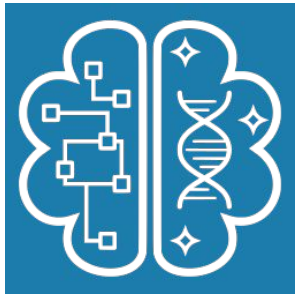
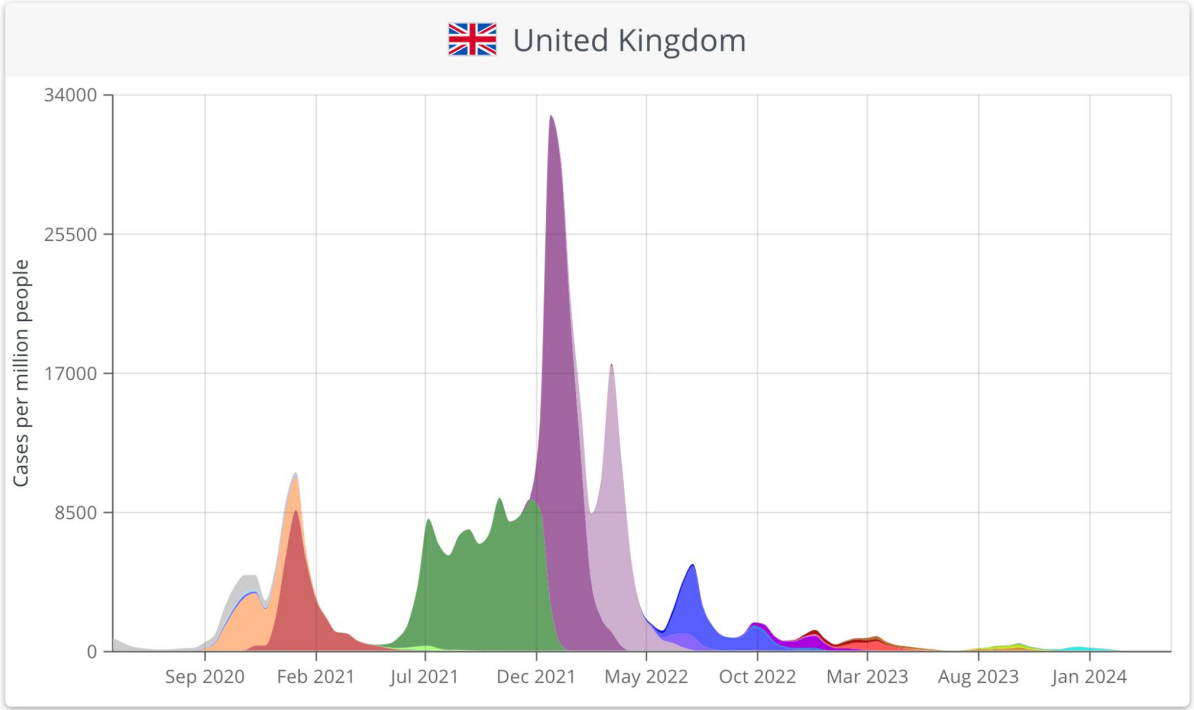


Loculus

A highly configurable, open-source platform for sharing genomic data

Our experience of the pandemic was driven by different genetic variants of the SARS-CoV-2 virus

