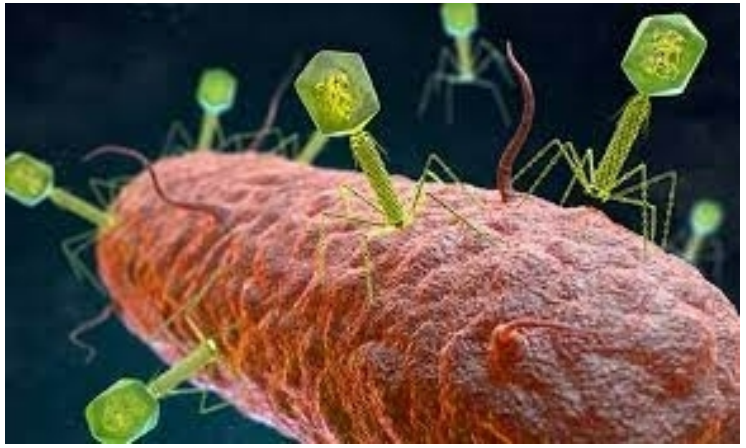
SAND2022-9558C

Sandia National Laboratories

Computational Discovery of Bacterial Immune Systems



Rohan Krishna, Bioinformatics Year-Round Intern

July 25, 2022, Mentorship by Catherine Mageeney

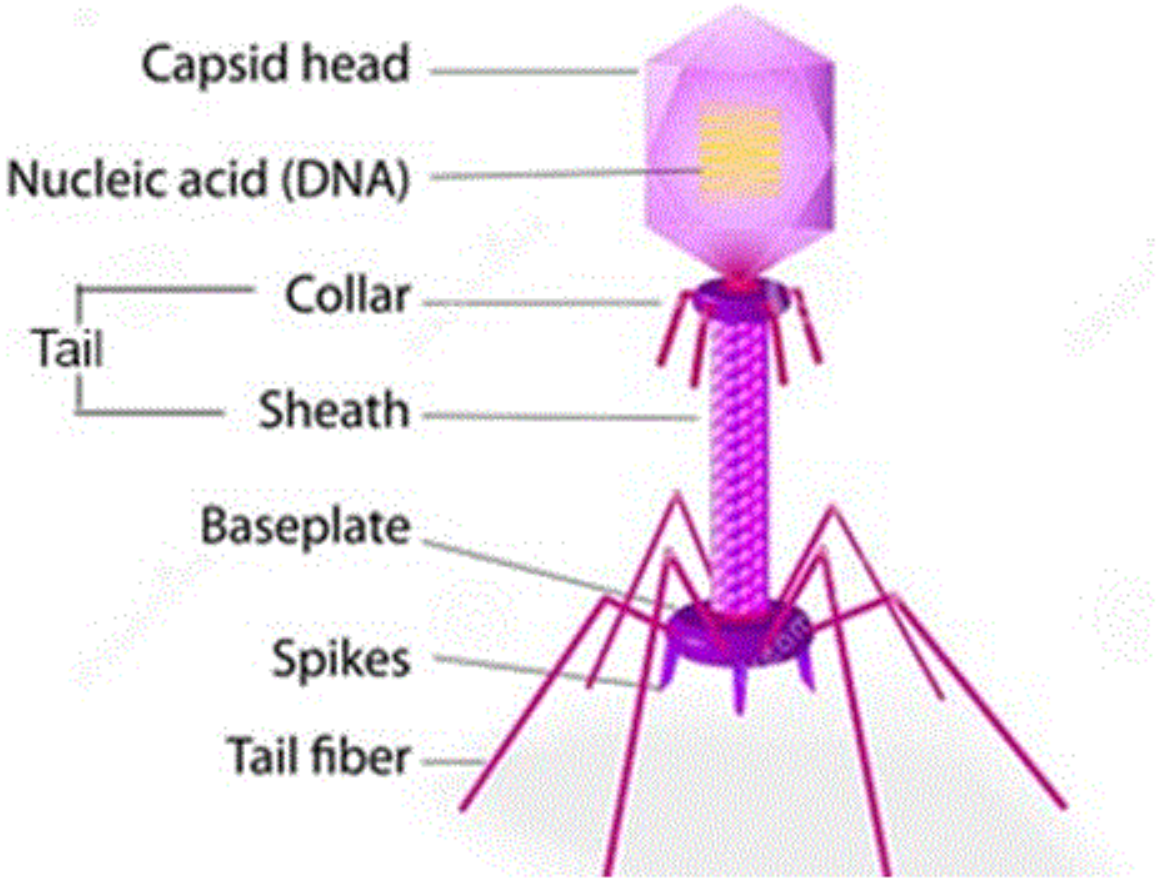
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Introduction

