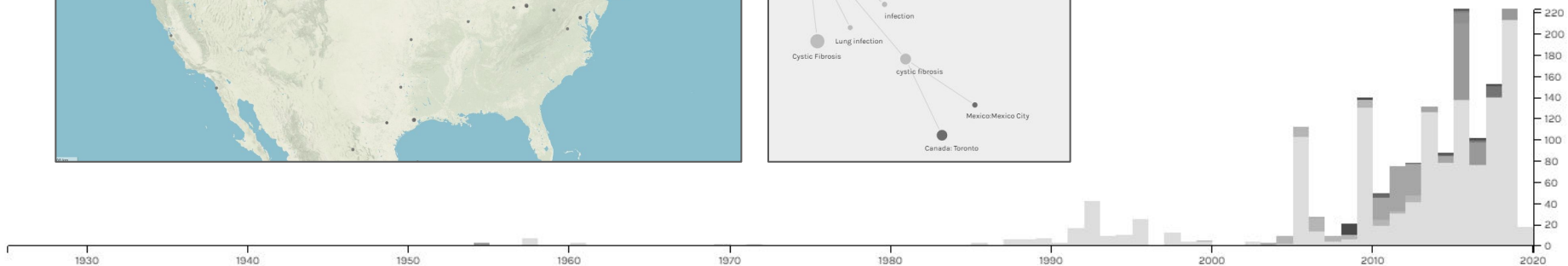
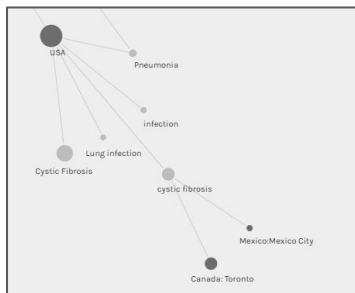
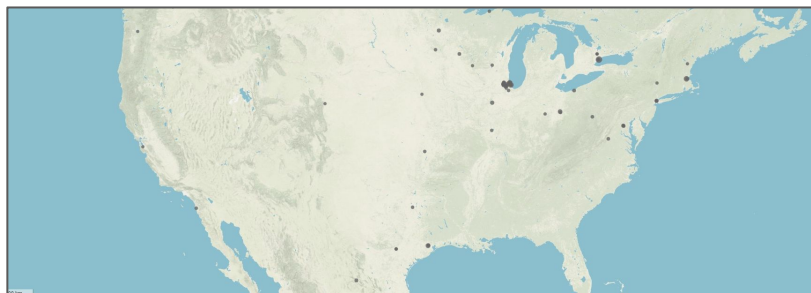


NCBImeta

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:	<i>Pseudomonas aeruginosa</i>
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 Resources How To Sign in to NCBI

Assembly Assembly "Pseudomonas aeruginosa"[Organism] Search Help

Advanced Browse by organism

Search results << First < Prev Page 1 of 246 Next > Last >>

Items: 1 to 20 of 4914

Filters activated: Latest, Exclude derived from surveillance project, Exclude anomalous. [Clear all](#) to show 4982 items.

ASM676v1

- Organism: **Pseudomonas aeruginosa** PAO1 (g-proteobacteria)
 Intraspecific name: Strain PAO1
 Submitter: PathoGenesis Corporation
 Date: 2006/07/07
 Assembly level: Complete Genome
 Genome representation: full
 RefSeq category: reference genome
 GenBank assembly accession: GCA_000006765.1 (**latest**)
 RefSeq assembly accession: GCF_000006765.1 (**latest**)
 IDs: 28348 [UID] 7568 [GenBank] 28348 [RefSeq]

Filters activated: Latest, Exclude derived from surveillance project, Exclude anomalous. [Clear all](#)



Assembly

A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

NCBI Web Interface

The Sequence Read Archive (SRA)