MOTIVATION



We can track the geographic dynamics of a virus

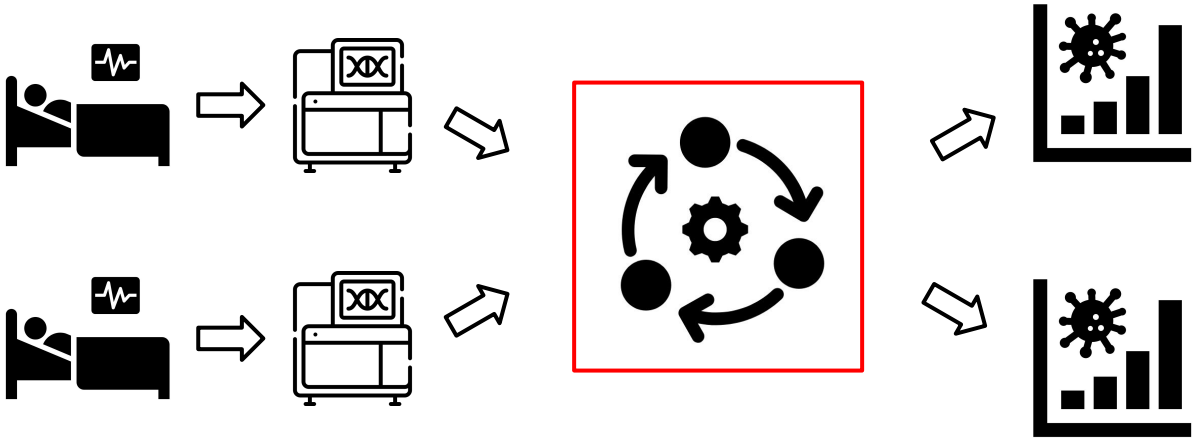
MOTIVATION

We can track the geographic distribution of a variant

MOTIVATION

We need a central hub that distributes virus data

MISSION



Genomic data helps understand the dynamics of a virus

MOTIVATION

- Understanding transmission dynamics:
  - How does it spread?
  - Almost impossible to trace with traditional means (e.g. contact tracing)
- Prediction of dynamics
  - To develop new vaccines
  - To predict case numbers (e.g. in case of a new variant)
  - To understand which countermeasures are effective

Loculus is infrastructure for real-time sharing of virus data

THE VISION

We are building Loculus, a general purpose software for virus genome databases that

- **enables accessible exchange** of virus data
- **facilitates fast querying** of virus data

Target users:

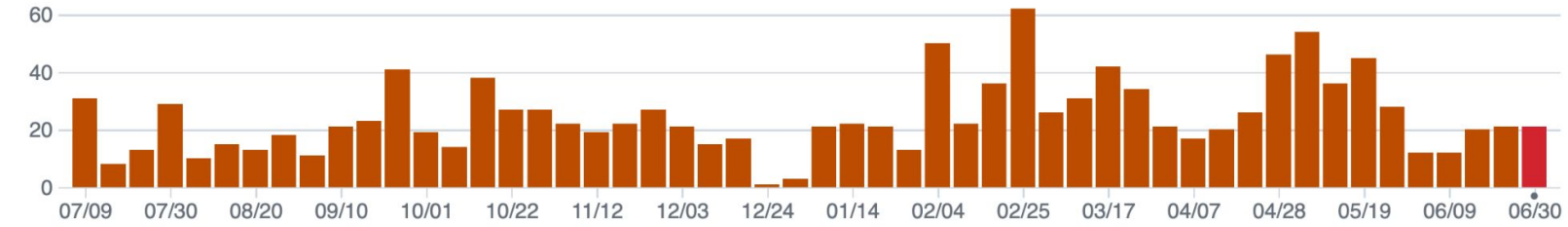
- Public health scientists
- Sequencing labs
- Research groups

