

Reference Sets in Viral Genomics:

Robert J. Gifford



What is a Reference Set?

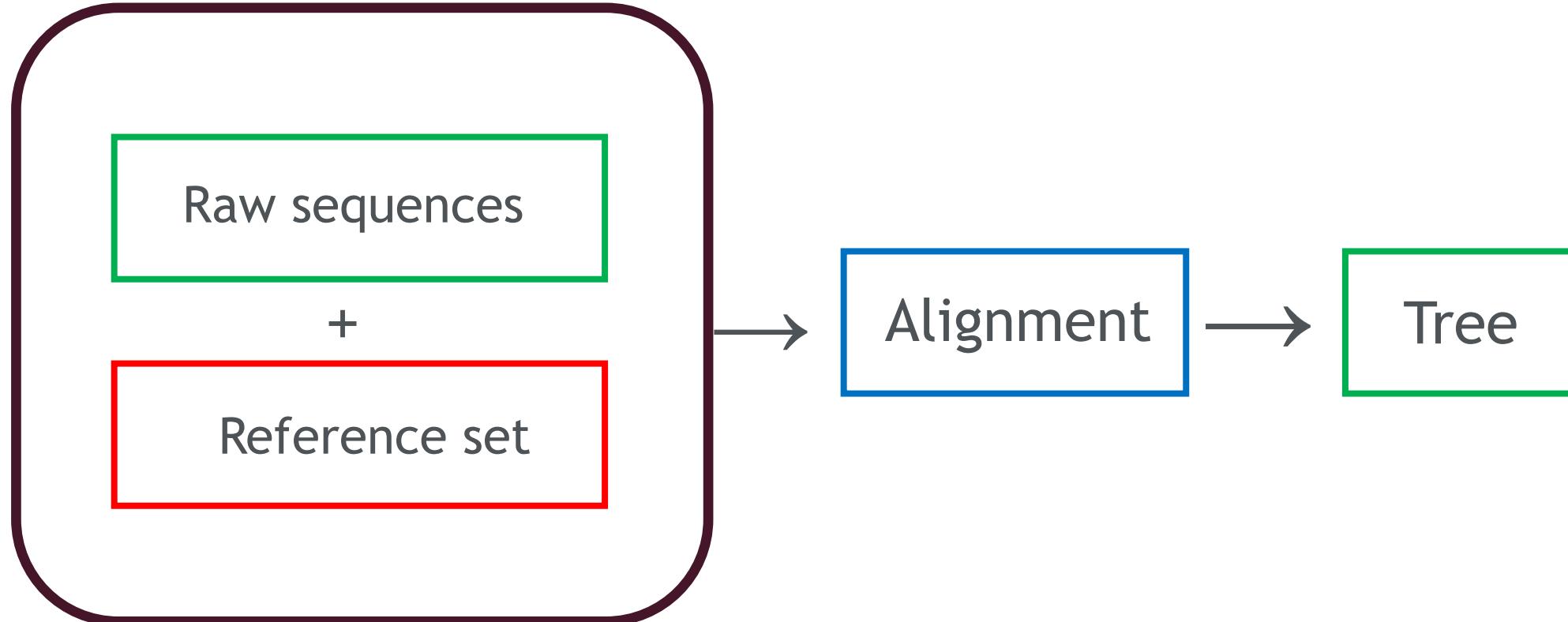


CERI
Centre for Epidemic
Response and Innovation

- 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?



