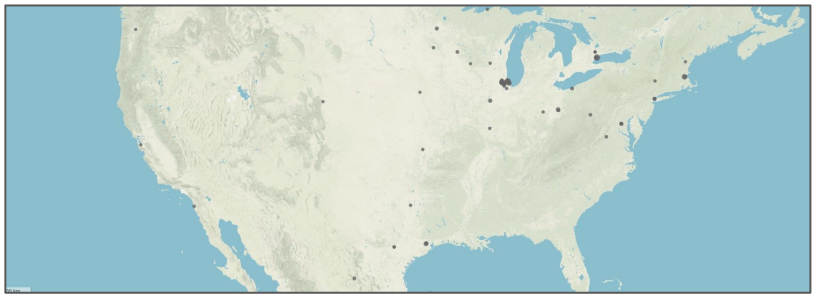
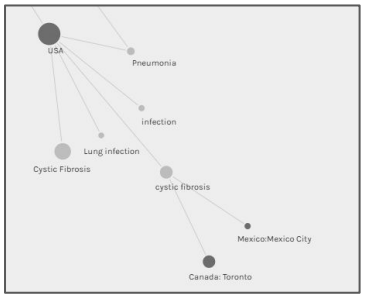
NCBImeta

Convert NCBI databases to SQLite and tabular format.

Katherine Eaton

September 16, 2019





Presentation Overview

Background

- NCBI Databases
- Power and Pitfalls
- Case Example
- Purpose

Program Overview

- Function
- Implementation
- Demo
- Output and Play

Publication Plan

- Why?
- When and Where?
- What's Left?

Background: NCBI

The National Centre for Biotechnology Information

Next Generation Sequencing

Due to advances in sequencing technology, online repositories are growing at unprecedented rates.
As a result, it can be challenging to search, filter, and explore that data effectively.

The Power

Organism: *Pseudomonas aeruginosa*
Disease: Cystic Fibrosis (CF)
Genomes: 4,914
NGS Data: 14,692

The Pitfalls

- How to organize ~20,000 records?
- Which criteria should be used for filtering?

The web-browser is suited to single-record viewing.

Bulk record retrieval is done with an **Application Programming Interface (API)** requiring knowledge of python/perl/R/etc.

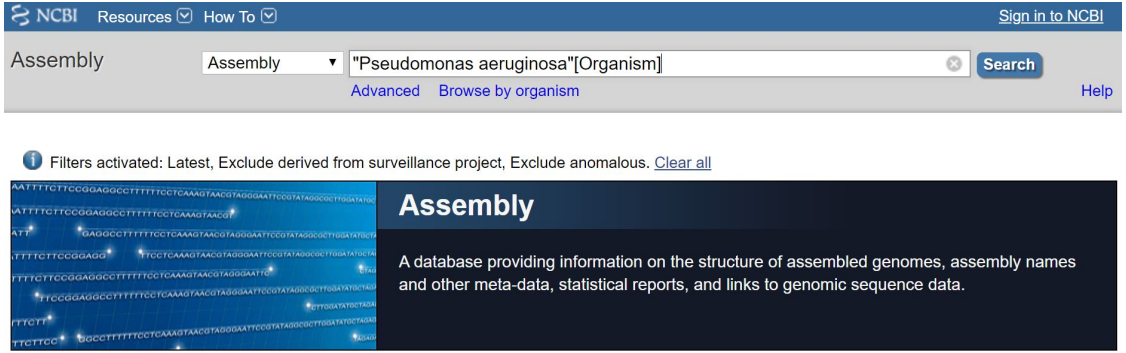
NCBI Web Interface

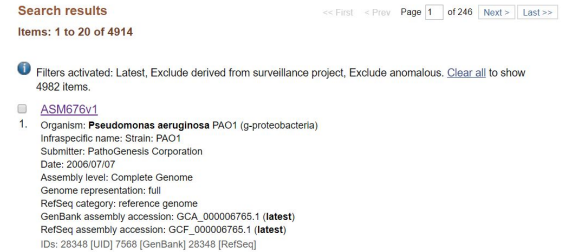
Assembled Genomes (Assembly)

Search and filter using an ***Entrez query***.

