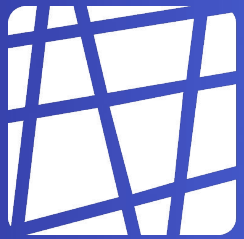
The Build Fellowship

BUILDFELLOWSHIP.COM



Open Avenues

Weekly Updates

- Please provide a quick update on either:
  - Something you did/saw this week that you thought was interesting
  - What you're looking forward to about this week's workshop

(Reminder - please have your cameras on if possible)

The Build Fellowship

OpenVenues

The Build Fellowship

Workshop 2
Data Exploration & Extraction

Sessions Overview

- Workshop 1 – Project Introduction & Setup
- **Workshop 2 – Genomic Data (A2 Assignment)**
- Workshop 3 – Data Analysis & Visualization (A3 Assignment)
- Workshop 4 – Featurization & Baseline Modeling (A4 Assignment)
- Workshop 5 – Model Training Approaches (Final Assignment Set)
- Workshop 6 – Model Tuning
- Workshop 7 – Performance Evaluation (Final Assignment Code/Testing Due)
- Workshop 8 – Results Presentation & Wrap up (Final Presentation Due)

What can be Measured?

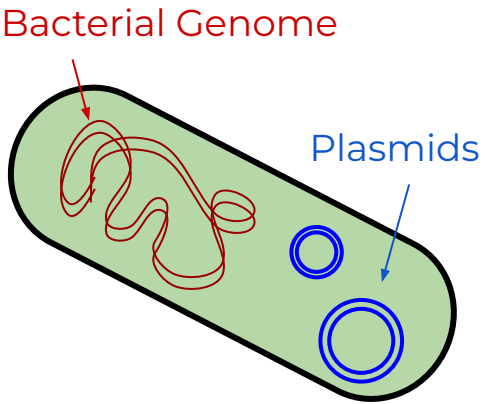
The Build Fellowship

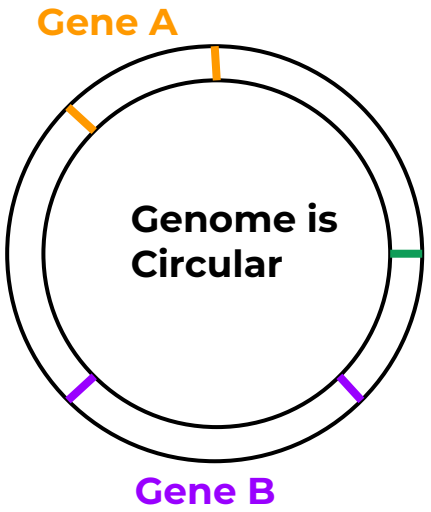
OpenAvenues FOUNDATION

Research Question

Remembering back to last week:

- Aiming to predict AMR directly from genomes
- Bacterial genomes are relatively small