Loculus will be published as open source software

LOCULUS DEVELOPMENT

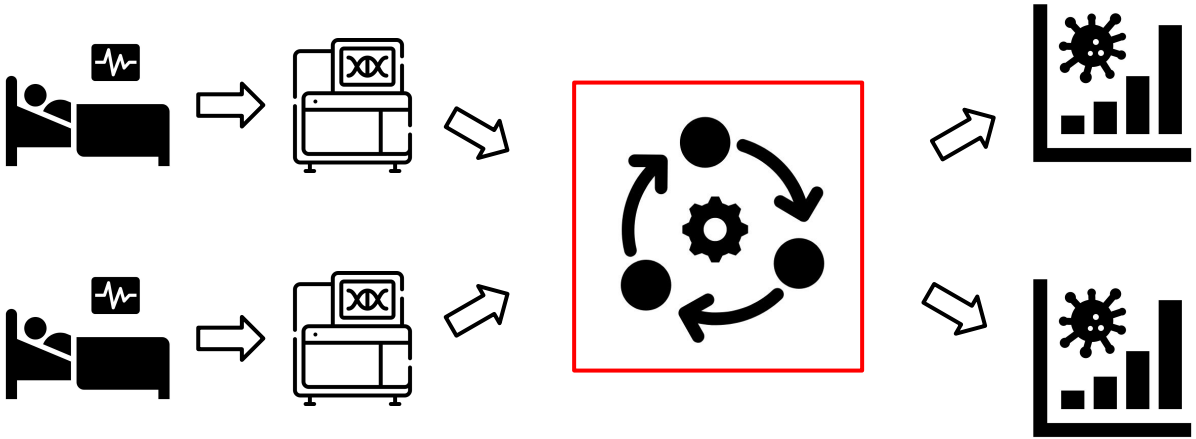
- Development is ongoing for one year
- Will be published as open source software
- Release planned during the course of the year

GitHub Commits over time



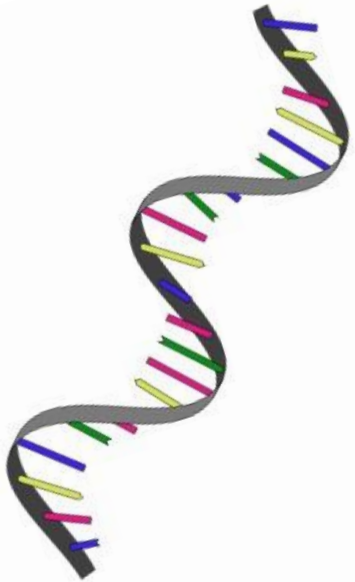
Loculus bridges between raw data and analysis tools

GENOMIC DATA



The sequencer yields unaligned nucleotide sequences

GENOMIC DATA



Sequence 1: TTTTCCGAACA…

Sequence 2: ATGTCTGAACA…

Sequence 3: TTGACCGAGCA…

Sequence 4: ATAACNATACT…

Sequence 5: TTTGNNCA…

Alignment matches the sequences to a reference genome

GENOMIC DATA

	1	2	3	4	5	6	7	8	9	10	11	…
Reference	T	T	G	T	C	C	G	A	A	C	A	…
Sequence 1	T	T	**T**	T	C	C	G	A	A	C	A	…
Sequence 2	**A**	T	G	T	C	**T**	G	A	A	C	A	…
Sequence 3	T	T	G	**A**	C	C	G	A	**G**	C	A	…
Sequence 4	**A**	T	**A**	**A**	C	**N**	**A**	**T**	A	C	**T**	…
Sequence 5	T	T	**-**	**-**	**-**	C	G	**N**	**N**	C	A	…

Nucleotide sequences can be translated to amino acids

GENOMIC DATA

- SARS-CoV-2:
  - 1 nucleotide sequence (aligned and unaligned)
  - 12 amino acid sequences (one for each protein)

Metadata contain additional information on the sequence

GENOMIC DATA

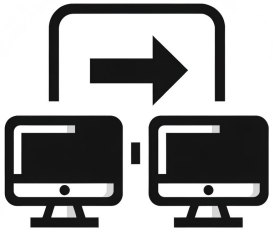
ID	region	country	age	…	Pango lineage	WHO clade	…	qcOverallScore	…
OQ907680	Asia	China	42		BF.38.3	Omicron		0.87	
OX968478	Europe	Switzerland	31		BE.7	Omicron		0.96	

There is no general purpose tool to share genomic data

CURRENT STATE

Lack of general purpose, **specialized tools**

There is no **reusable, open source software** to share and manage genomic data



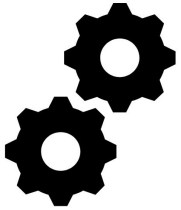
Limited, insufficient **API access** for automated analysis tools

