Automated Processing of Next-Gen Sequencing Data from Real-Time SARS-CoV-2 Surveillance

Riva A[1], Tagliamonte M[2,3], Rife-Magalis B[2,3], Marini S[2,3], Mavian C[2,3], Madore SJ[1], Salemi M[2,3]

[1]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL, USA; [2]Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA; [3]Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA.

UF | ICBR
your pursuit is what drives us.
BIOTECH.UFL.EDU

Introduction

Determining genome sequences of SARS-CoV-2 samples quickly and accurately is critical to real-time monitoring of the progression and evolution of the current pandemic.

FLACO is an integrated analysis and reporting pipeline to analyze samples collected as part of the large-scale genomic surveillance of SARS-CoV-2 cases under way at the University of Florida.

The FLACO database

The FLACO database is a relational database designed to store both metadata on analysis results for all samples processed by the pipeline.

Metadata fields include: sample name, date and location of collection, donor sex, age, and vaccination status (when known).

Analysis result fields include: total number of reads, reads mapped to SARS-CoV-2 and to human RNA controls, average and median SARS-CoV-2 genome coverage, fraction of spike protein covered, Pangolin lineage call.

Downstream analysis

The pipeline includes a command-line program to query the FLACO database and extract sample data for downstream phylogenetic analysis.

Samples can be filtered by geographical origin, vaccination status, range of sampling dates. Any combination of metadata and data fields can be queried

The tool is designed to be easily included in automated analysis scripts.

GISAID submission

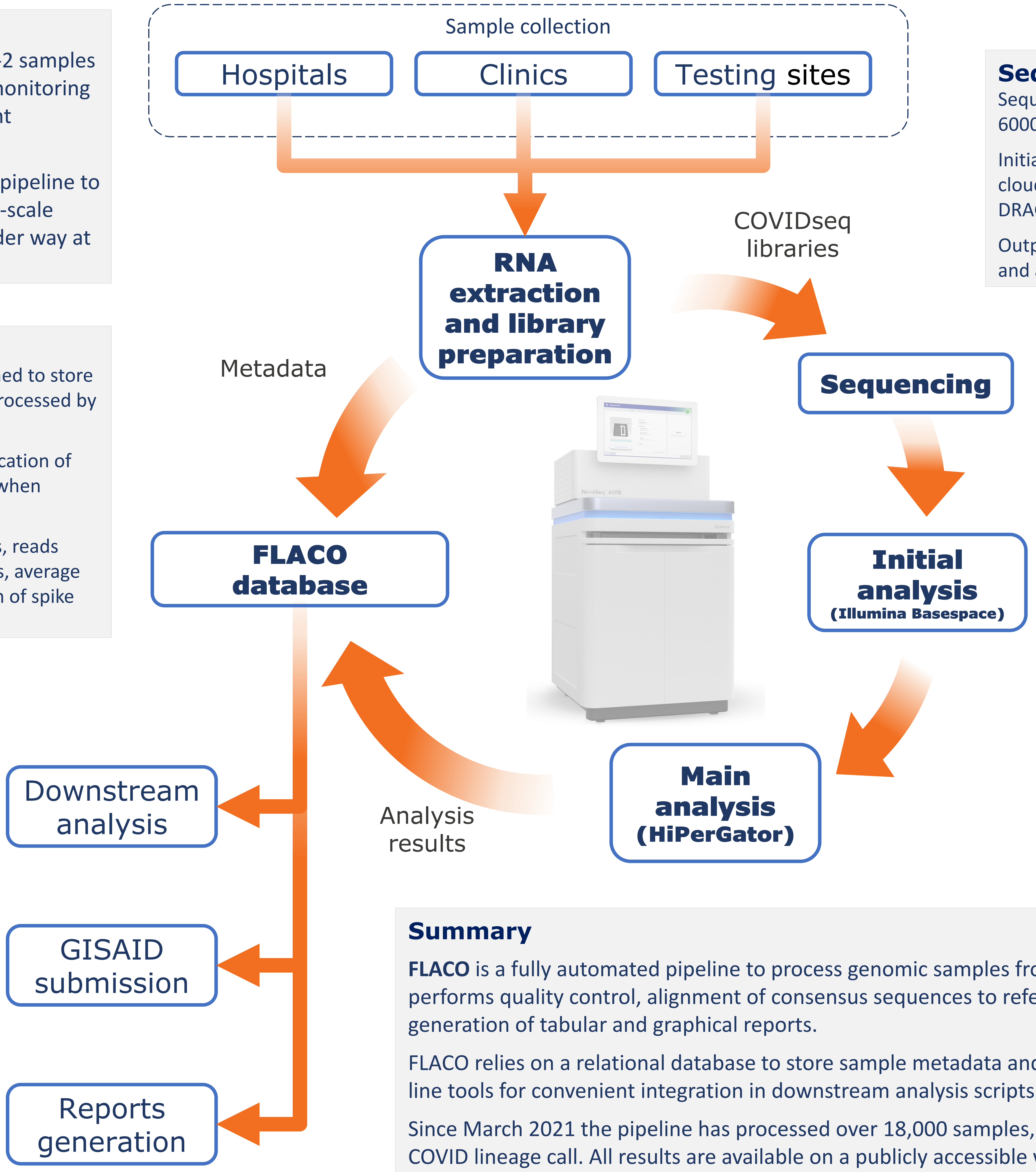
The pipeline extracts consensus sequences and metadata for successful samples in the format suitable for submission to GISAID.