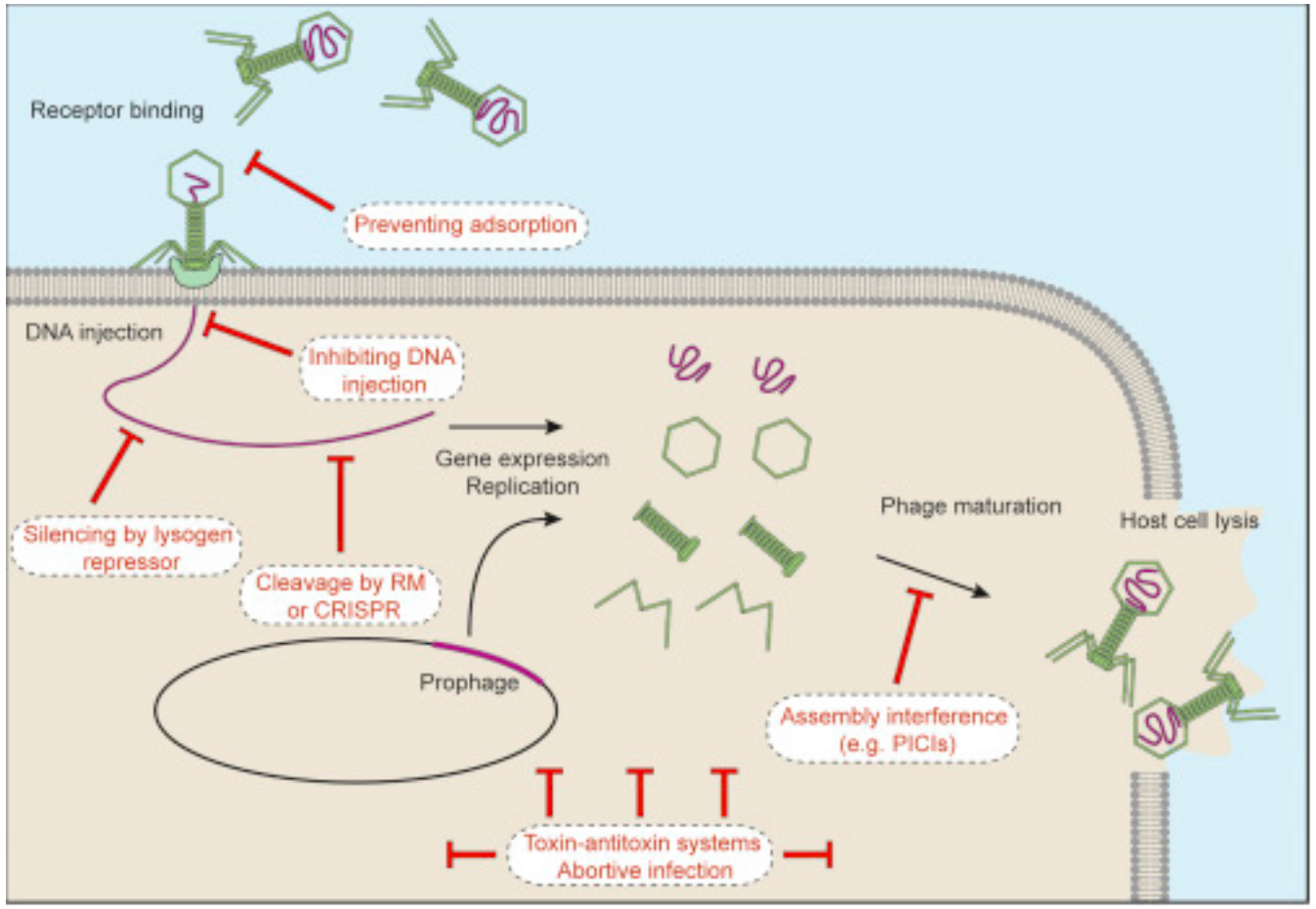
Introduction

- Bacteriophages are viruses for bacteria
  - Were first discovered in 1915 and scientists had found its potential to kill bacteria.
  - Only a small number of phages (primarily E.coli specific) were studied throughout the decades due to the comparative ease of antibiotics
- Applications:
  - Phage therapy
  - Genetic Exchange between bacteria
  - Manipulating microbiomes
- Recently, been a growing awareness in the amount of phages in bacteria dominated environments, but barely any research on how bacteria defend against them
  - With more knowledge on how bacteria combat phages, we are able to find potential treatments on bacterial infections and diseases through a natural approach using the idea of “predator prey” relationship.



Introduction (cont.)

- How does a bacteria combat against phages?
  - Receptor Proteins
    - Prevent binding/inhibit DNA injection
  - CRISPR/Restriction Enzymes
  - Assembly Interference
  - Toxin-antitoxin systems
  - Genomic Islands: part of a genome that codes for defense functions
  - Prophages: phage genome that exists as extrachromosomal plasmid within bacterial cell (defend against viral attacks)
- Overall goal is to map this bacterial defense genomic landscape.



Adapted from: Rostøl and Marraffini, 2019, Cell Host & Microbe

Methods

Workflow/Process

- Literature Search on Defense Mechanisms
- Compiling a list of known defense proteins, receptor proteins, proteins found from our search software.
- Creating a script that maps out location/size/type of all the proteins involved into an easy presentable visual.