NCBI SRA Run Selector ? ⚙️ 🔍 Log in to NIH

Filters List

- 1 DATASTORE provider
- 2 DATASTORE region
- 3 DATASTORE filetype
- 4 LibraryLayout
- 5 LibrarySource
- 6 Platform
- 7 AssemblyName
- 8 lane
- 9 product_part_number
- 10 research_project
- 11 analysis_type
- 12 INSDC_status
- 13 ENA_checklist
- 14 host_scientific_name
- 15 host_health_state
- 16 specific_host
- 17 geographic_location(country_and/or_sea_region)
- 18 ref_biomaterial
- 19 num_replicons
- 20 env_material
- 21 infection_side
- 22 ENA-FIRST-PUBLIC
- 23 ENA-LAST-UPDATE
- 24 geo_loc_name_country_continent
- 25 material_type
- 26 product_part_number
- 27 research_project
- 28 analysis_type

Common Fields

Consent: PUBLIC

Select

	Runs	Bytes	Bases	Download
Total	13316	6.43 Tb	11.02 T	RunInfo Table or Accession List
Selected	0	0	0	RunInfo Table or Accession List

Found 13,316 Items Search... 🔍 Clear < 1 1 267 >

<input checked="" type="checkbox"/>	Run	BioProject	BioSample	Assay Type	Center Name	Experiment	Instrument	LibraryLayout	LibrarySelection	LibrarySource	Organism
<input type="checkbox"/>	1 SRR057665	PRJNA40037	SAMN0092741	WGS	BI	SRX022099	454 GS FLX Titanium	PAIRED	RANDOM	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	2 SRR350633		SAMN00736150	WGS	JGI	SRX099643	Illumina Genome Analyzer II	SINGLE	RANDOM	GENOMIC	Pseudomonas aeruginosa UCSF15
<input type="checkbox"/>	3 SRR350634		SAMN00736152	WGS	JGI	SRX099642	Illumina Genome Analyzer II	SINGLE	RANDOM	GENOMIC	Pseudomonas aeruginosa UCSF15
<input type="checkbox"/>	4 SRR350635		SAMN00736151	WGS	JGI	SRX099644	Illumina Genome Analyzer II	SINGLE	RANDOM	GENOMIC	Pseudomonas aeruginosa UCSF15
<input type="checkbox"/>	5 SRR502988	PRJNA167366	SAMN00996515	AMPLICON	CSIR-NEIST	SRX148579	Ion Torrent PGM	SINGLE	size fractionation	GENOMIC	Pseudomonas aeruginosa N002
<input type="checkbox"/>	6 SRR1773001	PRJNA273663	SAMN03295100	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX852996	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	7 SRR1773022	PRJNA273663	SAMN03295100	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853019	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	8 SRR1773054	PRJNA273663	SAMN03295100	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853056	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	9 SRR1773135	PRJNA273663	SAMN03295101	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853092	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	10 SRR1773136	PRJNA273663	SAMN03295101	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853139	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	11 SRR1773137	PRJNA273663	SAMN03295102	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853141	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	12 SRR1773138	PRJNA273663	SAMN03295103	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853142	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	13 SRR1773139	PRJNA273663	SAMN03295104	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853143	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	14 SRR1773140	PRJNA273663	SAMN03295105	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853144	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa
<input type="checkbox"/>	15 SRR1773144	PRJNA273663	SAMN03295106	Tn-Seq	UNIVERSITY OF WASHINGTON	SRX853148	Illumina Genome Analyzer II	SINGLE	unspecified	GENOMIC	Pseudomonas aeruginosa

NCBI Databases



BioSample

Sample characteristics



PubMed

Publication status, title, authors



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
<i>Pseudomonas aeruginosa</i> a1	1953	Singapore	Scaffold	Illumina	1,546,237	7,676
<i>Pseudomonas aeruginosa</i> b2	1994	Canada	Complete	PacBio	5678	7,454
<i>Pseudomonas aeruginosa</i> c3	2011	Ireland	Contig	454	323	7,232
<i>Pseudomonas aeruginosa</i> d4	2012	Mali	Contig	Illumina	343,565	6,989
<i>Pseudomonas aeruginosa</i> e5	1934	France	Contig	Illumina	2,256,676	7,232

Program Overview

Create and explore a local database with [NCBImeta](#).

