



The Observable Galaxy

Stephen D. Shank
Alex Ostrovsky
Charles Hadley King
Sergei L. Kosakovsky Pond

July 17th, 2022



Visualization in JavaScript



- JavaScript has a rich ecosystem for UI and data visualization:
 - Low level libraries (D3)
 - High level frameworks (Vega, Vega-Lite)
 - Structural viewers (NGL)
 - Multiple sequence alignments (alignment.js)
 - Phylogenetic trees (phylotree.js)

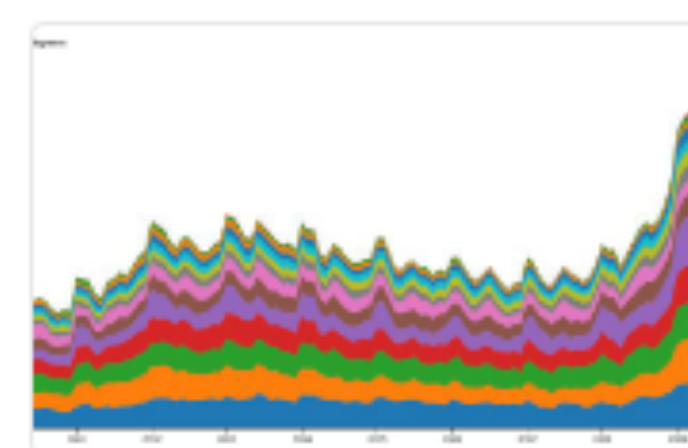


Observable

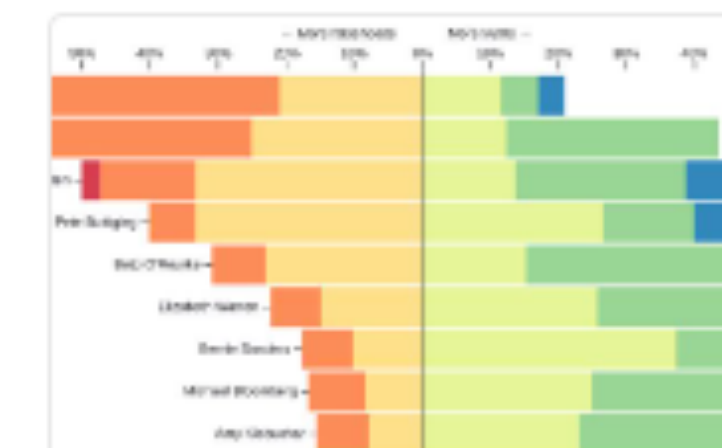
Data-Driven Documents



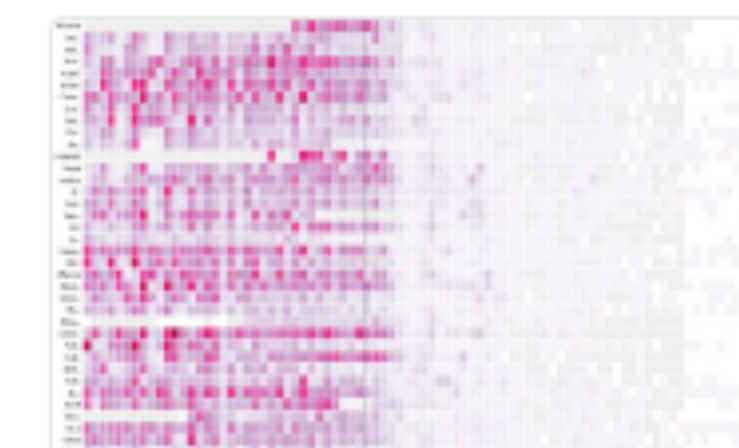
<https://observablehq.com/@d3/gallery>



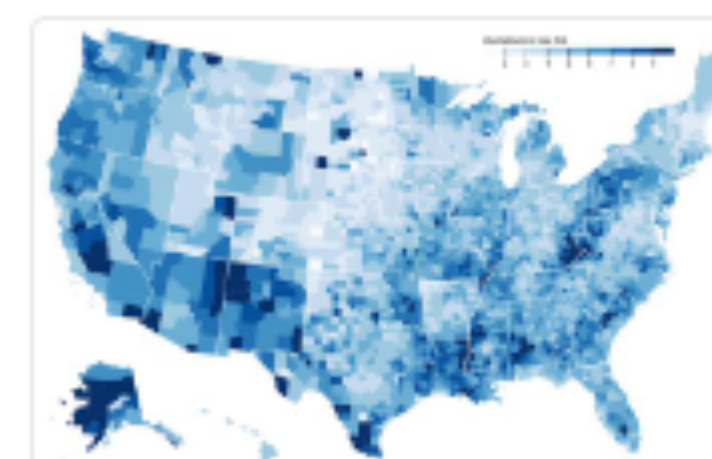
Stacked area chart



Diverging stacked bar chart



The impact of vaccines



Choropleth



Force-directed graph



Circle packing

- Feature rich notebook platform based on JavaScript
 - “Reactive dataflow” execution model
 - Rich, well abstracted, stateful UI
 - Fork/merge based collaboration
 - Import/export modularity
 - Markdown/LaTex support
 - Much more...

