

*Modernizing

nextstrain/lassa

Nextstrain build for Lassa virus



- Bedford Lab Meeting -

Jennifer Chang, Ph.D.

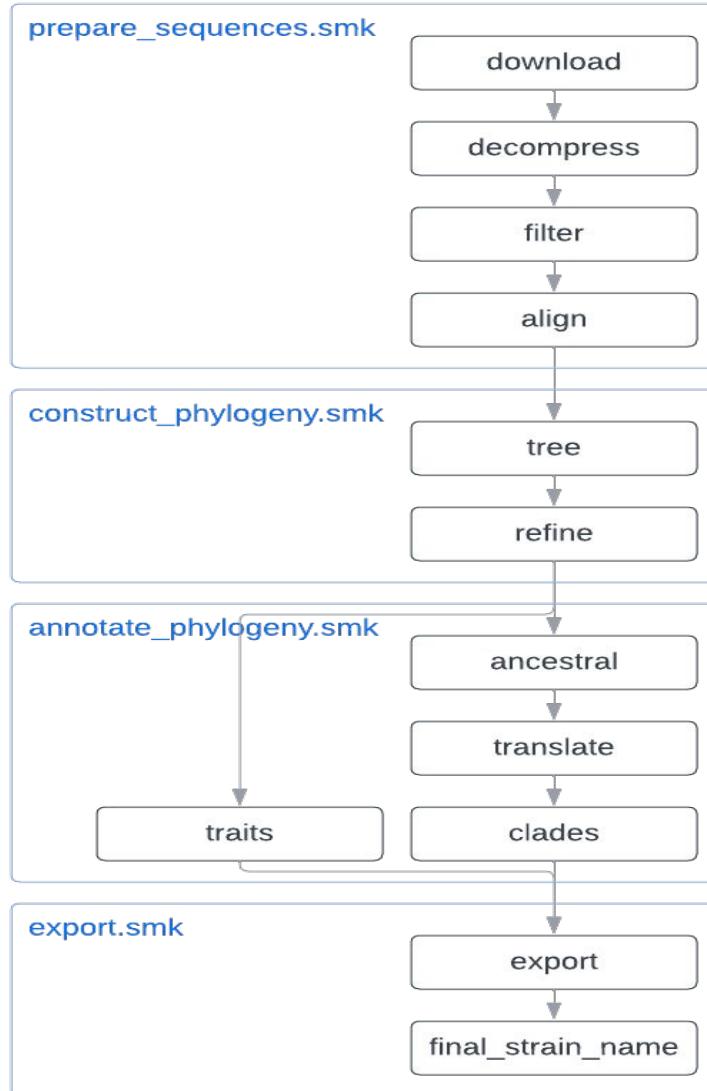
Bioinformatic Analyst III

Fred Hutchinson Cancer Center

Outline

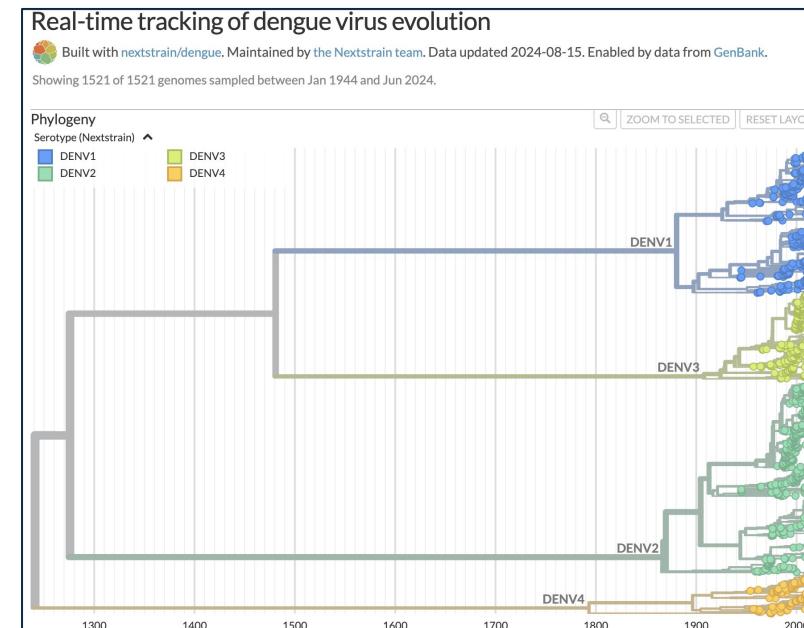
- **Lessons learned from modernizing Dengue workflows**
- About Lassa Virus
- Organizing meetings with Subject Matter Experts and feedback
- Thinking about external contributions
- Next steps