Reports generation

Full reports, formatted as web pages and as machine-readable tables, are automatically generated and uploaded to a public website. Reports are organized by sequencing run, and by geographical origin of the samples.

Four types of report are generated:

- Sequencing statistics (total number of reads, mapping to genome);
- Genome coverage statistics (average and median coverage, spike protein coverage) – Figure 1;
- Lineage frequency;
- Temporal distribution of lineages – Figure 2.



Sequencing and initial analysis

Sequencing is performed on an Illumina NovaSeq 6000, S1 or SP flowcell, in 2x100 configuration.

Initial analysis takes place on the Illumina Basespace cloud platform, using the “RNA Pathogen Detection” DRAGEN-accelerated app.

Output consists of a full-length consensus sequence and alignment statistics for each sample.

Main analysis

The main analysis pipeline run on UF’s HiPerGator cluster computer, and performs the following steps:

1. Extract quality control parameters and use them to filter out low-quality consensus sequences;
2. Align good-quality consensus sequences to SARS-CoV-2 reference sequence;
3. Assign lineages to samples using Pangolin.

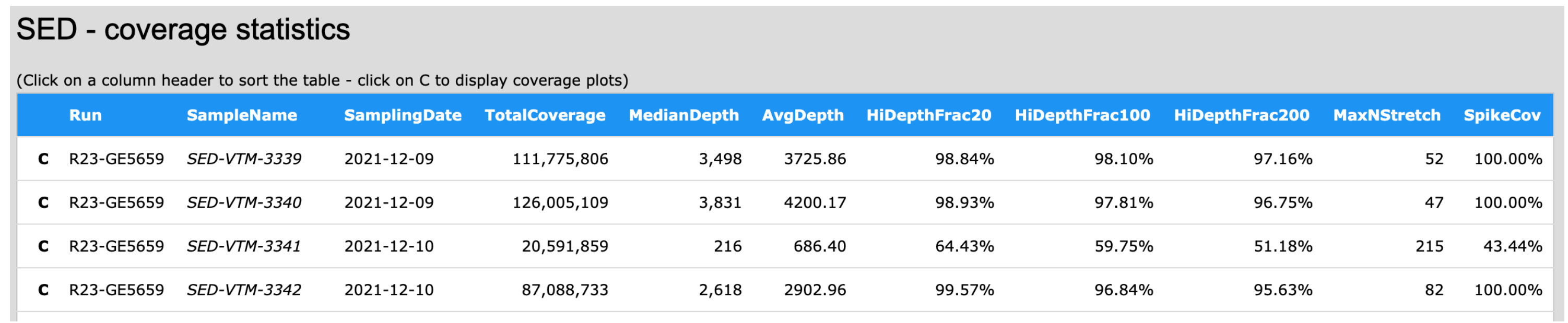
QC parameters and analysis results for each sample are stored in the FLACO database. The database structure allows storing multiple analysis results for each sample, in case it is sequenced more than once.

Summary

FLACO is a fully automated pipeline to process genomic samples from large-scale SARS-CoV-2 surveillance. It performs quality control, alignment of consensus sequences to reference genome, lineage assignment, and generation of tabular and graphical reports.

