Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

Reference Sets in Viral Genomics:

Robert J. Gifford

What is a Reference Set?

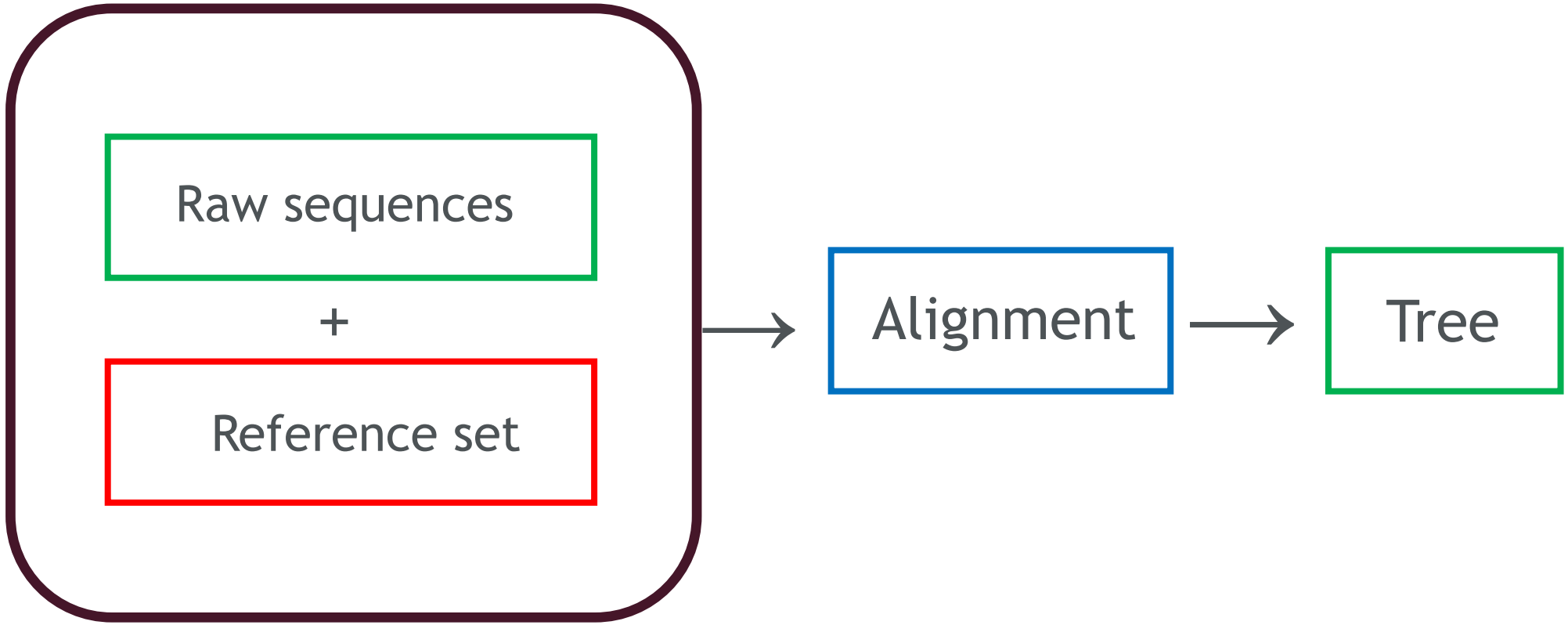
Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

- A *curated collection* of sequences chosen to anchor analysis.

- Provides **context** for:
  - Alignments
  - Phylogenetic trees
  - Genotyping and annotation
- Acts as a **baseline** against which new or unknown sequences are compared.

What is a Reference Set?



Why Do Reference Sets Matter?