Reactive dataflow

the sum is 12

```
text = md`the sum is sum`
```

sum = 12

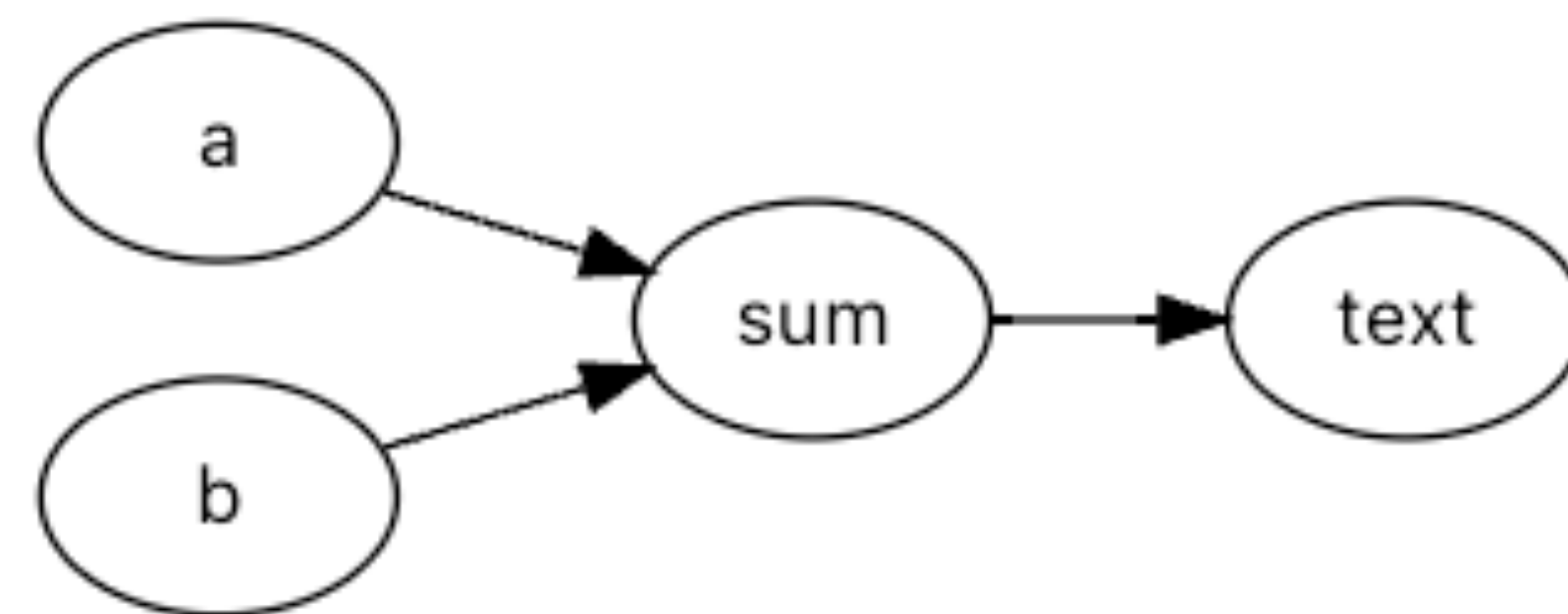
```
sum = a + b
```

a = 5

```
a = 5
```

b = 7

```
b = 7
```





- Feature rich notebook platform based on JavaScript
 - “Reactive dataflow” execution model
 - **Rich, well abstracted, stateful UI**
 - Fork/merge based collaboration
 - Import/export modularity
 - Markdown/LaTex support
 - Much more...

Stateful widgets

<https://observablehq.com/@observablehq/inputs>

Range

Pick a number. [Examples](#) > [API Reference](#) >

Favorite number

```
viewof n = Inputs.range([0, 255], {step: 1, label: "Favorite number"})
```

128

n



The Observable Galaxy



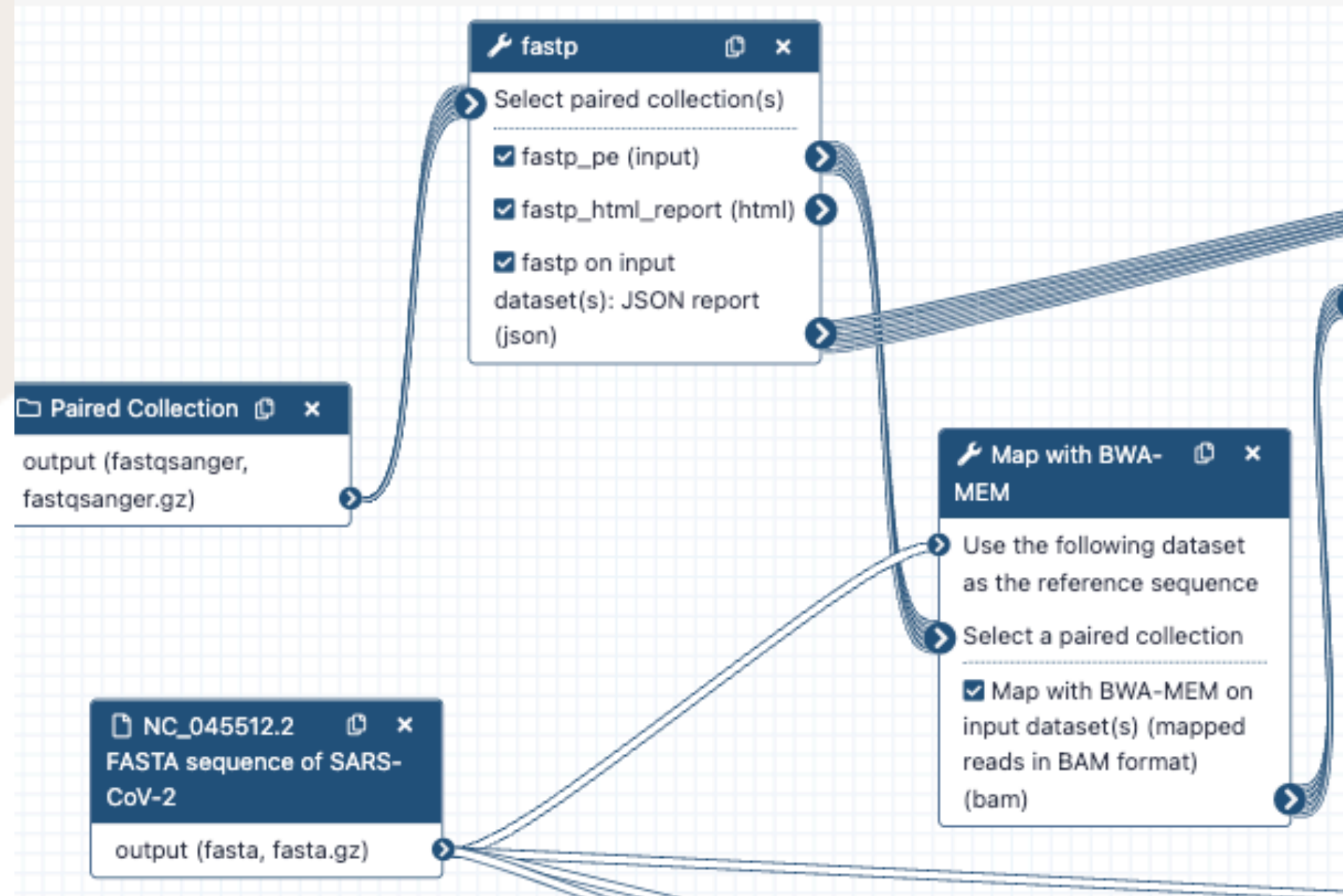
A datatype and display application to integrate Galaxy and Observable

```

accessions.txt
1 SRR17309643
2 SRR14460618
3 SRR14460619
4 SRR14460620
5 SRR14460621
6 SRR14460622
7 SRR14460623

```

COVID-19: variation analysis on WGS PE data (imported from uploaded file)