The Build Fellowship

OpenAvenues FOUNDATION

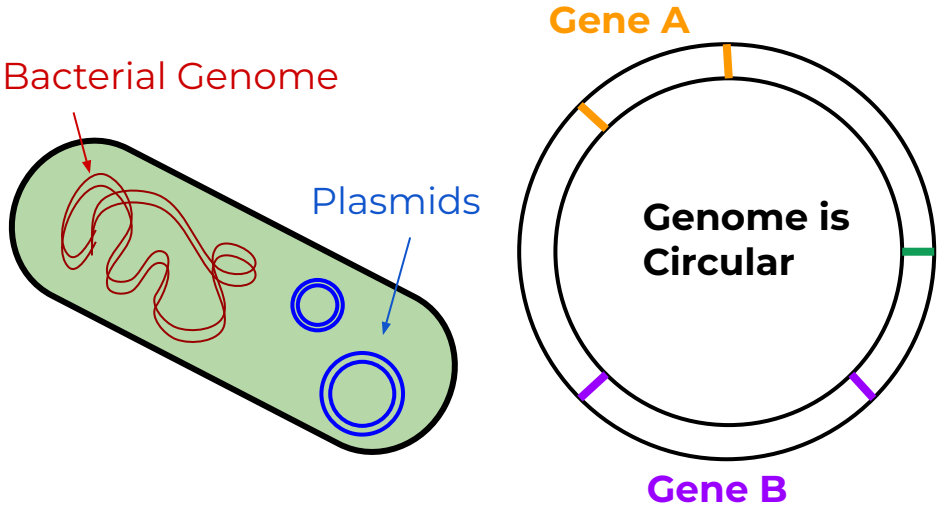
Research Question

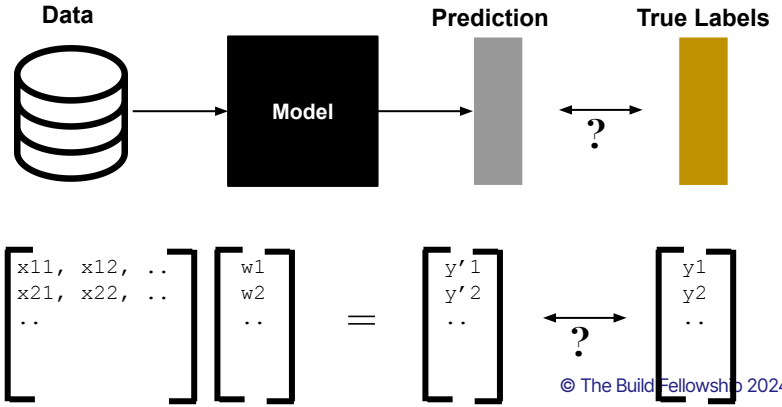
Remembering back to last week:

- Aiming to predict AMR directly from genomes
- Bacterial genomes are relatively small

For our supervised learning task we need two sets of information:

- Y labels: ground truths
- X data: feature matrix





The Build Fellowship

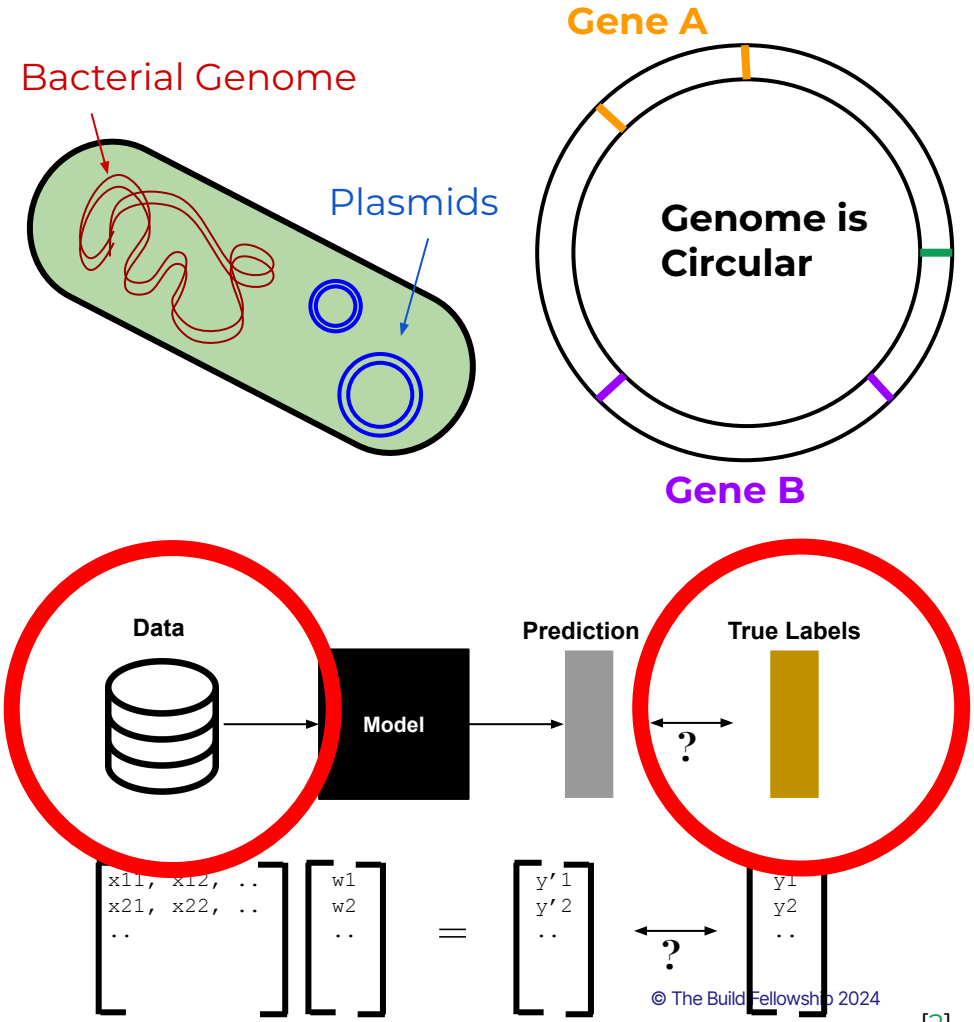
OpenAvenues FOUNDATION

Research Question

Remembering back to last week:
- Aiming to predict AMR directly from genomes
- Bacterial genomes are relatively small

For our supervised learning task we need two sets of information:
- Y labels: ground truths
- X data: feature matrix

Two main areas we need to investigate:
1. How do we measure antibiotic resistance?
2. What features can we extract from the bacterial genomes and how?



The Build Fellowship

OpenVenues FOUNDATION

AMR Profiling

The Build Fellowship

OpenAvenues FOUNDATION

Antimicrobial Resistance

There are a number of methods for testing AMR in bacteria but the one we're going to focus on is "Broth dilution"

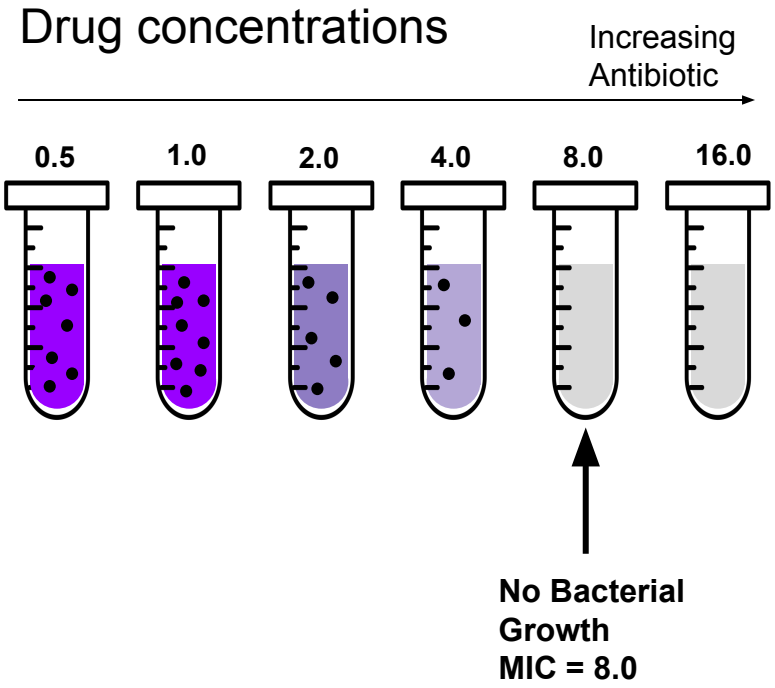
What concentration of antibiotic prevents bacteria from growing?