A platform to upload and download sequence data

KEY REQUIREMENTS

Upload data

Manage data



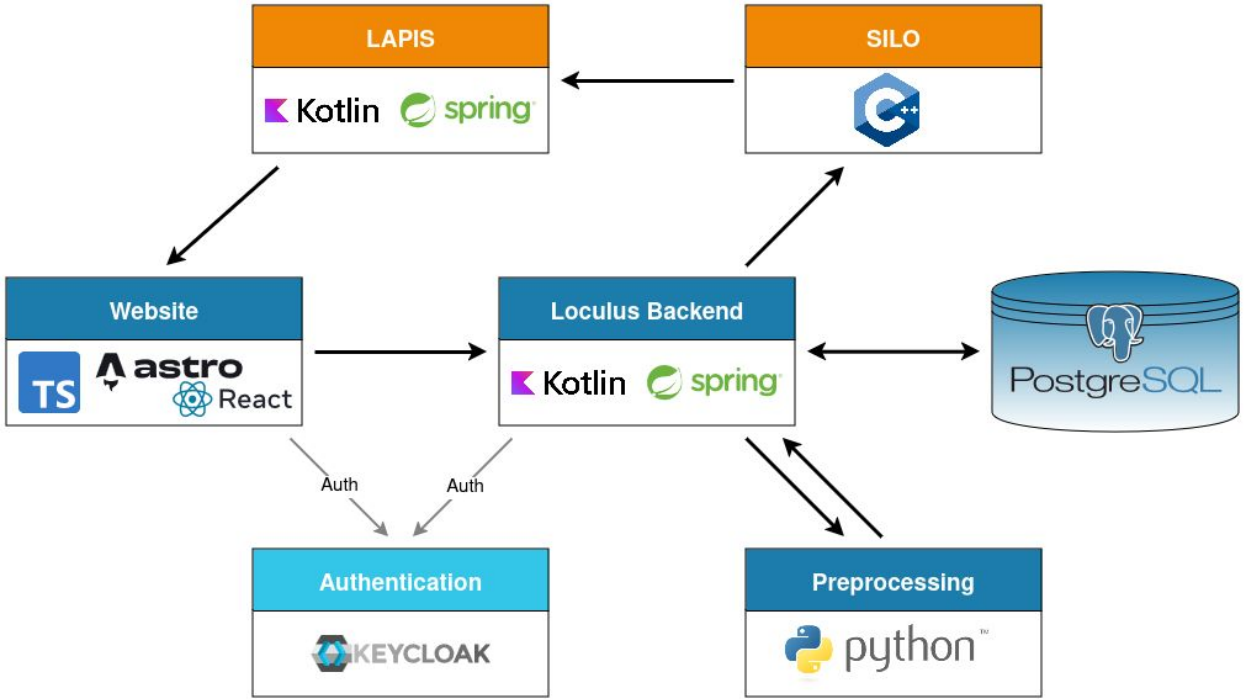
Preprocessing

Query data

Download data

The architecture with a visualization of the data flow

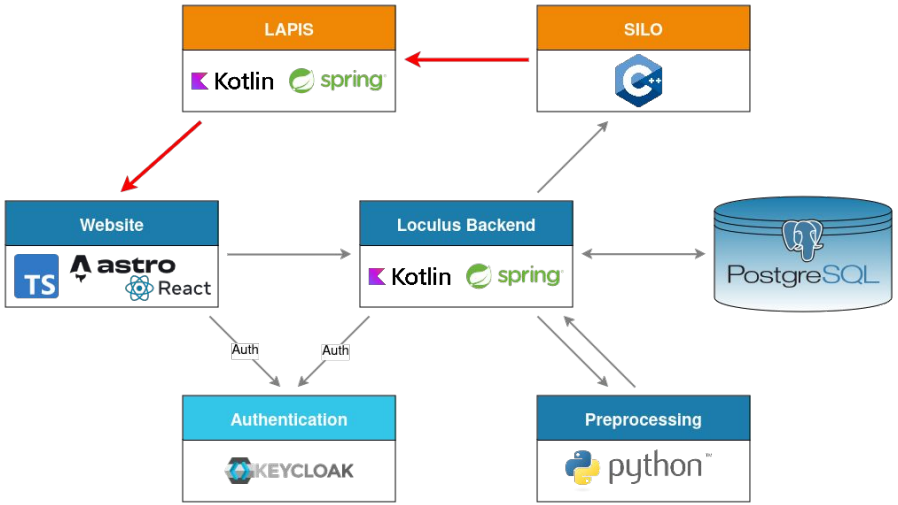
ARCHITECTURE OVERVIEW



SILO is a query engine optimized for genomic data

THE QUERY ENGINE

- SILO: Optimized for fast queries on the data:
  - “Get all sequences from the UK that have character ‘A’ at position 500 in the genome.”