FLACO relies on a relational database to store sample metadata and analysis results, and provides command-line tools for convenient integration in downstream analysis scripts.

Since March 2021 the pipeline has processed over 18,000 samples, of which about 13,000 received a valid COVID lineage call. All results are available on a publicly accessible website.



SED - coverage statistics

(Click on a column header to sort the table - click on C to display coverage plots)

	Run	SampleName	SamplingDate	TotalCoverage	MedianDepth	AvgDepth	HiDepthFrac20	HiDepthFrac100	HiDepthFrac200	MaxNStretch	SpikeCov
C	R23-GE5659	*SED-VTM-3339*	2021-12-09	111,775,806	3,498	3725.86	98.84%	98.10%	97.16%	52	100.00%
C	R23-GE5659	*SED-VTM-3340*	2021-12-09	126,005,109	3,831	4200.17	98.93%	97.81%	96.75%	47	100.00%
C	R23-GE5659	*SED-VTM-3341*	2021-12-10	20,591,859	216	686.40	64.43%	59.75%	51.18%	215	43.44%
C	R23-GE5659	*SED-VTM-3342*	2021-12-10	87,088,733	2,618	2902.96	99.57%	96.84%	95.63%	82	100.00%

Figure 1. Extract from report showing genome coverage statistics.

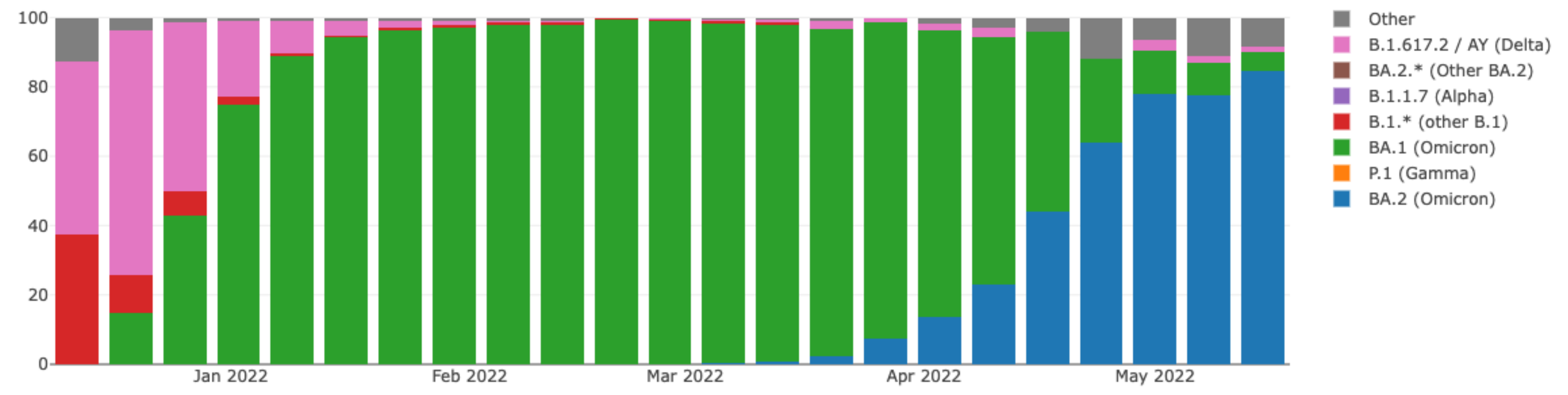


Figure 2. Plot showing frequency of COVID lineages over time.

Publications

1. Tagliamonte M *et al*. Rapid emergence and spread of SARS-CoV-2 gamma (P.1) variant in Haiti. *Clinical Infectious Diseases*, 2021. doi: 10.1093/cid/ciab736.
2. Rife-Magalis, B *et al*. Low-frequency variants in mildly symptomatic vaccine breakthrough infections presents a doubled-edged sword. *Journal of Medical Virology,* 2022. doi: 10.1002/jmv.27726.
3. Rife-Magalis B et al. Severe Acute Respiratory Syndrome Coronavirus 2 Delta Vaccine Breakthrough Transmissibility in Alachua County, Florida. *Clinical Infectious Diseases*, 2022. doi: 10.1093/cid/ciac197.

Acknowledgments

This work was supported by funding from the NSF, the Rockefeller Foundation, and UF Research.