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

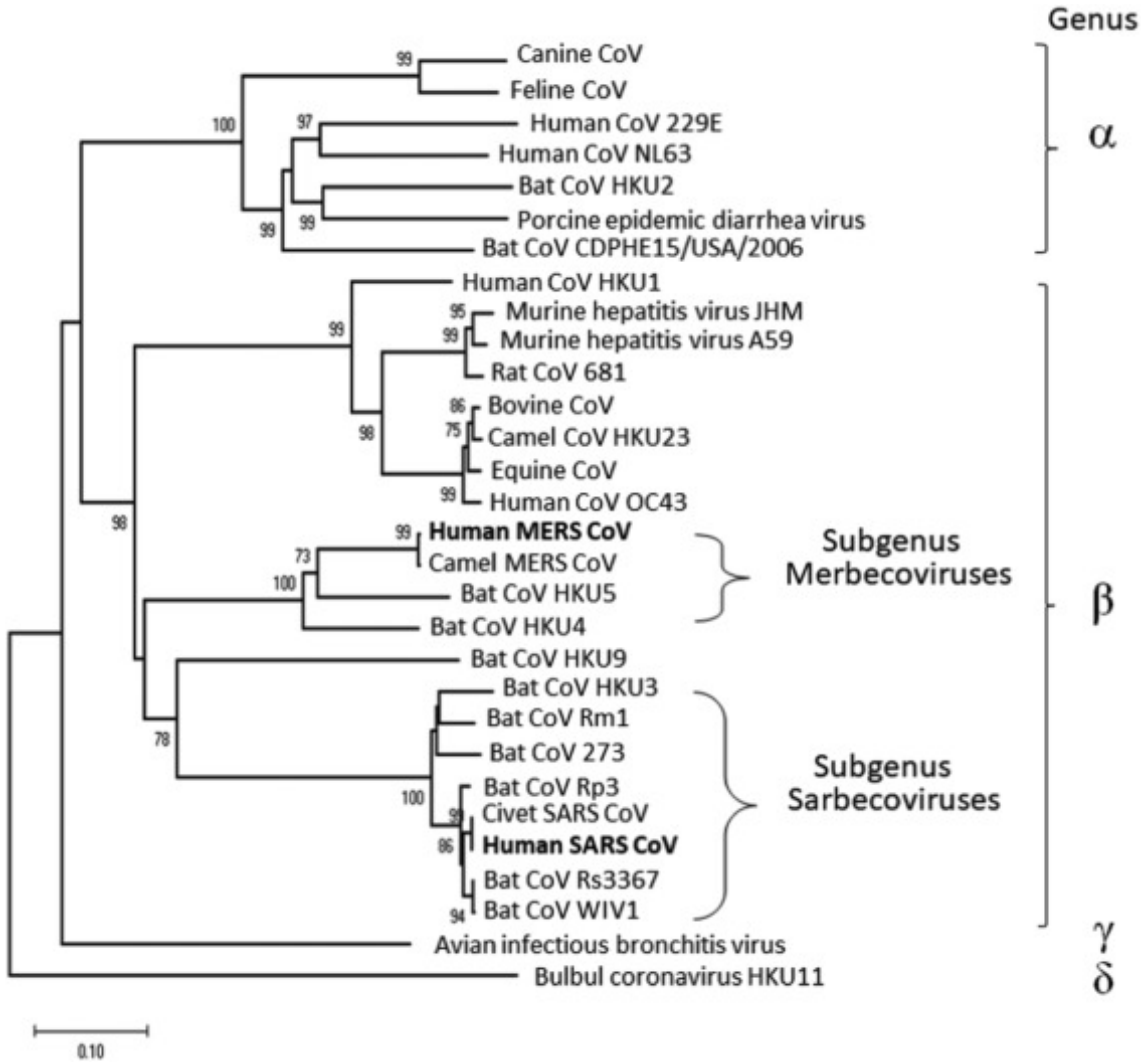
CERI
Centre for Epidemic Response and Innovation

- Shape the **accuracy of alignments**
  - bad references → misaligned regions.
- Influence **phylogenetic trees**
  - different reference sets can shift topology.
- Reduce **bias** by balancing representation.
- Provide **interpretability**
  - results only make sense relative to good context.

Why Do Reference Sets Matter?

