SUPPORTING A CLOUD-BASED BIOMEDICAL DATA INFRASTRUCTURE

National Center for Biotechnology Information

Advised by Carl Leubsdorf

CIVIC DIGITAL FELLOWSHIP

NIH

NIK MARDA
Stanford
CS + Political Science

DAVID FRANKEL
WashU
CS + Gender Studies

CURRENT INFRASTRUCTURE

- National Center for Biomedical Information (NCBI) has over 1.5 billion visits per year
- NCBI hosts valuable tools like PubMed, the Basic Local Alignment Search Tool, and the Sequence Read Archive
- NCBI currently hosts over 40 petabytes of data on internal servers
    - This data is expensive for the NCBI to host locally
    - Single-region access hinders researchers' use of modern computational tools

BASIC LOCAL ALIGNMENT SEARCH TOOL

- The Basic Local Alignment Search Tool (BLAST) finds regions of similarity between biological sequences
- The queried sequences are compared to sequence databases for similarity
- Researchers are not familiar with using these tools and data in the cloud
- Created engaging and effective tutorials for teaching these new techniques

CIVIC DIGITAL FELLOWSHIP

BASIC LOCAL ALIGNMENT SEARCH TOOL

We can use the dataframe to extract other information. In the next two cells, we'll check how many rows are in the table, and how many unique database matches we found.

In [7]:
```
print('There are {} alignments, with {} unique subject sequences'.format(blast_results.index.size,
                                                                          blast_results.sseqid.unique().size))

# generate descriptive statistics for numerical columns
blast_results.describe()
```

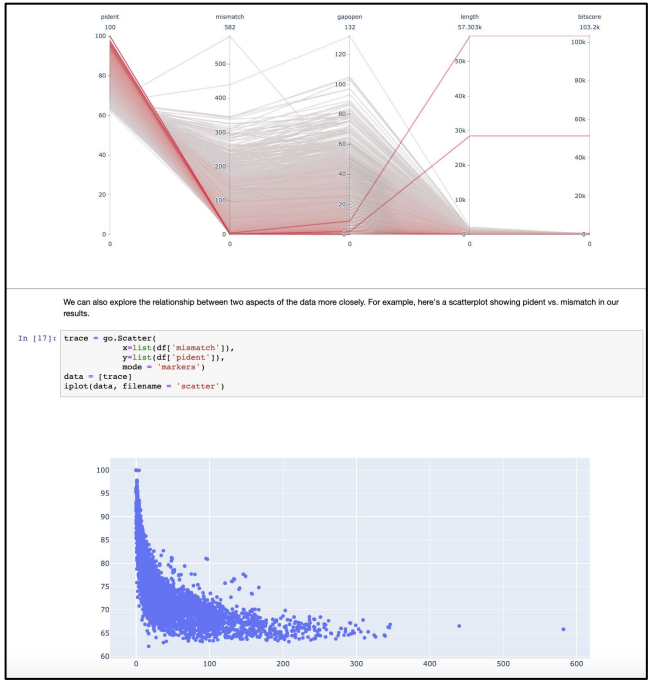
```
There are 81 alignments, with 77 unique subject sequences
```

Out[7]:

	pident	qcovs	length	mismatch	gapopen	qstart	qend	sstart	send	qframe	sframe	evalue	
count	81.000000	81.0	81.000000	81.000000	81.000000	81.000000	81.0	81.000000	81.000000	81.0	81.000000	81.0	8
mean	99.852704	100.0	812.888889	1.160494	0.024691	1.111111	813.0	44596.740741	44566.740741	1.0	-0.037037	0.0	149
std	0.128150	0.0	1.000000	1.005694	0.156150	1.000000	0.0	40848.854130	40827.196349	0.0	1.005540	0.0	
min	99.508000	100.0	804.000000	0.000000	0.000000	1.000000	813.0	964.000000	152.000000	1.0	-1.000000	0.0	147
25%	99.754000	100.0	813.000000	0.000000	0.000000	1.000000	813.0	15694.000000	16506.000000	1.0	-1.000000	0.0	149
50%	99.754000	100.0	813.000000	2.000000	0.000000	1.000000	813.0	33320.000000	32694.000000	1.0	-1.000000	0.0	149
75%	100.000000	100.0	813.000000	2.000000	0.000000	1.000000	813.0	53317.000000	54129.000000	1.0	1.000000	0.0	150
max	100.000000	100.0	813.000000	3.000000	1.000000	10.000000	813.0	193811.000000	192999.000000	1.0	1.000000	0.0	150

There were 81 rows in the table, but only 77 different database sequences (i.e., plasmids) were found. This indicates that some plasmids contained multiple copies of the AMR gene. To confirm this, we'll need to go back and take a look at the blast results. The next command will identify those plasmids with multiple BLAST matches and print them out.

