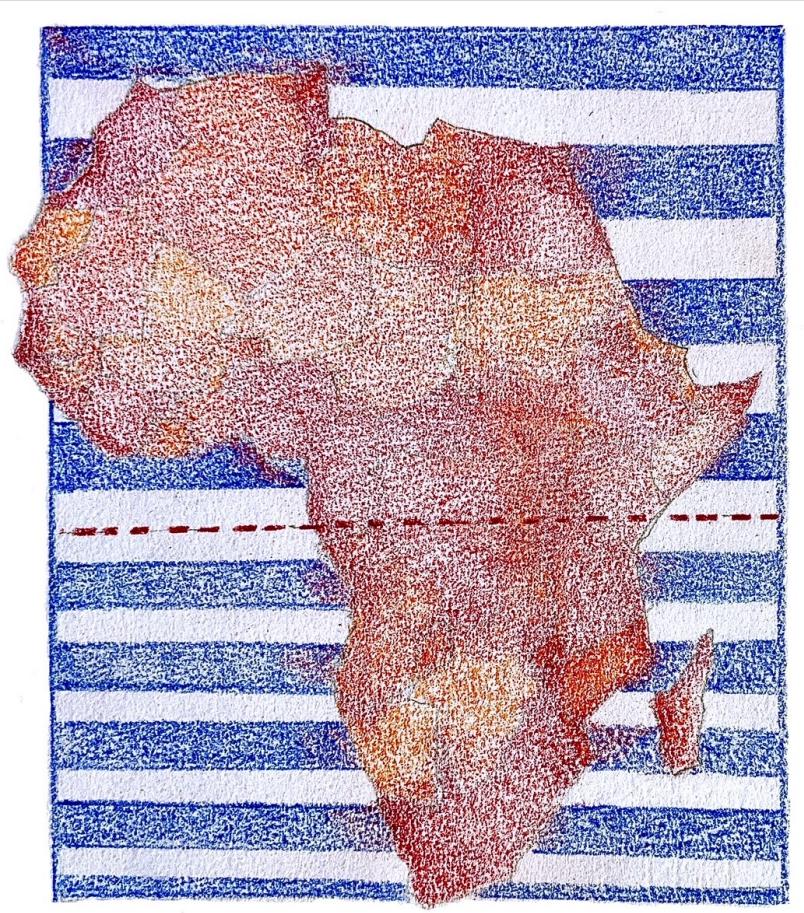




# Africa 75K analysis

## Figure 3 Methods

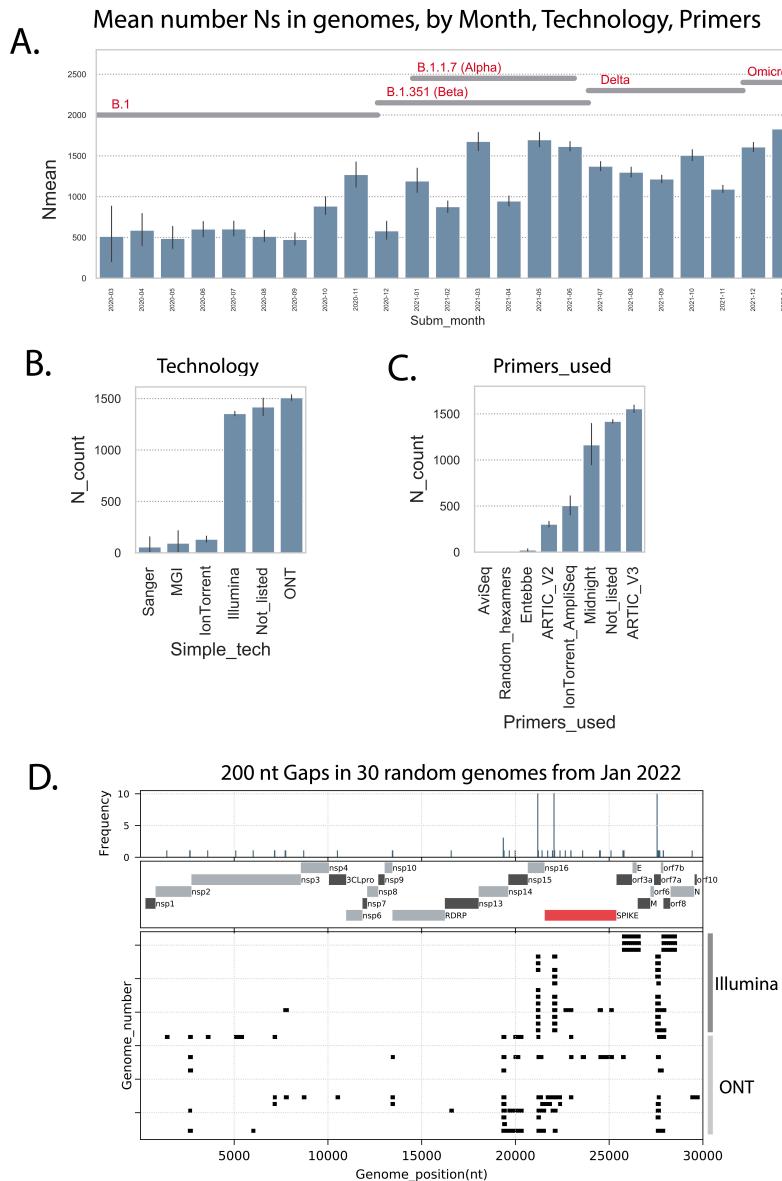
# Matthew Cotten MRC unit, Entebbe



MRC/UVRI and LSHTM Uganda Research Unit



**Uganda  
Virus  
Research  
Institute**



# 1. Retrieve sequence data from GISAID

The screenshot shows the GISAID EpiCoV™ search interface. At the top, there are tabs for Registered Users, EpiFlu™, EpiCoV™ (which is selected), and My profile. Below the tabs, there are links for EpiCoV™, Search, Downloads, and Upload. A message indicates the user is logged in as Matthew Cotten with a logout link.

The main area is titled "Search" with a "Reset filters" button. It includes search fields for EPI\_ISL ID, Virus name, Location (set to Africa), Host, Collection date range, Clade (all), Lineage, Substitutions, Variants, and a "Text Search" input field. To the right of these fields are several checkboxes for filtering: Complete, High coverage, Low coverage excluded, With patient status, Collection date complete, and Under investigation.

A table below lists 7 virus entries, each with a checkbox next to it:

|                                     | Virus name                 | Passage # | Accession ID     | Collection date | Submission date | Length | Host  | Location           | Originating |
|-------------------------------------|----------------------------|-----------|------------------|-----------------|-----------------|--------|-------|--------------------|-------------|
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114317/2022 | Original  | EPI_ISL_11116862 | 2022-01-25      | 2022-03-17      | 29,747 | Human | Africa / Kenya / K | KEMRI-We    |
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114305/2022 | Original  | EPI_ISL_11116860 | 2022-01-25      | 2022-03-17      | 29,753 | Human | Africa / Kenya / K | KEMRI-We    |
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114287/2022 | Original  | EPI_ISL_11116859 | 2022-01-24      | 2022-03-17      | 29,750 | Human | Africa / Kenya / K | KEMRI-We    |
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114273/2022 | Original  | EPI_ISL_11116858 | 2022-01-24      | 2022-03-17      | 29,462 | Human | Africa / Kenya / K | KEMRI-We    |