589: ObservableHQ on data 574

JavaScript Object Notation (JSON) format: **observablehq**, database: ?

```

/home/galaxy/galaxy/tools/development/stephenshank/exploring-sars-cov-2-intrahost-variation-individual-sampl,9b305a114b324ccf,
/home/galaxy/galaxy/database/objects/b0dde-49ec-a869-15c8eb455cf0.dat,
b1d0a8de1661cd49,

```

View results: ObservableHQ Dashboard

```

{
  "notebook": "stephenshank/exploring-sars-cov-2-intrahost-variation-individual-sampl,9b305a114b324ccf",
  "history_id": "9b305a114b324ccf",
  "payload_id": "b1d0a8de1661cd49",
  "summary_tsv": "2e09e426de1114e5"
}

```

https://observablehq.com/@stephenshank/exploring-sars-cov-2-intrahost-variation-descriptive?payload_id=035579cc588f5c92

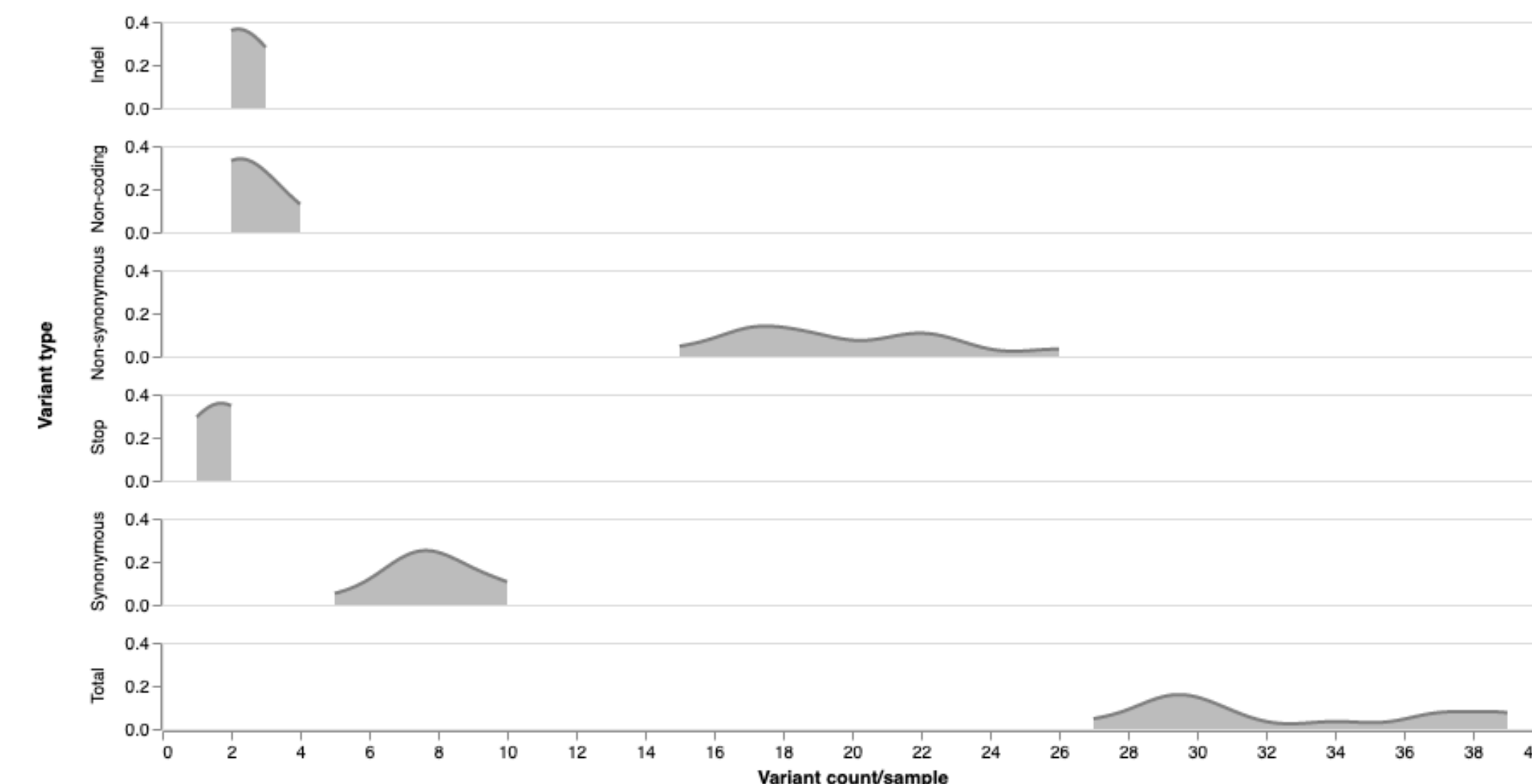
Exploring SARS-CoV-2 intrahost variation (descriptive)

This variant dataset contains information on **427** allelic variants from **12** samples, and comprises **135** unique allelic variants.

Allele frequency range for Figures 1, 2, and 3

50% ... 100%

Figure 1. Distribution of variants across samples (including 391 variants)





Galaxy - Datatype and tool



- ObservableHQ (OHQ) JSON datatype
- Target notebook (user/notebook name)
- Automatically extracts
 - history ID
 - OHQ JSON dataset ID (denoted **payload ID**)
 - variable number of key/dataset IDs pairs
- Associated tool
- Associated display application

ObservableHQ Select multiple datasets for analysis on the ObservableHQ platform

Username/notebook

Observable parameters

1: Observable parameters

Key for this object in the observable sheet

Input file for observable associated with above key

589: ObservableHQ on data 574

+ Insert Observable parameters

```
"notebook": "stephenshank/exploring-sars  
"history_id": "9b305a114b324ccf",  
"payload_id": "b1d0a8de1661cd49",  
"summary_tsv": "2e09e426de1114e5"
```



Galaxy - Datatype and tool



- ObservableHQ (OHQ) JSON datatype
- Target notebook (user/notebook name)
- Automatically extracts
 - history ID
 - OHQ JSON dataset ID (denoted **payload ID**)
 - variable number of key/dataset IDs pairs
- Associated tool
- Associated display application

ObservableHQ Select multiple datasets for analysis on the ObservableHQ platform

Username/notebook

stephenshank/exploring-sars-cov-2-intrahost-variation-descriptive

Observable parameters

1: Observable parameters

Key for this object in the observable sheet

summary_tsv

Input file for observable associated with above key

589: ObservableHQ on data 574

+ Insert Observable parameters

```
"notebook": "stephenshank/exploring-sars",  
"history_id": "9b305a114b324ccf",  
"payload_id": "b1d0a8de1661cd49",  
"summary_tsv": "2e09e426de1114e5"
```