In [8]:
```
blast_results[blast_results.duplicated('sseqid', False)]
```



CIVIC DIGITAL FELLOWSHIP

Examples from in-browser Jupyter Notebook tutorials

OTHER TUTORIALS AND WEBSITE

- Created tutorials for getting set up with Google and Amazon cloud services
- Improved readability and usability of Sequence Read Archive tutorials
- Conducted and incorporated feedback from multiple rounds of user testing
- Outlined and prioritized other tutorials to be built in the future
- Built web page to host tutorials

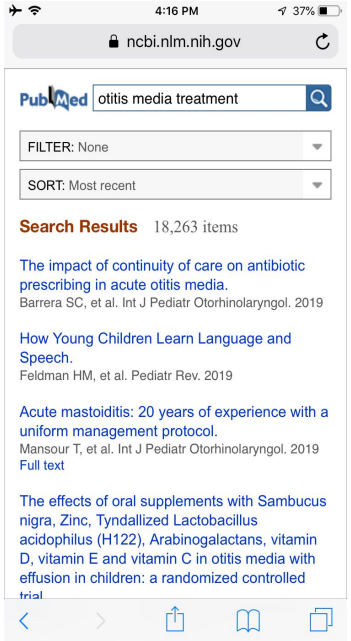
CIVIC DIGITAL FELLOWSHIP

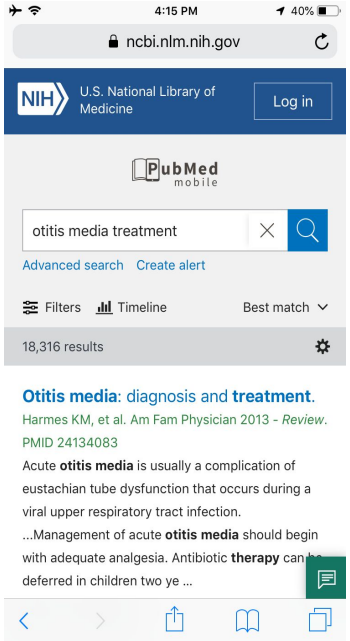
PUBMED LABS

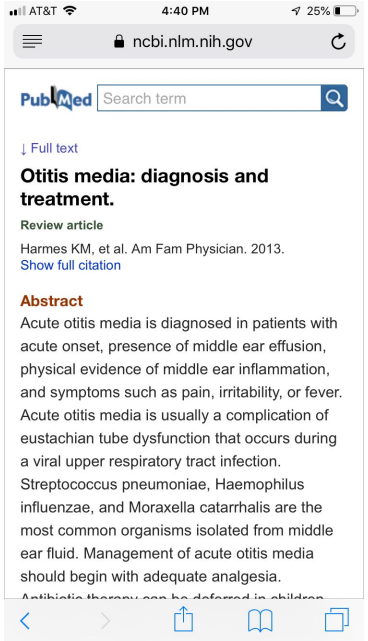
- PubMed Labs is a cloud-based redesign of PubMed that has
  - Cutting-edge search
  - Mobile-first design with article snippets
  - Cross-browser compatibility and responsive design
- Incorporated new features and functionality into PubMed Labs
  - Standardized functionality across platforms
  - Resolved accessibility barriers
  - Redesigned features around browser flaws

CIVIC DIGITAL FELLOWSHIP

PUBMED LABS

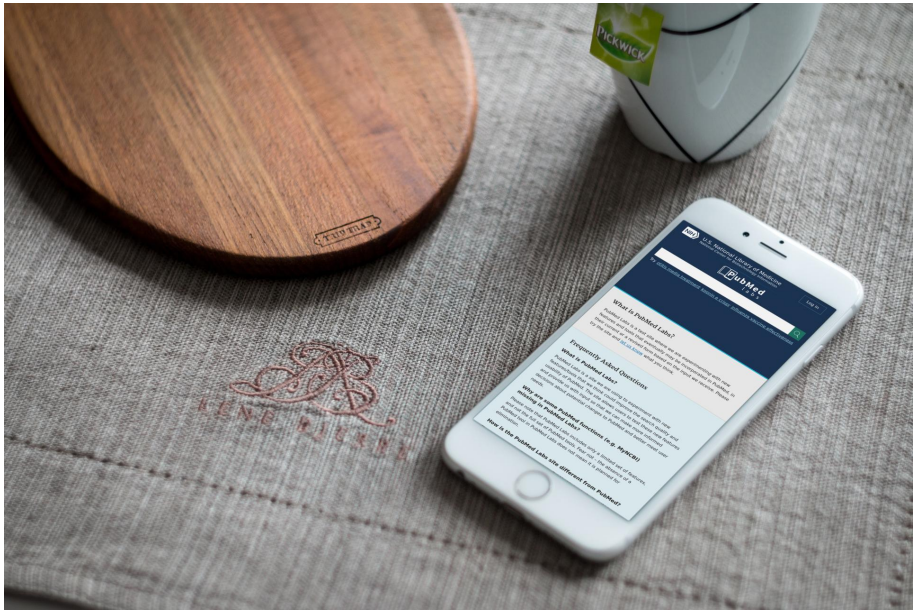






CIVIC DIGITAL FELLOWSHIP

PUBMED LABS



CIVIC DIGITAL FELLOWSHIP

THANK YOU

- Thank you to everyone at NCBI/NLM/NIH for their support!

CIVIC DIGITAL FELLOWSHIP