Spfy: bringing real-time, big data batch analyses of E. coli to SuperPhy

using containerization and graph-based data storage

Supervisor: Chad Laing
Student: Kevin Le

Kevin Le

Currently:

- BASc Software Engineering (uOttawa)
  - Focus on applied math
- Co-op @NML Lethbridge (8-months)

Previously Completed:

- BSc Neuroscience (Dalhousie)
  - Research: transgenic mouse models of Alzheimer’s, molecular neurosci.
  - Cryptography & network security

Software Background:

- Largely Python
- Linux
- Virtualization

Career Goals:

- Big-data companies / grad. studies in comp. sci

Bioinformatics Co-ops @Lethbridge Winter ‘17

Why work on *E. coli*?

- Within Canada it is estimated that over 4,000 hospitalizations and 100 deaths per year occur due to bacterial infections
- E. coli represents 33% of bacterial outbreaks from produce
- Incidence of ~30.3/100,000 people

Why work on *E. coli*?

Common Symptoms:

- Stomach cramps
- Diarrhea
- Vomiting

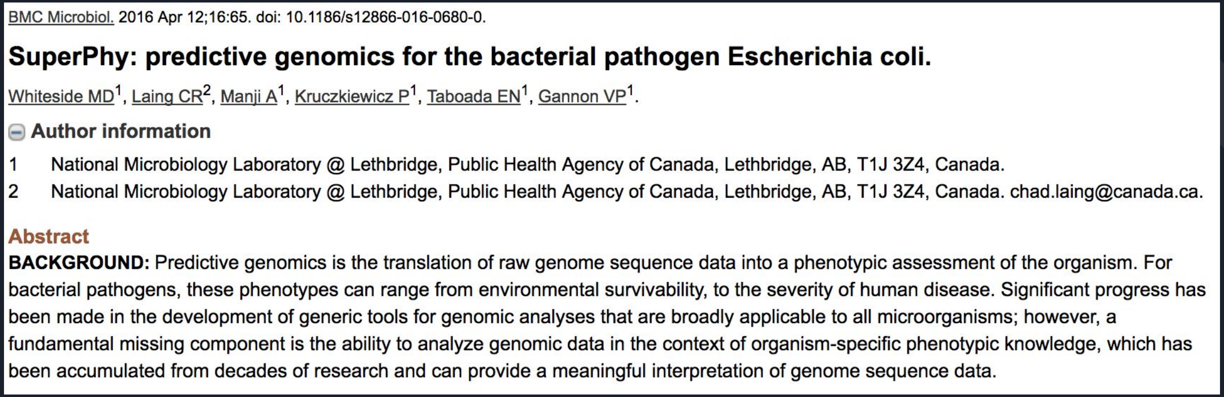
Complications:

- Hemolytic uremic syndrome (HUS)
- Can lead to kidney failure

Why work on *E. coli*?

- Traditional monitoring efforts used wet-lab methods for subtyping

WGS-Based Predictions



BMC Microbiol. 2016 Apr 12;16:65. doi: 10.1186/s12866-016-0680-0.

SuperPhy: predictive genomics for the bacterial pathogen Escherichia coli.

Whiteside MD[1], Laing CR[2], Manji A[1], Kruczkiewicz P[1], Taboada EN[1], Gannon VP[1].

⊖ Author information

1 National Microbiology Laboratory @ Lethbridge, Public Health Agency of Canada, Lethbridge, AB, T1J 3Z4, Canada.

2 National Microbiology Laboratory @ Lethbridge, Public Health Agency of Canada, Lethbridge, AB, T1J 3Z4, Canada. chad.laing@canada.ca.

Abstract

BACKGROUND: Predictive genomics is the translation of raw genome sequence data into a phenotypic assessment of the organism. For bacterial pathogens, these phenotypes can range from environmental survivability, to the severity of human disease. Significant progress has been made in the development of generic tools for genomic analyses that are broadly applicable to all microorganisms; however, a fundamental missing component is the ability to analyze genomic data in the context of organism-specific phenotypic knowledge, which has been accumulated from decades of research and can provide a meaningful interpretation of genome sequence data.

The Previous Version

What Worked:

- Pre-computed, predictive genomic analyses for Shiga toxin subtype, AMR genes, virulence factors
- Presence / absence of genomic regions and single nucleotide variants, for bacterial sub-groups

What Didn’t:

- No batch upload and get results in real-time
- Large scale storage and retrieval of results in real-time
- Easy deployment and replication

Goal: BIG data

Why big data?