NCBImeta

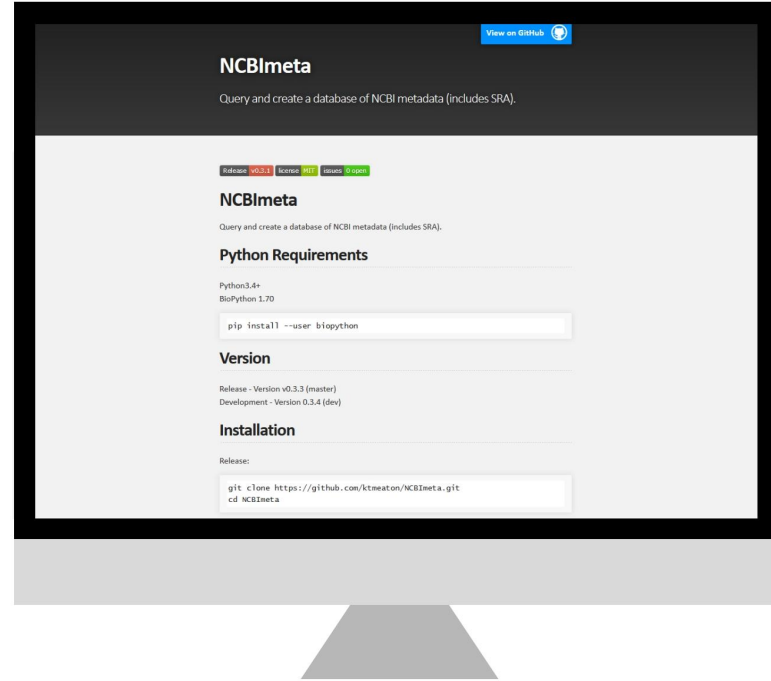
Convert NCBI databases to SQLite and tabular format.

