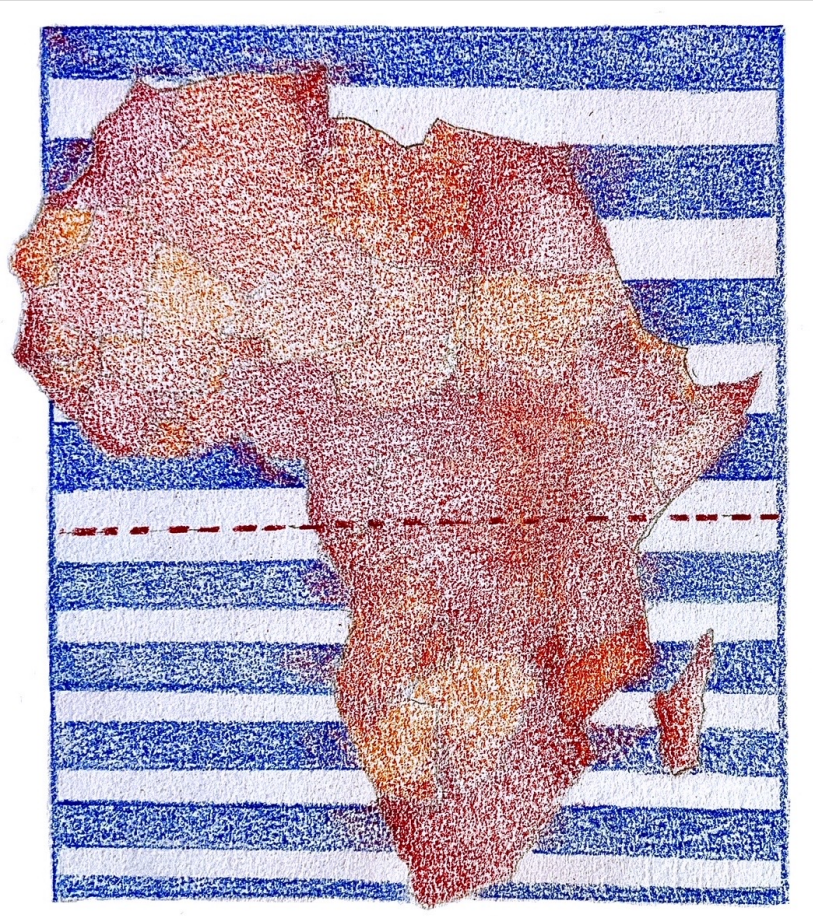
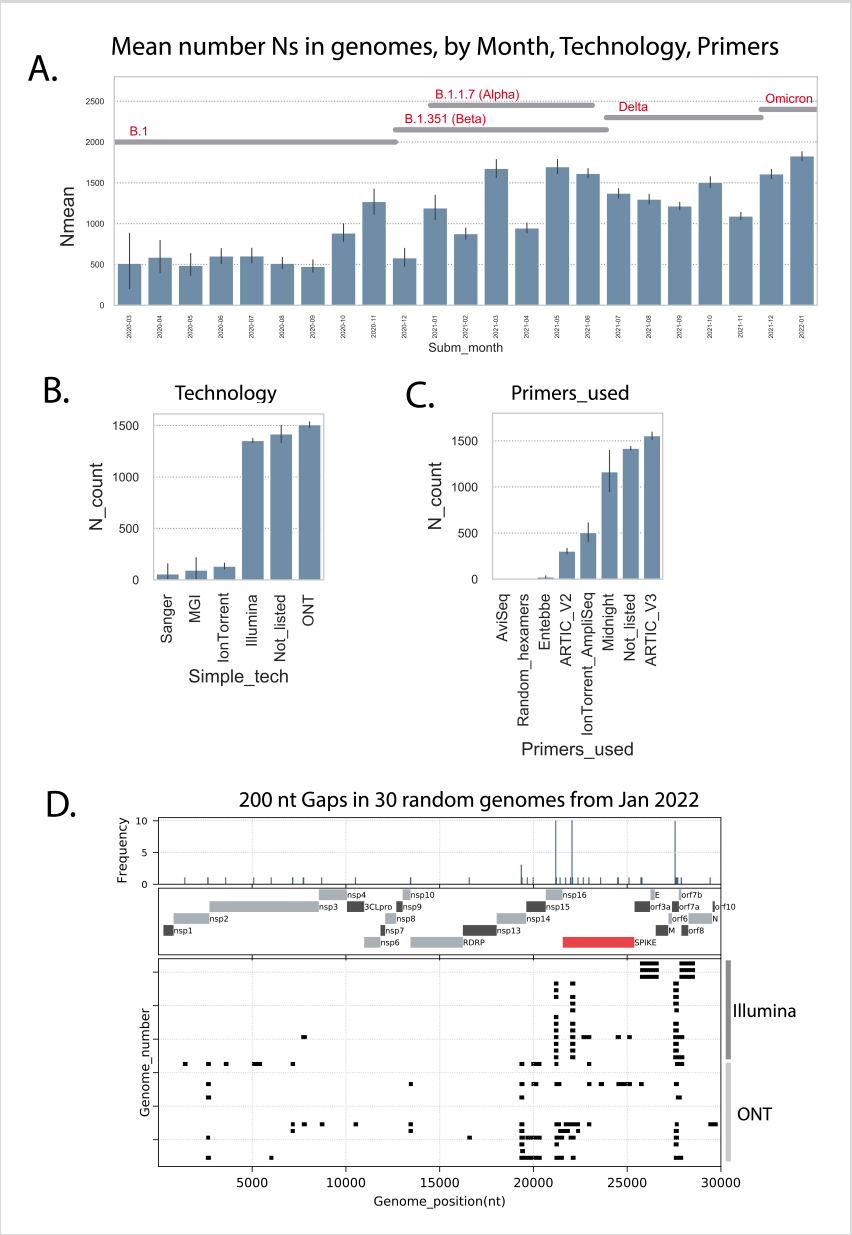
Africa 75K analysis
Figure 3 Methods