- LAPIS: Convenience API around SILO

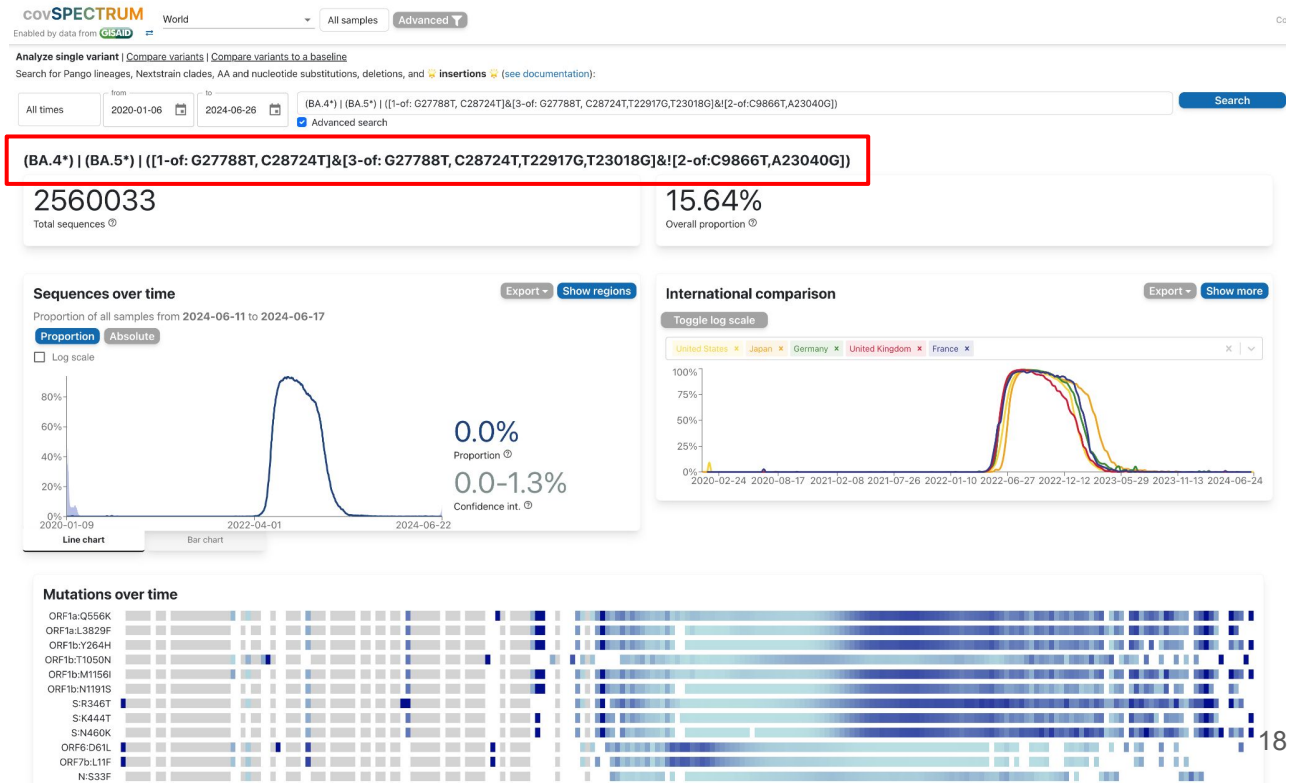


An example of a complicated query

THE QUERY ENGINE

This query filters

- 500 GB of data
- 16 Mio. rows
- 100.000 columns



SILO stores data in column-wise bitmaps for efficient queries

THE QUERY ENGINE

- C++
- Column-wise, in-memory data store
  - Time efficient computation of filters on the data
- Uses Roaring bitmaps
  - Very good compression (90 - 95% compression ratio)
- HTTP interface

