Search strategies for antimicrobial resistance associated genes

Data Science Workshop

Aug 7, 2018

Aram Avila-Herrera

LLNL-PRES-755360

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Antimicrobial resistance (AMR) numbers

- 2 million

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]
- 23,000

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]
- 23,000
  - deaths per year in US directly from AMR [1]

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]
- 23,000
  - deaths per year in US directly from AMR [1]
  - more from indirect complications [1]

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]
- 23,000
  - deaths per year in US directly from AMR [1]
  - more from indirect complications [1]
- 35 billion

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]
- 23,000
  - deaths per year in US directly from AMR [1]
  - more from indirect complications [1]
- 35 billion
  - dollars per year in costs to US households from AMR [2]

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

[2] Golkar *et al.*, *J Infect Dev Ctries* 2014

Antimicrobial resistance (AMR) numbers

- 2 million
  - AMR infections per year in US [1]
- 23,000
  - deaths per year in US directly from AMR [1]
  - more from indirect complications [1]
- 35 billion
  - dollars per year in costs to US households from AMR [2]
  - 20 billion per year in costs to US health care system [2]

[1] Antibiotic Resistance Threats in the United States 2013, *CDC*

[2] Golkar *et al.*, *J Infect Dev Ctries* 2014

What is antimicrobial resistance?

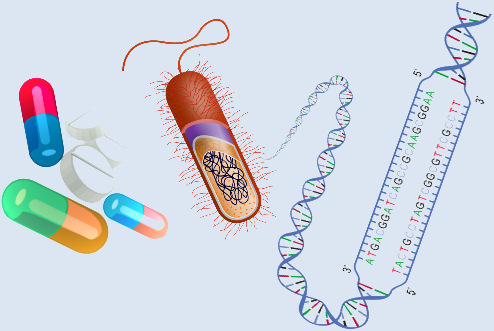
- When microbes such as bacteria are able to counteract antibiotics

What is antimicrobial resistance?

- When microbes such as bacteria are able to counteract antibiotics
- One mechanism:
  - **Genes** that encode drug-inactivating proteins are **acquired** or **evolved**

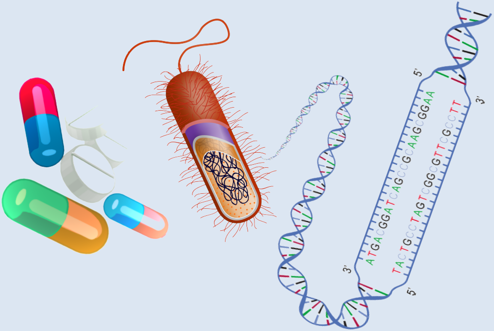
What is antimicrobial resistance?

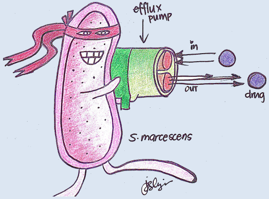
- When microbes such as bacteria are able to counteract antibiotics
- One mechanism:
  - **Genes** that encode drug-inactivating proteins are **acquired** or **evolved**



What is antimicrobial resistance?

- When microbes such as bacteria are able to counteract antibiotics
- One mechanism:
  - **Genes** that encode drug-inactivating proteins are **acquired** or **evolved**

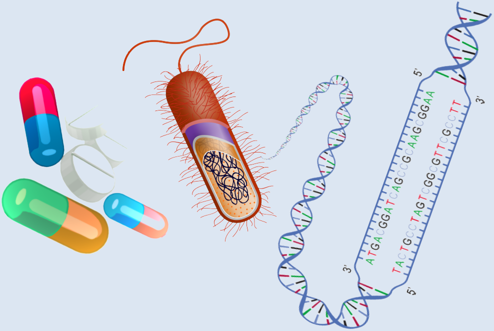


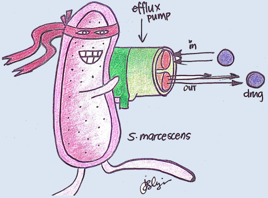


credit: Joseliya Embuscado

What is antimicrobial resistance?