Browse individual records by navigating web pages.

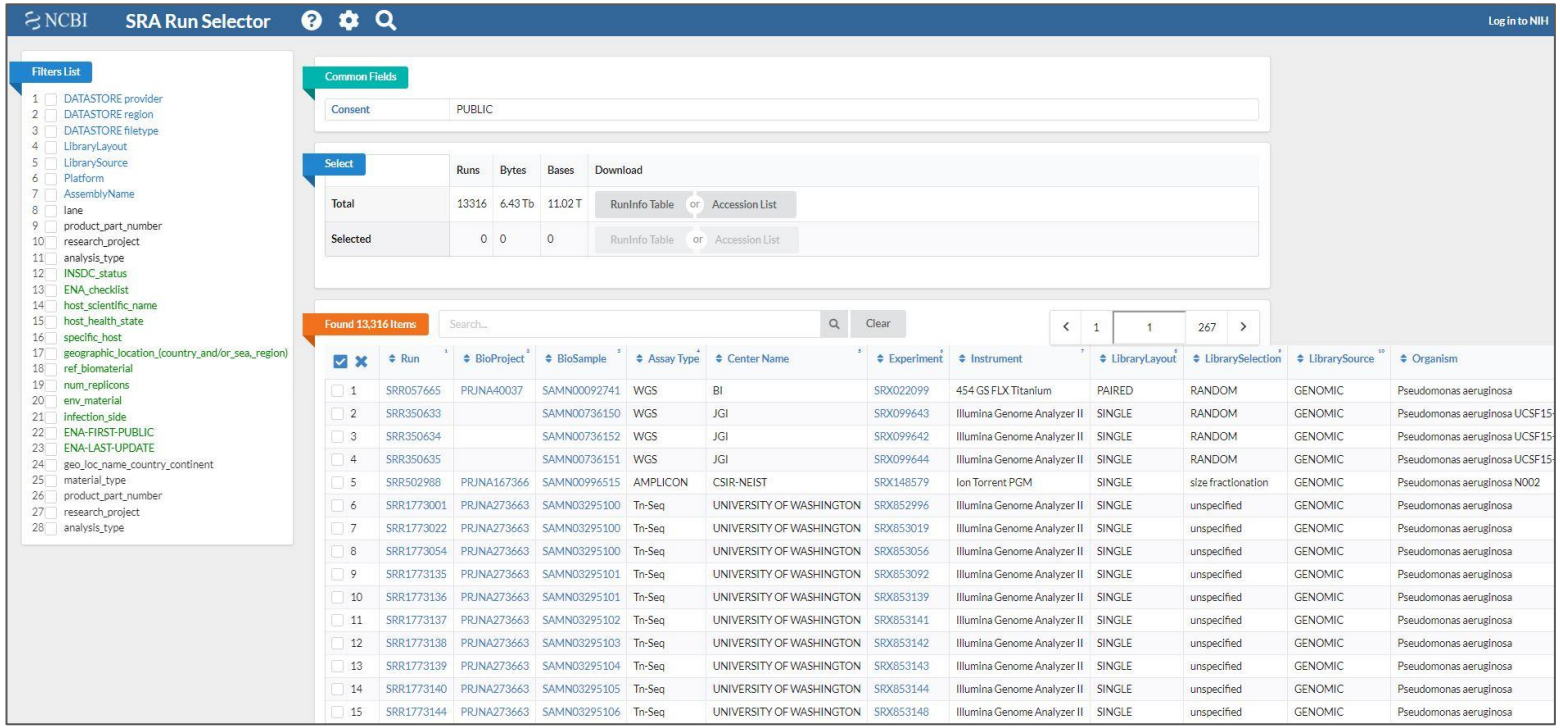
But how do you compare **multiple samples**?
Or information across **multiple databases**?