Spfy replicates *E.coli* related functions of traditional reference labs

- Performs the same tasks except using whole-genome sequencing (WGS) data

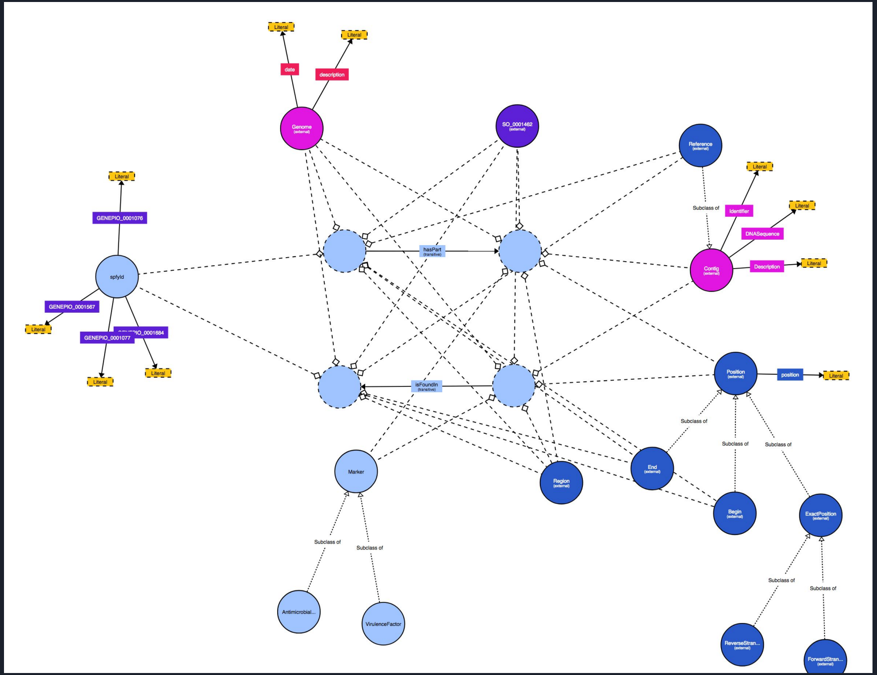
Spfy is also *online* & *in-silico*, incorporating species-specific knowledge for:

- Population-wide analysis
- Historical context
- Automatically linking new results