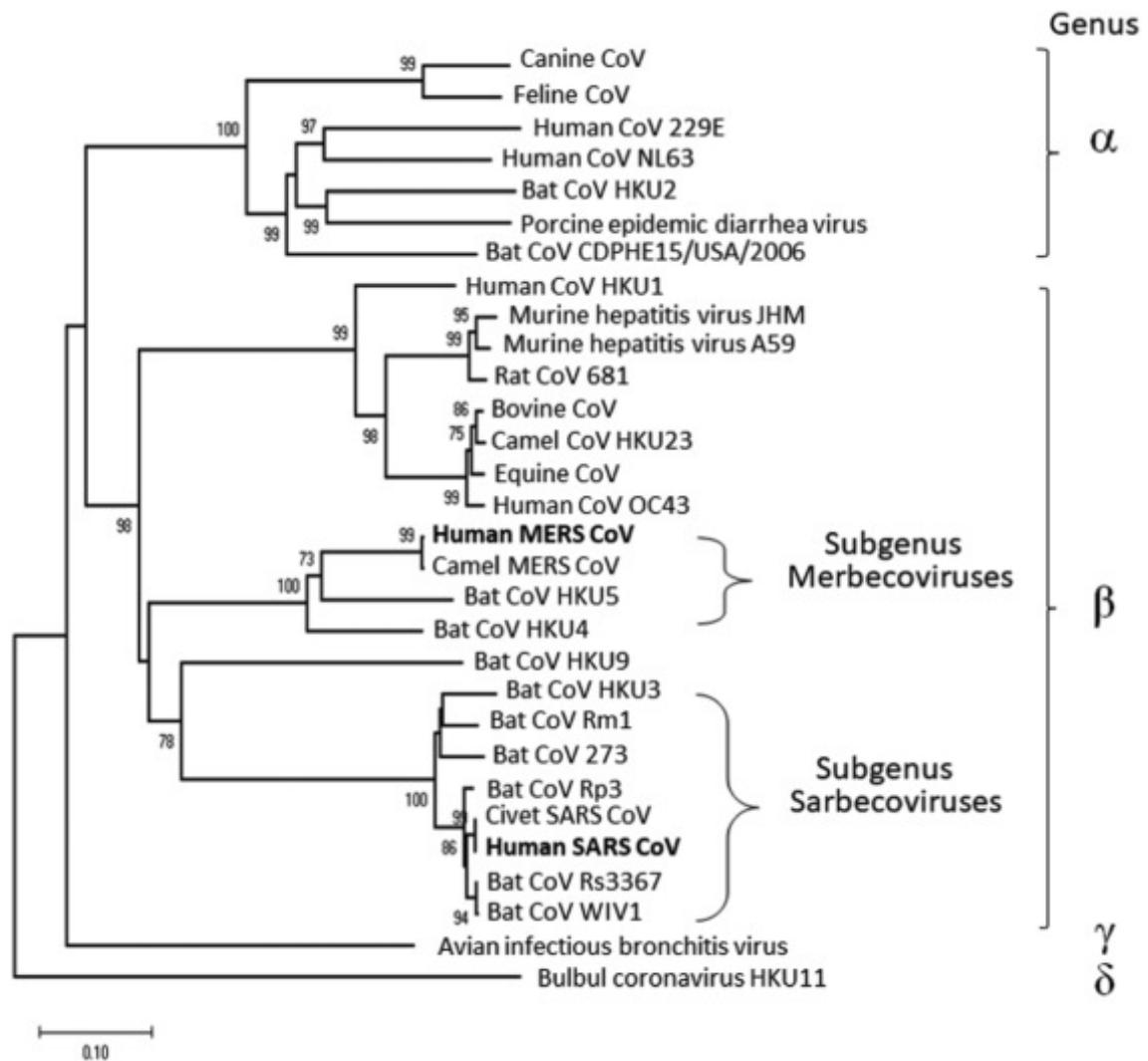
Why Do Reference Sets Matter?