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

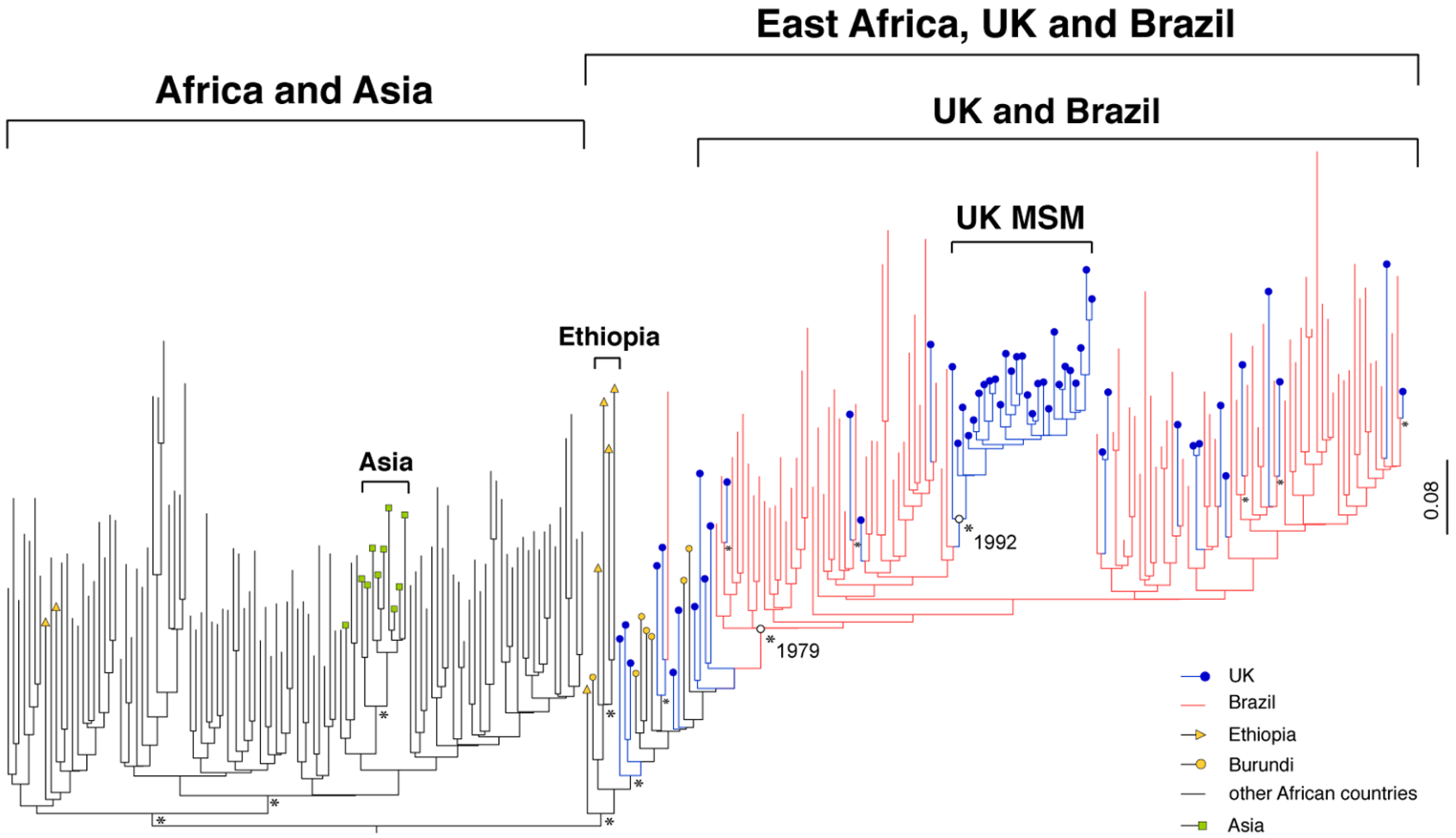
CERI
Centre for Epidemic Response and Innovation



Why Do Reference Sets Matter?

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation



Sources of Reference Sequences

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

GenBank

- **Strengths:**
  - Largest open-access virus database.
  - Long-term archive with >1M IAV sequences, >500k IBV.
- **Challenges:**
  - Genome segments stored separately → harder to reconstruct full genomes.
  - Inconsistent isolate metadata (host, location, date, clinical info).
  - Variable accuracy/annotation quality.
  - Limited taxonomy (often only subtype level, lineage rarely consistent).
  - Very large datasets → difficult to manage at scale.

