a smaller region than the L and S segments?
3. Does Lassa have vaccine strains that need annotation on the tree?
4. Are there recommended host categories for tree annotation?
5. Can we link the L and S segment by methods other than strain name?

- https://next.nextstrain.org/staging/lassa/l

Amino acid position 60 in the GPC



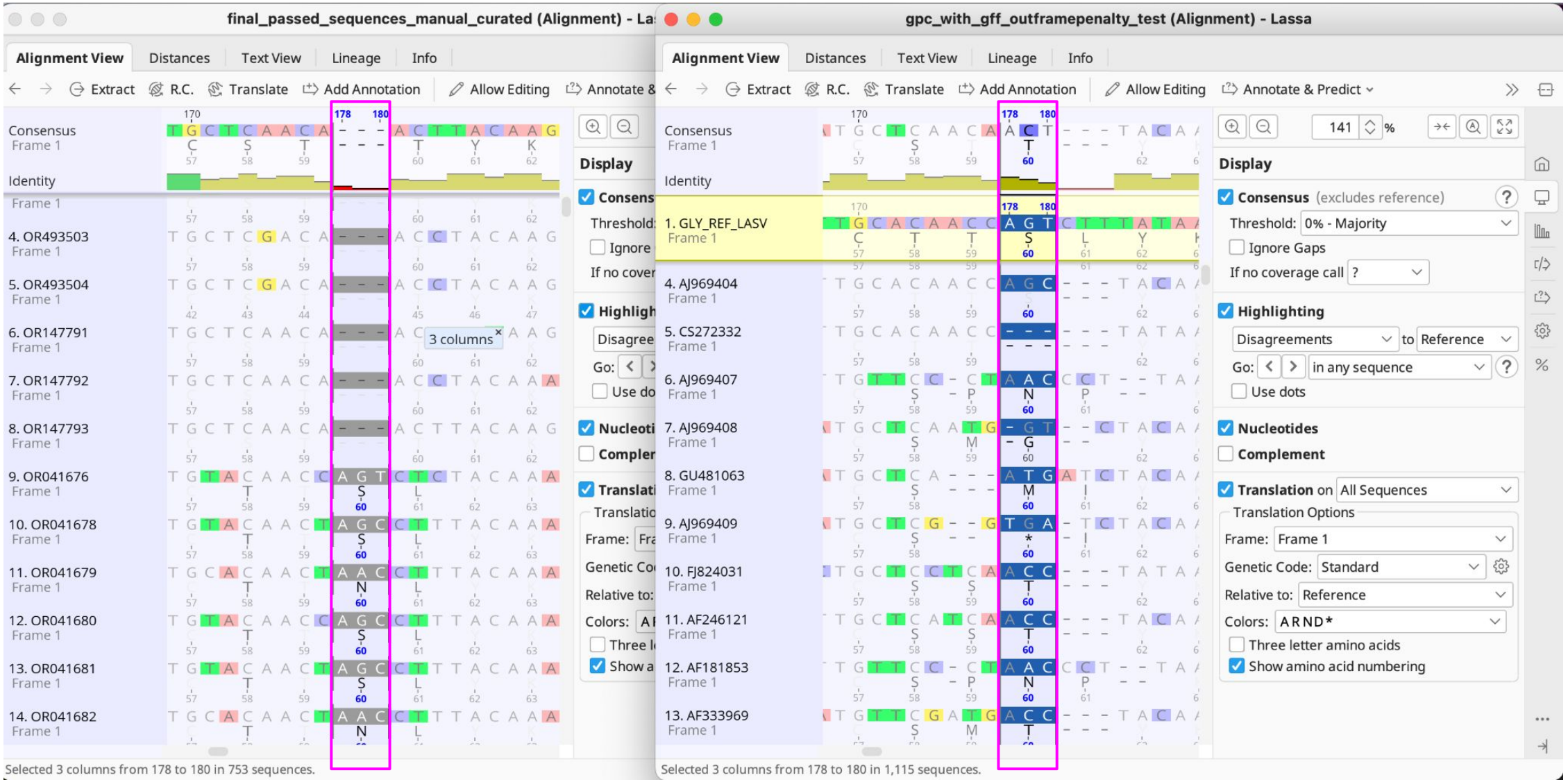


[Daudu et al, preprint, biorxiv]