- 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>
  </InfraspeciesList>
</Biosource>

```

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:	PAO1

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

NCBI XML



NCBImeta Table

Table: Assembly

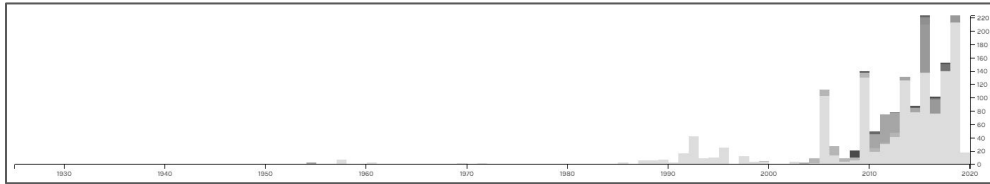
New Record

	AssemblyCoverage	AssemblyChromosomes	AssemblyContigCount	AssemblyContigN50	AssemblyContigL50	onChromosome	AssemblyReplicons	AssemblyScaffold	AssemblyScaffoldN	AssemblyScaffoldL	AssemblyTotalLength
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	127	0	90	195834	12	0	0	87	195834	12	6597606
2	29	0	231	60545	28	0	0	231	60545	28	5978076
3	101	0	102	180535	13	0	0	102	180535	13	7039774
4	56	0	128	160726	17	0	0	128	160726	17	6936405
5	66	0	132	143943	16	0	0	132	143943	16	6852952
6	85	0	147	141957	18	0	0	147	141957	18	7253347
7	35	0	105	170835	12	0	0	105	170835	12	6788260
8	63	0	70	273552	8	0	0	70	273552	8	6692355
9	57	0	82	169259	10	0	0	82	169259	10	6348901
10	68	0	120	209881	8	0	0	120	209881	8	7062394
11	56	0	170	162760	13	0	0	170	162760	13	6997845
12	53	0	153	155549	14	0	0	158	155549	14	6996224
13	85	0	131	163432	15	0	0	131	163432	15	6784185