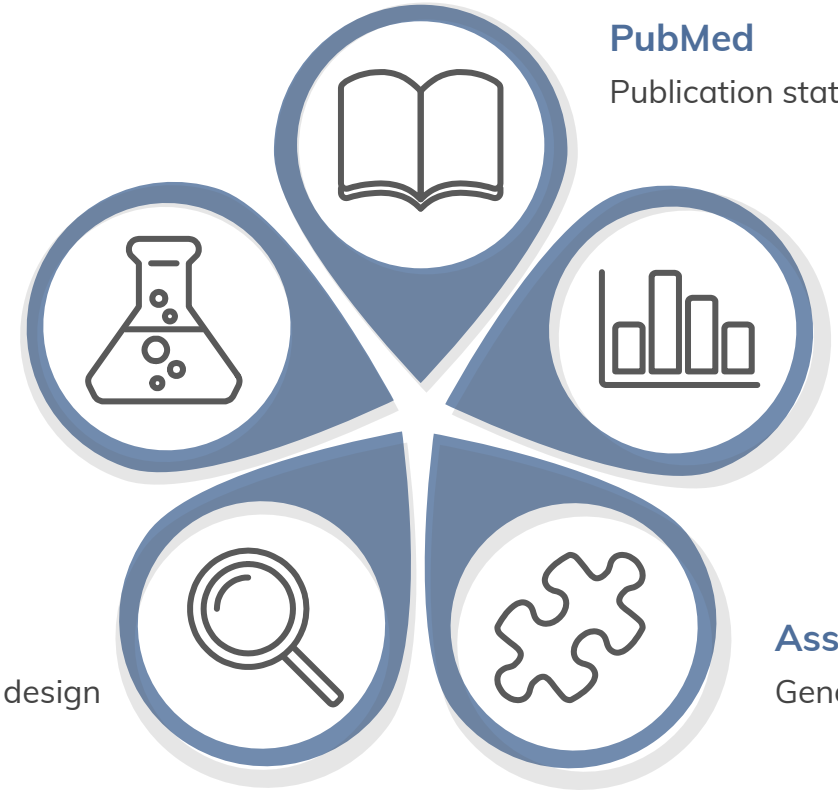


NCBI Web Interface

The Sequence Read Archive (SRA)



NCBI Databases



PubMed

Publication status, title, authors

BioSample

Sample characteristics

Nucleotide

Annotated feature counts

BioProject

Institution, study design

Assembly

Genome coverage, depth, quality

NCBI Tabular Database

Plus Another 100 columns...

Strain	Date	Location	Assembly Level	Sequencing Platform	Number of Reads	CDS
Pseudomonas aeruginosa a1	1953	Singapore	Scaffold	Illumina	1,546,237	7,676
Pseudomonas aeruginosa b2	1994	Canada	Complete	PacBio	5678	7,454
Pseudomonas aeruginosa c3	2011	Ireland	Contig	454	323	7,232
Pseudomonas aeruginosa d4	2012	Mali	Contig	Illumina	343,565	6,989
Pseudomonas aeruginosa e5	1934	France	Contig	Illumina	2,256,676	7,232

Program Overview

Create and **explore** a local database with **NCBImeta**.