Sources of Reference Sequences

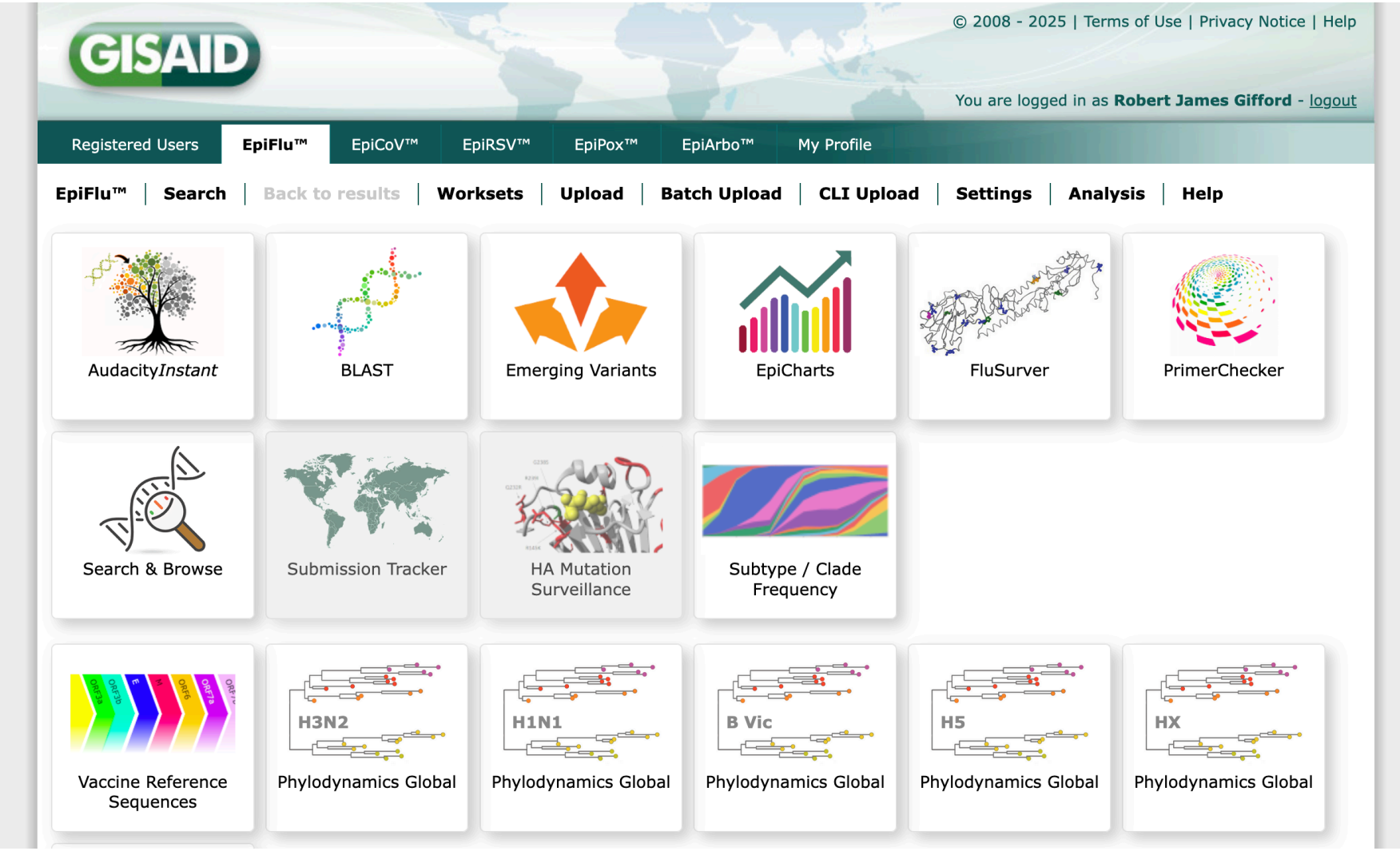
Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

GISAID

- **Strengths:**
  - Rich, curated metadata (host, date, location standardized).
  - Near-real-time coverage of influenza & SARS-CoV-2.
  - Lineage/phylogeny assignments often included.
  - Collaborative ethos → credit to data generators.
- **Challenges:**
  - Controlled access → must register, agree to terms.
  - Restricted redistribution of sequences.
  - Coverage focused on major pathogens (influenza, SARS-CoV-2) — not comprehensive.
  - Metadata occasionally inconsistent across submitters.