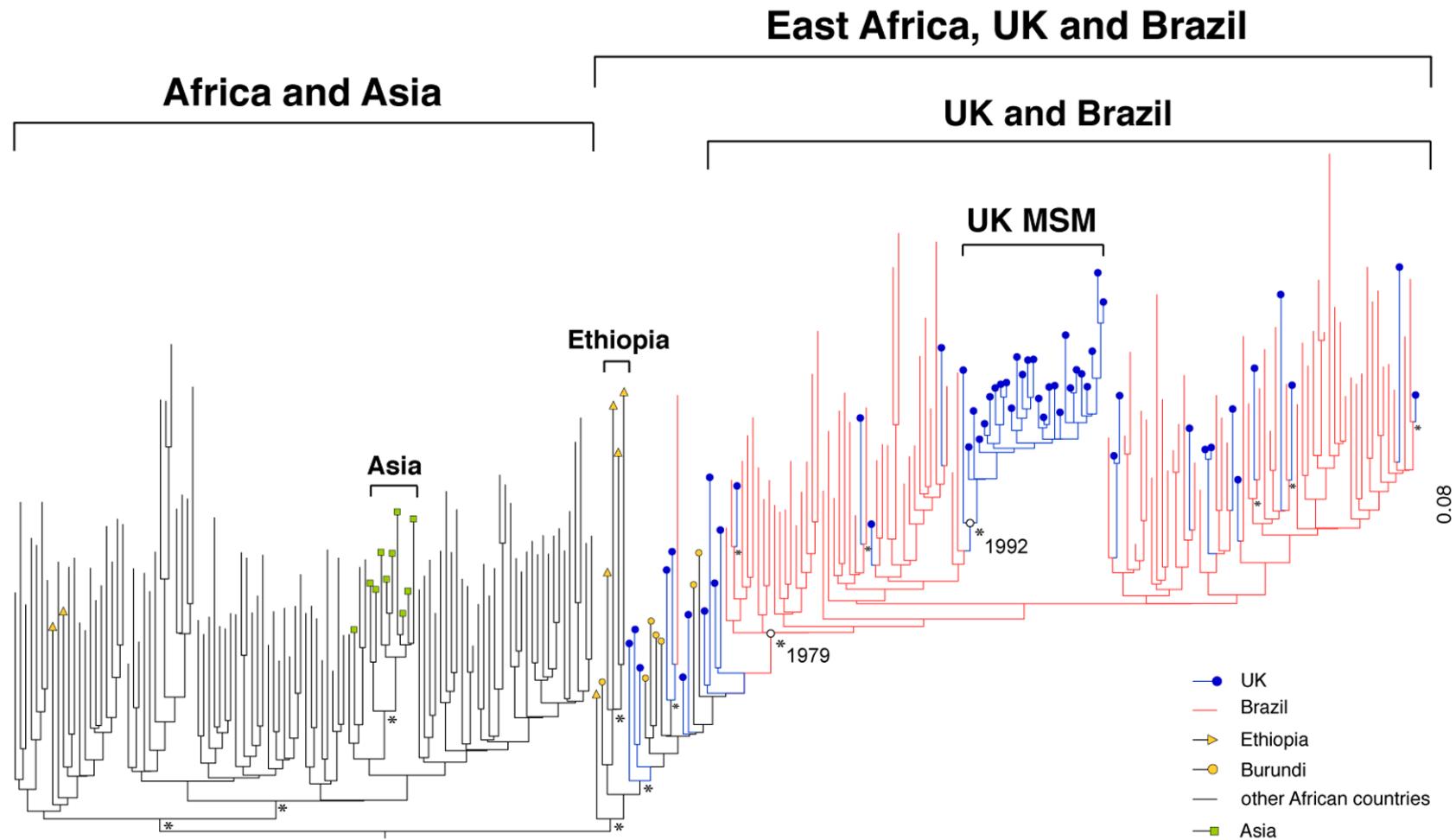
CERI
Centre for Epidemic
Response and Innovation

- 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?



Why Do Reference Sets Matter?