- When microbes such as bacteria are able to counteract antibiotics
- One mechanism:
  - **Genes** that encode drug-inactivating proteins are **acquired** or **evolved**



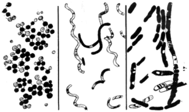


credit: Joseliya Embuscado

Detect these genes to:

- Prescribe the correct medicine, dosage
- Develop effective new drugs
- Learn about evolution

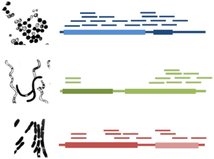
Gene detection: sequence and search



Microbial DNA is extracted from environment.

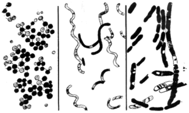
...**ACATATACG**...
...**CATATACGC**...
...**TAGACAT**...

DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.

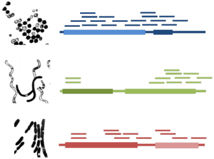
Gene detection: sequence and search



Microbial DNA is extracted from environment.

...ACATATACG...
...CATATACGC...
...TAGACAT...

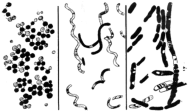
DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.

- String comparison with a few twists:

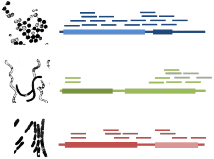
Gene detection: sequence and search



Microbial DNA is extracted from environment.

...ACATATACG...
...CATATACGC...
...TAGACAT...

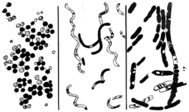
DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.

- String comparison with a few twists:
  - Millions of short reads per sample—ambiguity

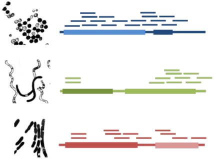
Gene detection: sequence and search



Microbial DNA is extracted from environment.

...ACATATACG...
...CATATACGC...
...TAGACAT...

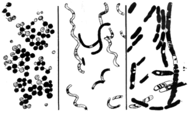
DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.

- String comparison with a few twists:
  - Millions of short reads per sample—ambiguity
  - Natural variation—mismatches not equally important

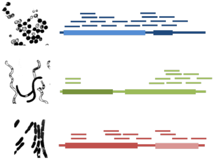
Gene detection: sequence and search



Microbial DNA is extracted from environment.

...**ACATATACG**...
...**CATATACGC**...
...**TAGACAT**...

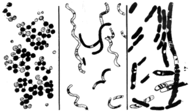
DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.

- String comparison with a few twists:
  - Millions of short reads per sample—ambiguity
  - Natural variation—mismatches not equally important

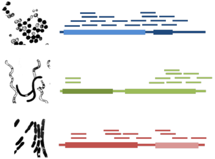
Gene detection: sequence and search



Microbial DNA is extracted from environment.

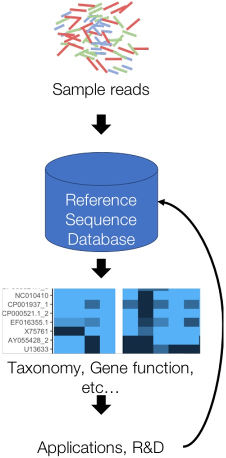
...**ACATATACG**...
...**CATATACGC**...
...**TAGACAT**...

DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.

- String comparison with a few twists:
  - Millions of short reads per sample—ambiguity
  - Natural variation—mismatches not equally important
- Fundamental step for R&D, applications



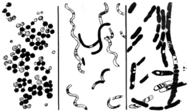
Sample reads

Reference Sequence Database

Taxonomy, Gene function, etc…

Applications, R&D

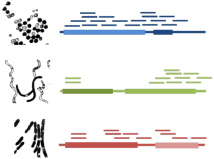
Gene detection: sequence and search



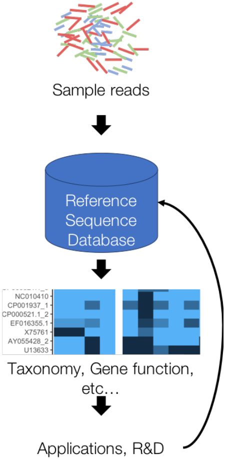
Microbial DNA is extracted from environment.

...**ACATATACG**...
...**CATATACGC**...
...**TAGACAT**...

DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.