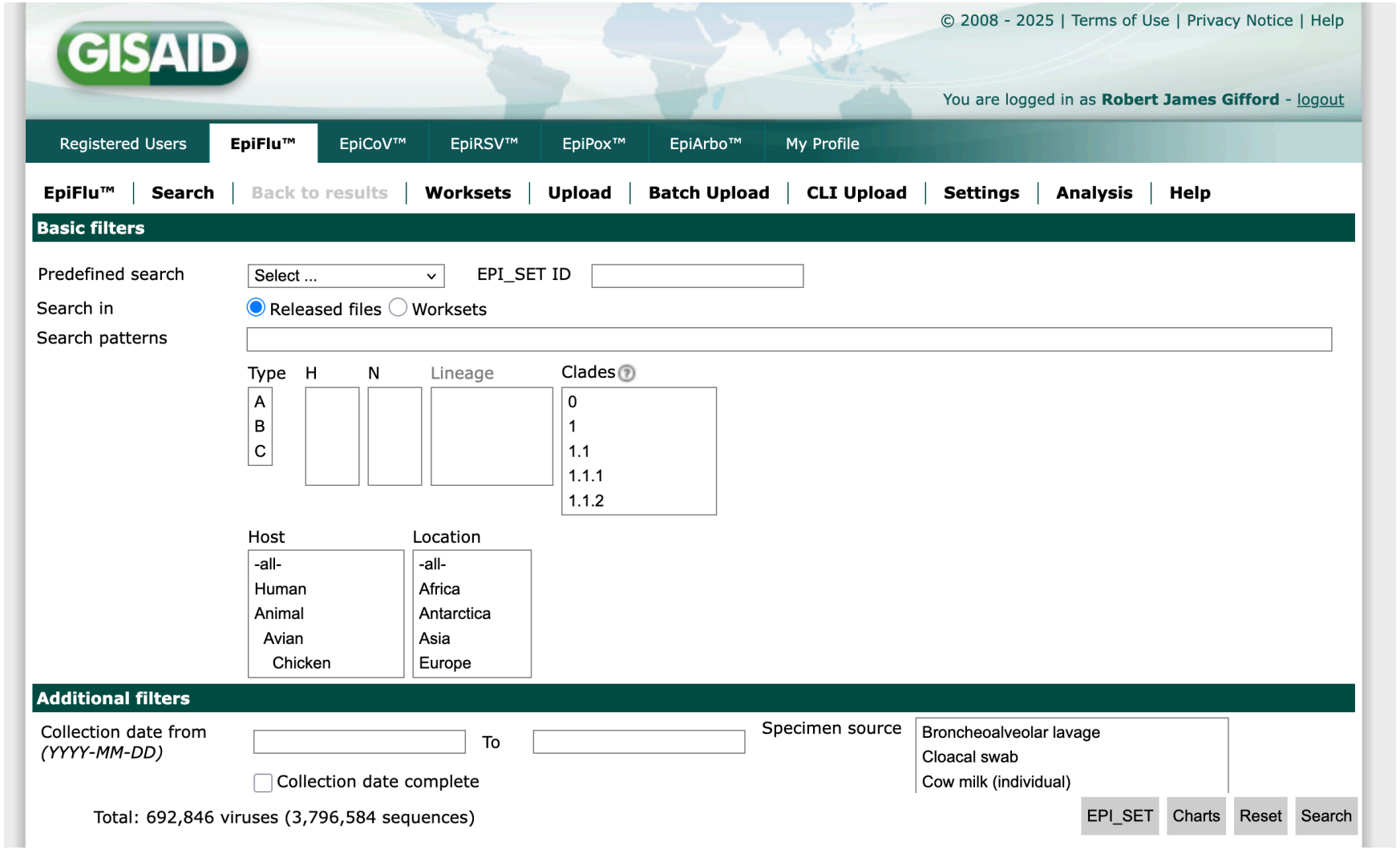
Sources of Reference Sequences



Sources of Reference Sequences

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation



Sources of Reference Sequences

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

NextStrain

- **Strengths**
  - Provides curated reference datasets for specific pathogens.
- **Challenges**
  - Reference sets tailored to specific builds → not comprehensive.
  - Coverage focused on major pathogens (e.g. Flu, SARS-CoV-2, Ebola).
  - Less suited for deep historical or lineage-wide reference curation.
  - Not a general-purpose sequence archive.