Sources of Reference Sequences



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



CERI
Centre for Epidemic
Response and Innovation

GISAID

- **Strengths:**

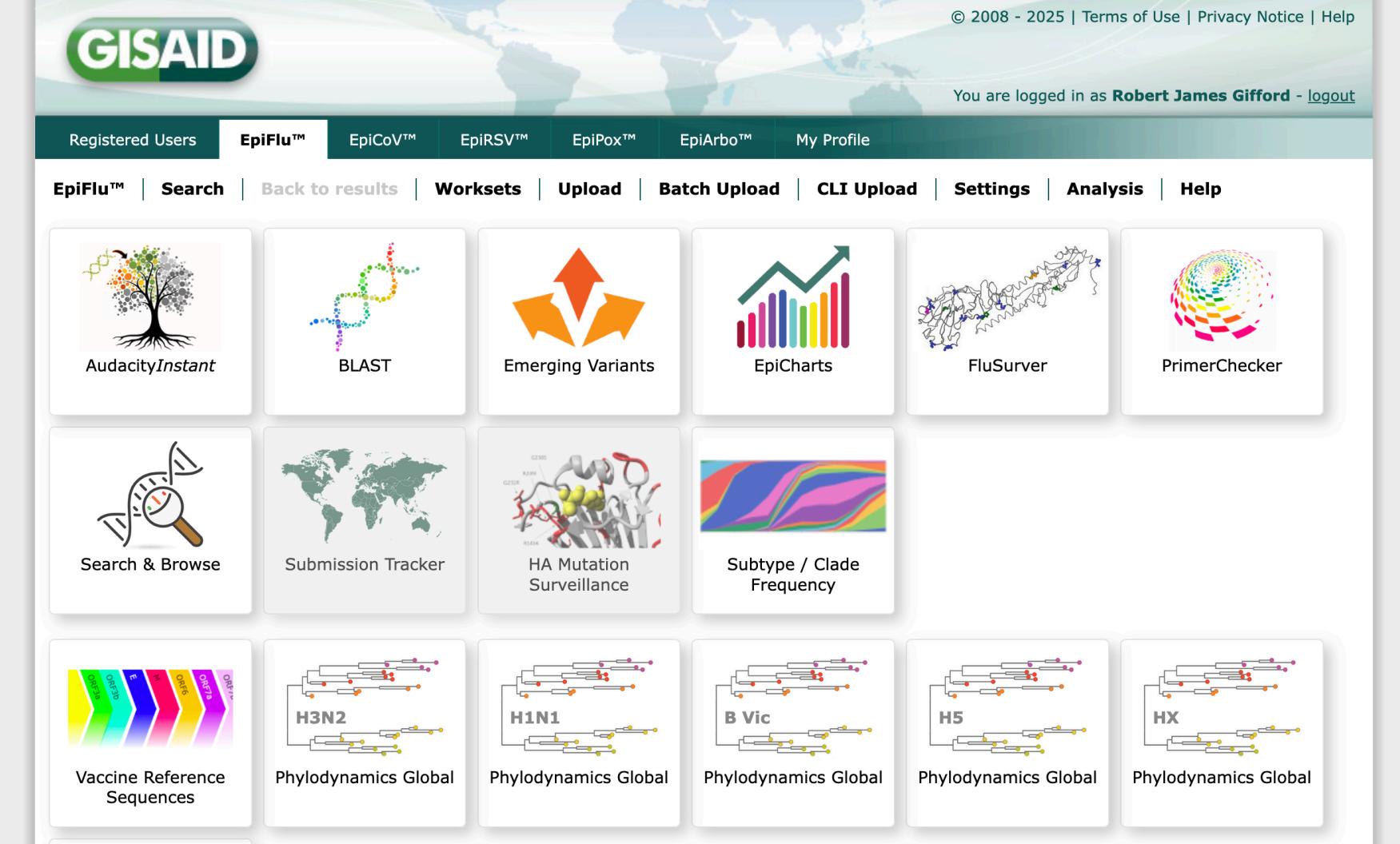
- Rich, curated metadata (host, date, location standardized).
- Near-real-time coverage of influenza & SARS-CoV-2.
- Lineage/phylogeny assignments often included.
- Collaborative ethos → credit to data generators.

- **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.