Sample reads

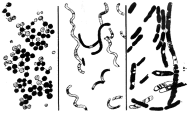
Reference Sequence Database

Taxonomy, Gene function, etc…

Applications, R&D

- String comparison with a few twists:
  - Millions of short reads per sample—ambiguity
  - Natural variation—mismatches not equally important
- Fundamental step for R&D, applications
  - 10b reads (1Tbp) of data per run x samples per day

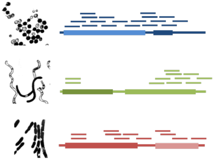
Gene detection: sequence and search



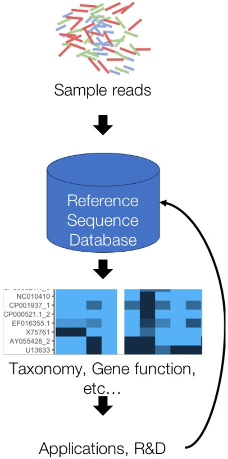
Microbial DNA is extracted from environment.

...**ACATATACG**...
...**CATATACGC**...
...**TAGACAT**...

DNA is fragmented and sequenced (*reads*).



Goal: Computationally map *reads* to known genes, and infer **biological functions**.



Sample reads

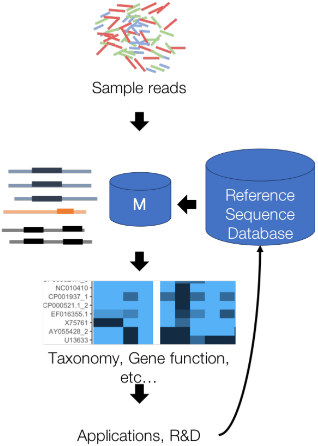
Reference Sequence Database

Taxonomy, Gene function, etc…

Applications, R&D

- String comparison with a few twists:
  - Millions of short reads per sample—ambiguity
  - Natural variation—mismatches not equally important
- Fundamental step for R&D, applications
  - 10b reads (1Tbp) of data per run x samples per day
  - Reference DBs can be large, are regularly updated

ShortBRED: addresses reference database size



Sample reads

M

Reference Sequence Database

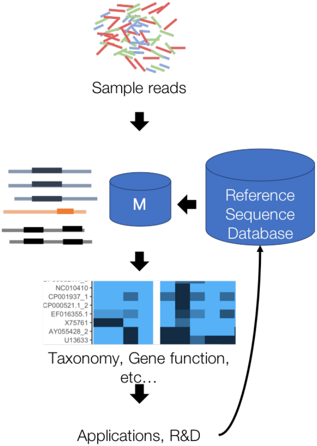
Taxonomy, Gene function, etc…

Applications, R&D

Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

- Identifies *marker* sequences in ref. DB

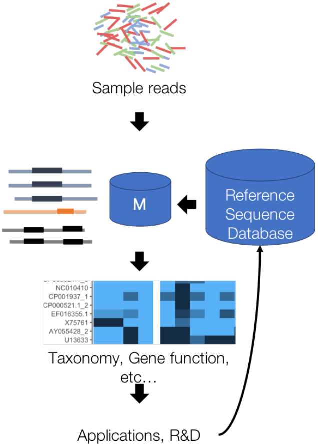
ShortBRED: addresses reference database size



Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

- Identifies *marker* sequences in ref. DB
- Searches sample reads against smaller **marker DB**

ShortBRED: addresses reference database size



Sample reads

M

Reference Sequence Database

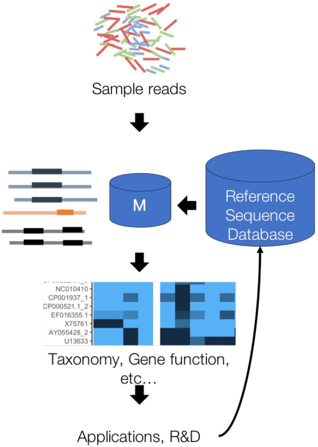
Taxonomy, Gene function, etc…

Applications, R&D

- Identifies *marker* sequences in ref. DB
- Searches sample reads against smaller **marker DB**

Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

ShortBRED: addresses reference database size

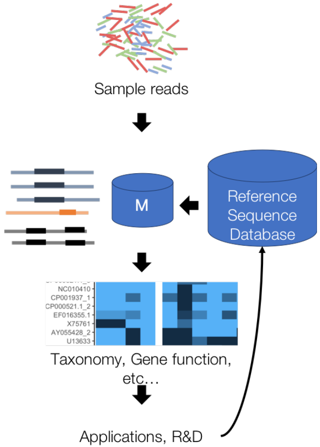


Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

- Identifies *marker* sequences in ref. DB
- Searches sample reads against smaller **marker DB**