Receptor Proteins
- a
- b

Additional Defense Proteins
- a
- b

TIGER software
- a
- b

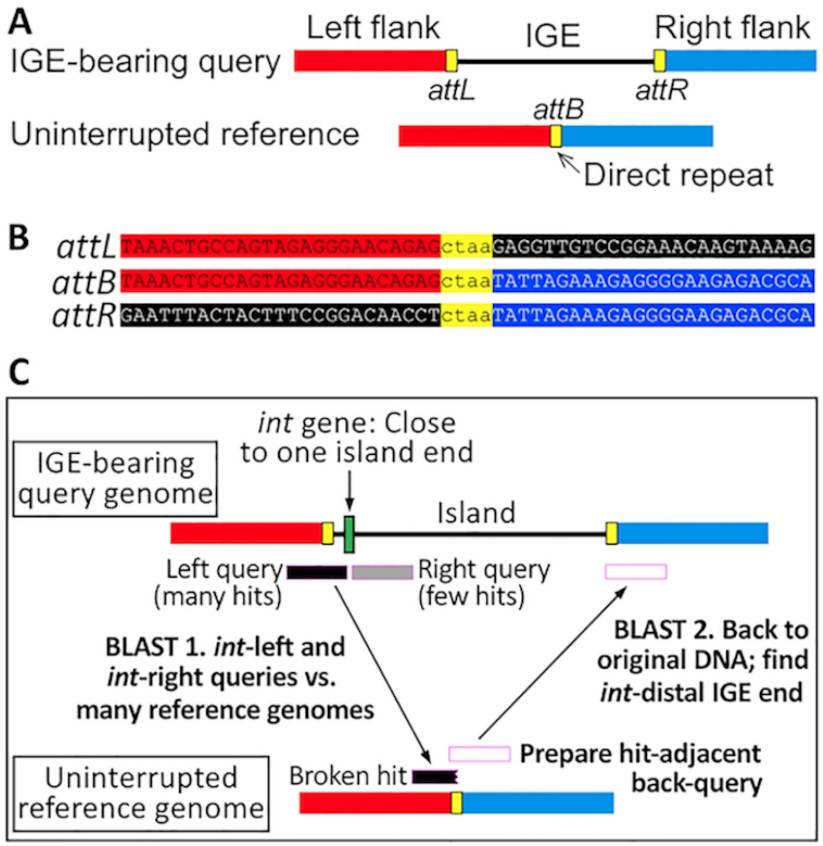
Islander software
- a
- b

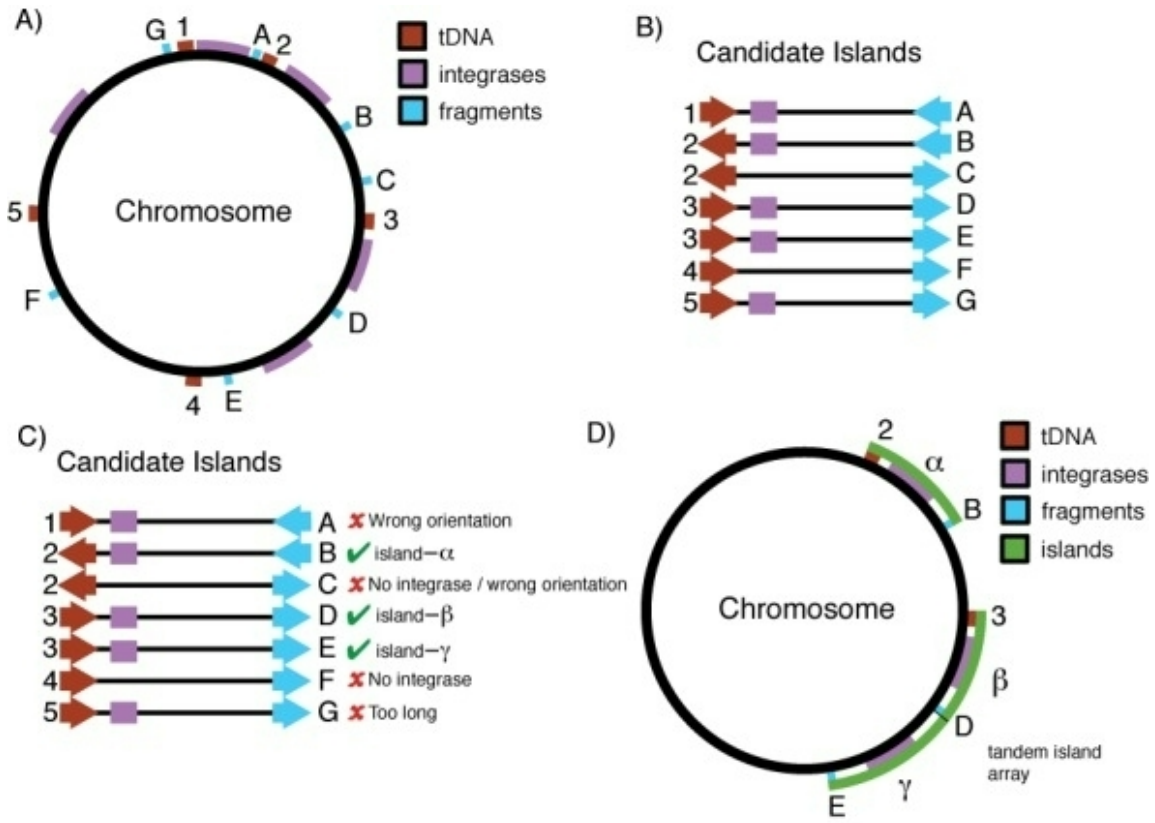
CRISPR CASCollect
- a
- b

PADLOC
- a
- b

Name	Software	Type	Direction	L	R	Info

Searching for Genomic Islands & Prophages

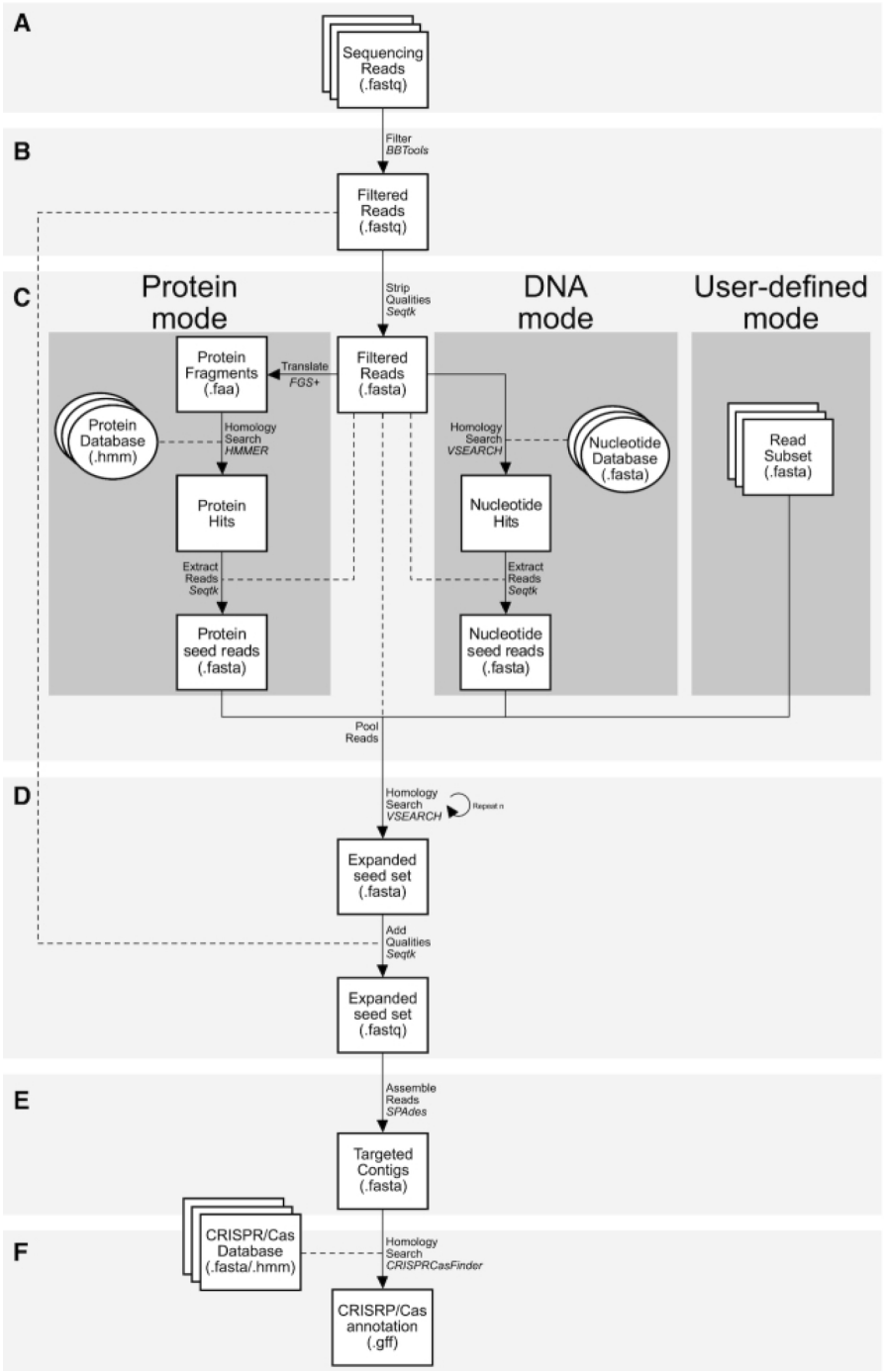




TIGER software mechanism; Mageeney CM, Lau BY, Wagner JM, Hudson CM, Schoeniger JS, Krishnakumar R, Williams KP. New candidates for regulated gene integrity revealed through precise mapping of integrative genetic elements. Nucleic Acids Res. 2020 May 7;48(8):4052-4065. doi: 10.1093/nar/gkaa156. PMID: 32182341; PMCID: PMC7192596.