Lessons learned from modernizing Dengue



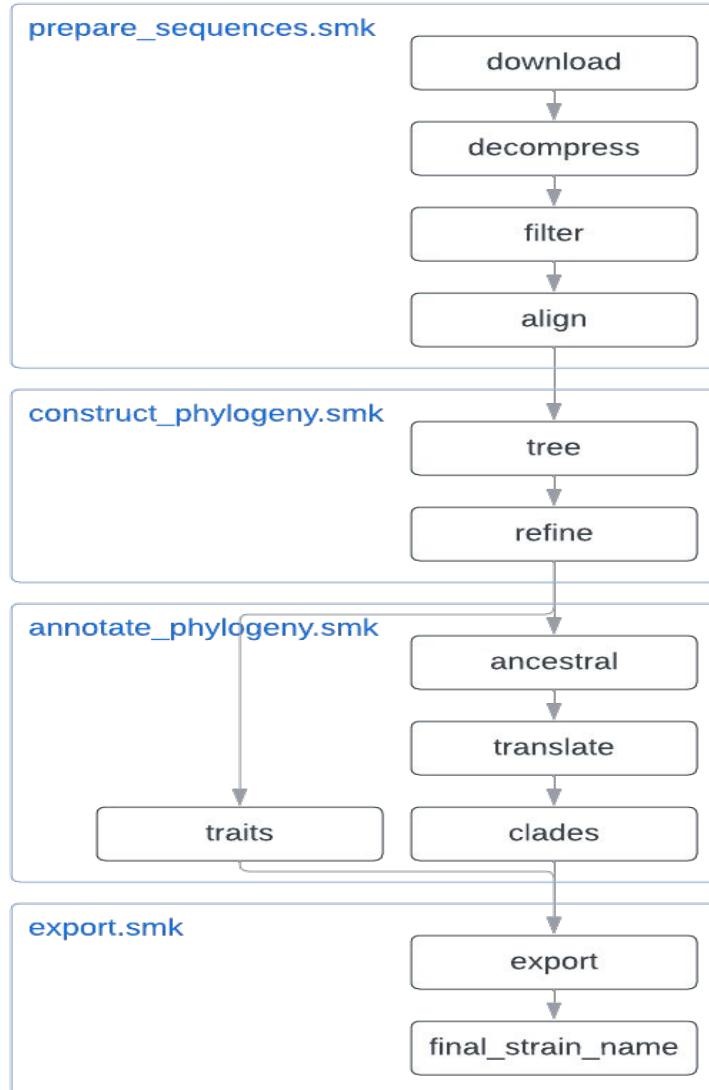
<https://github.com/nextstrain/dengue>

- 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](https://github.com/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](https://github.com/nextstrain/pathogen-repo-guide)

Lessons learned from modernizing Dengue

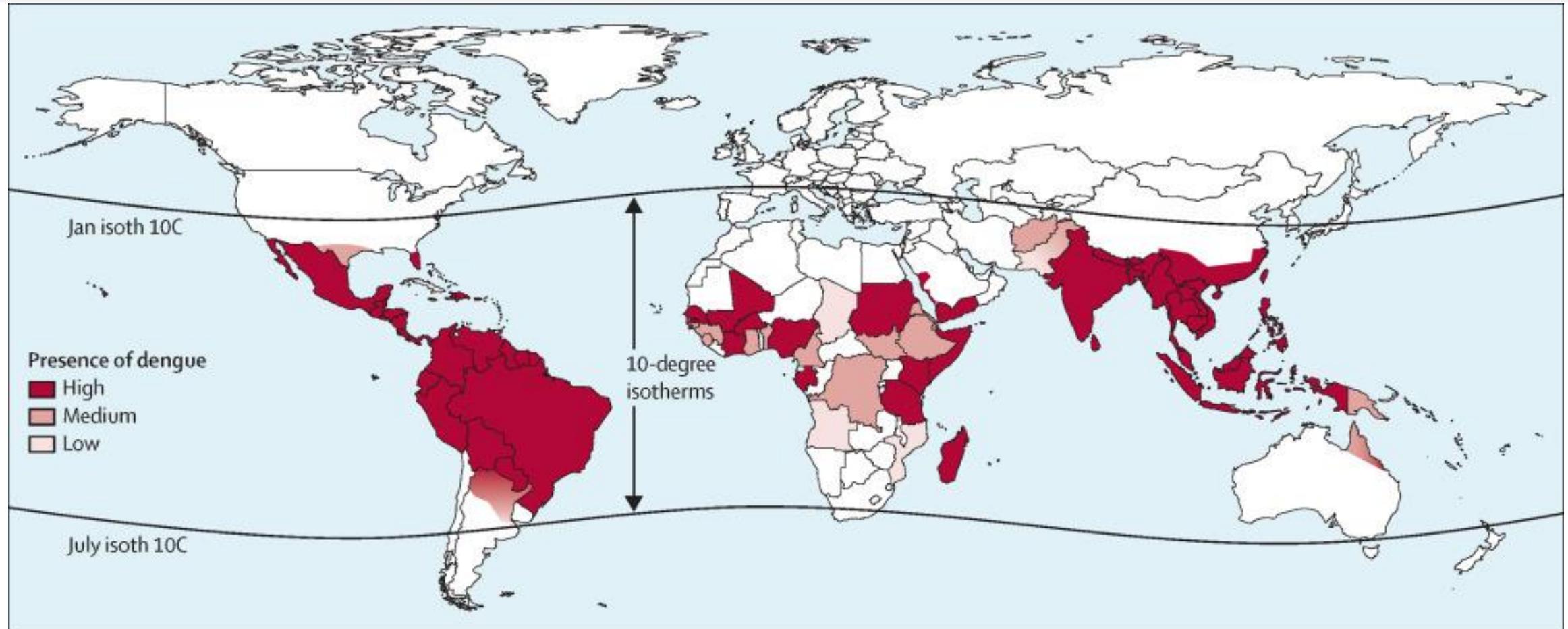
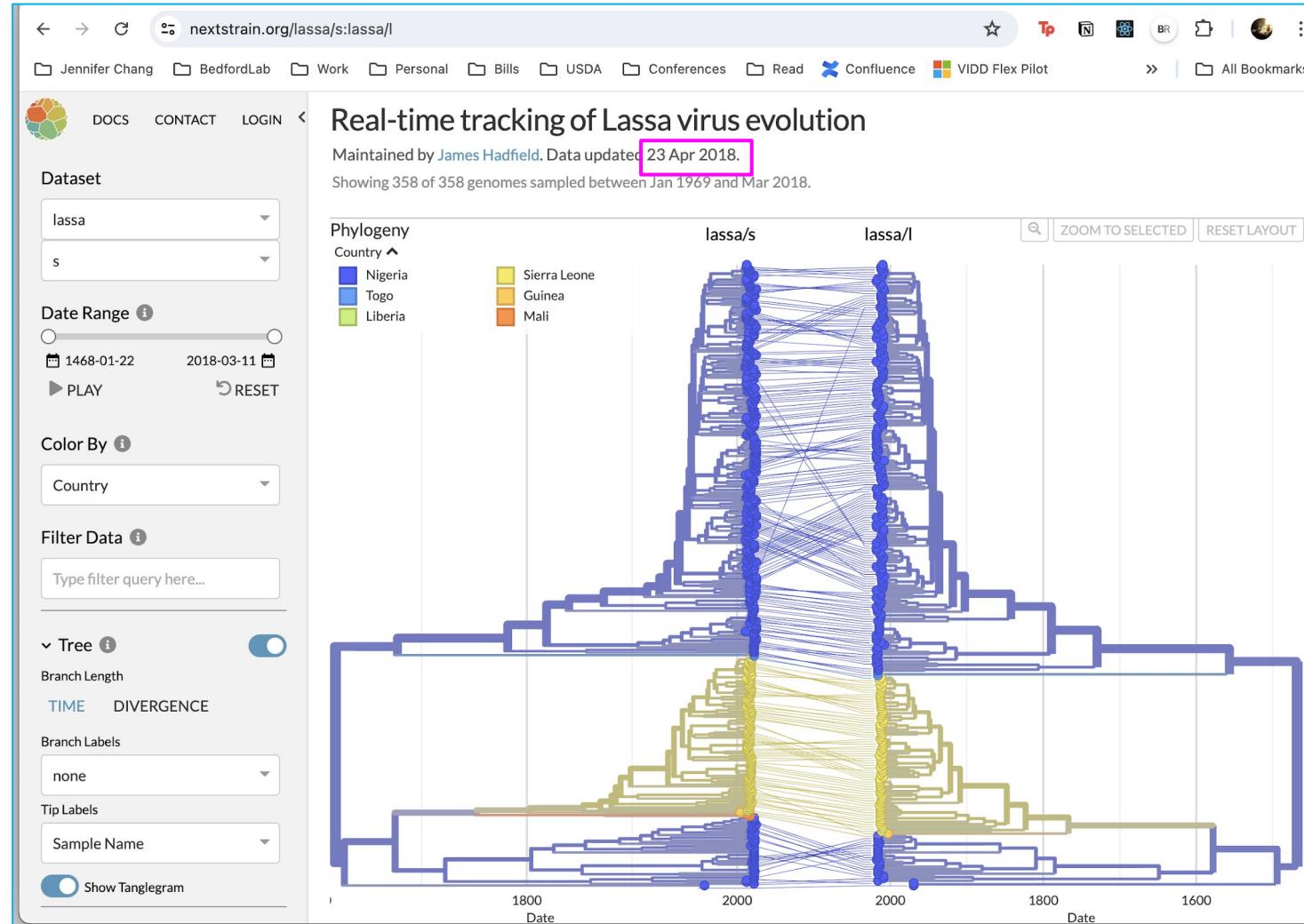