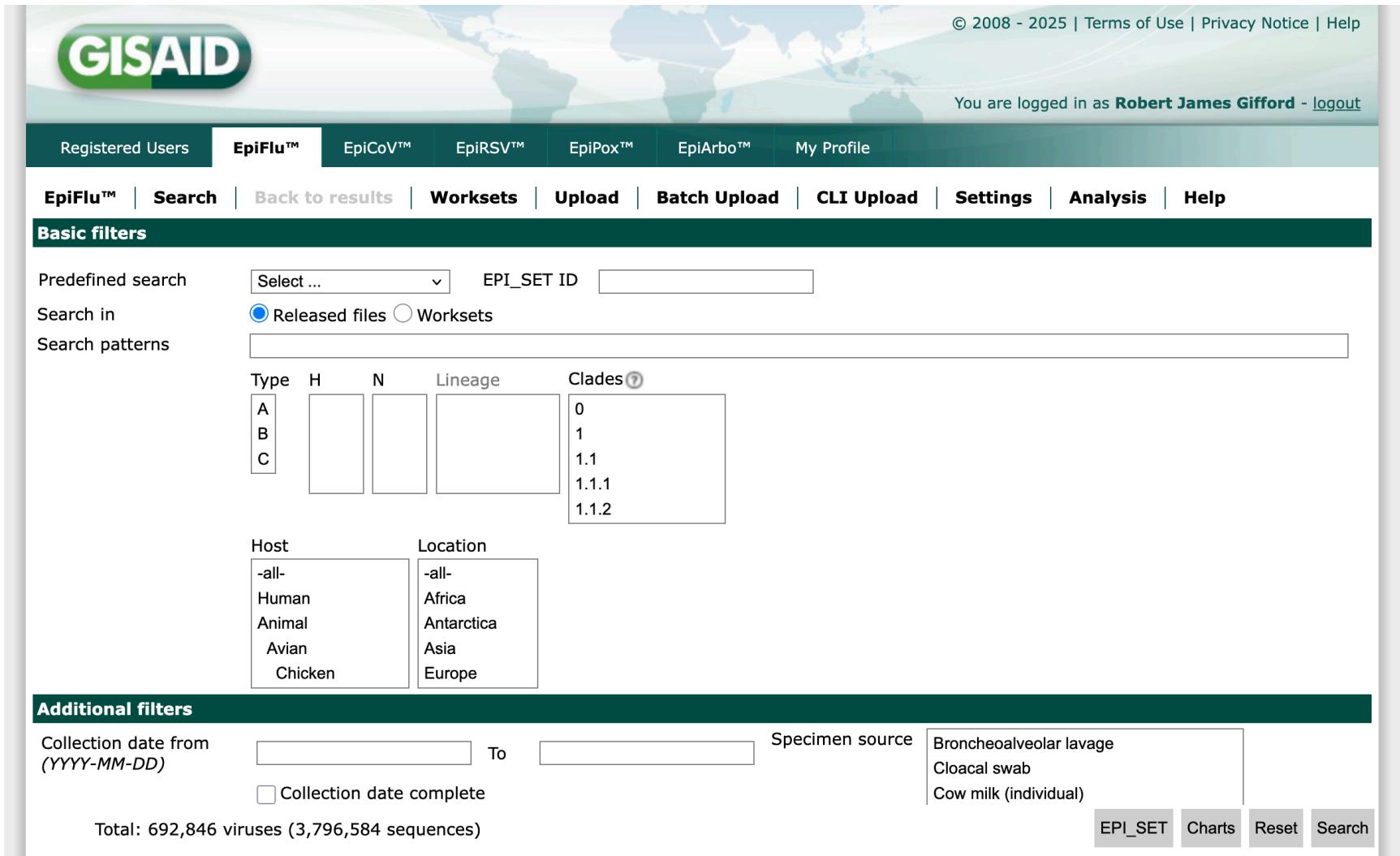


Sources of Reference Sequences



The screenshot shows the GISAID EpiFlu™ interface. At the top, there is a navigation bar with tabs for Registered Users, EpiFlu™ (which is active), EpiCoV™, EpiRSV™, EpiPox™, EpiArbo™, and My Profile. Below the navigation bar, there is a secondary navigation bar with links for EpiFlu™, Search, Back to results, Worksets, Upload, Batch Upload, CLI Upload, Settings, Analysis, and Help. The main content area is divided into three rows of tool icons. The first row contains AudacityInstant (tree icon), BLAST (DNA helix icon), Emerging Variants (upward arrow icon), EpiCharts (bar chart icon), FluSurver (protein structure icon), and PrimerChecker (colorful spiral icon). The second row contains Search & Browse (magnifying glass icon), Submission Tracker (world map icon), HA Mutation Surveillance (protein structure icon), and Subtype / Clade Frequency (wavy line icon). The third row contains Vaccine Reference Sequences (colorful arrows icon), Phydynamics Global (H3N2, H1N1, B Vic, H5, HX icons), and Phydynamics Global (Phydynamics Global icons).