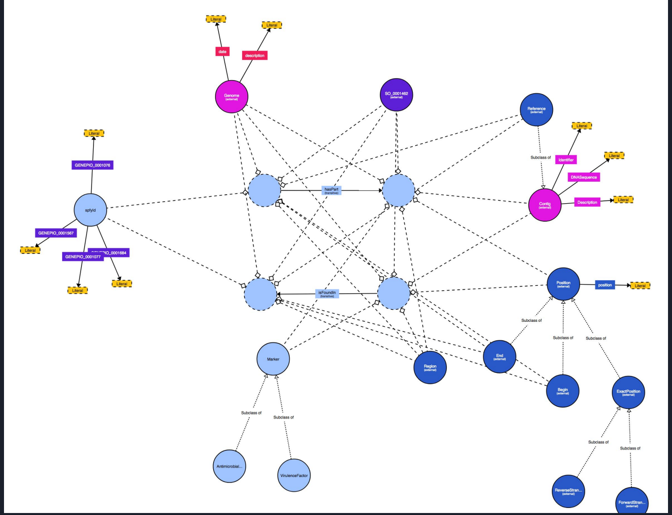
Main Approach

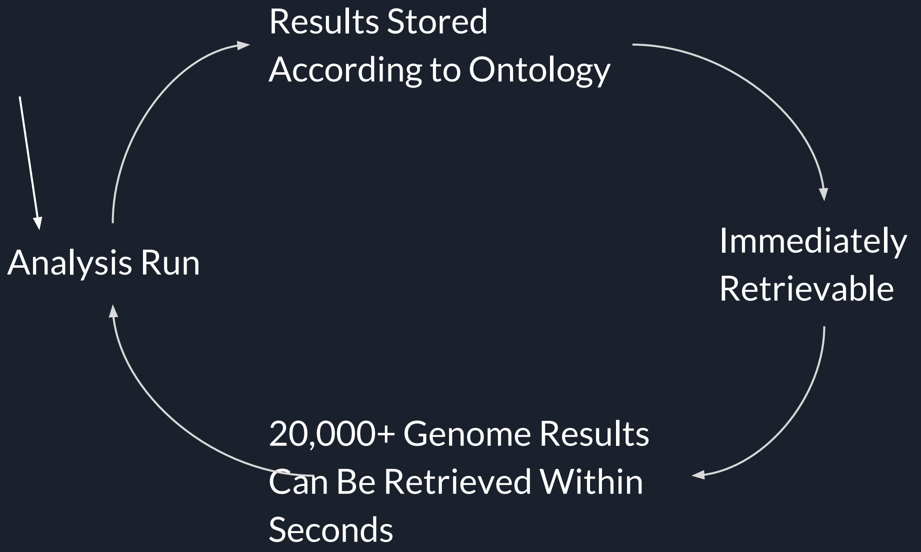
- Graph Database
- Parallelization

The Graph Database is the Core



The Graph Database is the Core





Results Stored According to Ontology

Immediately Retrievable

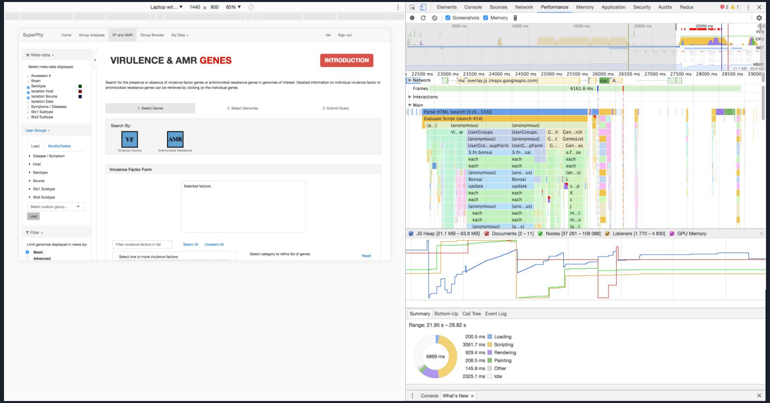
20,000+ Genome Results Can Be Retrieved Within Seconds

Analysis Run

The Web Platform

How do we get to results?

Flask
web development,
one drop at a time



Laptop wit... ▼ 1440 × 900 65% ▼

SuperPhy Home Group Analyses VF and AMR Group Browse My Data ▾ kle Sign out

Meta-data ▾

Select meta-data displayed:

- □ Accession #
- □ Strain
- ☑ Serotype
- ☑ Isolation Host
- ☑ Isolation Source
- □ Isolation Date
- □ Symptoms / Diseases
- □ Stx1 Subtype
- □ Stx2 Subtype

VIRULENCE & AMR GENES

INTRODUCTION

Search for the presence or absence of virulence factor genes or antimicrobial resistance genes in genomes of interest. Detailed information on individual virulence factor or antimicrobial resistance genes can be retrieved by clicking on the individual genes.

1. Select Genes 2. Select Genomes 3. Submit Query

Search By: VF Virulence Factors AMR Antimicrobial Resistance

Virulence Factor Form

Selected factors:

Elements Console Sources Network Performance Memory Application Security Audits Redux

☑ Screenshots ☑ Memory

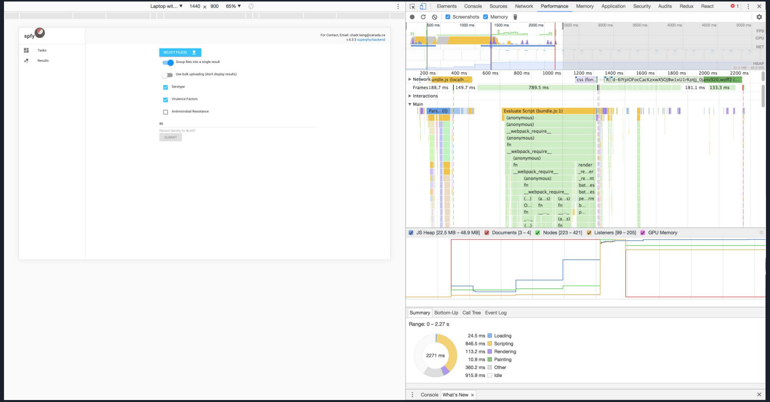
☑ JS Heap [21.1 MB – 63.8 MB] ☑ Documents [2 – 11] ☑ Nodes [37 261 – 109 088] ☑ Listeners [1 770 – 4 830] ☑ GPU Memory

Summary Bottom-Up Call Tree Event Log

Range: 21.95 s – 28.82 s

6869 ms

200.5 ms	Loading
3061.7 ms	Scripting
929.4 ms	Rendering
206.5 ms	Painting
145.9 ms	Other
2325.1 ms	Idle



Laptop wit... ▼ 1440 × 900 65% ▼

spfy

Tasks
Results

For Contact, Email: chadr.laing@canada.ca
v.4.3.3 superphy/backend

SELECT FILE(S)

Group files into a single result

Use bulk uploading (don't display results)

- ☑ Serotype
- ☑ Virulence Factors
- □ Antimicrobial Resistance

90
Percent Identity for BLAST

SUBMIT

Elements Console Sources Network Performance Memory Application Security Audits Redux React