NCBImeta

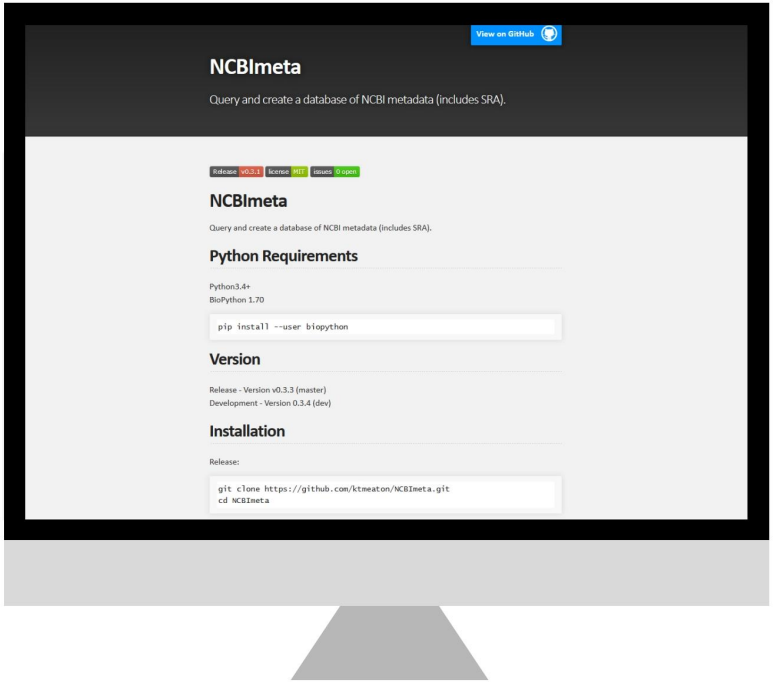
Convert NCBI databases to SQLite and tabular format.

- Command Line Interface
- Python 3 + Biopython
- Input: Single Configuration File

Optional Scripts:

- Integrate your own metadata text files.
- Export SQLite to txt.
- Master Join all tables.

★ **Coming Soon**: Conda Packaging



https://ktmeaton.github.io/NCBImeta/

NCBImeta Workflow

Set Up the Config File

1. Databases to search.
2. Query terms to use.
3. Column/fields to retrieve.

Execute Optional Scripts

1. Add custom metadata (ex. Date)
2. Export to text files.
3. Join into a mega table.

Execute the Program

python3 NCBImeta.py --config params.config

Explore!

Examine the database contents with Excel, DB Browser, CLI.