Steps:

1. Generate multiple concentrations of antibiotic in individual tubes (doubling dilutions)
2. Add bacteria into each tube
3. Monitor growth over time and determine what concentration stops growth

This concentration is the:

Minimum Inhibitory Concentration



The Build Fellowship

OpenAvenues FOUNDATION

Resistance Overview

So we have a Minimum Inhibitory Concentration (MIC)

Resistance and Susceptibility is defined by governing bodies (in the US this is the NCBI)

CLSI publishes the M100 each year - this free resource contains a translation between MIC for each bacterial species/drug to a canonical resistance call:

- R = Resistant to the drug, not recommended for use
- I = Intermediate, shows some resistance, not recommended for use
- S = Susceptible, drug will be effective against organism

Often R & I are grouped together into “Not Susceptible”

Example Table from M100

Antimicrobial Agent	Interpretive Categories and MIC Breakpoints, µg/mL				Comments
	S	SDD	I	R	
CEPHEMS (PARENTER IV. Please refer to Glossary I.) (Continued)					
Cefotetan*	≤ 16	–	32^	≥ 64	
Cefoxitin	≤ 8	–	16^	≥ 32	
Cefuroxime (parenteral)	≤ 8	–	16^	≥ 32	See comment (14).
Ceftazidime	≤ 4	–	8^	≥ 16	See comment (14).
Cefamandole*	≤ 8	–	16^	≥ 32	See comment (14).
Cefmetazole*	≤ 16	–	32^	≥ 64	**(19)** Insufficient new data exist to reevaluate breakpoints listed here.
Cefonicid*	≤ 8	–	16^	≥ 32	See comment (14).

[1]

The Build Fellowship

OpenAvenues FOUNDATION

Genomic Sequencing

The Build Fellowship

Open Avenues FOUNDATION

QUIZ TIME !?

What data are we trying to measure with sequencing?

a) The number of cells in the organism
b) The amino acid sequences and proteins of the organism
c) The size of the bacteria from smallest to largest
d) The nucleotide sequences of the organism

The Build Fellowship

OpenAvenues

Long Read vs Short Read

Sequencing technologies have evolved rapidly over the past few decades and continue to do so now.

Currently two state of the art methods for sequencing genomes split into two categories:

1. Short read (e.g. Illumina [2])
2. Long read (e.g. ONT [3])

Steps are similar:

3. Chemically extract and amplify the DNA of the bacteria
4. Prepare the DNA for sequencing (insertion of adaptors and other synthetics sequences)
5. Run through the sequencer and record the resulting bases

We'll focus on short read data for this project as it's more available and common

Illumina Sequencer

[4]

ONT Sequencer

The Build Fellowship

OpenAvenues FOUNDATION

Short Read Data

Short read sequencers build segments of of the DNA that are inherently small in length (typically 50-300 nucleotides)

This is very small compared to our bacterial genome (2-6 million nucleotides).

We're going to end up with millions of tiny chunks of DNA

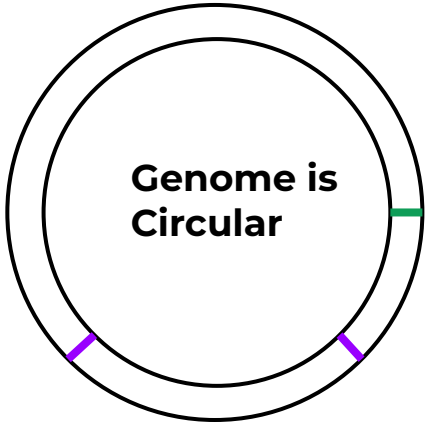
FASTQ:

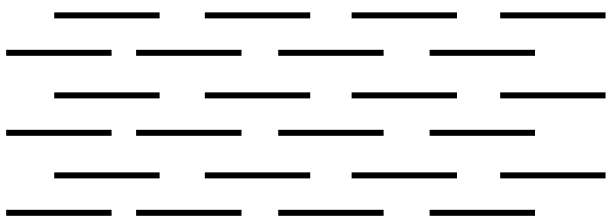
```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAAC ... GATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFF ... FBFFFFFFFFFFF7FFFF<FF
```

Sequence name
DNA sequence
Quality line break
Quality scores

[6]

FASTA:
FASTA is identical to the above but without the quality information





DNA "Soup" - Where are the Genes?