☑ Screenshots ☑ Memory

☑ JS Heap [22.5 MB – 48.9 MB] ☑ Documents [3 – 4] ☑ Nodes [223 – 421] ☑ Listeners [99 – 205] ☑ GPU Memory

Summary Bottom-Up Call Tree Event Log

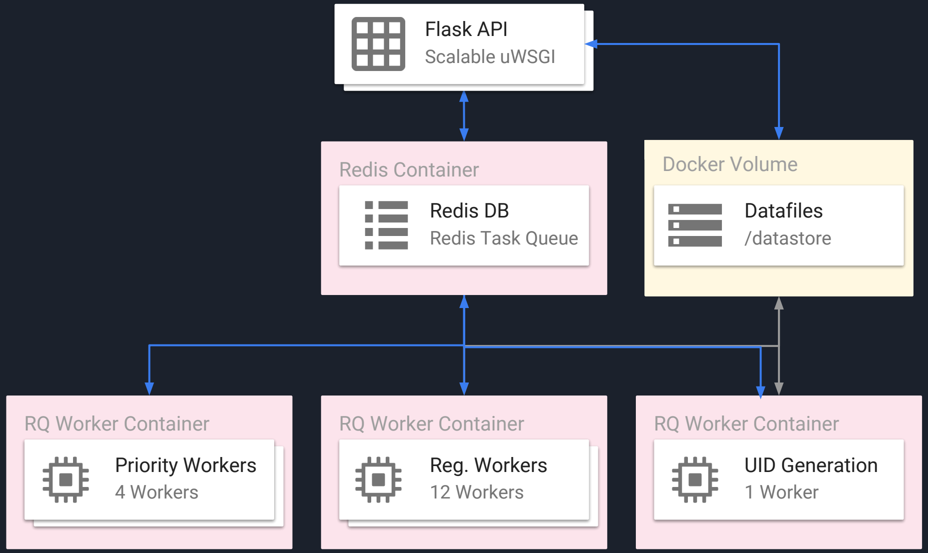
Range: 0 – 2.27 s

2271 ms

24.5 ms	Loading
846.5 ms	Scripting
113.2 ms	Rendering
10.9 ms	Painting
360.2 ms	Other
915.9 ms	Idle

Major Change:

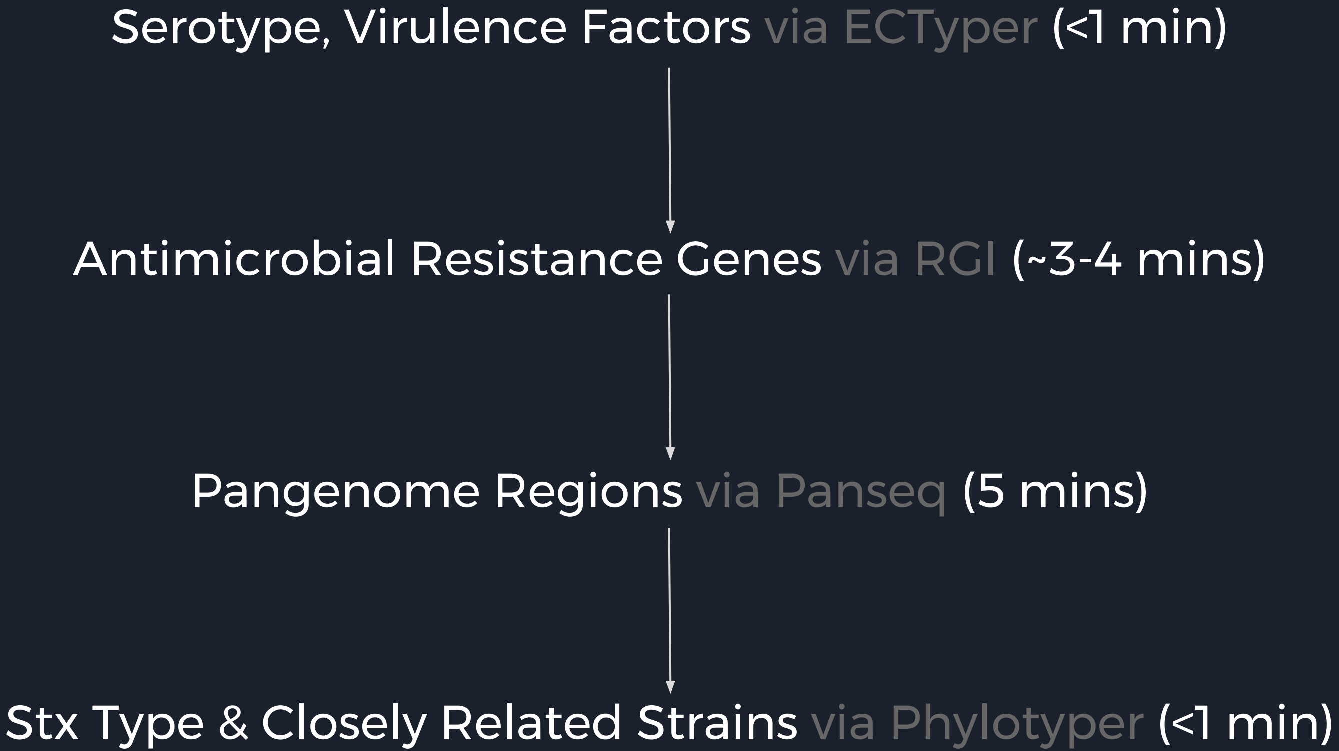
task queues



Dozen+ Parallel Task Workers

What can we do with this?

given a new whole genome sequencing result…



Serotype, Virulence Factors via ECTyper (<1 min)

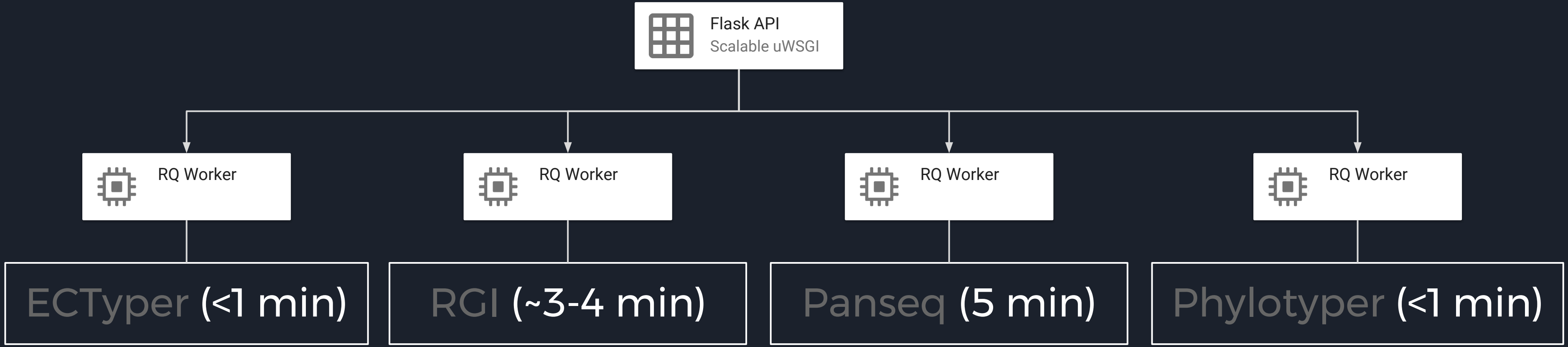
Antimicrobial Resistance Genes via RGI (~3-4 mins)

Pangenome Regions via Panseq (5 mins)

Stx Type & Closely Related Strains via Phylotyper (<1 min)