Figure 1: The global dengue burden, 2014 (Guzman and Harris, The Lancet, 2015)

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



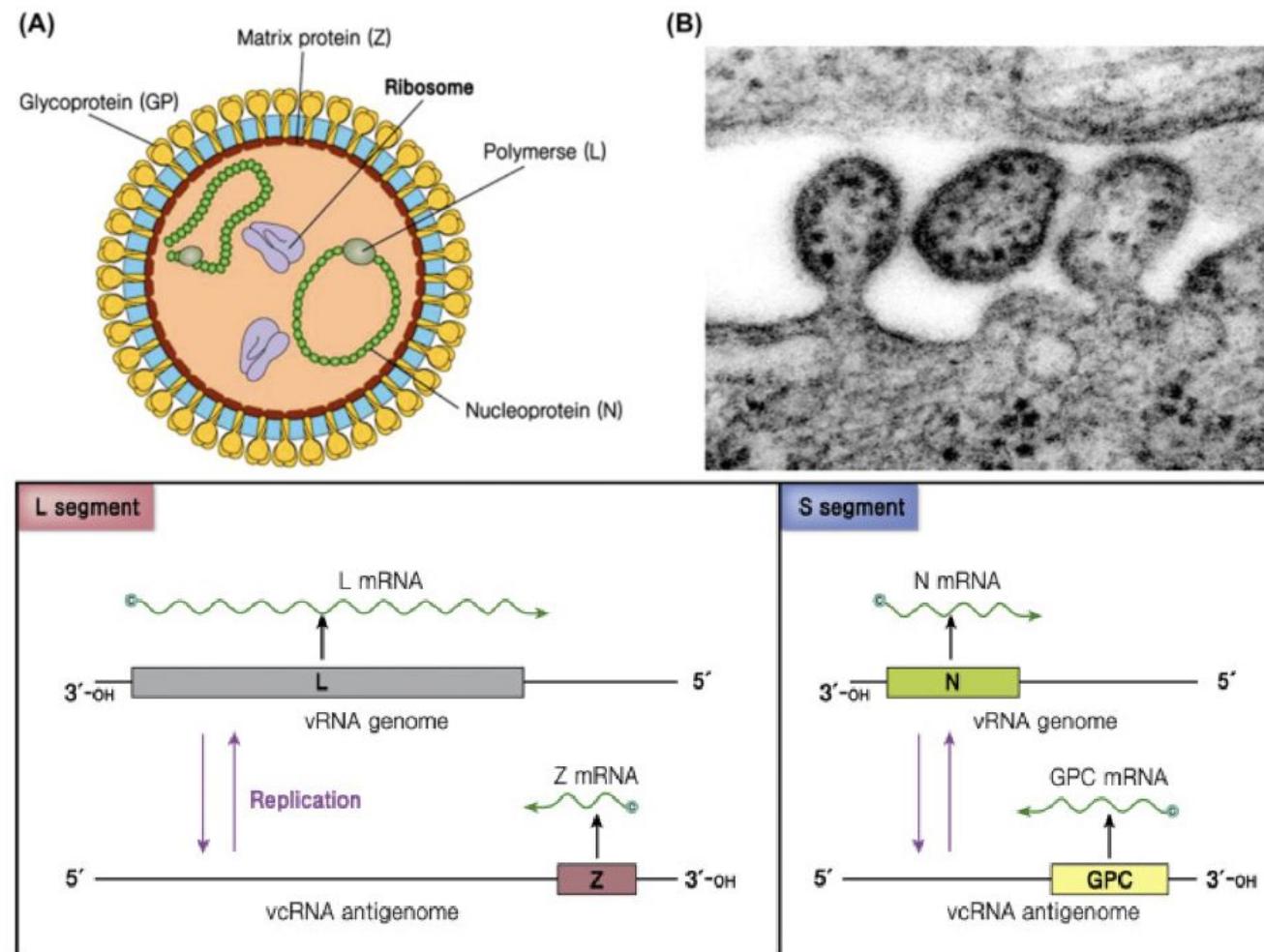
- Update the dataset
- Adhere to the pathogen-repo-guide
- Move and refactor ‘phylogenetic’ workflows into subdirectories
- Connect GitHub action automation
- Have a team of SME (subject-matter-experts) to check the validity of the site and make suggestions
- Do not “scoop” people’s work, provide a general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