Islander software mechanism; Hudson CM, Lau BY, Williams KP. Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. Nucleic Acids Res. 2015 Jan;43(Database issue):D48-53. doi: 10.1093/nar/gku1072. Epub 2014 Nov 5. PMID: 25378302; PMCID: PMC4383910.

Searching for CRISPR/Cas9 Defense



CRISPR CasCollect workflow;

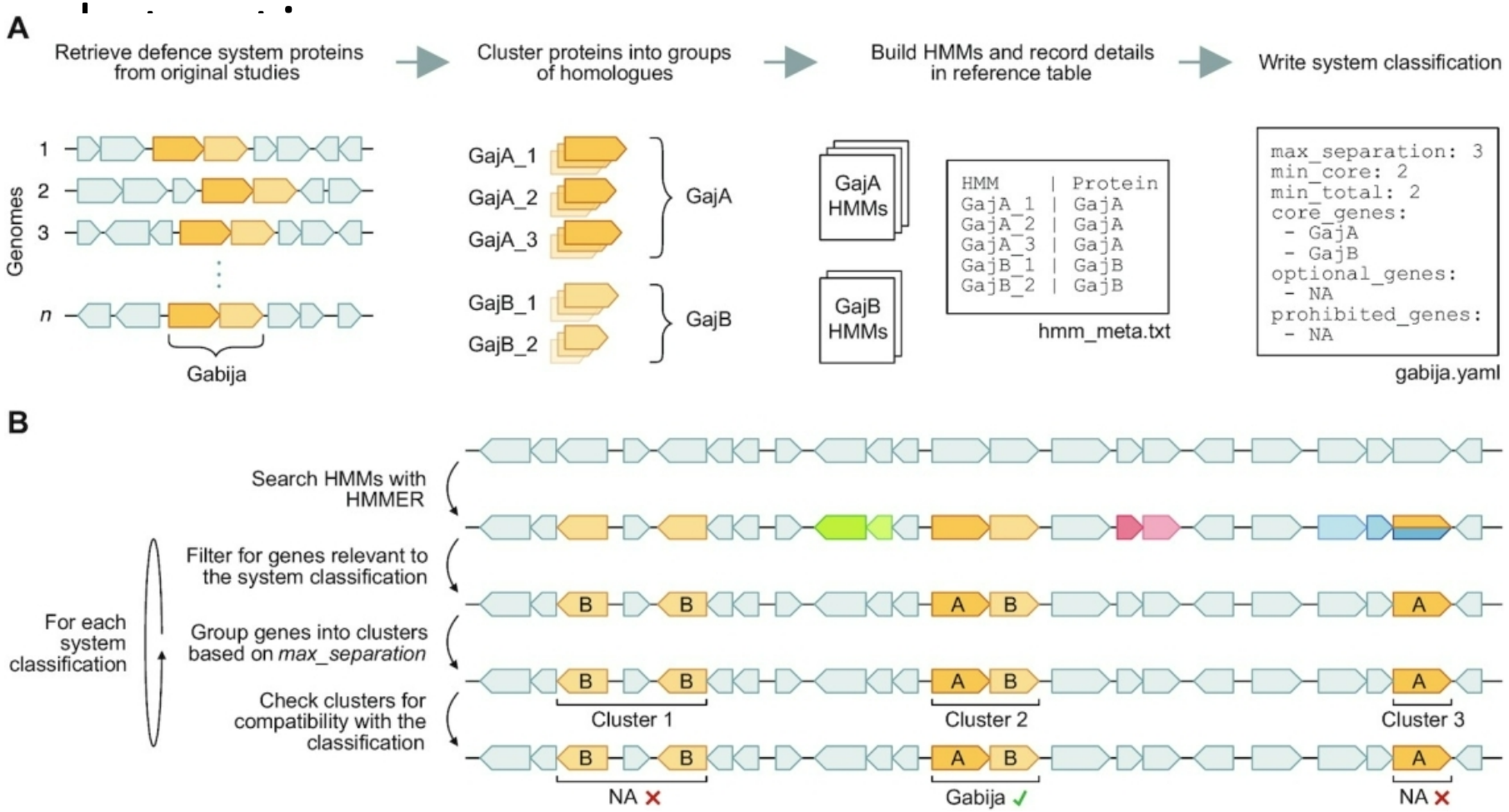
Podlevsky JD, Hudson CM, Timlin JA, Williams KP. CasCollect: targeted assembly of CRISPR-associated operons from high-throughput sequencing data. NAR Genom Bioinform. 2020 Sep 3;2(3):lqaa063. doi: 10.1093/nargab/lqaa063. PMID: 33575613; PMCID: PMC7671303.

What is an HMM?