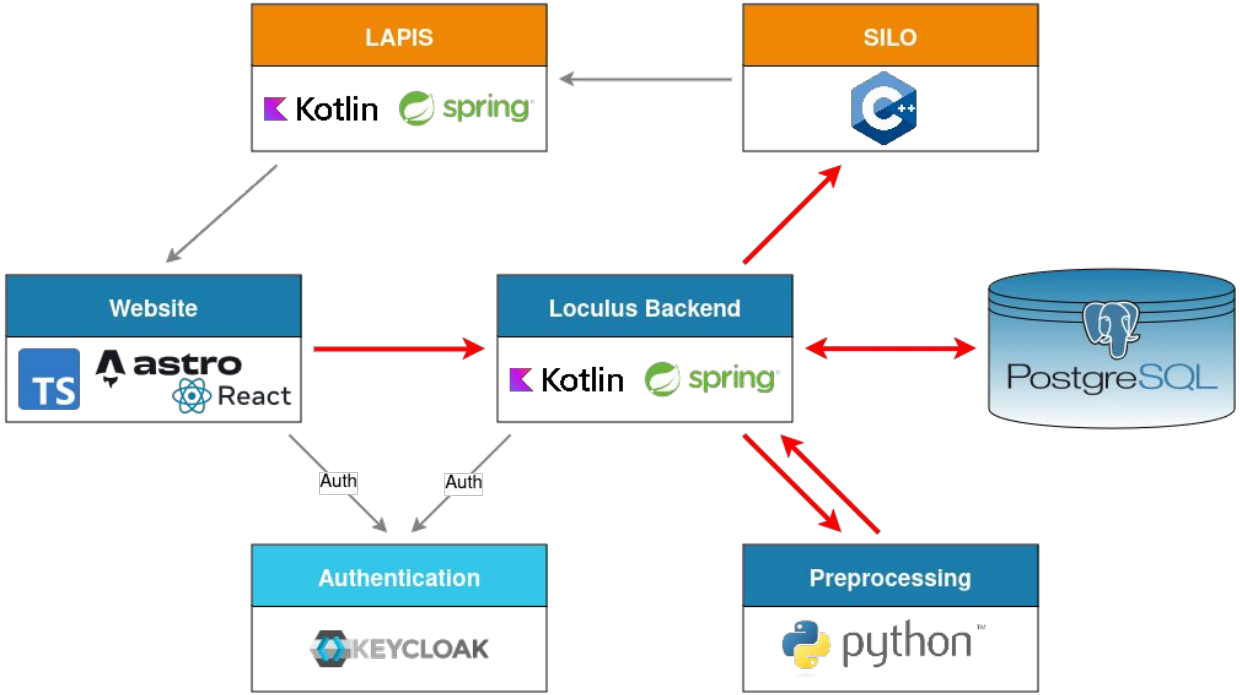
Nucleotide in genome	**1**	**2**	**3**	**4**	**5**
Sequence 1	T	T	T	T	C
Sequence 2	A	T	G	T	C
Sequence 3	T	T	G	A	C
Sequence 4	A	T	A	A	C

Is nucleotide a T?	**1**	**2**	**3**	**4**	**5**
Sequence 1	1	1	1	1	0
Sequence 2	0	1	0	1	0
Sequence 3	1	1	0	0	0
Sequence 4	0	1	0	0	0