About Lassa Virus

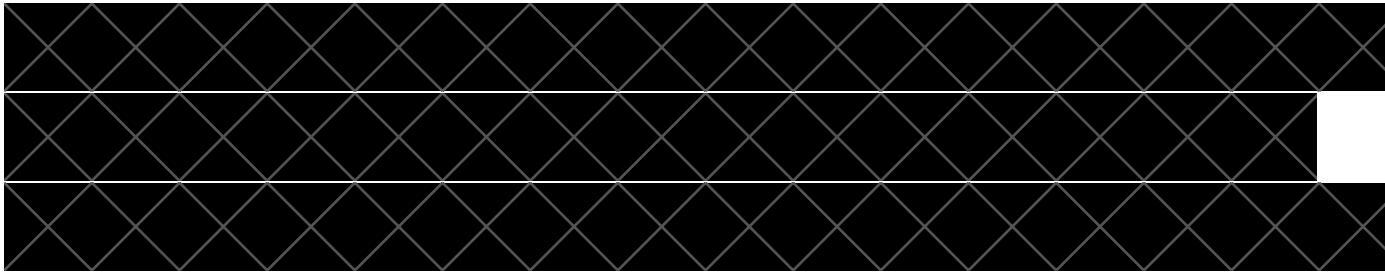
- Biosafety Level 4 (BSL-4)
- Causes severe disease with high mortality rates (15-20% in hospitalized patients)
- Spread by rats found in parts of West Africa although person-to-person transmission can occur
- No licensed vaccine
- Treatment options limited
- Stable as an aerosol, remains infectious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

Background

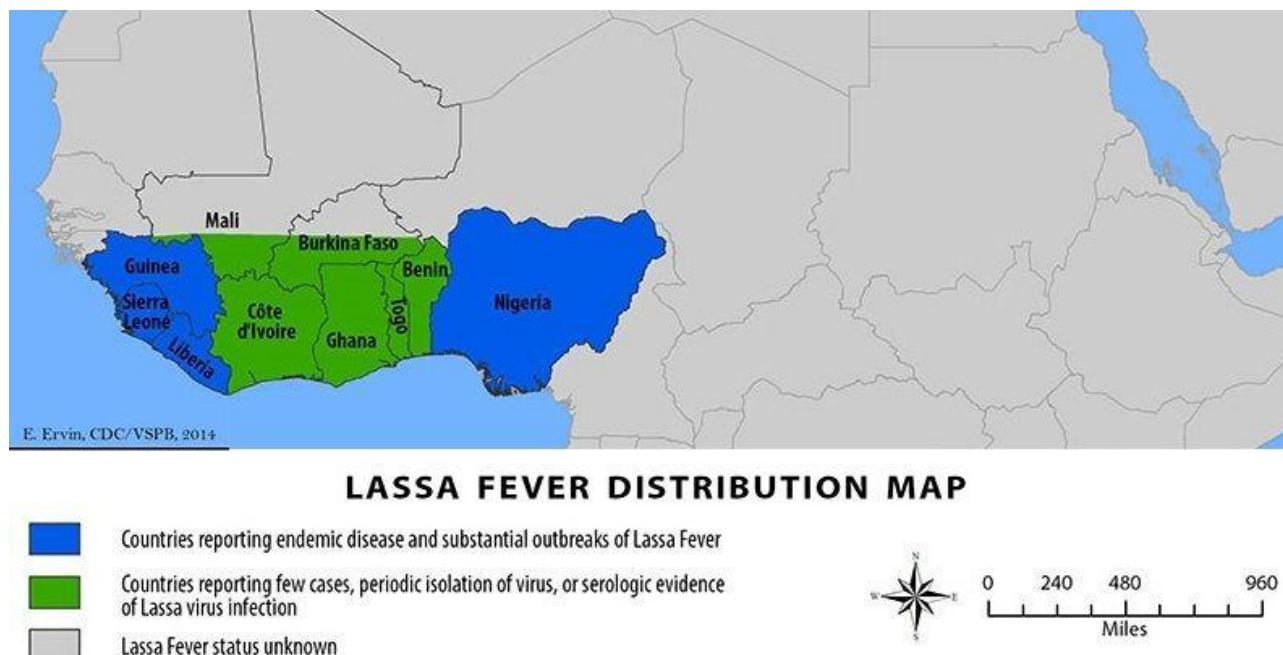
Lassa has two segments "L" and "S" from Chapter 16 of "Molecular Virology of Human Pathogenic Viruses" by Wang-Shick Ryu



List of potential SMEs



- Other suggested contacts?