- Hidden Markov Models – a consensus model for proteins
- Predicts protein domains and locations on an amino acid sequence
- Can also find the protein family which this new protein sequence belongs to.

```
HMMER3/f [3.1b2 | February 2015]
NAME  abiU
LENG  566
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Fri Feb 25 16:01:43 2022
NSEQ  97
EFFN  0.861420
CKSUM 401435747
STATS LOCAL MSV      -11.9794  0.69746
STATS LOCAL VITERBI  -12.8539  0.69746
STATS LOCAL FORWARD   -6.1972  0.69746
HMM          A        C        D        E        F        G        H        I        K
   V        W        Y
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.63128  4.53121  2.89439  2.60613  3.19734  3.11781  3.78420  2.70982  2.60217
2.64969  4.77198  3.26226
          2.68619  4.42230  2.77524  2.73128  3.46358  2.40517  3.72499  3.29359  2.67745
2.98523  4.58482  3.61508
          0.68844  1.37814  1.40408  0.31977  1.29578  0.00000        *
      1   2.25343  4.45732  3.05154  2.57148  3.71540  3.24955  3.66966  3.09709  2.54499
2.72287  5.07469  3.79620      4 a - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741
2.98518  4.58477  3.61503
          0.03245  3.84003  4.56237  0.61958  0.77255  0.72914  0.65841
      2   2.57778  4.78009  2.85082  2.43011  4.08751  3.36316  3.60039  3.29681  1.80042
3.15598  5.30133  3.99445      5 k - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741
2.98518  4.58477  3.61503
          0.03245  3.84003  4.56237  0.61958  0.77255  0.44648  1.02131
      3   2.58357  4.55611  3.12248  2.59042  3.67105  3.40269  3.68014  3.07413  2.22336
2.79845  5.09505  3.81406      6 k - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741
2.98518  4.58477  3.61503
          0.02636  4.04487  4.76722  0.61958  0.77255  0.48419  0.95762
_
```

Searching using HMM based homologue



PADLOC software mechanism; Leighton J Payne, Thomas C Todeschini, Yi Wu, Benjamin J Perry, Clive W Ronson, Peter C Fineran, Franklin L Nobrega, Simon A Jackson, Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types, *Nucleic Acids Research*, Volume 49, Issue 19, 8 November 2021, Pages 10868–10878, https://doi.org/10.1093/nar/gkab883

Additional Defense Systems/ Receptors

Proteins found from Lit.

Defense Proteins
- a
- b
- c
- d
- e
- f
- g

Known HMMs

```
HMMER3
LENG
ALF
RF
MM
CONS
DATE
HMM          A    B    C    D    E    F    G    H
                -    -    -    -    -    -    -    -
```

Known GFF File

Name	Software	Type	Dir	L	R	Info
P1	prodigal	CDS.	+	1	10	ID=12,pfam = a
P1	prodigal	CDS.	+	11	50	ID=12,pfam = b
P1	prodigal	CDS.	+	51	70	ID=12,pfam = y
P1	prodigal	CDS.	+	71	90	ID=12,pfam = d
P1	prodigal	CDS.	+	91	100	ID=12,pfam = z
P1	prodigal	CDS.	+	101	110	ID=12,pfam = f
P1	prodigal	CDS.	+	111	140	ID=12,pfam = c
P1	prodigal	CDS.	+	141	160	ID=12,pfam = g
P1	prodigal	CDS.	+	161	180	ID=12,pfam = x

Proteins found from Lit.

Receptor Proteins
- a
- b
- c
- d
- e
- f
- g

Create HMMS for defense proteins that are not made yet.

Search through Pfam database to see any matching proteins and output a domtble with all hits.

P1	.	.	+	1	10	protein = a
P1	.	.	+	11	50	protein = b
P1	.	.	+	71	90	protein = d
P1	.	.	+	101	110	protein = f
P1	.	.	+	111	140	protein = c
P1	.	.	+	141	160	protein = g

New list in GFF format with matching proteins from known file and lit.

Results

Down-selection of Bacteria for Testing

- This prediction software is used to inform wet lab experiments
- Criteria for wet lab experimentation
  - Relevance to DOE programs
  - Few phages currently identified
  - Few mechanisms of resistance predicted