Galaxy - Datatype and tool



- ObservableHQ (OHQ) JSON datatype
- Target notebook (user/notebook name)
- Automatically extracts
 - **history ID**
 - OHQ JSON dataset ID (denoted **payload ID**)
 - variable number of key/dataset IDs pairs
- Associated tool
- Associated display application

ObservableHQ Select multiple datasets for analysis on the ObservableHQ platform

Username/notebook

Observable parameters

1: Observable parameters

Key for this object in the observable sheet

Input file for observable associated with above key

589: ObservableHQ on data 574

+ Insert Observable parameters

```
"notebook": "stephenshank/exploring-sars  
"history_id": "9b305a114b324ccf",  
"payload_id": "b1d0a8de1661cd49",  
"summary_tsv": "2e09e426de1114e5"
```




Galaxy - Datatype and tool



- ObservableHQ (OHQ) JSON datatype
- Target notebook (user/notebook name)
- Automatically extracts
 - history ID
 - **OHQ JSON dataset ID (denoted payload ID)**
 - variable number of key/dataset IDs pairs
- Associated tool
- Associated display application

ObservableHQ Select multiple datasets for analysis on the ObservableHQ platform

Username/notebook

Observable parameters

1: Observable parameters

Key for this object in the observable sheet

Input file for observable associated with above key

589: ObservableHQ on data 574

+ Insert Observable parameters

```
"notebook": "stephenshank/exploring-sars  
"history_id": "9b305a114b324ccf",  
"payload_id": "b1d0a8de1661cd49",  
"summary_tsv": "2e09e426de1114e5"
```



Galaxy - Datatype and tool



- ObservableHQ (OHQ) JSON datatype
- Target notebook (user/notebook name)
- Automatically extracts
 - history ID
 - OHQ JSON dataset ID (denoted **payload ID**)
 - **variable number of key/dataset IDs pairs**
- Associated tool
- Associated display application

ObservableHQ Select multiple datasets for analysis on the ObservableHQ platform

Username/notebook

stephenshank/exploring-sars-cov-2-intrahost-variation-descriptive

Observable parameters

1: Observable parameters

Key for this object in the observable sheet

summary_tsv

Input file for observable associated with above key

589: ObservableHQ on data 574

+ Insert Observable parameters

```
"notebook": "stephenshank/exploring-sars  
"history_id": "9b305a114b324ccf",  
"payload_id": "b1d0a8de1661cd49",  
"summary_tsv": "2e09e426de1114e5"
```

Observable - Handling payloads

- Utilizes Galaxy's REST API
- Utilities for
 - autofetching data
 - fetching payloads
 - linking to associated histories
 - fetching associated datasets
- Aspiring for reproducible, “clear box” biological big data analysis

<https://observablehq.com/@stephenshank/observable-galaxy-mre>

```
import { autofetcher } from "@stephenshank/the-observable-galaxy"
```

```
datasets = ▶ Object {sample: Promise, another_sample: Promise}
```

```
datasets = await autofetcher(default_payload_id)
```

https://observablehq.com/@stephenshank/the-observable-galaxy?payload_id=04623fca6e71f19d

```
observablehq_json = ▶ Object {notebook: "ste|
```

```
observablehq_json = await payload_fetcher()
```

The Observable Galaxy

Associated history: <https://galaxy.hyphy.org/u/stephenshank/h/the-observable-galaxy>

```
markdown_banner(observablehq_json)
```

Observable - Handling payloads

- Utilizes Galaxy's REST API
- Utilities for
 - autofetching data
 - fetching payloads
 - linking to associated histories
 - fetching associated datasets
- Aspiring for reproducible, “clear box” biological big data analysis

[https://observablehq.com/
@stephenshank/observable-galaxy-mre](https://observablehq.com/@stephenshank/observable-galaxy-mre)

```
import { autofetcher } from "@stephenshank/the-observable-galaxy"
```

```
datasets = ▶ Object {sample: Promise, another_sample: Promise}  
datasets = await autofetcher(default_payload_id)
```

https://observablehq.com/@stephenshank/the-observable-galaxy?payload_id=04623fca6e71f19d

```
observablehq_json = ▶ Object {notebook: "ste|  
observablehq_json = await payload_fetcher()
```

The Observable Galaxy

Associated history: <https://galaxy.hyphy.org/u/stephenshank/h/the-observable-galaxy>

```
markdown_banner(observablehq_json)
```

Observable - Handling payloads

- Utilizes Galaxy's REST API
- Utilities for
 - autofetching data
 - fetching payloads
 - linking to associated histories
 - fetching associated datasets
- Aspiring for reproducible, “clear box” biological big data analysis

<https://observablehq.com/@stephenshank/observable-galaxy-mre>

```
import { autofetcher } from "@stephenshank/the-observable-galaxy"
```

```
datasets = ▶ Object {sample: Promise, another_sample: Promise}
```

```
datasets = await autofetcher(default_payload_id)
```

https://observablehq.com/@stephenshank/the-observable-galaxy?payload_id=04623fca6e71f19d

```
observablehq_json = ▶ Object {notebook: "ste|
```

```
observablehq_json = await payload_fetcher()
```

The Observable Galaxy

Associated history: <https://galaxy.hyphy.org/u/stephenshank/h/the-observable-galaxy>

```
markdown_banner(observablehq_json)
```



Why a payload?