- Updates –> rebuild markers

ShortBRED: addresses reference database size

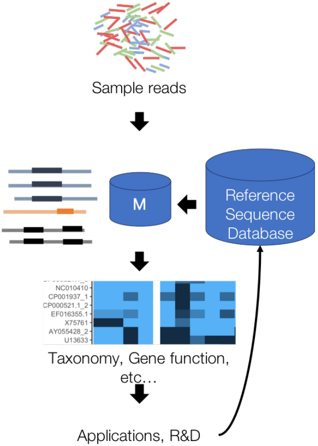


Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

- Identifies *marker* sequences in ref. DB
- Searches sample reads against smaller **marker DB**

- Updates –> rebuild markers
- Miss hits outside of the markers

ShortBRED: addresses reference database size



Sample reads

M

Reference Sequence Database

Taxonomy, Gene function, etc…

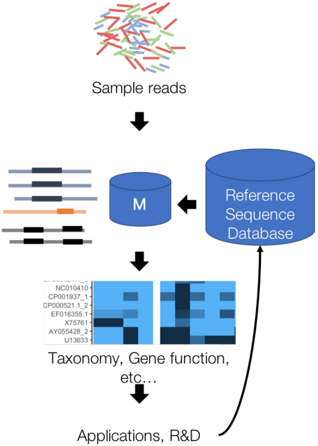
Applications, R&D

Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

- Identifies *marker* sequences in ref. DB
- Searches sample reads against smaller **marker DB**

- Updates –> rebuild markers
- Miss hits outside of the markers
  - is the whole gene present?

ShortBRED: addresses reference database size



Sample reads

M

Reference Sequence Database

Taxonomy, Gene function, etc…

Applications, R&D

Kaminski *et al.*, *PloS Comp Bio* 2015; (Huttenhower lab)

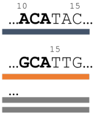
- Identifies *marker* sequences in ref. DB
- Searches sample reads against smaller **marker DB**

- Updates –> rebuild markers
- Miss hits outside of the markers
  - is the whole gene present?
  - limited mutation analysis

Our approach: De Bruijn graph based data structure

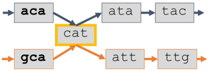
Antibiotic	Gene
Beta-lactam	G1
Erythromycin	G2
…	…

Resistance annotation



Gene sequences

⬌



k-mer	(Gene, Position),…
aca	(G1, 10)
cat	(G1, 11), (G2, 13)
…	…

Expert curated reference

k-mer index for fast searching

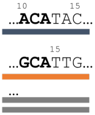
- *Whole* ref. seqs. represented as sliding window substrings

Pearce, Ames, Zemla, Allen

Our approach: De Bruijn graph based data structure

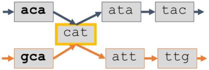
Antibiotic	Gene
Beta-lactam	G1
Erythromycin	G2
...	...

Resistance annotation



Gene sequences

Expert curated reference



k-mer	(Gene, Position),...
aca	(G1, 10)
cat	(G1, 11), (G2, 13)
...	...

k-mer index for fast searching

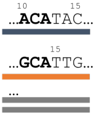
- *Whole* ref. seqs. represented as sliding window substrings
- Search algo. based on short exact matches (see also *Salmon*, *kallisto*)

Pearce, Ames, Zemla, Allen

Our approach: De Bruijn graph based data structure

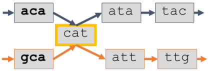
Antibiotic	**Gene**
Beta-lactam	G1
Erythromycin	G2
…	…

Resistance annotation



Gene sequences

Expert curated reference



k-mer	(Gene, Position),…
aca	(G1, 10)
cat	(G1, 11), (G2, 13)
…	…

k-mer index for fast searching

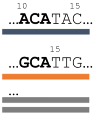
- *Whole* ref. seqs. represented as sliding window substrings
- Search algo. based on short exact matches (see also *Salmon*, *kallisto*)
- Score: breadth, weighted by extended matches