Bitmaps

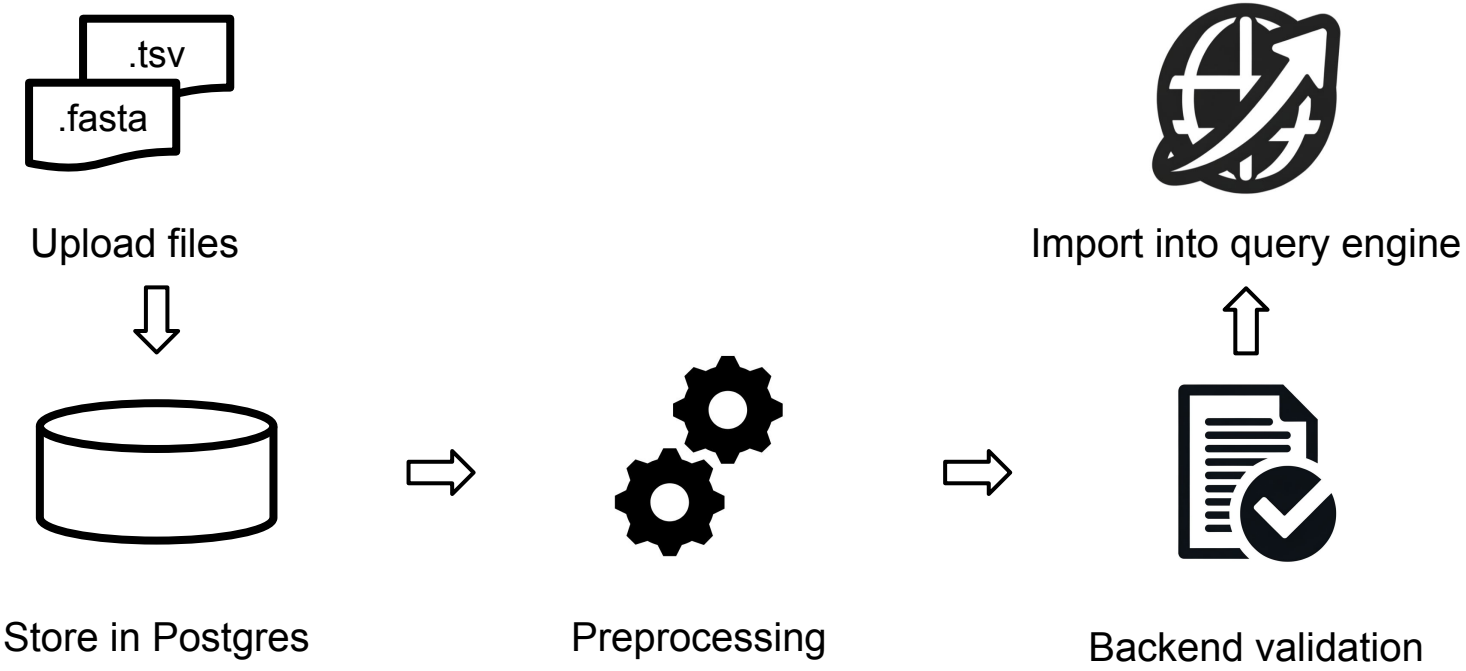
The submission process is a key component of Loculus code