1 + 4 + 5 + 1 = ~11 min

Recall: Parallel Task Queues



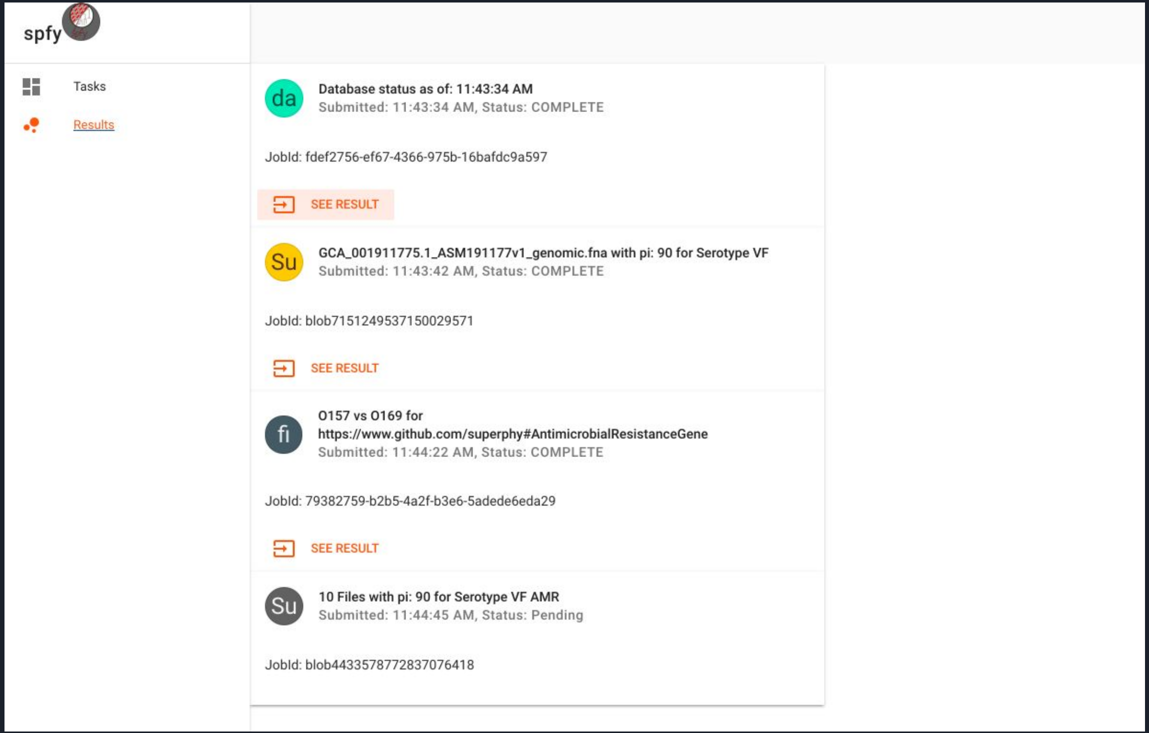
Not affected by the size of the database

Group Comparisons

Ask: what Pangenome Regions are over-represented among O157 strains over O53 strains?

Simplicity:

Google's Material Design



da

Database status as of: 11:43:34 AM

Submitted: 11:43:34 AM, Status: COMPLETE

JobId: fdef2756-ef67-4366-975b-16bafdc9a597

Su

GCA_001911775.1_ASM191177v1_genomic.fna with pi: 90 for Serotype VF

Submitted: 11:43:42 AM, Status: COMPLETE

JobId: blob7151249537150029571

fi

O157 vs O169 for https://www.github.com/superphy#AntimicrobialResistanceGene

Submitted: 11:44:22 AM, Status: COMPLETE

JobId: 79382759-b2b5-4a2f-b3e6-5adede6eda29

Su

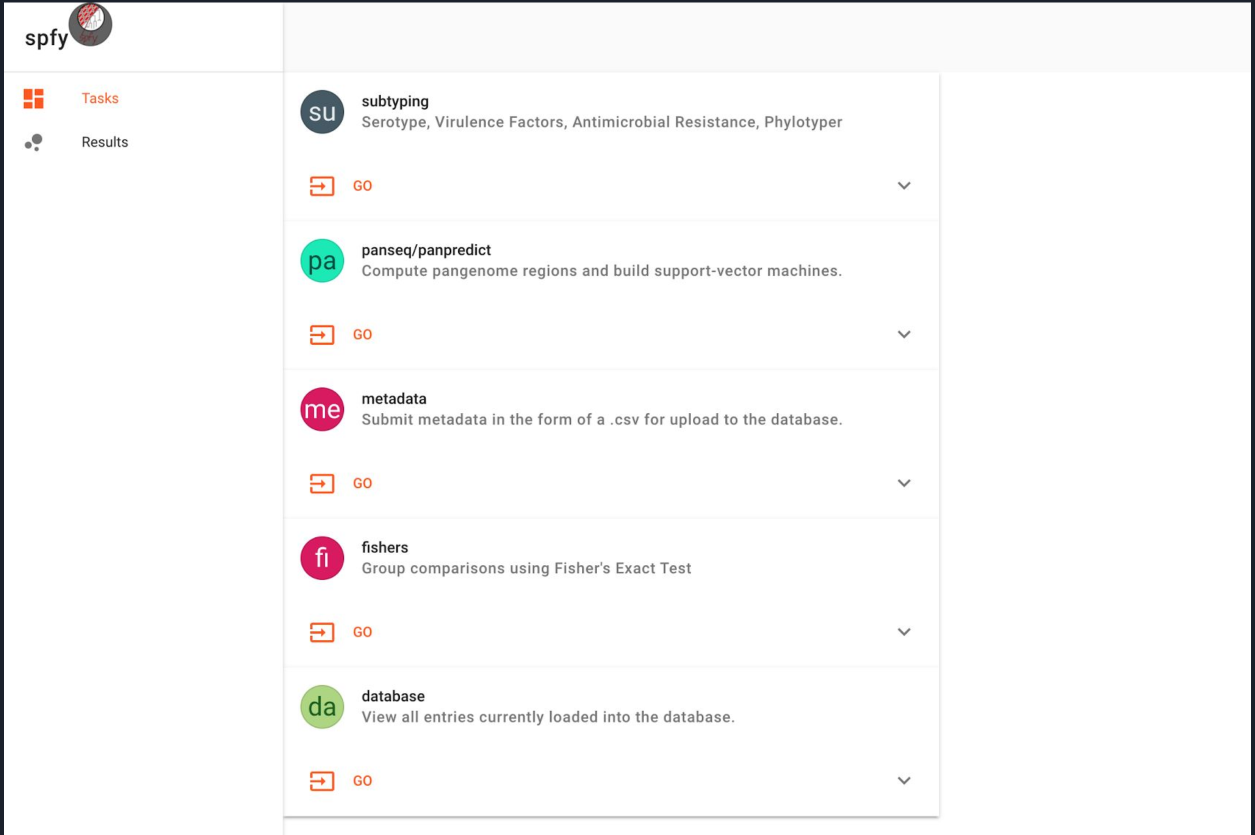
10 Files with pi: 90 for Serotype VF AMR