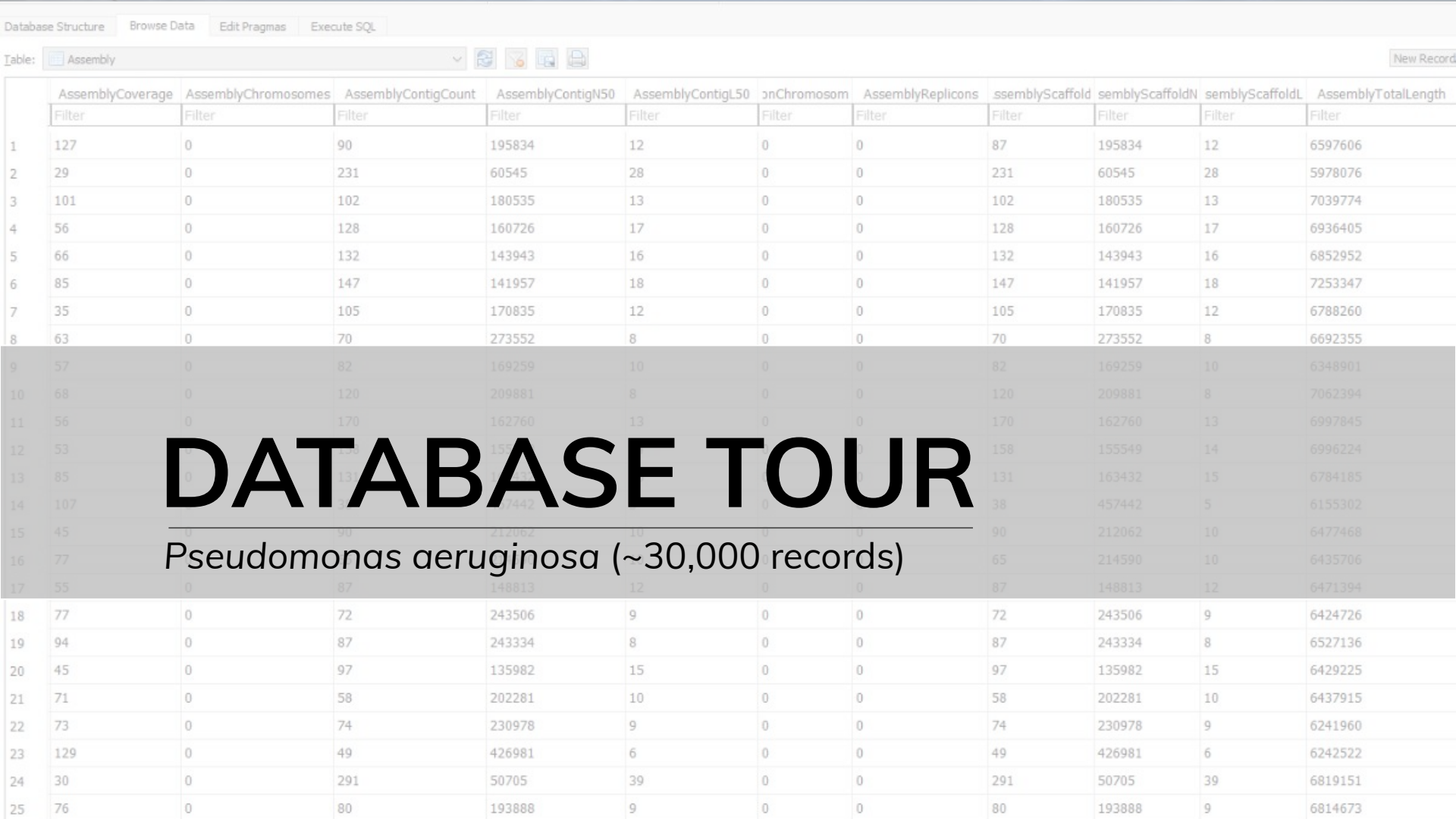
<?xml version="1.0"?>
- <DocumentSummary uid="28348">
<RsUid>28348</RsUid>
<GbUid>7568</GbUid>
<AssemblyAccession>GCF_000006765.1</AssemblyAccession>
<LastMajorReleaseAccession>GCF_000006765.1</LastMajorReleaseAccession>
<LatestAccession/>
<ChainId>6765</ChainId>
<AssemblyName>ASM676v1</AssemblyName>
<UCSCName/>
<EnsemblName/>
<Taxid>208964</Taxid>
<Organism>Pseudomonas aeruginosa PAO1 (g-proteobacteria)</Organism>
<SpeciesTaxid>287</SpeciesTaxid>
<SpeciesName>Pseudomonas aeruginosa</SpeciesName>
<AssemblyType>haploid</AssemblyType>
<AssemblyClass>haploid</AssemblyClass>
<AssemblyStatus>Complete Genome</AssemblyStatus>
<WGS/>
- <GB_BioProjects>
- <Bioproj>
<BioprojectAccn>PRJNA331</BioprojectAccn>
<BioprojectId>331</BioprojectId>
</Bioproj>
</GB_BioProjects>
<GB_Projects/>
- <RS_BioProjects>
- <Bioproj>
<BioprojectAccn>PRJNA57945</BioprojectAccn>
<BioprojectId>57945</BioprojectId>
</Bioproj>
</RS_BioProjects>
<RS_Projects/>
<BioSampleAccn>SAMN02603714</BioSampleAccn>
<BioSampleId>2603714</BioSampleId>
- <Biosource>
- <InfraspeciesList>
- <Infraspecie>
<Sub_type>strain</Sub_type>
<Sub_value>PAO1</Sub_value>
</Infraspecie>

NCBI XML

Assembly Accession:	GCF_000006765.1
Assembly Name:	ASM676v1
Taxid:	280964
Species Taxid:	287
Species Name:	Pseudomonas aeruginosa
Assembly Type:	haploid
Assembly Status:	Complete Genome
RefSeq BioProject:	PRJNA331
Genbank BioProject:	PRJNA57945
BioSample:	SAMN02603714
Strain:	PA01

Available fields are defined by the **NCBI API** and are provided as a list for the user to select from.