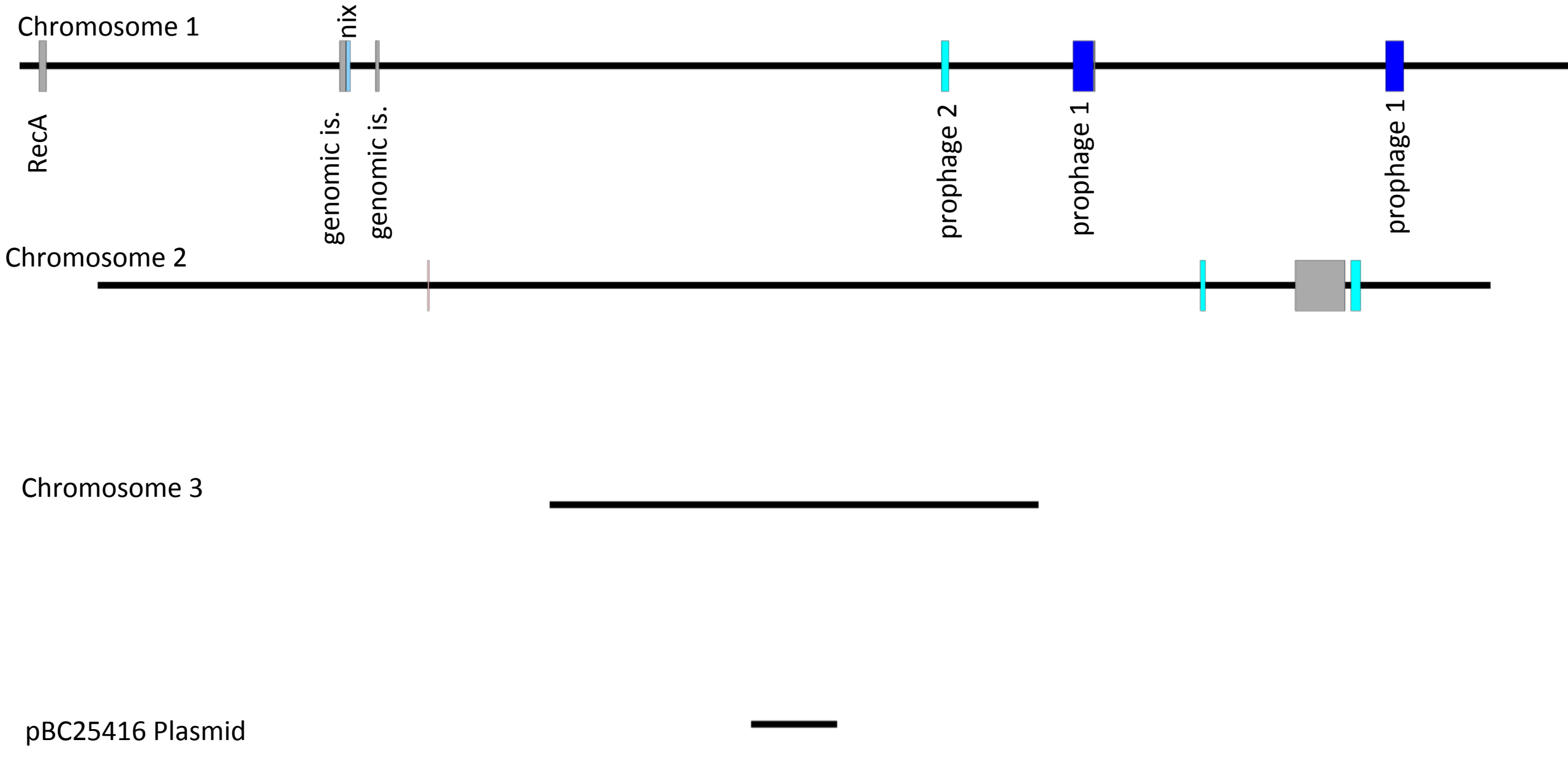
- *Burkholderia cepacia*
  - Soil-dwelling and human opportunistic pathogen
  - 42 phages isolated, sandia has a small collection of these
  - Previous predictions included prophages, CRISPR systems, restriction modification systems.