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

Outline

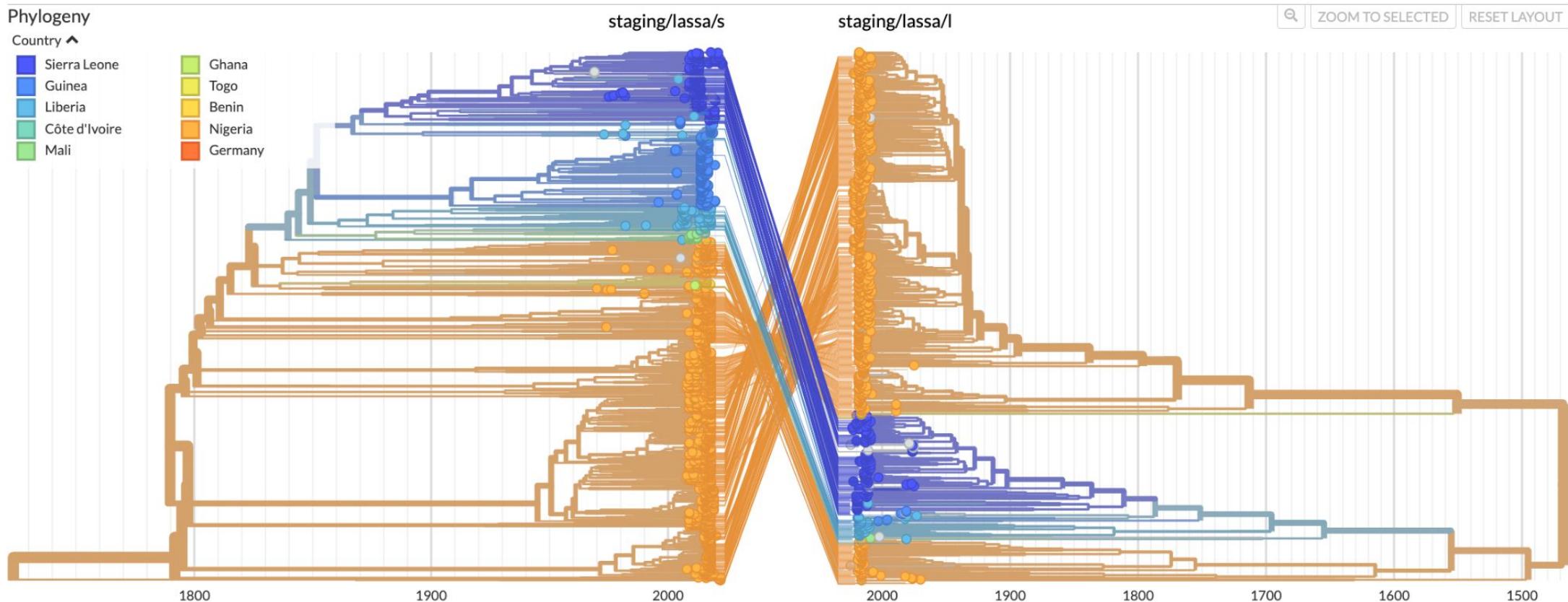
- Lessons learned from modernizing Dengue workflows
- About Lassa Virus
- **Organizing meetings with Subject Matter Experts and feedback**
 - [REDACTED]
- Thinking about external contributions
- Next steps

Update the Lassa trees

Real-time tracking of Lassa virus evolution

Maintained by the Nextstrain team. Data updated 2024-08-02. Enabled by data from GenBank.

Showing 1138 of 1138 genomes sampled between Dec 1968 and Aug 2024



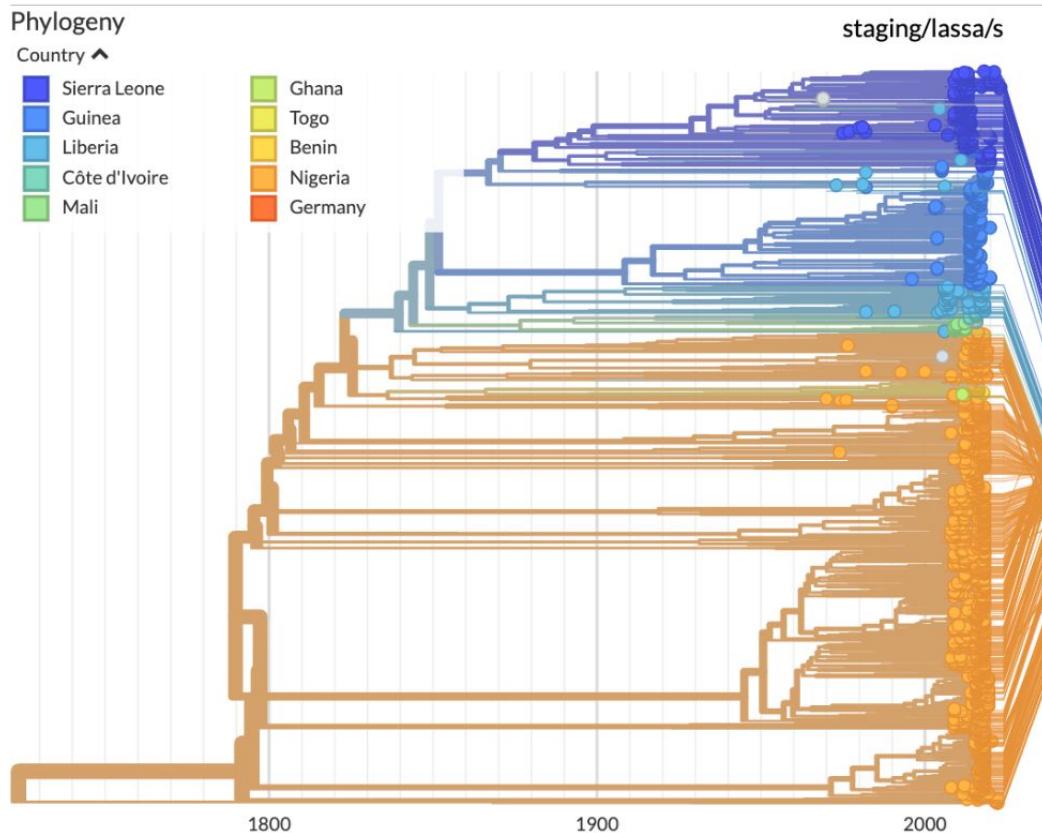
- <https://next.nextstrain.org/staging/lassa/l>

Update the Lassa trees

Real-time tracking of Lassa virus evolution

Maintained by the Nextstrain team. Data updated 2024-08-02. Enabled by data from GenBank.

Showing 1138 of 1138 genomes sampled between Dec 1968 and Aug 2024.