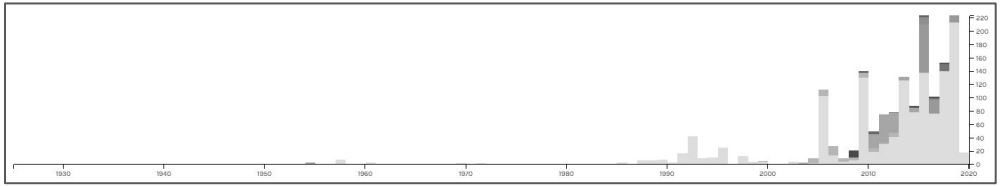
NCBImeta Table



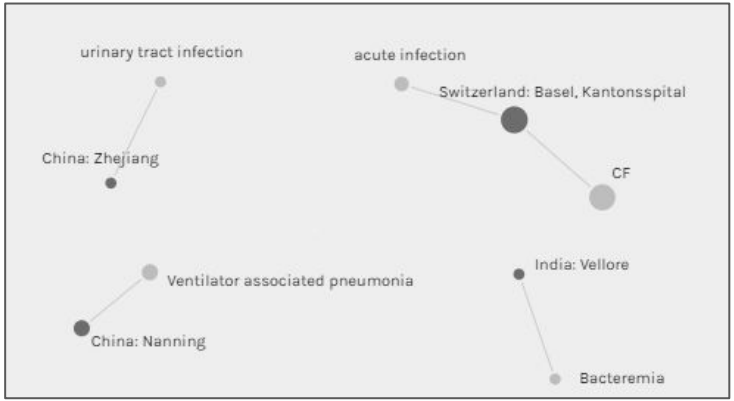
DATABASE TOUR

Pseudomonas aeruginosa (~30,000 records)