The Build Fellowship

OpenAvenues FOUNDATION

Where can we Find Data?

The Build Fellowship

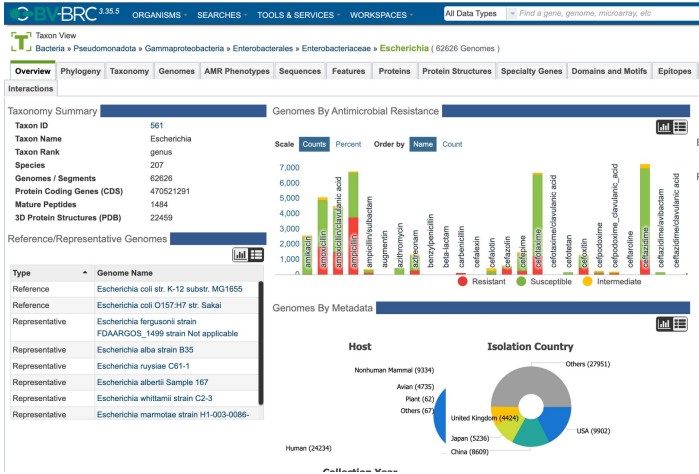
OpenAvenues FOUNDATION

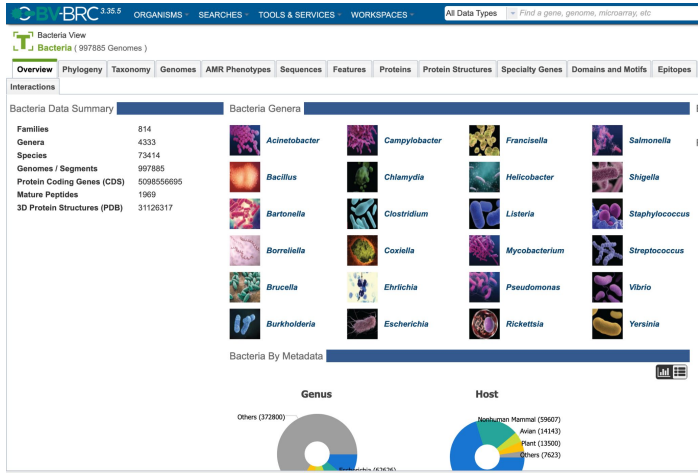
Public Data Sources

Bacteria and Viral Bioinformatics Resource Center (BV-BRC) [7]

Online database of bacteria (and viruses) with large amount of metadata around each sample

Importantly many strains come with AMR test results





Assembled Data

- For this course we're going to be working with "Assembled" data
- Assembly is a bioinformatic technique to take the millions of short reads and align them to themselves
- Using overlapping regions the reads can be stitched back together
- Still not a full single circular genome
- Fewer pieces to work with and higher quality!

The Build Fellowship

OpenAvenues FOUNDATION

Data Preprocessing

Genes & Alignment

A perfectly assembled bacterial genomes can be anywhere between 2 million and 6 million nucleotides

That's a very long string!

We certainly could try working with the whole sequence but it might prove computationally difficult

Instead we can try to subset the genome down to just areas we care about

In this case: **Known Resistance Genes**

This will introduce biases - we are throwing away much of the genomic sequences but it will allow us to achieve a predictive model in a reasonable time frame

If we had a list of important genes (toy example):

CGTATAACTGA

GGAACAGCGGT

Can we search for them in each sample?

CGTATAACTGA GGAACAGCGGT

AACGTATAACTGATCGGAACAGCGGTA

Assembled Sample Genomes (toy example)

The Build Fellowship

OpenAvenues FOUNDATION

CARD

So where can we gather our set of important genes?