Sources of Reference Sequences

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

PathoPlexus

- **Strengths**
  - Open-source, non-profit platform for *all human viral pathogens*.
  - Integrates with Nexstrain, INSDC (NCBI/ENA/DDBJ).
  - Community-driven governance, transparent policies.
  - Offers SeqSets and APIs for filtering/search.
  - Flexible data-sharing: immediate open or embargoed.
- **Challenges**
  - Still emerging → limited pathogen coverage so far.
  - Focused on human pathogens, not broader host range.
  - Curation depth varies by virus.
  - Dependent on community participation for sustainability.

Sources of Reference Sequences

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

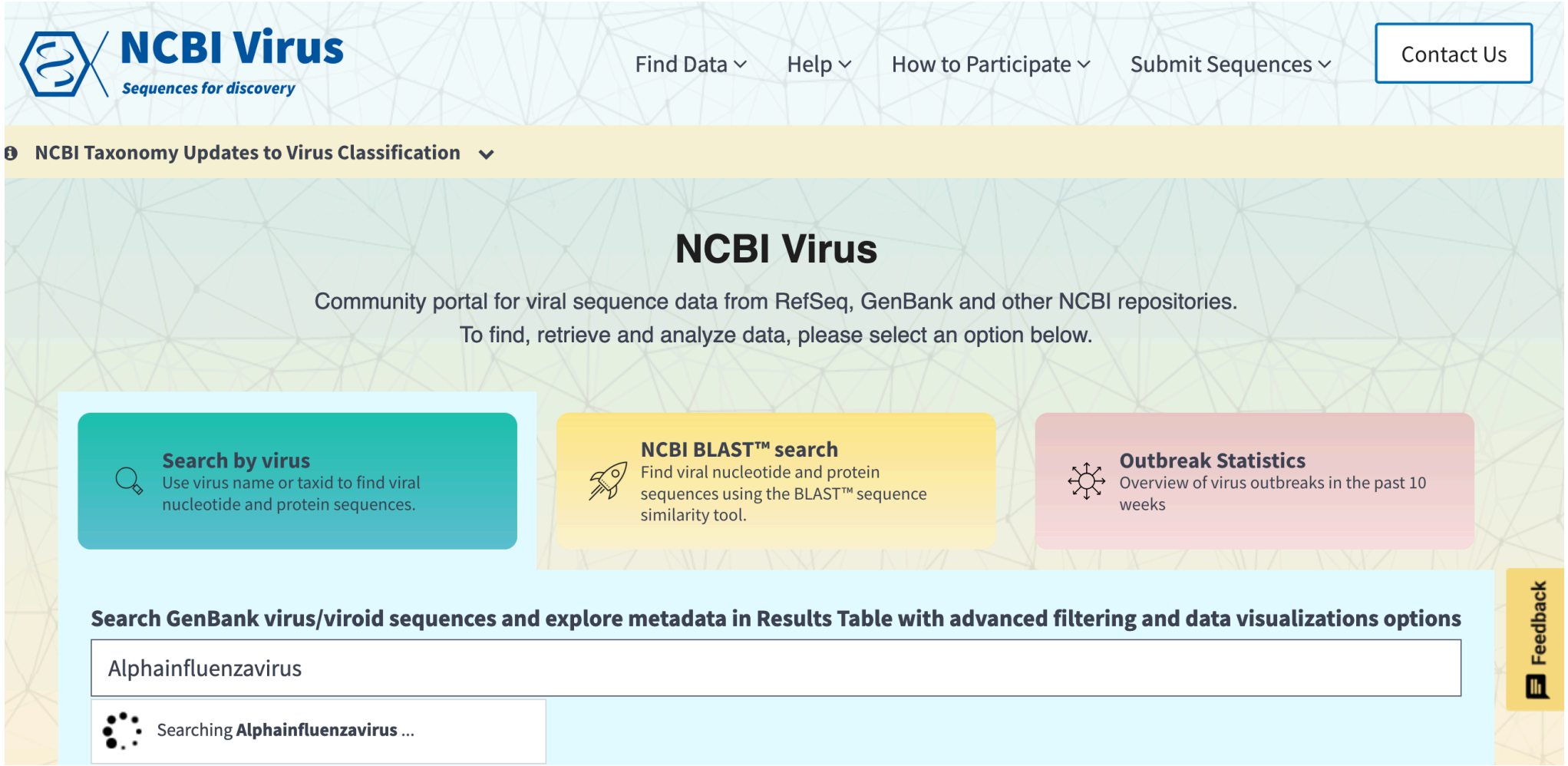
GLUE

- **Strengths**
  - Openly available isolate-organised table (iav_nuccore_isolates.tsv).
  - Metadata cleaned, standardised, and cross-checked using GLUE.
  - Links isolates to all 8 genome segments → easy accession retrieval.
  - Simple to filter/subsample with Unix tools.
  - Fits directly into Nextclade/Augur/Auspice workflows.
- **Challenges**
  - Not leveraging GLUE's full interactive framework (command layer, modules).
  - Updates depend on periodic refresh, not real-time feeds.
  - Coverage only for influenza (not all viruses).
  - Some metadata fields less detailed than GISAID (e.g. passage history).