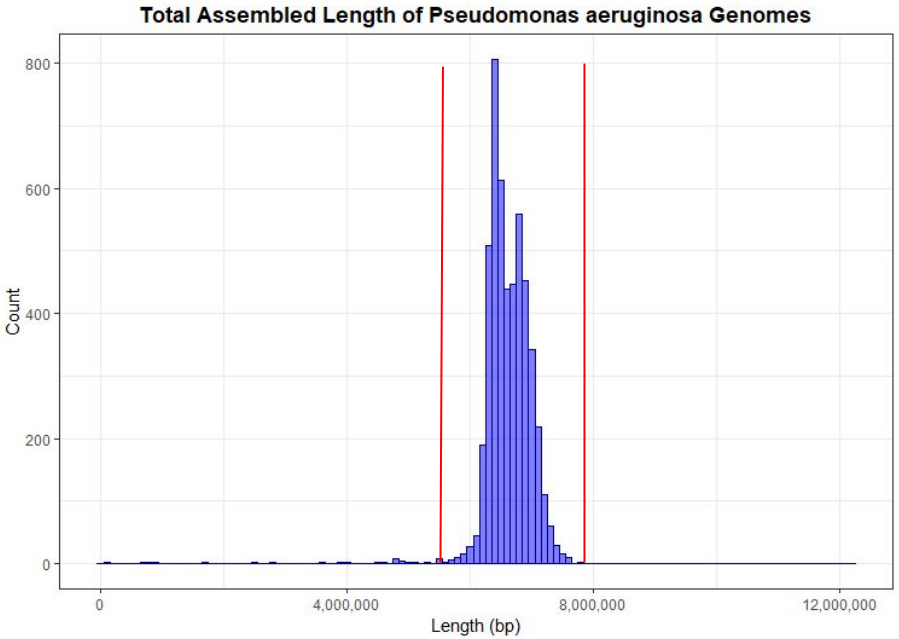
Playing Around

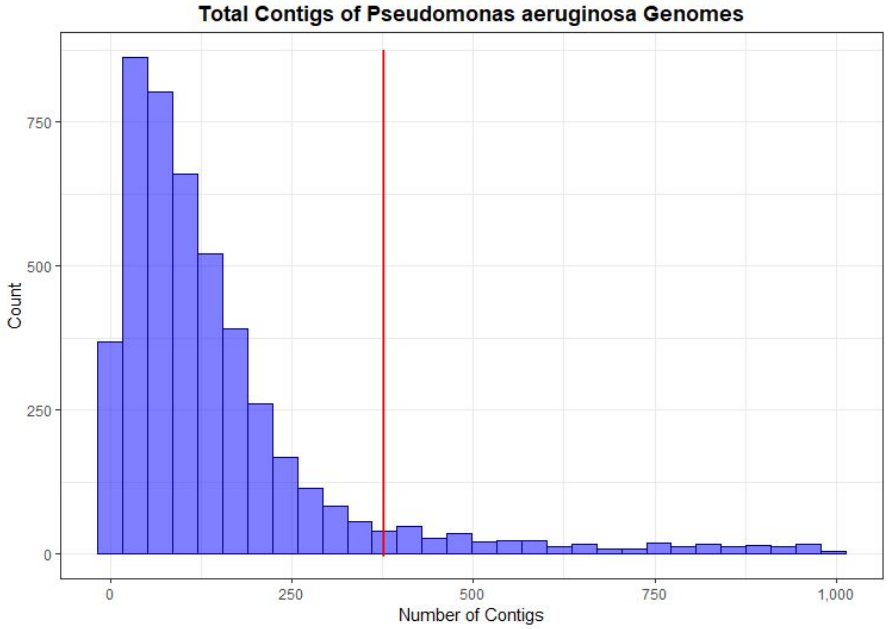


Where?
When?
What Disease?