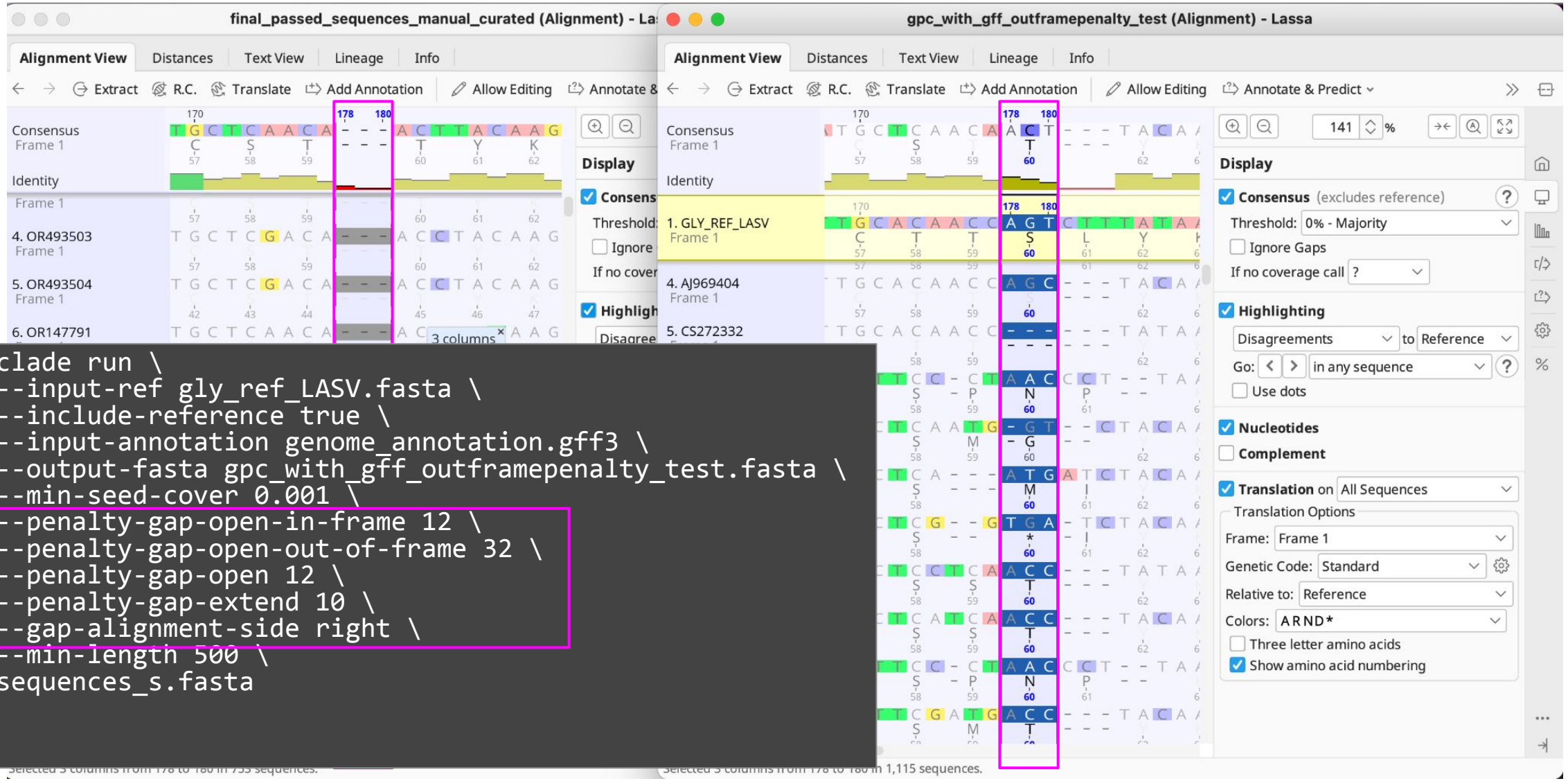
Our phylogenetic trees

"disrupt the functioning of all known anti-GPC antibodies (Robinson et al., 2016)."

A.A. pos. 60 in the GPC - fix codon alignment



A.A. pos. 60 in the GPC - fix codon alignment




```
nextclade run \
    --input-ref gly_ref_LASV.fasta \
    --include-reference true \
    --input-annotation genome_annotation.gff3 \
    --output-fasta gpc_with_gff_outframepenalty_test.fasta \
    --min-seed-cover 0.001 \
    --penalty-gap-open-in-frame 12 \
    --penalty-gap-open-out-of-frame 32 \
    --penalty-gap-open 12 \
    --penalty-gap-extend 10 \
    --gap-alignment-side right \
    --min-length 500 \
    sequences_s.fasta
```

[redacted] Questions

More details:

- Reference: based on
  - lassa_l.gb (Pinneo-NIG-1969)
  - lassa_s.gb (Nig08_04)
- Alignment: Augur align (mafft)
  - masking of GC regions
- Tree: built using IQTree (default params)
- Refine: using TreeTime
- [redacted] seemed particularly interested in the criteria for inclusion in the analysis

- https://next.nextstrain.org/staging/lassa/l

Outline

- Lessons learned from modernizing Dengue workflows
- About Lassa Virus
- Organizing meetings with Subject Matter Experts and feedback
- **Thinking about external contributions**
- Next steps

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

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

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

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
- We can collaborative submit and go through the PR process

Reviewer

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

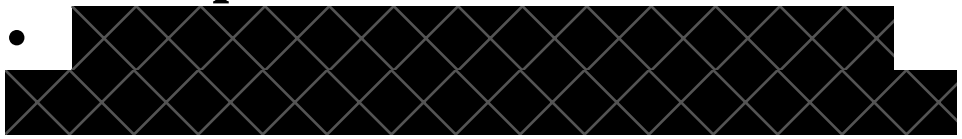
- I will email out an update with the live build
- If possible, get response if the trees look acceptable or not within a week (or set email that you are on vacation)

SME Contributor

- Option 3: SME Contributor
  - Responsibilities:
    - Regularly summarize recent Lassa 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

Outline

- Lessons learned from modernizing Dengue workflows
- About Lassa Virus
- Organizing meetings with Subject Matter Experts and feedback
- Thinking about external contributions
- **Next steps**
  - 

Outline

- Lessons learned from modernizing Dengue workflows
- About Lassa Virus
- Organizing meetings with Subject Matter Experts and feedback
- Thinking about external contributions
- **Next steps**
  - Creating slides and onboarding for an Active Code Contributor
  - Test run with [redacted] to contribute to the Lassa Repo
  - Use slides and onboarding for [redacted] work with [redacted]