For sharable, seamless integration and enhanced productivity



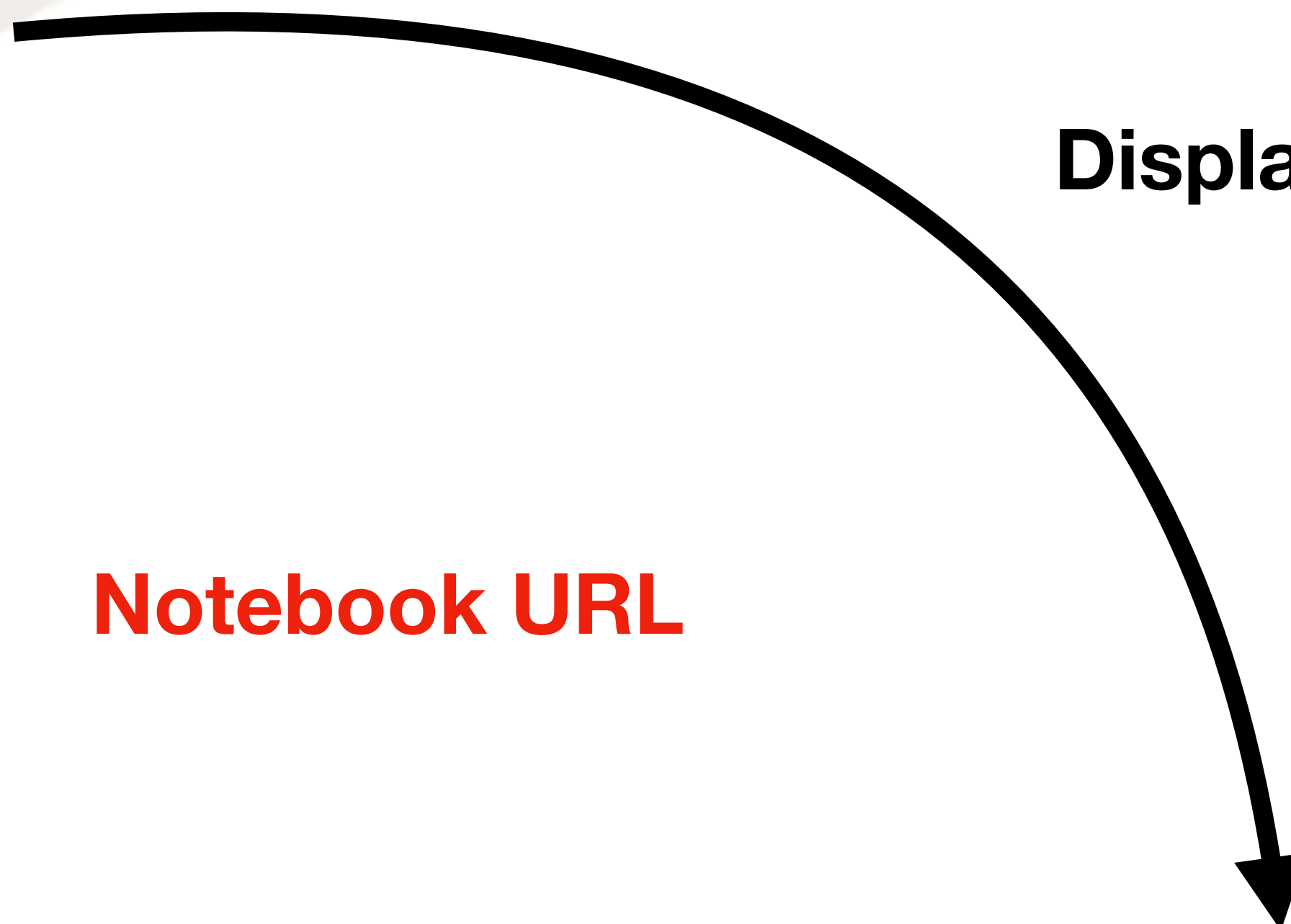
◎ Observable



Why a payload?



For sharable, seamless integration and enhanced productivity



Display Application

Notebook URL

Observable



Why a payload?



For sharable, seamless integration and enhanced productivity



Fetch data
Fetch history information

History ID

Link back to history

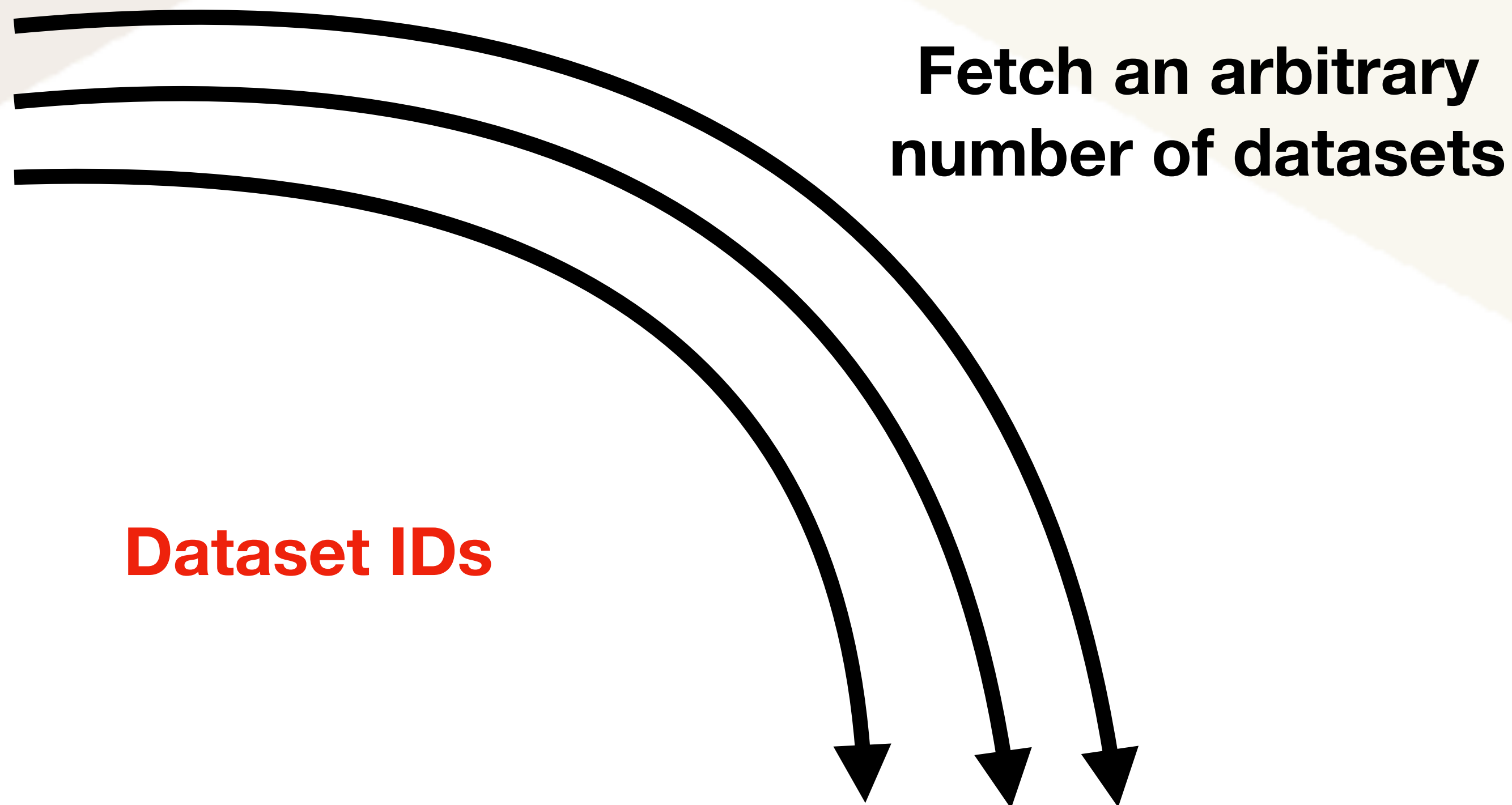
 **Observable**



Why a payload?