Establishing Filtering Parameters

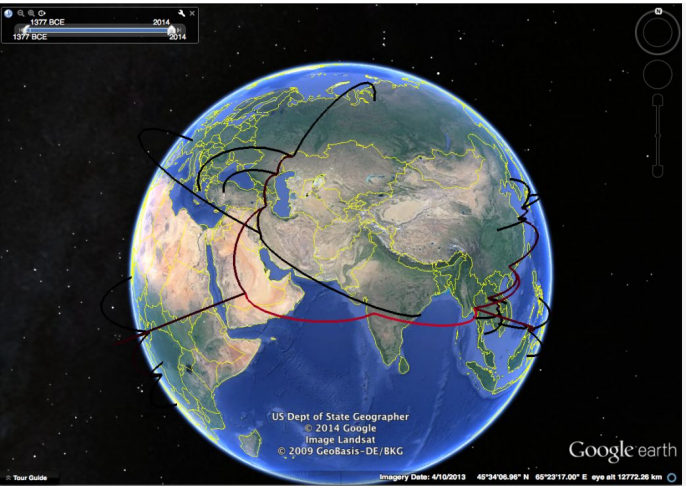


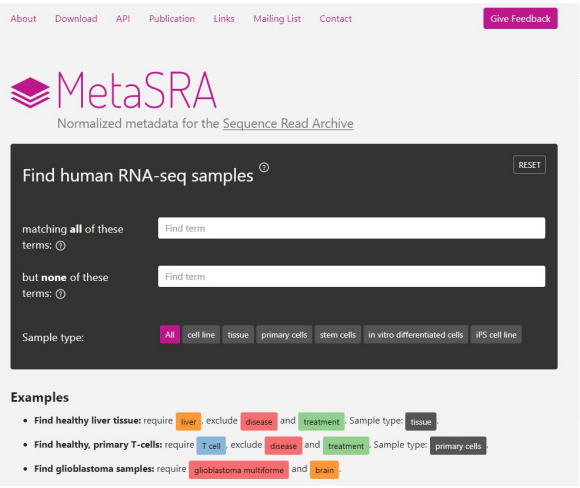


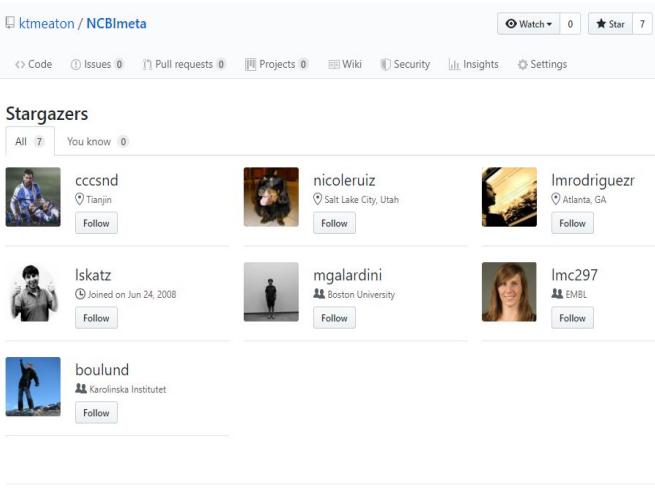
Publication Plan

Create and **explore** a local database with **NCBImeta**.

Publication Value







Metadata-Driven Analysis

- BEAST (Dates, Location)
- NextStrain (Pub Info)
- NGS as Big Data

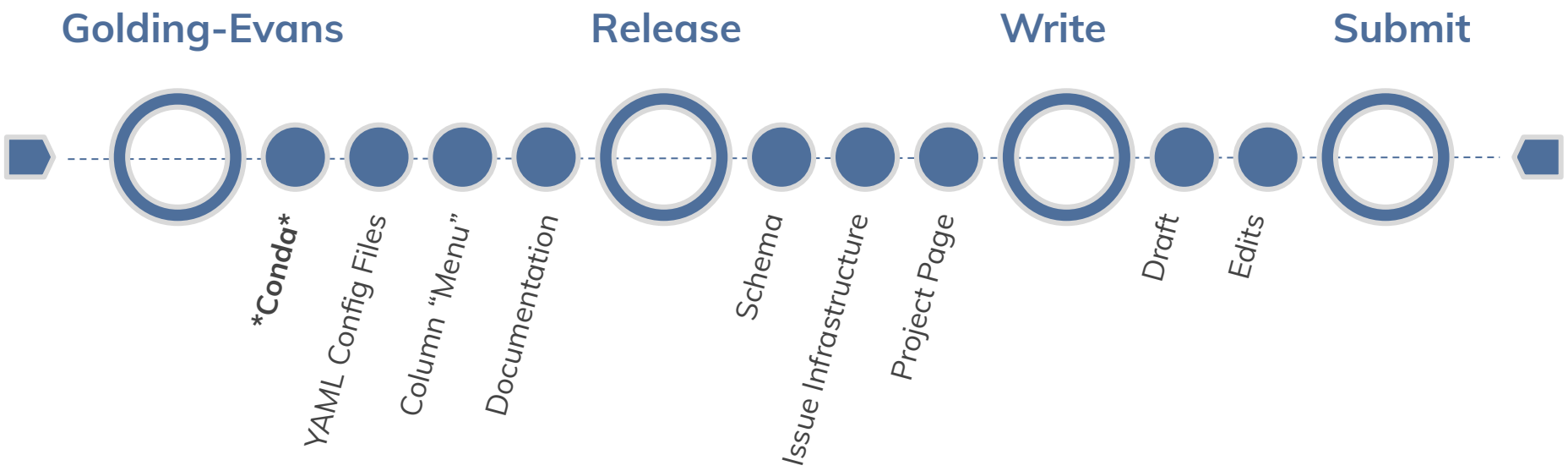
Existing NCBI Tools

- SRAdb (2013)
- MetaSRA (2017)
- pysradb (2019)

Repository Activity

- (Some) User Interaction

Publication Roadmap



Conclusion

I aimed to create a tool that could **save time,** enhance **data filtering,** support **visualization**, and promote **discovery** of global research trends.

Acknowledgements

The Poinar Lab

\+ **Ravneet Sidhu** and **Dirk Hackenberger**

The Golding Lab

- Brian Golding
- George Long
- Zachery Dickson

MacDATA INSTITUTE

Lewis & Ruth Sherman Centre for Digital Scholarship