Comprehensive Antibiotic Resistance Database (CARD) [8]

Two main types of resistance genes:

1. Presence/Absence (if exists then R)
2. Variant (if mutation in sequence then R)

Comes in FASTA format:

- Header = Gene name + ID
- Body = DNA string (or Protein string)

We'll work with the DNA FASTA format

Presence/Absence Gene

Protein DNA

```
>gb|BAM16262.1|+|AAC(2')-IIa [Burkholderia glumae]
MKDRSHDDSMAEVCRNTSENHWLKTDYRTLFRLCPDGRIERENDPDCSPGPRFWLAGCSEGNVFGVRADVPDDIALKLEELASVEPPFTP
PAIPKHLERYLSLLGSDGPVTHDLGLIYELPHAQQYPSKARLIGSGSEEGESLMQSWAEDRVPEALFELGFREVADFWTPWCAAVVDGEV
ASIAFAARLADAGAELGLVTAKAFRGQGFAAAATAGWSRLSALRSRTLFYSTDRDNISSQRVAARLGLRLRGASLRISRA
```

Variant Gene (variants highlighted in red)

Published Variants:

PMID: 29091182 S83F D87G

Protein DNA

```
>gb|AAC75291.1|-|Escherichia coli gyrA with mutation conferring resistance to triclosan [Escherichia coli str. K-12 substr.
MG1655]
MSDLAREITPVNIEEELKSSYLDYAMSVIVGRALPDVRDGLKPVHRRVLYAMNVLGNDWN
KAYKKSARVVGDVIGKYHPHGDSAVYDTIVRMAQPFSLRYMLVDGQGNFGSIDGDSAAAM
RYTEIRLAKIAHELMADLEKETVDFVDNYDGTEKIPDVMPTKIPNLLVNGSSGIAVGMAT
NIPPHNLTEVINGCLAYIDDEDISIEGLMEHIPGPDFPTAAIINGRRGIEEAYRTGRGKV
YIRARAEVEVDAKTGRETIIVHEIPYQVNKARLIEKIAELVKEKRVEGISALRDESDKDG
MRIVIEVKRDAVGEVVLNNLYSQTQLQVSFGINMVALHHGQPKIMNLKDIIAAFVRHRRE
VVTRRTIFELRKARDRAHILEALAVALANIDPIIELIRHAPTPAEAKTALVANPWQLGNV
AAMLERAGDDAARPEWLEPEFGVRDGLYYLTEQQAQAILDLRLQKLTGLEHEKLLDEYKE
LLDQIAELLRILGSADRLMEVIREELELVREQFGDKRRTEITANSADINLEDLITQEDVV
VTLSHQGYVKYQPLSEYEAQRRGGKGKSAARIKEEDFIDRLLVANTHDHILCFSSRGRVY
SMKVYQLPEATRGARGRPIVNLLPLEQDERITAILPVTEFEEGVKVFMATANGTVKKTVL
TEFNRLRTAGKVAIKLVDGDELIGVDLTSGEDEVMLFSAEGKVVRFKESSVRAMGCNTTG
VRGIRLGEGDKVVSLIVPRGDGAILTATQNGYGKRTAVAEYPTKSRATKGVISIKVTERN
GLVVGAVQVDDCDQIMMITDAGTLVRTRVSEISIVGRNTQGVILIRTAEDENVVGLQRVA
EPVDEEDLDTIDGSAAEGDDEIAPEVDVDDEPEEE
```

The Build Fellowship

OpenAvenues FOUNDATION

Workshop 2 Data Exploration

The Build Fellowship

OpenAvenues FOUNDATION

References

[1]:https://em100.edaptivedocs.net/Login.aspx?_ga=2.87309720.1855123465.1712506668-1848165892.1711849752

[2]:https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html#:~:text=The%20Illumina%20semiconductor%20sequencing%20method,high%20data%20accuracy%20of%20SBS.

[3]:https://nanoporetech.com/platform/technology

[4]:https://www.illumina.com/systems/sequencing-platforms/nextseq-1000-2000.html

[5]:https://nanoporetech.com/products/sequence/promethion

[6]:https://gencoded.com/index.php/2020/05/20/fastq-format-an-overview/

[7]:https://www.bv-brc.org/

[8]:https://card.mcmaster.ca/

The Build Fellowship

OpenAvenues FOUNDATION