Example Islander Software Output GFF File for Bce248

```
CP012981.1	island_finder	genomic_island	782973	799865	9.34643923420159e-09	+	.	ID=Bce248.17K;trna=trna.5;int_site=35;int_site_type=a_site;trna_dupe=+1;tRN
;tRNA_aa=K;A_site=35;J_site=65;questionable=IPD:-2.10;qStart=28;qEnd=76;hitStart=799859;hitEnd=799907;percent_id=100.000;bit_score=89.7;size=16893;group=3prime.trna.5[1i][2i];segm
ta_int=153;foreign=3.0326874545507;housekeep=7.14424437847642;hypoth=0.573764638674664;delta_GC=-0.08325;dinuc=0.13918125;intList=Y-Int.2,;intCoords=798331..799713,;pot_ints=Y-In
;project=genome;division=Bacteria;phylum=Proteobacteria;order=Betaproteobacteria;class=Burkholderiales;family=Burkholderiaceae;genus=Burkholderia;species=Burkholderia cepacia;org=
ode=11;
CP012981.1	island_finder	genomic_island	799866	810395	1.76589877054795e-11	+	.	ID=Bce248.11K;trna=trna.5;int_site=35;int_site_type=a_site;trna_dupe=+1;tRN
;tRNA_aa=K;A_site=35;J_site=65;questionable=IPD:-2.10;qStart=32;qEnd=76;hitStart=810393;hitEnd=810437;percent_id=100.000;bit_score=82.4;size=10530;group=3prime.trna.5[1i][2i];segm
ta_int=205;foreign=10.7635845420156;housekeep=12.1056185029081;hypoth=0.37932019423022;delta_GC=-0.10845;dinuc=0.13811875;intList=Y-Int.3,;intCoords=800071..801153,;pot_ints=Y-Int
project=genome;division=Bacteria;phylum=Proteobacteria;order=Betaproteobacteria;class=Burkholderiales;family=Burkholderiaceae;genus=Burkholderia;species=Burkholderia cepacia;org=B
de=11;
CP012981.1	island_finder	genomic_island	1045395	1048308	1.43991113000688e-10	+	.	ID=Bce248.3R;trna=trna.8;int_site=35;int_site_type=a_site;trna_dupe=+1;tRNA
g;tRNA_aa=R;A_site=35;J_site=65;questionable=;qStart=32;qEnd=76;hitStart=1048306;hitEnd=1048350;percent_id=100.000;bit_score=82.4;size=2914;group=3prime.trna.8,B2[1i],3;segment=1;
nt=317;foreign=1.8160461528638 6;housekeep=6.08507179658309;hypoth=0.157097972007997;delta_GC=-0.1566;dinuc=0.1576375;intList=Y-Int.4,;intCoords=1045712..1045978,;pot_ints=Y-Int.4;
=Bacteria;phylum=Proteobacteria;order=Betaproteobacteria;class=Burkholderiales;family=Burkholderiaceae;genus=Burkholderia;species=Burkholderia cepacia;org=Burkholderia cepacia ATC
CP012981.1	island_finder	genomic_island	2256567	2265008	3.35201760357244e-10	+	.	ID=Bce248.8R;trna=trna.36;int_site=36;int_site_type=a_site;trna_dupe=+1;tRN
rg;tRNA_aa=R;A_site=36;J_site=66;questionable=;qStart=29;qEnd=77;hitStart=2265002;hitEnd=2265050;percent_id=97.959;bit_score=84.2;size=8442;group=3prime.trna.36[1i];segment=1;orig
83;foreign=3.70837748788633;housekeep=16.0593437179831;hypoth=0.573764638674664;delta_GC=-0.0922000000000001;dinuc=0.10469375;intList=Y-Int.5,;intCoords=2263453..2264826,;pot_ints
7;project=genome;division=Bacteria;phylum=Proteobacteria;order=Betaproteobacteria;class=Burkholderiales;family=Burkholderiaceae;genus=Burkholderia;species=Burkholderia cepacia;org
code=11;
CP012981.1	island_finder	genomic_island	2472095	2629574	0.581143152660045	+	.	ID=Bce248.157V;trna=trna.38;int_site=66;int_site_type=j_site;trna_dupe=+1;t
=Val;tRNA_aa=V;A_site=36;J_site=66;questionable=;qStart=47;qEnd=72;hitStart=2629556;hitEnd=2629580;percent_id=88.462;bit_score=30.1;size=157480;group=3prime.trna.38[1i][2i];segmen
ta_int=1283;foreign=0.376663387706586;housekeep=1.70539751692625;hypoth=0.244219184129209;delta_GC=-0.02825;dinuc=0.02209375;intList=Y-Int.6,;intCoords=2627213..2628292,;pot_ints=
-2674605;project=genome;division=Bacteria;phylum=Proteobacteria;order=Betaproteobacteria;class=Burkholderiales;family=Burkholderiaceae;genus=Burkholderia;species=Burkholderia cepa
594;gencode=11;
CP012981.1	island_finder	genomic_island	2629575	2674599	3.34308640129644e-06	+	.	ID=Bce248.45V;trna=trna.38;int_site=66;int_site_type=j_site;trna_dupe=+1;tR
Val;tRNA_aa=V;A_site=36;J_site=66;questionable=;qStart=47;qEnd=72;hitStart=2674581;hitEnd=2674605;percent_id=88.462;bit_score=30.1;size=45025;group=3prime.trna.38[1i][2i];segment=
_int=99;foreign=4.3057225526 7888;housekeep=10.4987035681059;hypoth=0.611643426553452;delta_GC=-0.0169;dinuc=0.0372562500000001;intList=Y-Int.7,;intCoords=2629674..2630699,;pot_int
81-2674605;project=genome;division=Bacteria;phylum=Proteobacteria;order=Betaproteobacteria;class=Burkholderiales;family=Burkholderiaceae;genus=Burkholderia;species=Burkholderia ce
83594;gencode=11;
```

Bce248 = Burkholderia cepacia UCB 717

Genomic Map of *Burkholderia cepacia* UCB 717



Future Plans/Conclusion

Conclusions

- Created a software that generates HMMs from protein sequence in a FASTA file and searches the Pfam sequence database for similar sequence matches, which helps identify newly discovered defense proteins.

- Created a script to generate a GFF file output of just receptor protein
  - Tested on *Burkholderia cepacia* UCB717

Future Plans

- Creating a wrapper script, combining all the proteins from our different softwares into a single output file
- Genomic map annotation of bacterial defense proteins.
- Large scale application (run software on hundreds of bacterial species rather than just 2-3)

THANK YOU