Pearce, Ames, Zemla, Allen

Our approach: De Bruijn graph based data structure

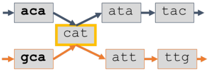
Antibiotic	Gene
Beta-lactam	G1
Erythromycin	G2
…	…

Resistance annotation



Gene sequences

Expert curated reference



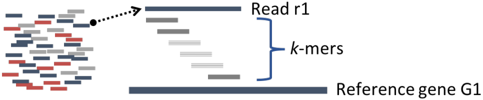
k-mer	(Gene, Position),…
aca	(G1, 10)
cat	(G1, 11), (G2, 13)
…	…

k-mer index for fast searching

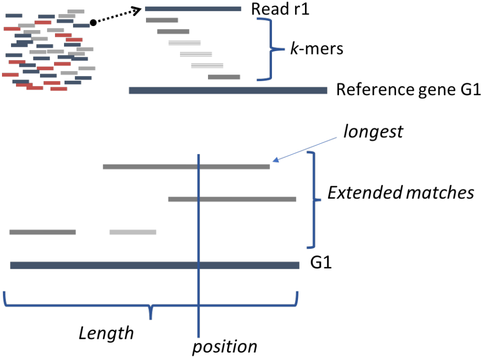
- *Whole* ref. seqs. represented as sliding window substrings
- Search algo. based on short exact matches (see also *Salmon*, *kallisto*)
- Score: breadth, weighted by extended matches
- Eventually: add ML-based signatures, multiple encodings (ARmo)

Pearce, Ames, Zemla, Allen

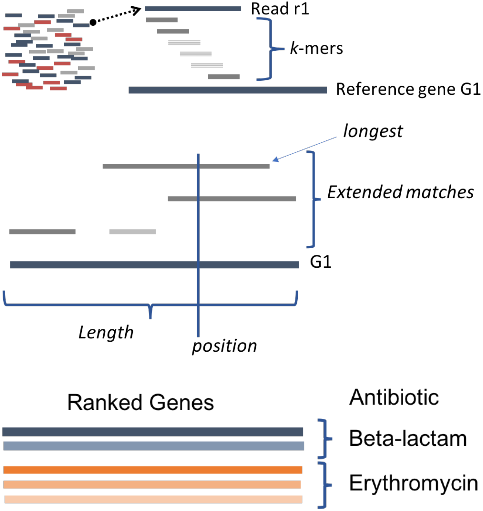
Gene scoring: Maximum exact match lengths as weights



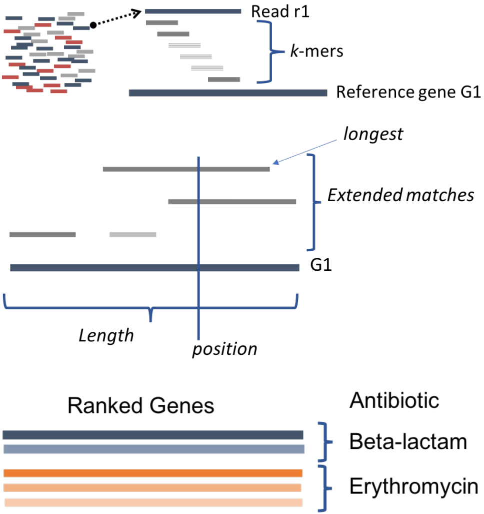
Gene scoring: Maximum exact match lengths as weights



Gene scoring: Maximum exact match lengths as weights



Gene scoring: Maximum exact match lengths as weights



Aggregate gene scores by antibiotic

Antibiotic	Score
Beta-lactam	1.00
Erythromycin	0.88
…	…

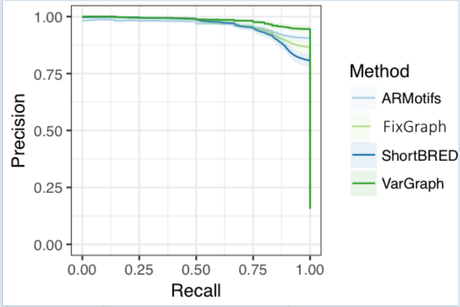
Predicting AMR presence and drug class: preliminary results

- Simulated 500 samples with AMR to multiple drug classes

Predicting AMR presence and drug class: preliminary results

- Simulated 500 samples with AMR to multiple drug classes

AMR presence-absence



(proportion correctly predicted)