THE SUBMISSION PROCESS



The uploaded data is prepared for the query engine

THE SUBMISSION PROCESS



The preprocessing enriches the data with useful information

THE SUBMISSION PROCESS

Align sequences, compute mutations and insertions

	1	2	3	4
Reference	T	T	G	T
Sequence 1	A	T	A	A
Sequence 2	T	T	-	-

Translate to amino acids

… A T G T T T G T T …

M F V

Compute derived metadata

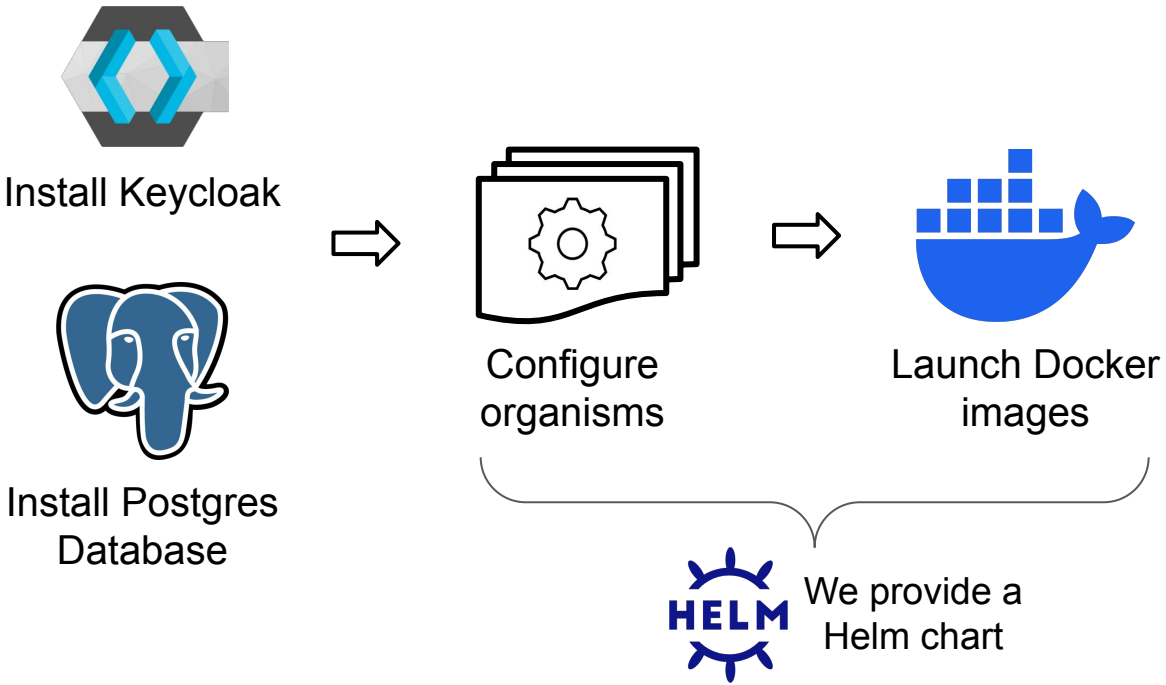
Pango lineage	WHO clade
BF.38.3	Omicron

Validate data

Report errors and warnings

The Loculus deployment should be a easy as possible

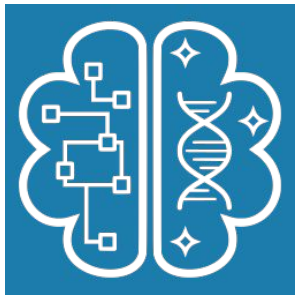
INSTALL LOCULUS



Demonstration of our test deployment

LIVE DEMONSTRATION

Live Demo



We would like to add more features to Loculus

FUTURE WORK

- Sharing of raw data
  - Sequencer actually yields a lot of sequence fragments
  - Open question: How do we analyse raw data?
- Crowd Curation
  - Users should be able to mark errors and suggest corrections on the data.
- Federation
  - Make Loculus instances communicate with each other.
  - Example: The lab instance could communicate with the uni instance or the “Swiss” instance.

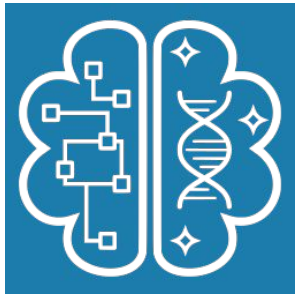
We learned a lot in a fun project

CONCLUSION AND LEARNINGS

- Software development in scientific area
  - Small and very agile environment
- Using Astro
  - Thinking in static content again
  - Reactive “Islands”
- Developing open source software has its own challenges
- Many languages and frameworks
  - Possible due to clear separation
- Configurable and flexible code is hard
- Helm templating downsides
- ArgoCD: Previews for all pull requests

Acknowledgements

- ETH Zürich
- University of Basel
- Francis Crick Institute
- Swiss TPH
- University of Toronto
- TNG



Thank you for the attention!

Any questions?