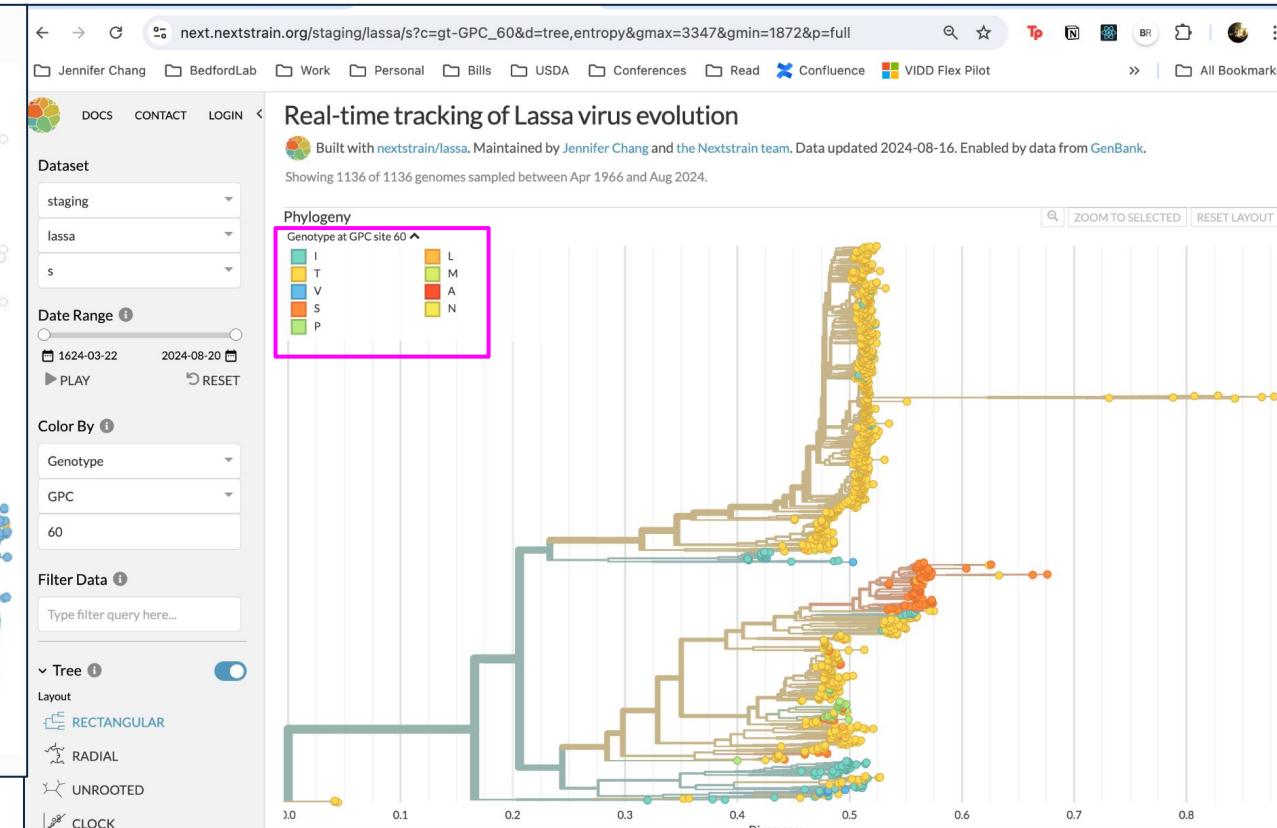
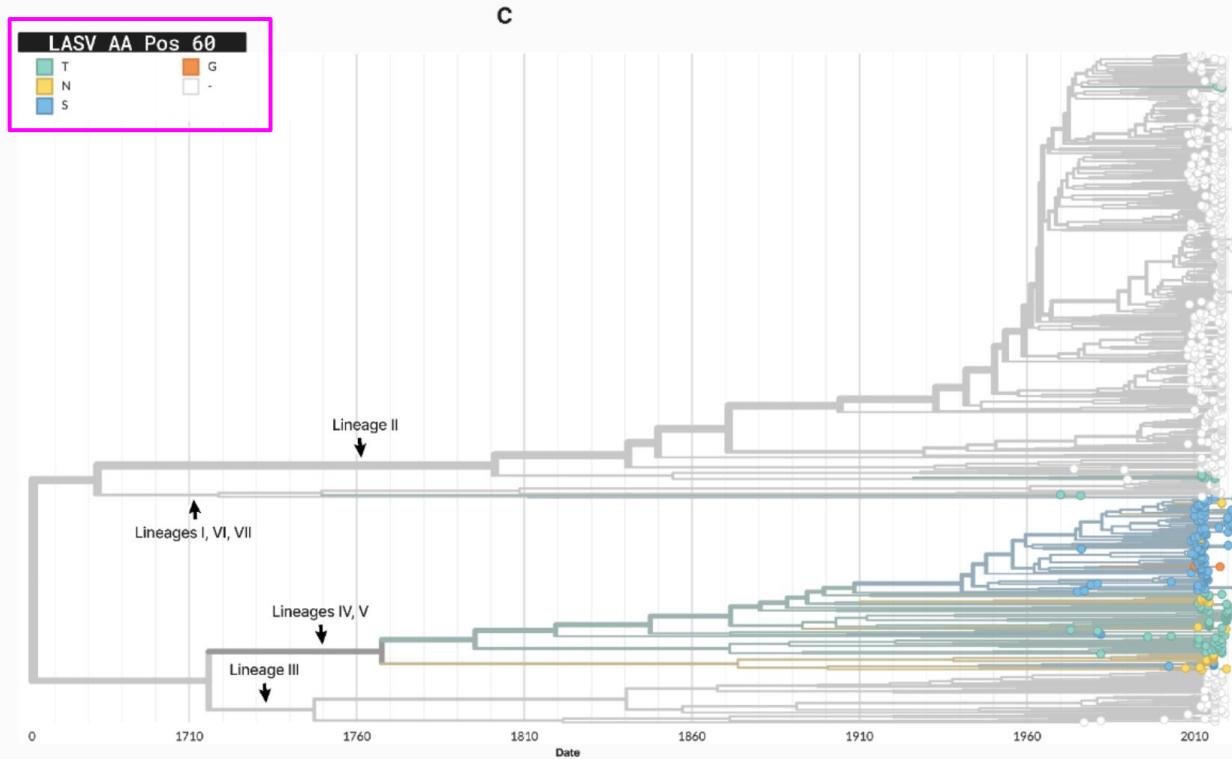