Sources of Reference Sequences



The screenshot shows the GISAID EpiFlu™ search interface. At the top, there are tabs for Registered Users, EpiFlu™ (which is selected), EpiCoV™, EpiRSV™, EpiPox™, EpiArbo™, and My Profile. Below the tabs, a banner displays the GISAID logo, a world map, and links to © 2008 - 2025 | Terms of Use | Privacy Notice | Help. It also shows the user is logged in as Robert James Gifford with a logout link.

The main search area has a green header "Basic filters". It includes fields for Predefined search (dropdown menu "Select ...") and EPI_SET ID (text input field). A radio button group for "Search in" shows "Released files" (selected) and "Worksets". A search pattern input field is present. Below these are dropdown menus for Type (A, B, C), H (Host: Human, Animal, Avian, Chicken), N (Location: -all-, Africa, Antarctica, Asia, Europe), Lineage (Host: -all-, Human, Animal, Avian, Chicken), and Clades (0, 1, 1.1, 1.1.1, 1.1.2).

At the bottom, there is an "Additional filters" section with fields for Collection date from (YYYY-MM-DD) and To (text input fields), a checkbox for Collection date complete, and a dropdown for Specimen source (Bronchoalveolar lavage, Cloacal swab, Cow milk (individual)).

At the very bottom, there are buttons for EPI_SET, Charts, Reset, and Search, along with a total count of 692,846 viruses (3,796,584 sequences).

Sources of Reference Sequences

NextStrain

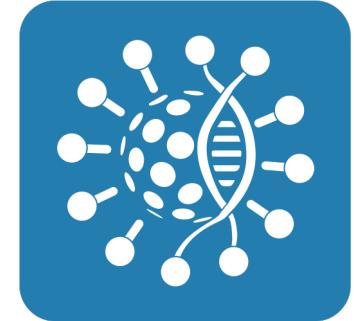
- **Strengths**
 - Provides curated reference datasets for specific pathogens.
- **Challenges**
 - Reference sets tailored to specific builds → not comprehensive.
 - Coverage focused on major pathogens (e.g. Flu, SARS-CoV-2, Ebola).
 - Less suited for deep historical or lineage-wide reference curation.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



Sources of Reference Sequences

PathoPlexus

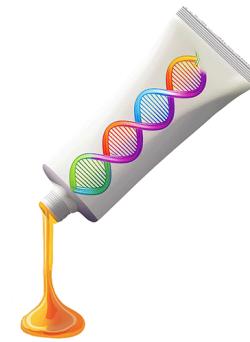
- **Strengths**
 - Open-source, non-profit platform for *all human viral pathogens*.
 - Integrates with Nexstrain, INSDC (NCBI/ENA/DDBJ).
 - Community-driven governance, transparent policies.
 - Offers SeqSets and APIs for filtering/search.
 - Flexible data-sharing: immediate open or embargoed.
- **Challenges**
 - 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

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



CERI
Centre for Epidemic
Response and Innovation

The screenshot shows the NCBI Virus homepage with a light blue background featuring a white geometric mesh pattern. At the top left is the NCBI Virus logo with the tagline "Sequences for discovery". At the top right are navigation links: Find Data, Help, How to Participate, Submit Sequences, and Contact Us. Below the header is a yellow banner with the text "NCBI Taxonomy Updates to Virus Classification". The main title "NCBI Virus" is centered above a subtitle: "Community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. To find, retrieve and analyze data, please select an option below." Three search options are displayed in colored boxes: a teal box for "Search by virus" (using a magnifying glass icon), a yellow box for "NCBI BLAST™ search" (using a rocket ship icon), and a pink box for "Outbreak Statistics" (using a sunburst icon). Below these is a section titled "Search GenBank virus/viroid sequences and explore metadata in Results Table with advanced filtering and data visualizations options". A search bar contains the text "Alphainfluenzavirus", and a progress indicator shows "Searching Alphainfluenzavirus ...". On the far right is a vertical "Feedback" button.

NCBI Virus

Community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories.

To find, retrieve and analyze data, please select an option below.

Search by virus
Use virus name or taxid to find viral nucleotide and protein sequences.

NCBI BLAST™ search
Find viral nucleotide and protein sequences using the BLAST™ sequence similarity tool.