For sharable, seamless integration and enhanced productivity



Dataset IDs

Observable



Why a payload?



For sharable, seamless integration and enhanced productivity



[https://observablehq.com/
@\\$USERNAME/\\$NOTEBOOK
?payload_id=\\$PAYLOAD_ID](https://observablehq.com/@$USERNAME/$NOTEBOOK?payload_id=$PAYLOAD_ID)

 **Observable**





Workflows and collections



- Integrates with workflows:
 - Multiple OHQ JSONs per workflow
 - Multiple datasets per OHQ JSON
- WIP integration with collections, either:
 - One OHQ JSON per element
 - One table with IDs from each element in collection

Final output of a workflow

Screenshot of a workflow output window titled "Spike/antibody complex". It shows "Observable parameters 1" with an input file for observable associated with above key. A checkbox labeled "observable_json (observablehq)" is checked.

Dataset IDs of a collection

Screenshot of a dataset viewer window titled "704: Collapse Collection on data 703, data 702, and others". It shows 12 lines of data in a tabular format. The data is as follows:

1	2
SRR14460618	f472d6d1ebb25495
SRR14460619	0c5d7fa180a56458
SRR14460620	26418eee447800da
SRR14460621	5b423150ebdf1795
SRR14460622	fbb1485fd41df3df

Screenshot of a collection viewer window titled "21: ObservableHQ on collection 12". It shows a list with 2 items:

- SRR17309642
- SRR17309643

A collection of OHQ JSONs



SARS-CoV-2 Variant Calling



Productizing published workflows...

https://observablehq.com/@stephenshank/exploring-sars-cov-2-intrahost-variation-individual-sampl?payload_id=b1d0a8de1661cd49

Exploring SARS-CoV-2 intrahost variation (individual samples)

Table 1. Summary of all allelic variants found in the 12 samples. Mean AF : mean frequency of the variant in samples where it found; $\geq 50\%$, 10–50%, $< 10\%$: number of AVs within this AF range in the sample. Subpopulations counts the number of AF clusters with more than one AV based on OPTICS clustering.

Sample	AVs	Mean AF	$\geq 50\%$	10-50%	$< 10\%$
SRR14460623	52	0.719	37	1	14
SRR14460619	41	0.931	39		2
SRR14460620	41	0.927	39		2
SRR17309643	39	0.956	37	2	
SRR14460621	35	0.958	34		1
SRR14460618	33	0.895	30	1	2
SRR14460624	33	0.917	31		2
SRR14460622	31	0.925	29		2
SRR14460626	31	0.949	30		1
SRR14460627	31	0.919	29		2
SRR14460628	31	0.919	29		2
SRR14460625	29	0.918	27		2

Select a sample to explore further

SRR14460618

Correspondence | [Published: 29 September 2021](#)

Ready-to-use public infrastructure for global SARS-CoV-2 monitoring

[Wolfgang Maier](#), [Simon Bray](#), [Marius van den Beek](#), [Dave Bouvier](#), [Nathan Coraor](#), [Milad Miladi](#), [Babita Singh](#), [Jordi Rambla De Argila](#), [Dannon Baker](#), [Nathan Roach](#), [Simon Gladman](#), [Frederik Coppens](#), [Darren P. Martin](#), [Andrew Lonie](#), [Björn Grüning](#) ✉, [Sergei L. Kosakovsky Pond](#) ✉ & [Anton Nekrutenko](#) ✉

[Nature Biotechnology](#) **39**, 1178–1179 (2021) | [Cite this article](#)

Link	Workflow
WorkFlowHub DockStore	Illumina ARTIC: Variant analysis from ampliconic data produced with ARTIC protocol v1, v2, v3, or v4, or any alternative primer scheme. ILL-AMP
WorkFlowHub DockStore	Oxford Nanopore ARTIC: Variant analysis from ampliconic data produced with ARTIC protocol v1, v2, v3, or v4, or any alternative primer scheme. ONT-AMP



SARS-CoV-2 Variant Calling



Figure 1. Distribution of allele frequencies for SRR14460618 and its smoothed kernel density

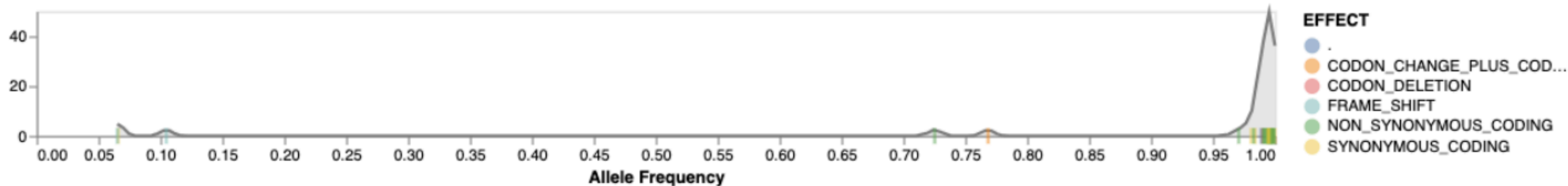
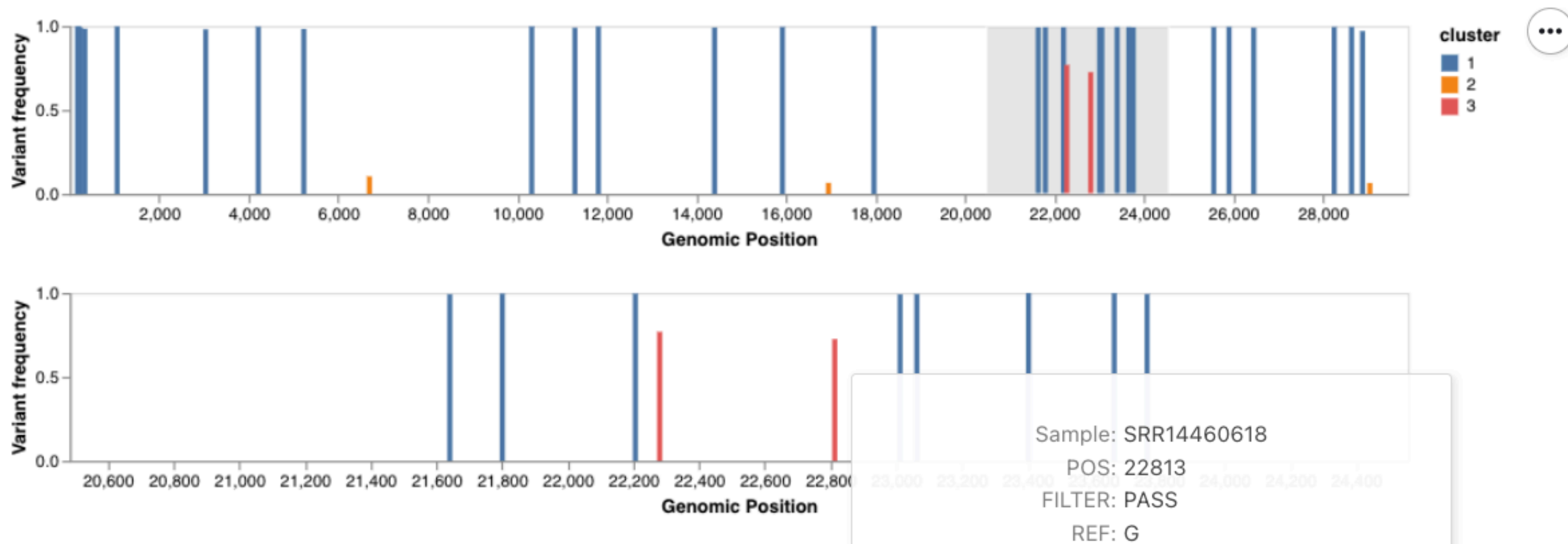