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

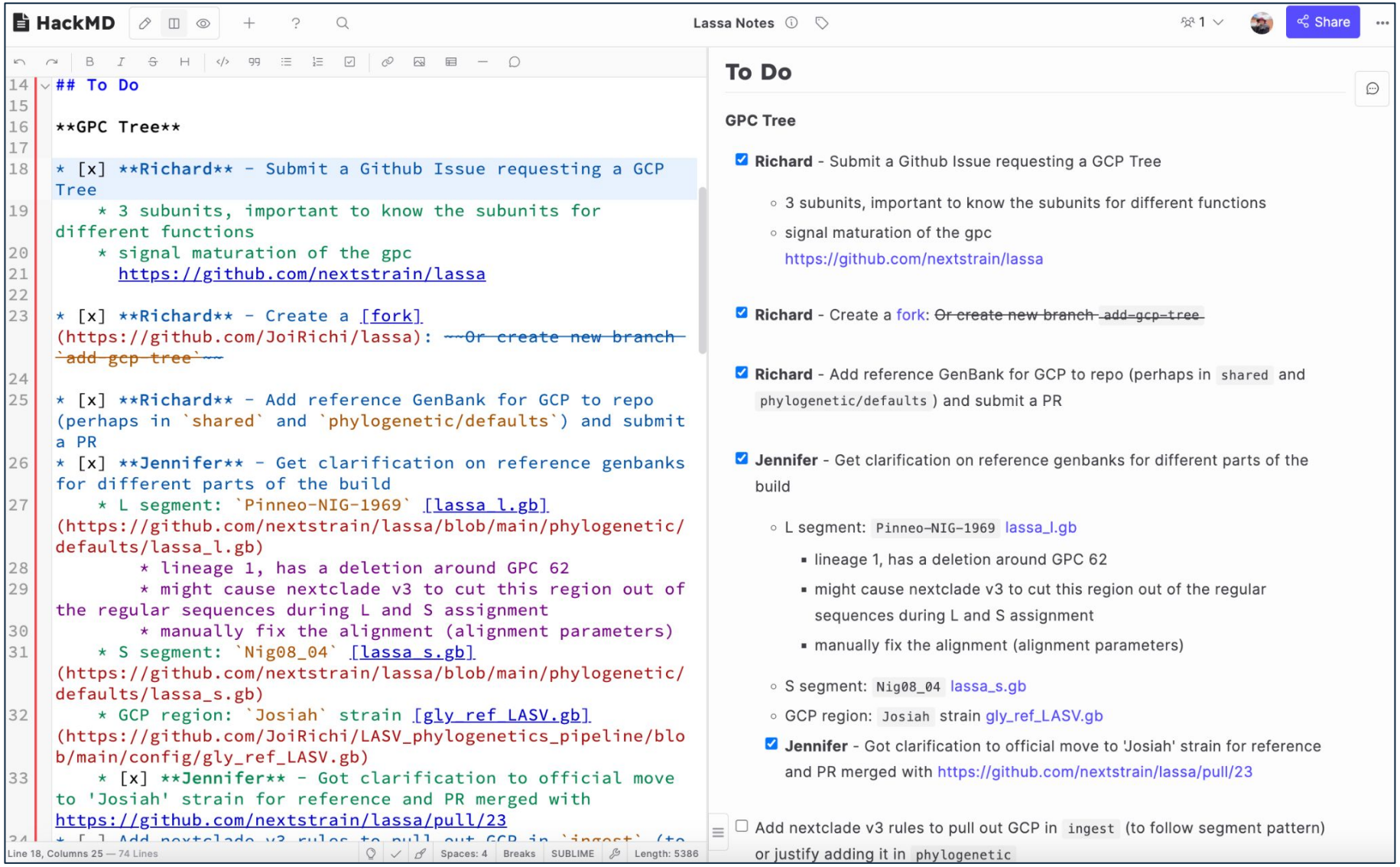
A high degree of comments on PR

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

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

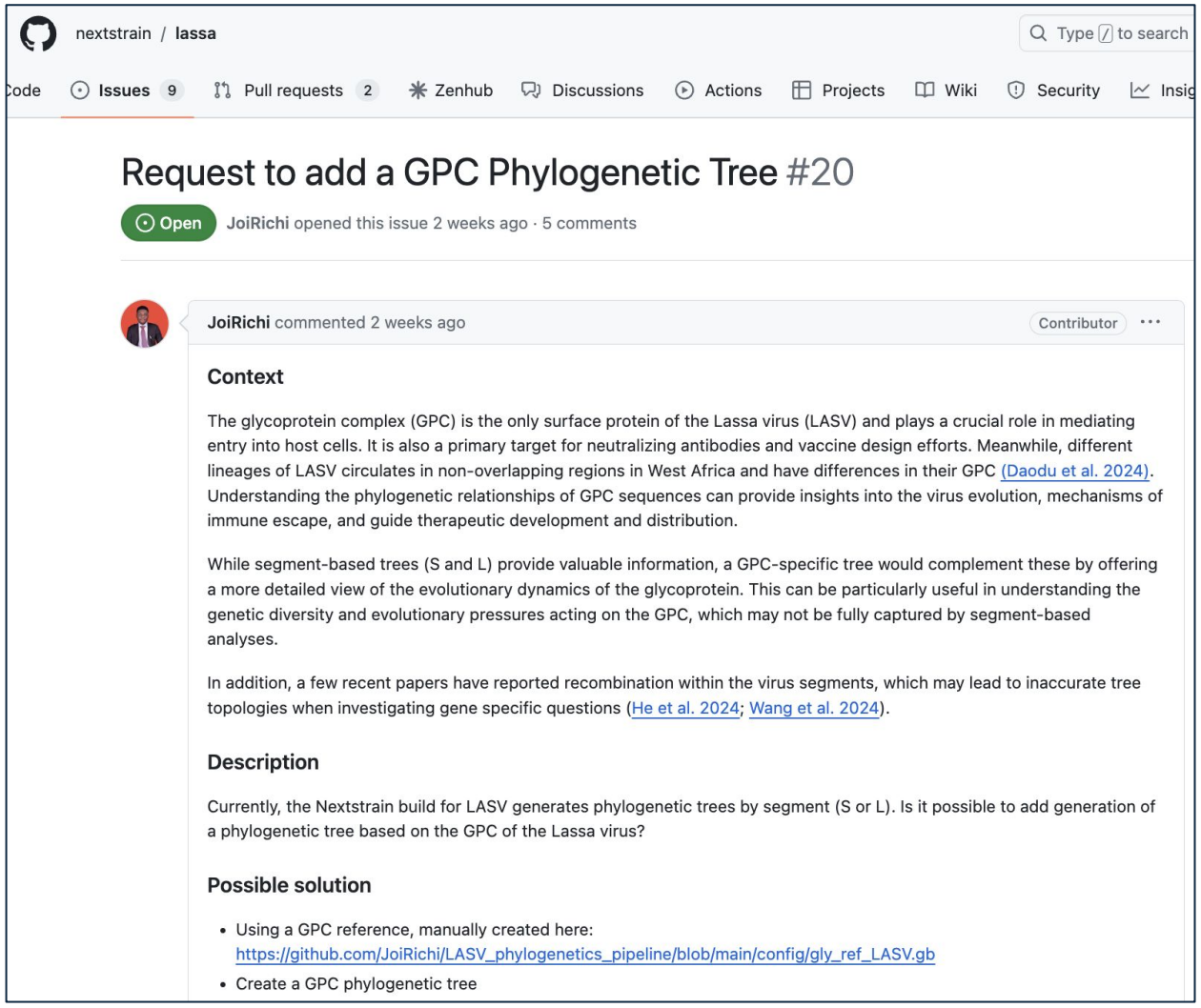