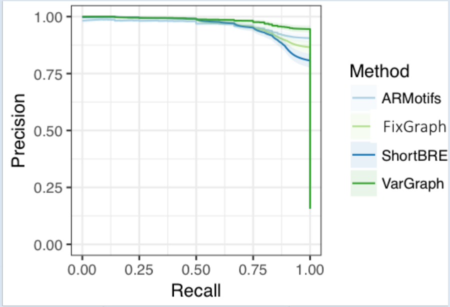
(proportion detected)

Predicting AMR presence and drug class: preliminary results

- Simulated 500 samples with AMR to multiple drug classes

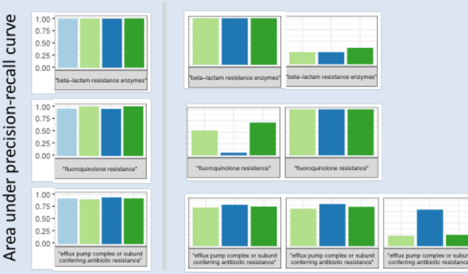
AMR presence-absence

(proportion correctly predicted)



(proportion detected)

AMR drug class



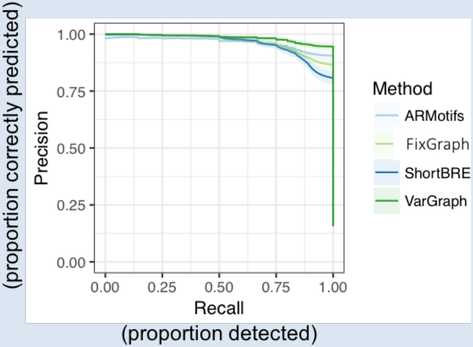
Broadly defined → Narrowly defined

Avila-Herrera, Pearce, Ames, Zemla, Allen, *manuscript in prep*

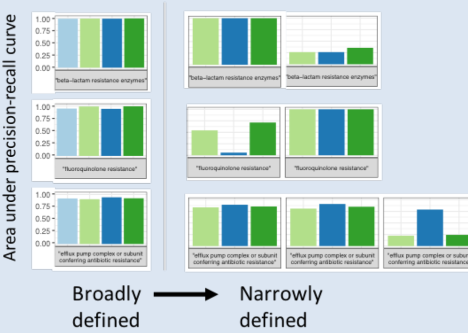
Predicting AMR presence and drug class: preliminary results

- Simulated 500 samples with AMR to multiple drug classes

AMR presence-absence



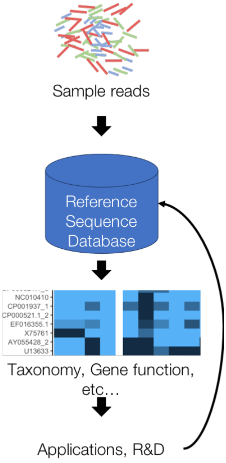
AMR drug class



Avila-Herrera, Pearce, Ames, Zemla, Allen, *manuscript in prep*

- `//TODO: larger standard data set, test scoring functions`

Search tools are critical in modern molecular biology



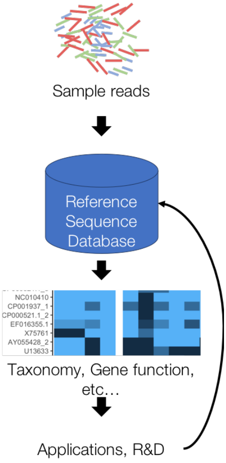
Sample reads

Reference Sequence Database

Taxonomy, Gene function, etc…

Applications, R&D

Search tools are critical in modern molecular biology



Sample reads

Reference Sequence Database

Taxonomy, Gene function, etc…

Applications, R&D

Goal: release tool to compbio community

- Accelerate basic research
- Enable continuous biosurveillance of AMR crisis
- Work towards timely precision medical diagnostics

Acknowledgements

Graph search project

- Jonathan Allen
- Roger Pearce
- Sasha Ames
- Adam Zemla

Related AMR projects

- Marisa Torres
- Nisha Mulakken
- Elizabeth Vitalis
- Nicholas Be
- Crystal Jaing
- Tom Slezak

Fin

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.