Questions for the Subject Matter Experts (SMEs)

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

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

Our phylogenetic trees

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

final_passed_sequences_manual_curated (Alignment) - Lassa

Alignment View Distances Text View Lineage Info

Consensus Frame 1

170	G	C	T	C	A	A	C	A	178	180	A	C	T	T	A	C	A	A
57	C	S	T	-	-	-	-	-	60	61	Y	T	T	62	K	-	-	-
58	-	-	-	-	-	-	-	-	61	62	-	-	-	-	-	-	-	-
59	-	-	-	-	-	-	-	-	62	-	-	-	-	-	-	-	-	-

Identity

Frame 1

4. OR493503

5. OR493504

6. OR147791

7. OR147792

8. OR147793

9. OR041676

10. OR041678

11. OR041679

12. OR041680

13. OR041681

14. OR041682

Selected 3 columns from 178 to 180 in 753 sequences.

Display

Consensus

Threshold: 0% - Majority

Ignore Gaps

If no coverage call ?

Highlight Disagreements

Go: < >

Use dots

Nucleotides

Complement

Translation

Frame: Frame 1

Genetic Code: Standard

Relative to: Reference

Colors: ARND*

Three letter amino acids

Show amino acid numbering

Selected 3 columns from 178 to 180 in 1,115 sequences.

Alignment View Distances Text View Lineage Info

Consensus Frame 1

170	T	G	C	T	C	A	A	C	A	178	180	A	C	T	T	A	C	A	A
57	T	G	C	T	C	A	A	C	A	60	T	S	L	Y	62	K	-	-	-
58	-	-	-	-	-	-	-	-	61	62	-	-	-	-	-	-	-	-	
59	-	-	-	-	-	-	-	-	62	-	-	-	-	-	-	-	-	-	

Identity

Frame 1

1. GLY_REF_LASV

4. AJ969404

5. CS272332

6. AJ969407

7. AJ969408

8. GU481063

9. AJ969409

10. FJ824031

11. AF246121

12. AF181853

13. AF333969

Selected 3 columns from 178 to 180 in 1,115 sequences.

Display

Consensus (excludes reference)

Threshold: 0% - Majority

Ignore Gaps

If no coverage call ?

Highlighting

Disagreements to Reference

Go: < > in any sequence

Use dots

Nucleotides

Complement

Translation on All Sequences

Translation Options

Frame: Frame 1

Genetic Code: Standard

Relative to: Reference

Colors: ARND*

Three letter amino acids

Show amino acid numbering

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

final_passed_sequences_manual_curated (Alignment) - Lassa

gpc_with_gff_outframepenalty_test (Alignment) - Lassa

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

Selected 5 columns from 178 to 180 in 755 sequences.

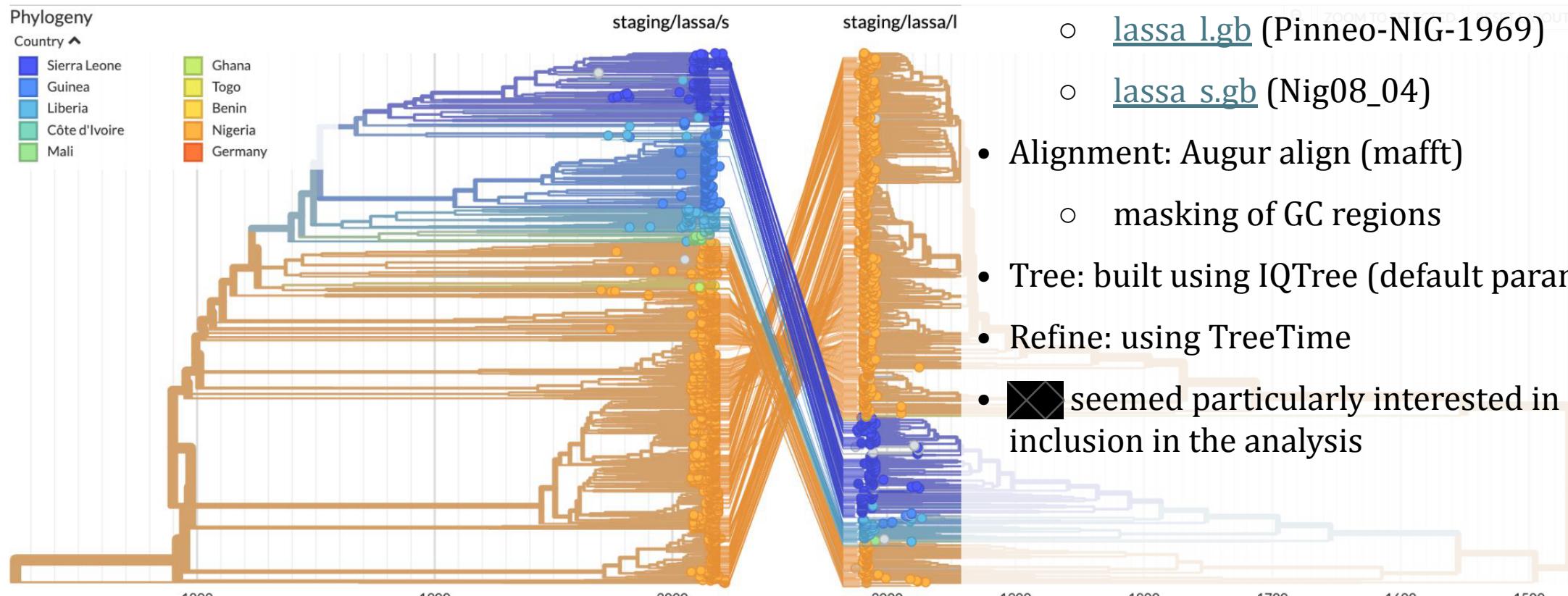
Selected 5 columns from 178 to 180 in 1,115 sequences.