14	107	0	7442	457442	5	0	0	38	457442	5	6155302
15	45	0	90	212062	10	0	0	90	212062	10	6477468
16	77	0	72	243506	9	0	0	65	214590	10	6435706
17	55	0	87	148813	12	0	0	87	148813	12	6471394
18	77	0	72	243506	9	0	0	72	243506	9	6424726
19	94	0	87	243334	8	0	0	87	243334	8	6527136
20	45	0	97	135982	15	0	0	97	135982	15	6429225
21	71	0	58	202281	10	0	0	58	202281	10	6437915
22	73	0	74	230978	9	0	0	74	230978	9	6241960
23	129	0	49	426981	6	0	0	49	426981	6	6242522
24	30	0	291	50705	39	0	0	291	50705	39	6819151
25	76	0	80	193888	9	0	0	80	193888	9	6814673

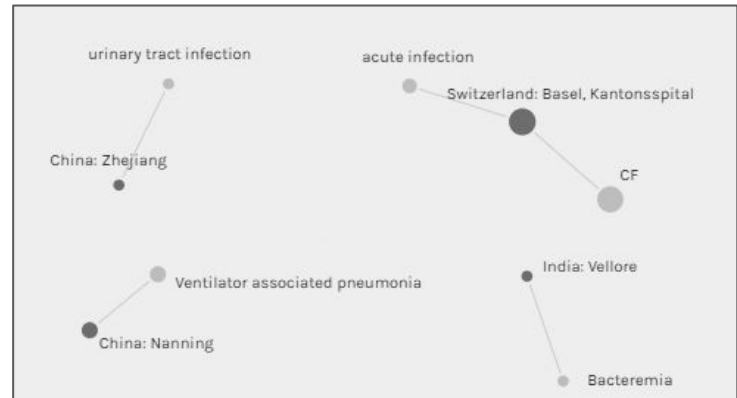
DATABASE TOUR

Pseudomonas aeruginosa (~30,000 records)

Playing Around



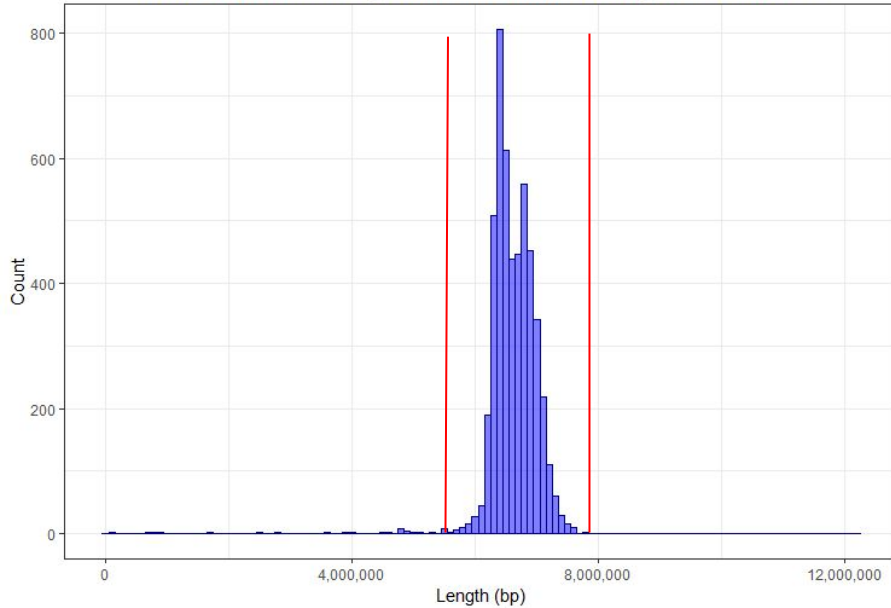
Where?
When?
What Disease?



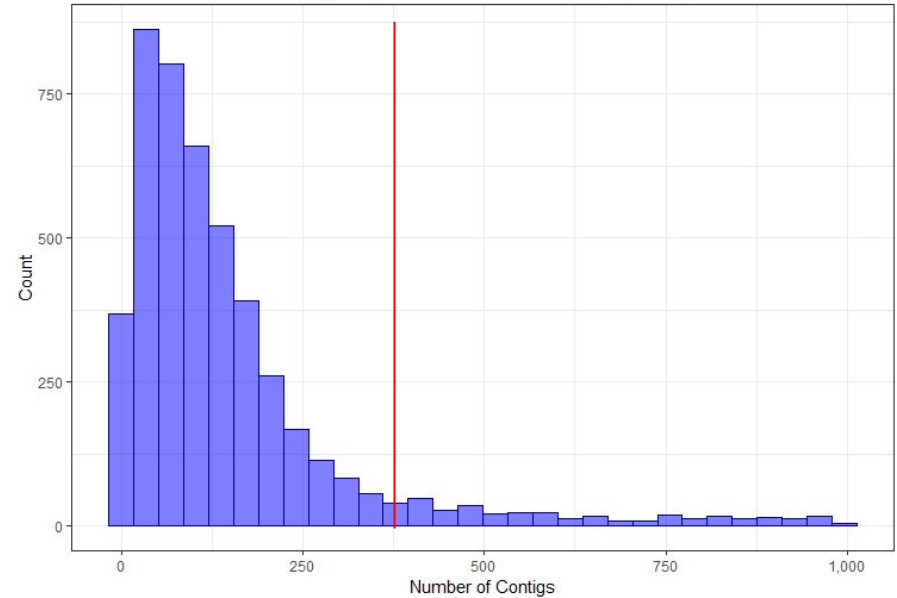
Establishing Filtering Parameters



Total Assembled Length of *Pseudomonas aeruginosa* Genomes



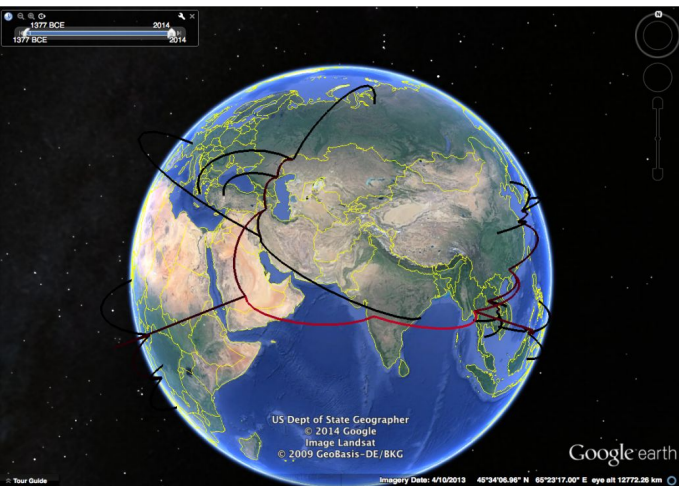
Total Contigs of *Pseudomonas aeruginosa* Genomes



Publication Plan

Create and explore a local database with [NCBImeta](#).

Publication Value



MetaSRA
Normalized metadata for the Sequence Read Archive

Find human RNA-seq samples

matching **all** of these terms:

but **none** of these terms:

Sample type: All cell line tissue primary cells stem cells in vitro differentiated cells iPSC cell line

Examples

- Find healthy liver tissue: require **liver** exclude **disease** and **treatment** Sample type: **tissue**
- Find healthy, primary T-cells: require **T cell** exclude **disease** and **treatment** Sample type: **primary cells**
- Find glioblastoma samples: require **glioblastoma multiforme** and **brain**

ktmeaton / NCBImeta

Code Issues Pull requests Projects Wiki Security Insights Settings

Stargazers

All 7 You know

- cccsnd Tianjin Follow
- nicoleruiz Salt Lake City, Utah Follow
- Imrodriguezr Atlanta, GA Follow
- lskatz Joined on Jun 24, 2008 Follow
- mgalardini Boston University Follow
- lmc297 EMBL Follow
- boulund Karolinska Institutet Follow

Metadata-Driven Analysis

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

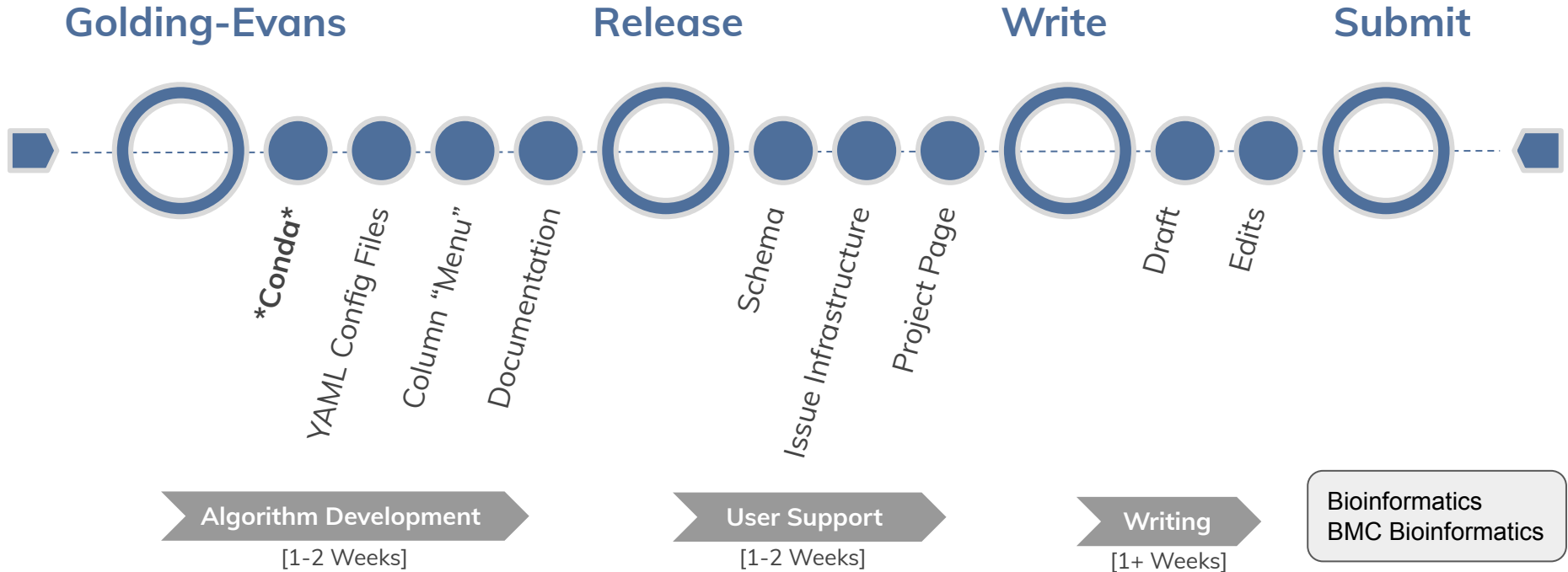
Existing NCBI Tools

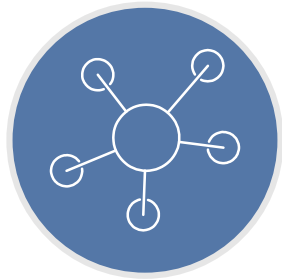
- SRAdb (2013)
- MetaSRA (2017)
- pysradb (2019)

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



Lewis & Ruth
Sherman Centre
for Digital Scholarship