Virus Reference Sequences

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation



GenBank Influenza Virus Entry

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

```
REFERENCE   3  (bases 1 to 890)
  CONSRTM   The NIAID Influenza Genome Sequencing Consortium
  TITLE     Direct Submission
  JOURNAL   Submitted (01-AUG-2005) on behalf of TIGR/Wadsworth-NYSDOH/NCBI,
            National Center for Biotechnology Information, NIH, Bethesda, MD
            20894, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from CY002068.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..890
                     /organism="Influenza A virus (A/New York/392/2004(H3N2))"
                     /mol_type="viral cRNA"
                     /strain="A/New York/392/2004"
                     /serotype="H3N2"
                     /isolation_source="gender:M; age:16y"
                     /host="Homo sapiens"
                     /db_xref="taxon:335341"
                     /segment="8"
                     /lab_host="RhMK 1 passage(s)"
                     /geo_loc_name="USA: Tompkins County, NY"
                     /collection_date="21-Dec-2004"
     gene            27..864
                     /gene="NEP"
                     /locus_tag="FLUAVH3N2_s8p1"
                     /gene_synonym="NS2"
                     /db_xref="GeneID:3655157"
     CDS             join(27..56,529..864)
                     /gene="NEP"
```

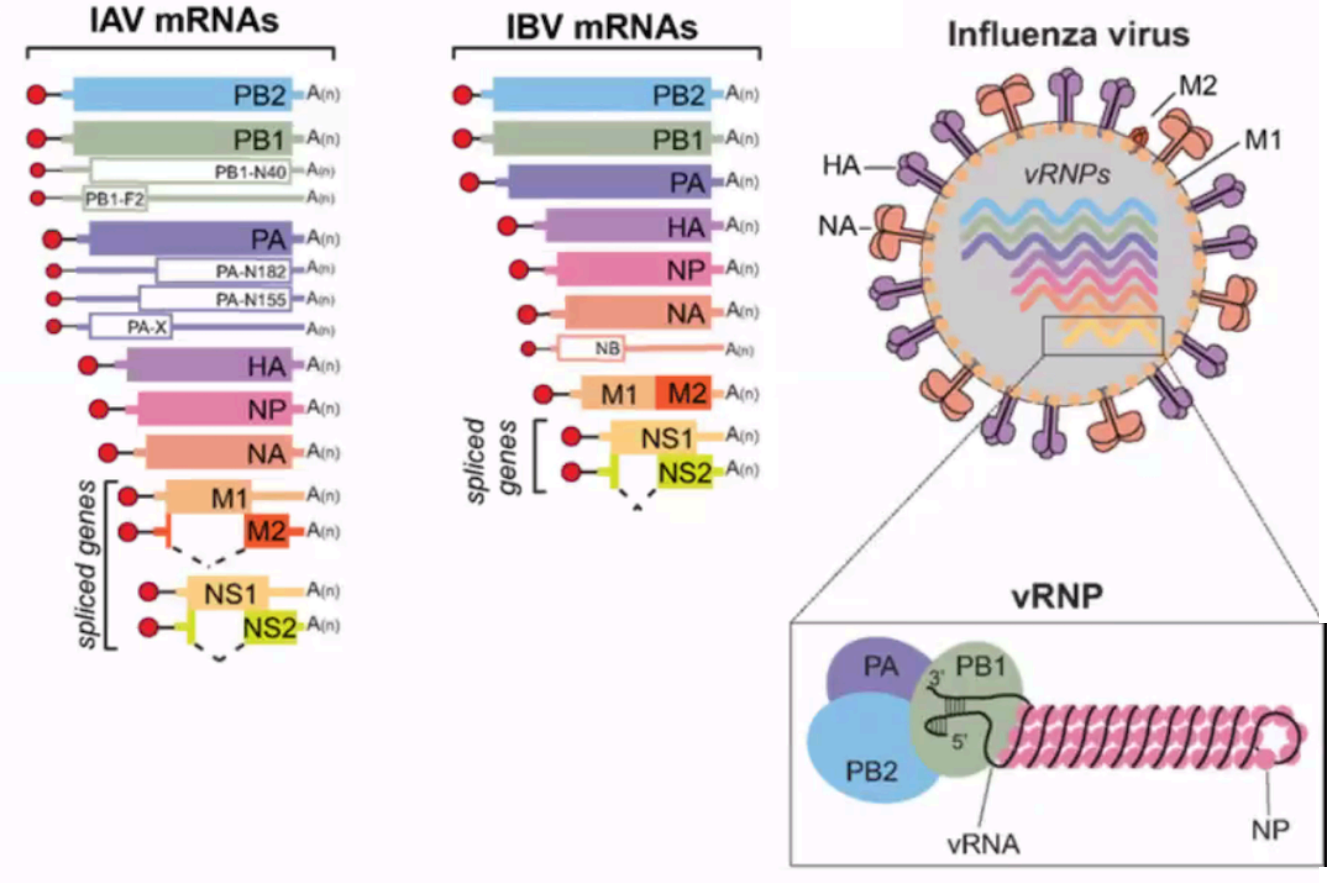
GenBank Influenza Virus Entry

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

- Each genome segment = separate record.
- Isolate details partly in strain name, partly in extra fields.
- Metadata (host, location, date) often inconsistent or missing.
- For influenza, formatting is *more standardised* than for many viruses — but still patchy.

Influenzavirus Genomes



Selecting Reference Sets (Web)

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

- **GenBank:**
  - Searchable, but limited filters.
  - Hard to select by segment or isolate fields.
- **GISAID:**
  - Advanced filters (host, date, clade, geography).
  - Better for curated sets — but export restrictions.
- **Nextstrain:**
  - Provides curated, ready-to-use reference sets.
  - Useful as examples/templates, not always comprehensive.

Selecting Reference Sets (CLI)

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