Outbreak Statistics
Overview of virus outbreaks in the past 10 weeks

Feedback

GenBank Influenza Virus Entry



CERI
Centre for Epidemic
Response and Innovation

REFERENCE 3 (bases 1 to 890)
CONSRM 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](#)"
join(27..56,529..864)
/gene="NEP"

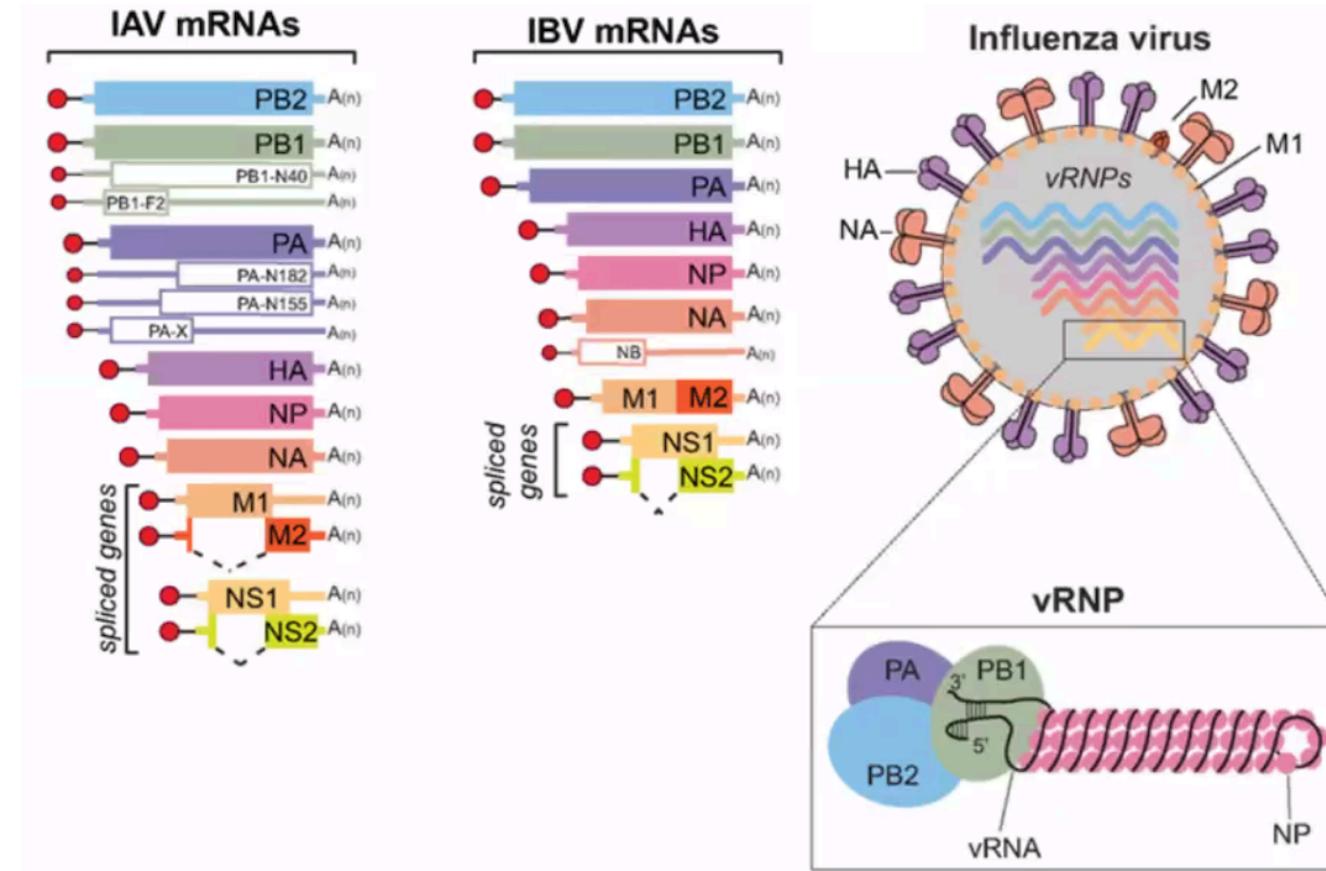
GenBank Influenza Virus Entry



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)



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)



- 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?



- **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



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



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



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



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.**

OVER TO YOU!!



Photo by Stefan Els