Figure 2. Allelic variants found in SRR14460618 and their inferred population clusters

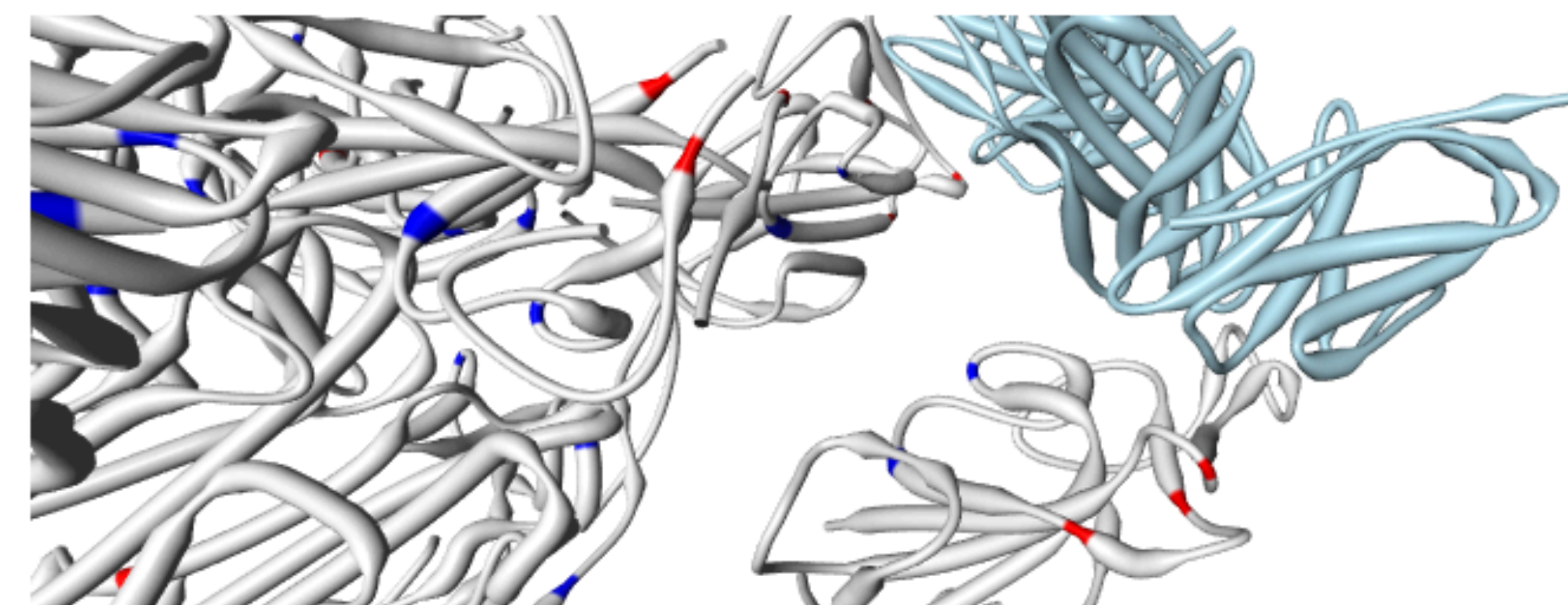


- HyPhy's FEL on 8000 ViPR genomes
- Map, compress, build tree, build alignment
- Rigorous statistical tests for non-neutral (diversifying/purifying) evolution
- Show selected sites on SARS-CoV-2 spike in complex with antibody
- Applications to vaccine design

https://observablehq.com/@stephenshank/selection-analysis-on-spike-antibody-complex-for-sars-cov-2?payload_id=01b92b54eb4f71c7

Selection analysis on spike/antibody complex for SARS-CoV-2 vaccine design

Using HyPhy's FEL method (Fixed Effect Likelihood), mapping selected sites to PDB 6XCM, with **positively selected sites in red**, **negatively selected sites in dark blue**, and **the antibody in light blue**.



Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection

Sergei L. Kosakovsky Pond and Simon D. W. Frost
Antiviral Research Center, University of California San Diego

- Currently requires **EITHER**
 - sharing history **OR**
 - using API key
- Enabling CORS
- Admits small attack vector
- Single use tokens would be preferred
- Exploring full embedding

The Observable Galaxy

Some utilities to help integrate [the Galaxy project](#) with the [ObservableHQ platform](#).

There is a Galaxy tool to generate an "ObservableHQ JSON" datatype. This requires a notebook URL (such as this one), specified via `user_name/notebook_name`, and a variable number of datasets. The tool will automatically extract the associated history ID, as well as all dataset IDs from all inputs. It then creates a [display application](#) link to the notebook with the dataset ID for the ObservableHQ JSON available as a GET parameter in the URL. With [CORS](#) enabled, ObservableHQ can fetch this JSON using [Galaxy's REST API](#), and then subsequently fetch all associated datasets for consumption by the notebook.

Some of the utilities currently require an API key, which can be specified below.

API key	<input type="text" value="....."/>
---------	------------------------------------

Fetcher

Fetch the associated ObservableHQ JSON from the `payload_id` GET parameter.

The `base_url` argument is optional, and currently defaults to <https://galaxy.hyphy.org>.