Submitted: 11:44:45 AM, Status: Pending

JobId: blob4433578772837076418

Card Based

Analysis Modules



su **subtyping**
Serotype, Virulence Factors, Antimicrobial Resistance, Phylotyper

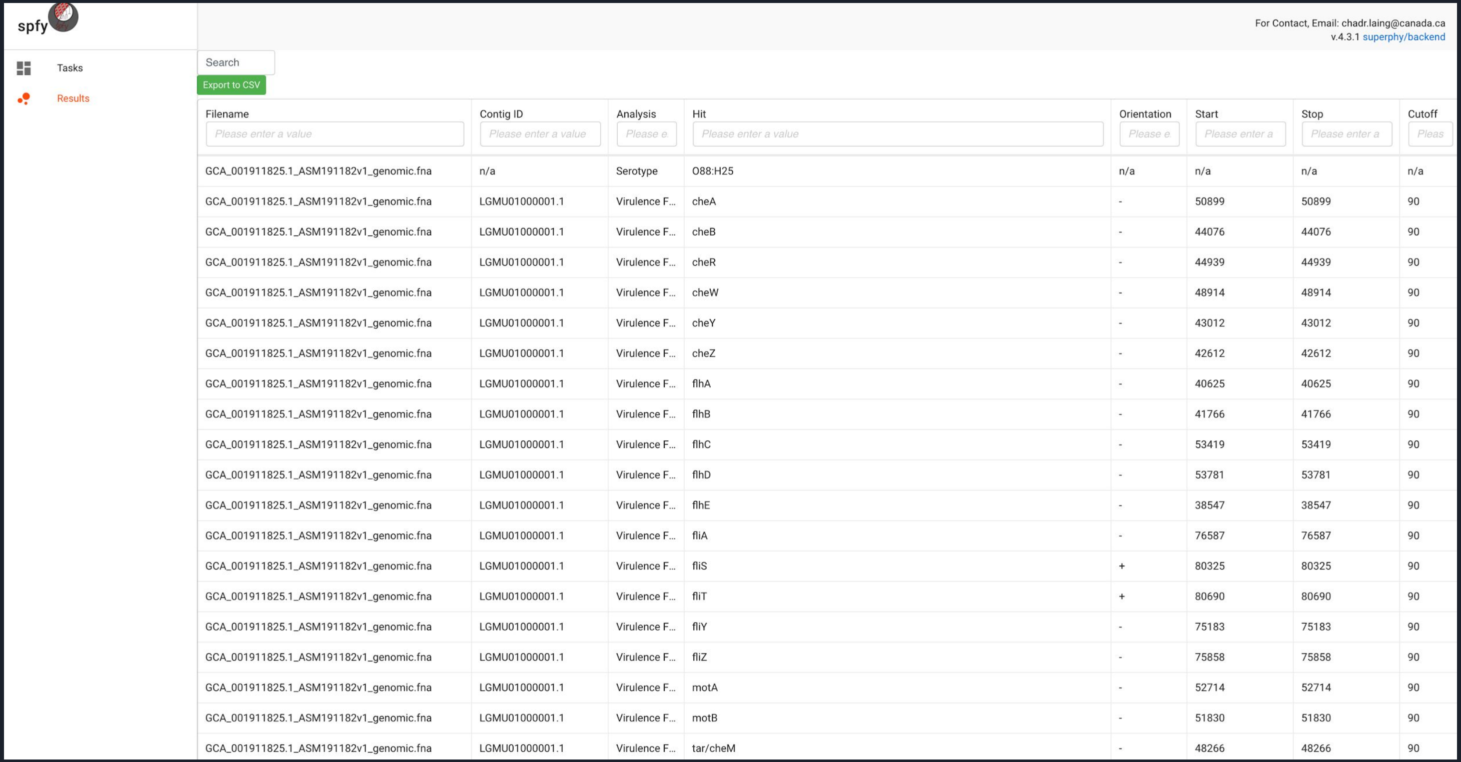
pa **panseq/panpredict**
Compute pangenome regions and build support-vector machines.

me **metadata**
Submit metadata in the form of a .csv for upload to the database.

fi **fishers**
Group comparisons using Fisher's Exact Test

da **database**
View all entries currently loaded into the database.