- **Entrez Tools (NCBI):**
  - esearch + efetch → programmatic access.
  - Reproducible queries (e.g. by host, segment, date).

- **Nextstrain Augur:**
  - Higher-level filtering & downsampling.
  - Integrates metadata + sequence selection.

Why Use Docker?

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

- **Consistency:** same environment across all computers.
- **Portability:** runs on Linux, Mac, Windows.
- **Reproducibility:** ensures identical results for everyone.
- **Convenience:** no dependency hell (prepackaged software).

Quality Control of Reference Sets

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

Nextclade

- Remove **low-quality or incomplete** sequences.
- Check for **frame shifts, stop codons, sequencing errors.**
- Confirm **correct genotyping** (e.g. HA subtype for influenza).
- Tools: **Nextclade** (genotyping + QC in one step).

Alignment & Tree Building

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

- **Multiple Sequence Alignment (MSA):**
  - Align sequences against reference set.
  - Anchors homologous sites across genomes.
- **Tree Building:**
  - Infer evolutionary relationships from aligned data.
  - Reference set provides stable “scaffold” for new sequences.
- **Tools (Nextstrain, Docker):**
  - Alignment → `augur align`
  - Phylogeny → `augur tree + augur refine`

Key Points

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

- Reference sets = **foundation** of viral sequence analysis.
- Good sets are: **representative, high-quality, balanced, fit-for-purpose.**
- Sources differ:
  - **GenBank** = comprehensive but messy.
  - **GISAID** = curated but controlled.
  - **Pathoplexus/NextStrain/GLUE** = pre-curated, niche
- Web portals = **exploration**.
- Command line + Docker = **reproducibility**.
- **QC, alignment, and tree building** depend on having a robust reference set.

Now It’s Your Turn

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI Centre for Epidemic Response and Innovation

We’ll walk through:

- **Downloading sequences** (GenBank + GISAID web).
- **Cleaning and preparing sequences & metadata**
- **Running QC** with Nextclade.
- **Filtering** with Augur.
- **Building and viewing** a tree with Nextstrain.

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

CERI
Centre for Epidemic Response and Innovation

OVER TO YOU!!

Photo by Stefan Els