Future work



- Demonstration at 10:20 A.M. on Wednesday, 7/20
- Better collection integration
- Better API usage
- Develop more utilities
- Embed in Galaxy
- Nothing special about ObservableHQ
 - Target arbitrary URLs
 - Build similar utility libraries

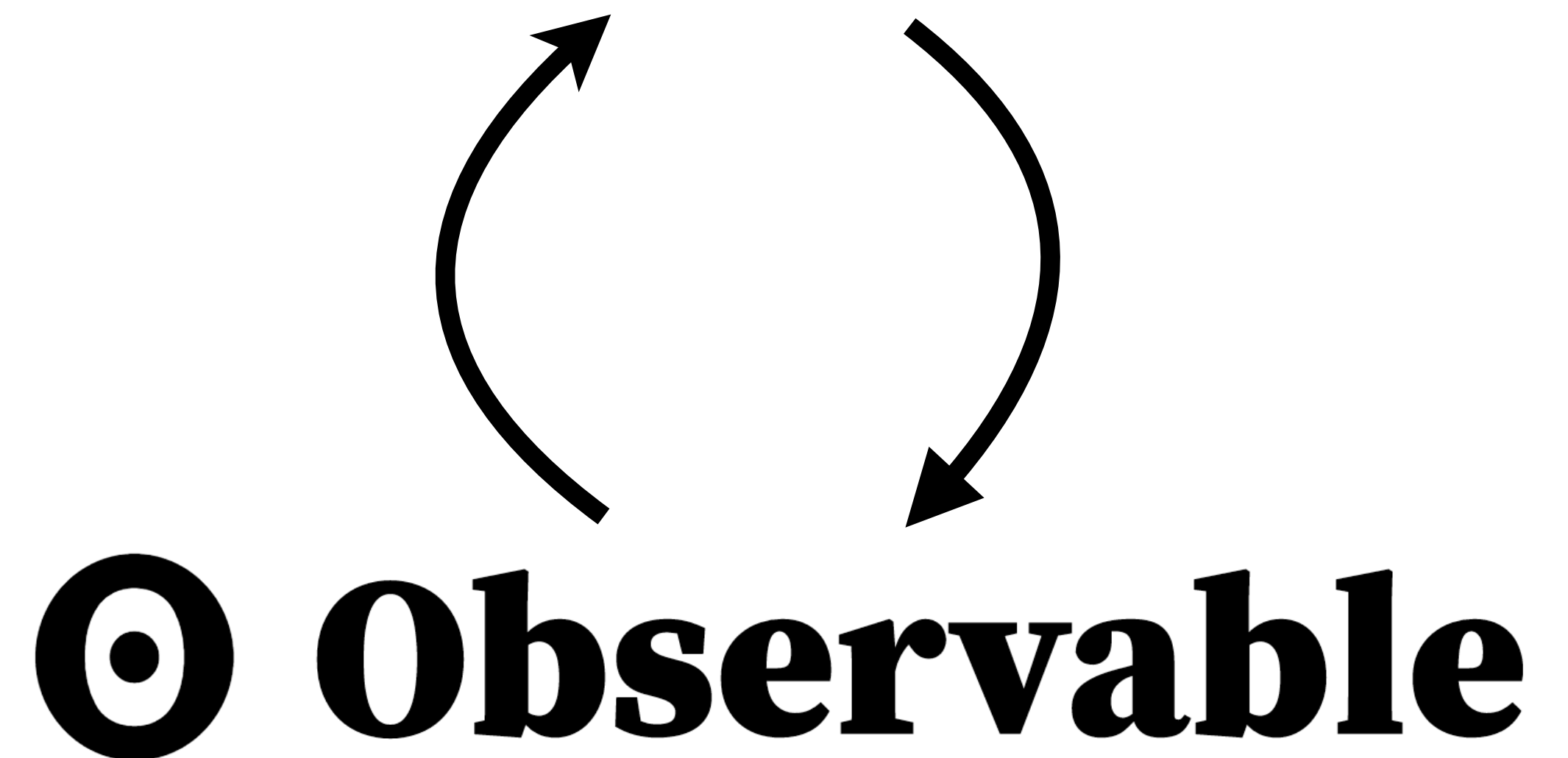


Summary



- Integrated ObservableHQ with Galaxy
- Created an associated datatype, tool, and display application
- Available on Github and test toolshed
- Integrated with
 - multiple charting frameworks
 - a rich structural viewer
 - multiple sequence alignments
 - phylogenetic trees
- Productized existing workflows

https://github.com/stephenshank/the_observable_galaxy





Acknowledgements



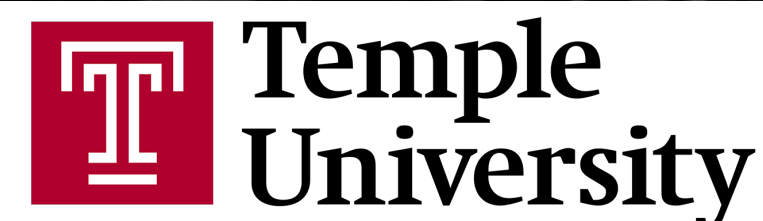
Sergei L. Kosakovsky Pond
Principal Investigator
Department of Biology
Temple University



Temple University
Alexander G. Lucaci
Jordan D. Zehr
Steven Weaver



Anton Nekrutenko
Principal Investigator
Department of Biochemistry and Molecular Biology
Penn State University



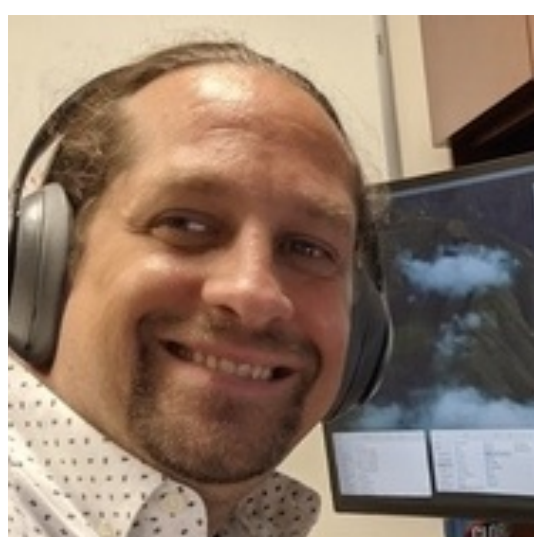
Galaxy team
Marius van den Beek
Wolfgang Maier



George Washington University
Raja Mazumder
Stephanie Singleton



Alex Ostrovsky
Senior Programming Analyst
Krieger School of Arts and Sciences
Johns Hopkins University



Charles Hadley King
Senior Research Associate
Department of Biochemistry and Molecular Medicine
George Washington University