Familiar Tabular Results for Download

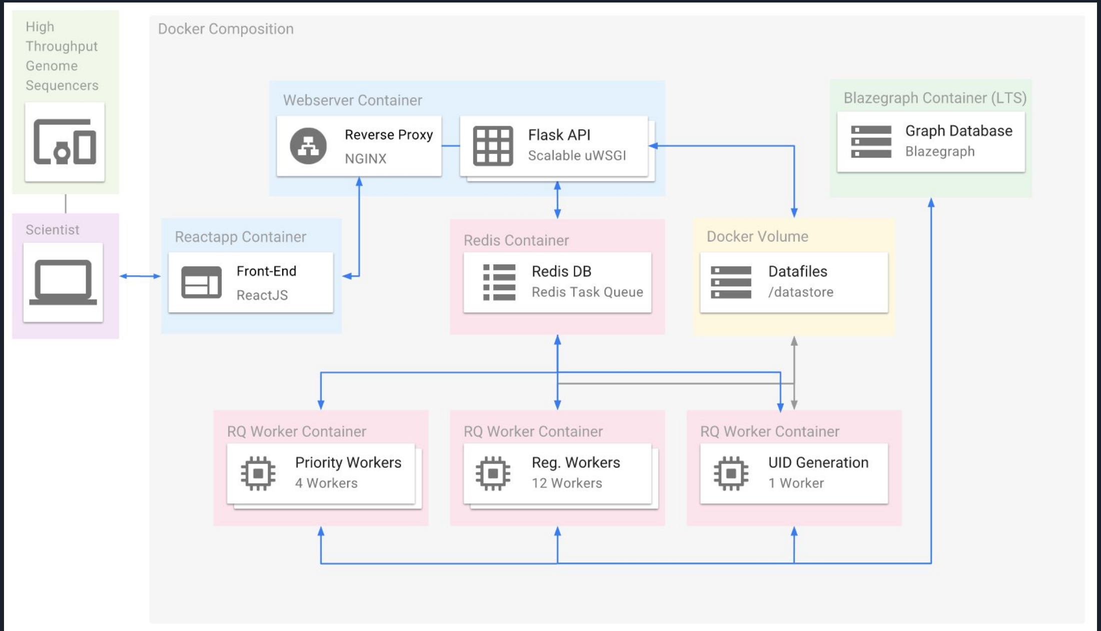


For Contact, Email: chadr.laing@canada.ca
v.4.3.1 superphy/backend

Filename	Contig ID	Analysis	Hit	Orientation	Start	Stop	Cutoff
GCA_001911825.1_ASM191182v1_genomic.fna	n/a	Serotype	O88:H25	n/a	n/a	n/a	n/a
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheA	-	50899	50899	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheB	-	44076	44076	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheR	-	44939	44939	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheW	-	48914	48914	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheY	-	43012	43012	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheZ	-	42612	42612	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhA	-	40625	40625	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhB	-	41766	41766	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhC	-	53419	53419	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhD	-	53781	53781	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhE	-	38547	38547	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliA	-	76587	76587	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliS	+	80325	80325	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliT	+	80690	80690	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliY	-	75183	75183	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliZ	-	75858	75858	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	motA	-	52714	52714	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	motB	-	51830	51830	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	tar/cheM	-	48266	48266	90

Orchestration:

Docker, Docker-Compose



Platform Architecture

Used Spfy to compute & store results for ~20,000 genomes

❖ One-Time Cost: Results are <u>permanently</u> stored for future comparisons - “Big Data”

What was done

What didn't work in Superphy:

- No batch upload and get results in real-time
- Large scale storage and retrieval of results in real-time
- Easy deployment and replication

How it was solved in Spfy:

- Modern website and use of task queues
- Graph database
- Docker

Where do we go from here?

- More analysis modules
- Data Visualization
- Machine Learning

Thanks!

E.coli team @ Lethbridge

- Chad Laing
- Matt Whiteside
- Vic Gannon

Campy team @ Lethbridge

- Eduardo Taboada
- Dillon Barker

Questions?