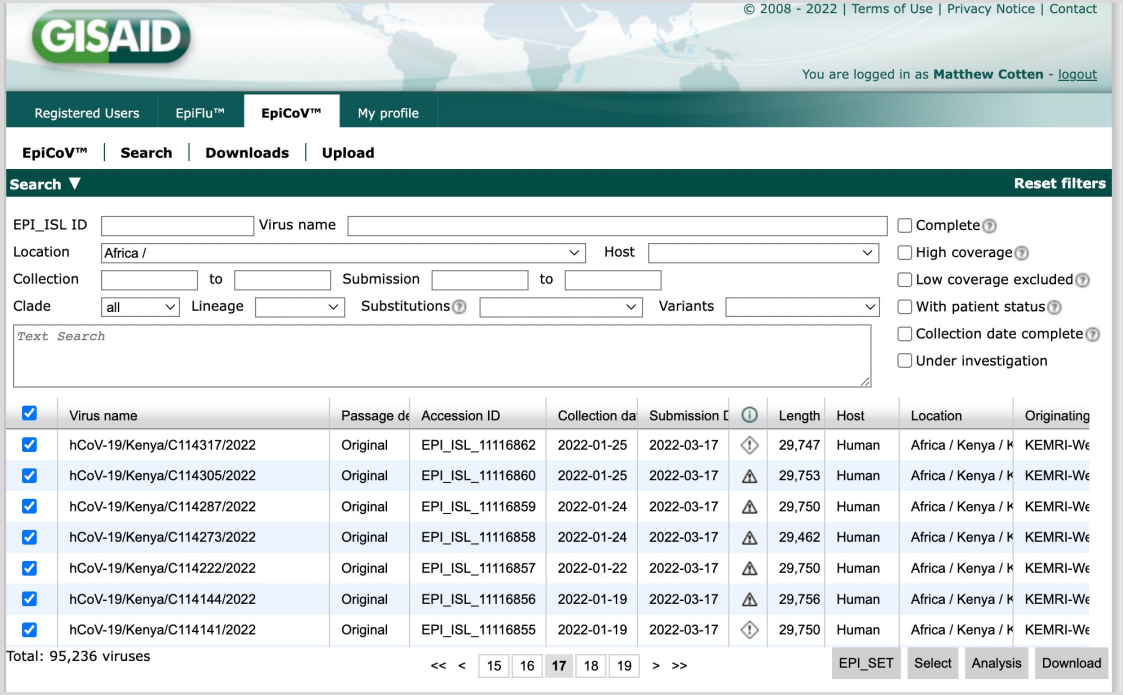
MRC/UVRI and LSHTM Uganda Research Unit

Matthew Cotten
MRC unit, Entebbe





1. Retrieve sequence data from GISAID



10k limit per download

Download in subsets

(e.g. adjust submission date range to capture about10,000 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