HackMD to coordinate with Richard Daudu



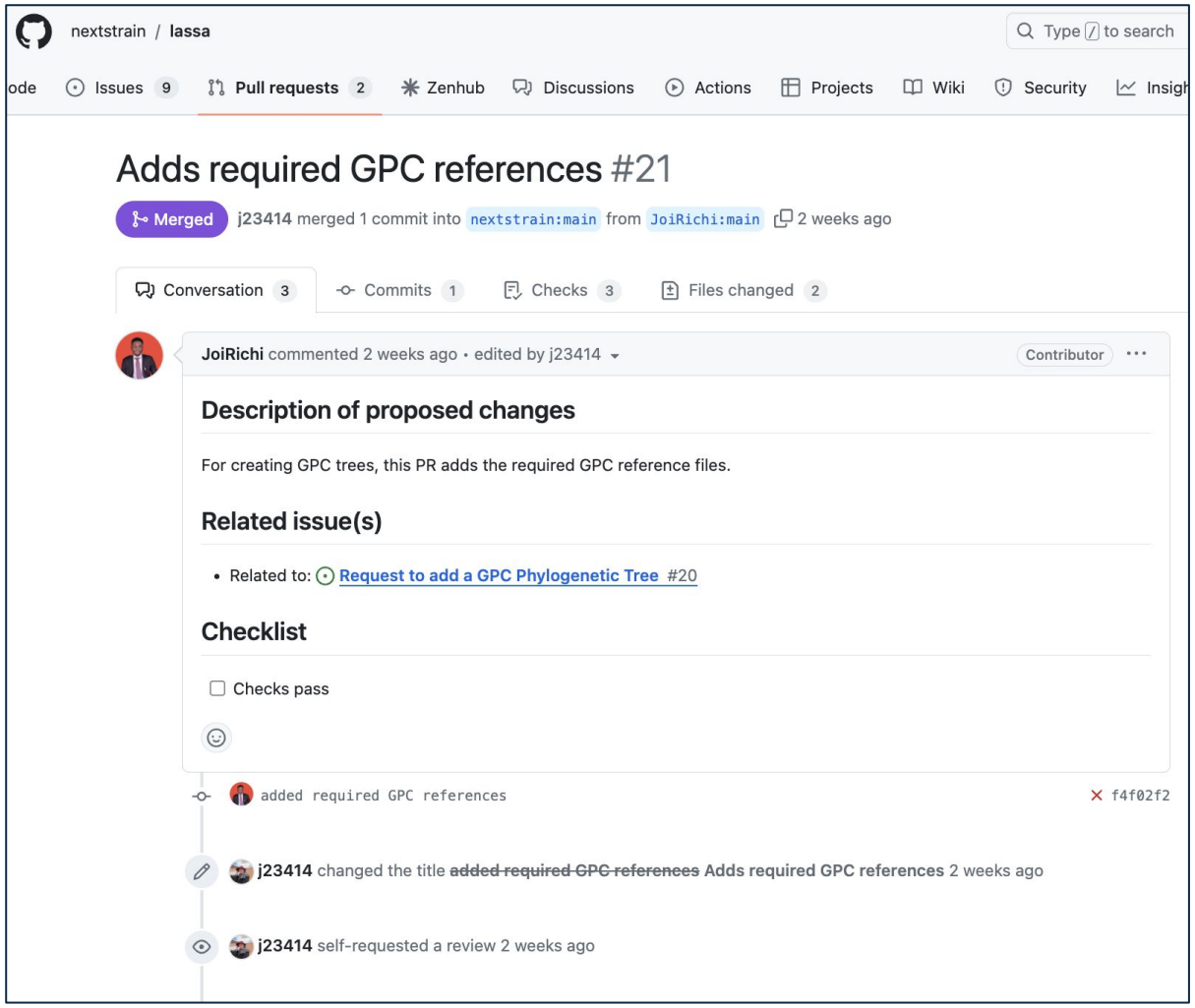
link

HackMD to coordinate with Richard Daudu

Submit a GitHub Issue



Submit a GitHub PR



Acknowledgements

- Nextstrain Team
- Bedford Lab
- SMEs:
  - Richard Daudu
  - [redacted]
  - [redacted]