Questions

Real-time tracking of Lassa virus evolution

Maintained by the Nextstrain team. Data updated 2024-08-02. Enabled by data from GenBank.

Showing 1138 of 1138 genomes sampled between Dec 1968 and Aug 2024.



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

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
- seemed particularly interested in the criteria for inclusion in the analysis

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**
 - [REDACTED]

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 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

The image shows a side-by-side comparison of two note-taking interfaces. On the left is HackMD, a code editor-style interface with a dark theme. On the right is Lassa Notes, a clean white interface with a sidebar.

HackMD (Left):

```
14 ## To Do
15
16 **GPC Tree**
17
18 * [x] **Richard** - Submit a Github Issue requesting a GCP Tree
    * 3 subunits, important to know the subunits for different functions
    * signal maturation of the gpc
        https://github.com/nextstrain/lassa
19
20
21
22
23 * [x] **Richard** - Create a \[fork\] (https://github.com/JoiRichi/lassa): —Or create new branch `add gcp tree`
24
25 * [x] **Richard** - Add reference GenBank for GCP to repo (perhaps in `shared` and `phylogenetic/defaults`) and submit a PR
26 * [x] **Jennifer** - Get clarification on reference genbanks for different parts of the build
    * L segment: `Pinneo-NIG-1969` \[lassa\_l.gb\] (https://github.com/nextstrain/lassa/blob/main/phylogenetic/defaults/lassa\_l.gb)
        * lineage 1, has a deletion around GPC 62
        * might cause nextclade v3 to cut this region out of the regular sequences during L and S assignment
            * manually fix the alignment (alignment parameters)
    * S segment: `Nig08_04` \[lassa\_s.gb\] (https://github.com/nextstrain/lassa/blob/main/phylogenetic/defaults/lassa\_s.gb)
        * GCP region: `Josiah` strain \[gly\_ref\_LASV.gb\] (https://github.com/JoiRichi/LASV\_phylogenetics\_pipeline/blob/main/config/gly\_ref\_LASV.gb)
            * [x] **Jennifer** - Got clarification to official move to 'Josiah' strain for reference and PR merged with https://github.com/nextstrain/lassa/pull/23
27
28
29
30
31
32
33
34
```

Lassa Notes (Right):

To Do

GPC Tree

- Richard** - Submit a Github Issue requesting a GCP Tree
 - 3 subunits, important to know the subunits for different functions
 - signal maturation of the gpc
 - <https://github.com/nextstrain/lassa>
- Richard** - Create a [fork](#): ~~Or create new branch `add gcp tree`~~
- Richard** - Add reference GenBank for GCP to repo (perhaps in `shared` and `phylogenetic/defaults`) and submit a PR
- Jennifer** - Get clarification on reference genbanks for different parts of the build
 - L segment: `Pinneo-NIG-1969 lassa_l.gb`
 - lineage 1, has a deletion around GPC 62
 - might cause nextclade v3 to cut this region out of the regular sequences during L and S assignment
 - manually fix the alignment (alignment parameters)
 - S segment: `Nig08_04 lassa_s.gb`
 - GCP region: `Josiah` strain `gly_ref_LASV.gb`
- Jennifer** - Got clarification to official move to 'Josiah' strain for reference and PR merged with <https://github.com/nextstrain/lassa/pull/23>

Add nextclade v3 rules to pull out GCP in `ingest` (to follow segment pattern) or justify adding it in `phylogenetic`

[link](#)

HackMD to coordinate with Richard Daudu

Submit a GitHub Issue

The screenshot shows a GitHub issue page for the repository 'nextstrain / lassa'. The title of the issue is 'Request to add a GPC Phylogenetic Tree #20'. It was opened by 'JoiRichi' 2 weeks ago and has 5 comments. A comment from 'JoiRichi' is visible, providing context about the glycoprotein complex (GPC) of the Lassa virus and its role in evolution and immune escape. The issue is currently open.

Context

The glycoprotein complex (GPC) is the only surface protein of the Lassa virus (LASV) and plays a crucial role in mediating entry into host cells. It is also a primary target for neutralizing antibodies and vaccine design efforts. Meanwhile, different lineages of LASV circulates in non-overlapping regions in West Africa and have differences in their GPC ([Daodu et al. 2024](#)). Understanding the phylogenetic relationships of GPC sequences can provide insights into the virus evolution, mechanisms of immune escape, and guide therapeutic development and distribution.

While segment-based trees (S and L) provide valuable information, a GPC-specific tree would complement these by offering a more detailed view of the evolutionary dynamics of the glycoprotein. This can be particularly useful in understanding the genetic diversity and evolutionary pressures acting on the GPC, which may not be fully captured by segment-based analyses.

In addition, a few recent papers have reported recombination within the virus segments, which may lead to inaccurate tree topologies when investigating gene specific questions ([He et al. 2024](#); [Wang et al. 2024](#)).

Description

Currently, the Nextstrain build for LASV generates phylogenetic trees by segment (S or L). Is it possible to add generation of a phylogenetic tree based on the GPC of the Lassa virus?

Possible solution

- Using a GPC reference, manually created here:
https://github.com/JoiRichi/LASV_phylogenetics_pipeline/blob/main/config/gly_ref_LASV.gb
- Create a GPC phylogenetic tree

Submit a GitHub PR

The screenshot shows a GitHub pull request page for the repository 'nextstrain / lassa'. The title of the PR is 'Adds required GPC references #21'. It was merged by 'j23414' 2 weeks ago. The PR has 3 conversations, 1 commit, 3 checks, and 2 files changed. A comment from 'JoiRichi' is visible, describing the proposed changes as adding required GPC reference files for creating GPC trees. The PR is related to issue #20.

Description of proposed changes

For creating GPC trees, this PR adds the required GPC reference files.

Related issue(s)

- Related to: [Request to add a GPC Phylogenetic Tree #20](#)

Checklist

Checks pass

j23414 added required GPC references f4f02f2

j23414 changed the title [added-required-GPC-references](#) Adds required GPC references 2 weeks ago

j23414 self-requested a review 2 weeks ago

Acknowledgements

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