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114222/2022 | Original  | EPI_ISL_11116857 | 2022-01-22      | 2022-03-17      | 29,750 | Human | Africa / Kenya / K | KEMRI-We    |
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114144/2022 | Original  | EPI_ISL_11116856 | 2022-01-19      | 2022-03-17      | 29,756 | Human | Africa / Kenya / K | KEMRI-We    |
| <input checked="" type="checkbox"/> | hCoV-19/Kenya/C114141/2022 | Original  | EPI_ISL_11116855 | 2022-01-19      | 2022-03-17      | 29,750 | Human | Africa / Kenya / K | KEMRI-We    |

Total: 95,236 viruses

Pagination controls: << < 15 16 17 18 19 > >>

Action buttons: EPI\_SET, Select, Analysis, Download

10k limit per download

Download in subsets

(e.g. adjust submission date range to capture about 10,000 genomes)

## 2. Retrieve associated data (sequencing methods, primers)

- Dates and Location
- Gene Sequences (FASTA)
- Input for the Augur pipeline
- Patient status metadata
- Sequencing technology metadata
  
- Acknowledgement table (< 500)

Acknowledgement of over 500 entries click here

← need Dates and Location table

← need Seq technology table

### 3. Include relevant data in fasta IDs:

fasta id format:

Acc-no | Country | Collection Date | Submission Date |  
Sample ID | Technology | Lineage

### 4. Count Ns, gaps (python scripts, output csv table)

[CountNs\\_in\\_seq\\_py3.py](#)

```
for record in SeqIO.parse(open(input_genome_file, "rU"),
"fasta"):
    this_sequence = (str(record.seq)).upper()
    this_id = str(record.id)
    N_count = this_sequence.count("N")
    f1 = open(outprefix+'_countNs.csv', 'a')#make empty
file
    print(this_id+', '+str(N_count), file=f1)
    f1.close()
print('That\'s All Folks!')
```

### 5. To visualize: Open table in Jupyter notebook

[Africa\\_to\\_28Jan22\\_Ncounts\\_versionGH.ipynb](#)

Suggestions:

Keep good notes:

- Source of data

- Full commands to run script

- Output

- Everything in one directory

- Intelligent file names

Write scripts assuming you will need to rerun next week with new data.

### Acknowledgements

My VT Phan, Dan Lule

Wellcome Trust, MRC for funding

GISAID and all investigators for sharing data

Test data, code available here:

[https://github.com/mlcotten13/AFRICA\\_75K\\_analysis](https://github.com/mlcotten13/AFRICA_75K_analysis)