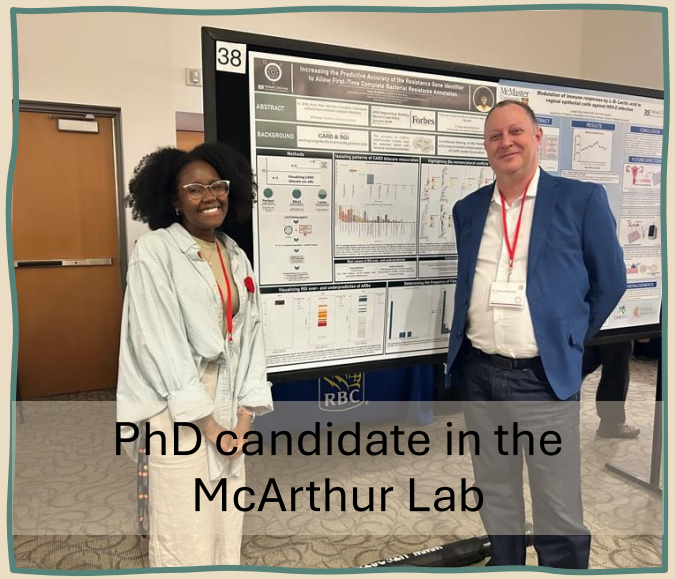
Karyn Mukiri

Predicting the Total Resistome

(hopefully)

2 tl;dr



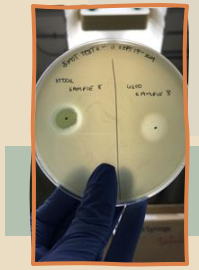
PhD candidate in the McArthur Lab

Experienced a very non-linear career path

I ***really*** like fish (I have 5!)

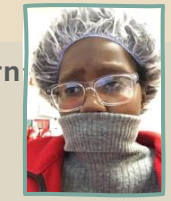
3 … How did we get here?

B. Tech, Biotechnology (undergraduate)
Class of 2019



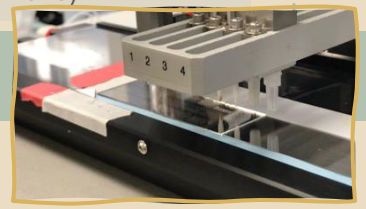
Research Assistant
Dr. Zeinab Hosseinidoust
Chemical Engineering
(co-op, 2017)

Quality Assurance Intern
Coca-Cola
(co-op, 2016)



Lab Tech. Assistant
Biotech. Undergrad Labs
($$$, 2018)

Capstone Project: *Biosynthesis of Silver Nanoparticles and Applications in Biosensor Technology*
("thesis", 2018)



Biochemistry & Biomedical Sciences (grad. school)
2021

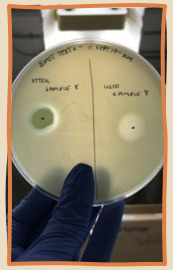
?

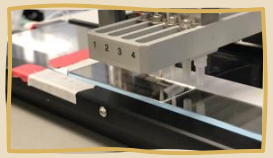
?

Post-grad work
2019

4 … How did we get here?

B. Tech, Biotechnology (undergraduate)
Class of 2019

Biochemistry & Biomedical Sciences (grad. school)
2021

?



QA & Manufacturing Intern
Coca-Cola (2019)

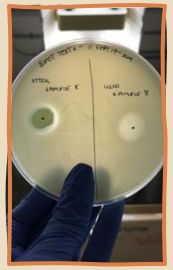
Desperately Applying to Grad. School
2019-2021

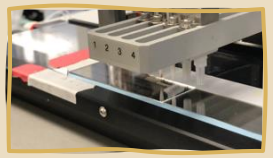
Post-grad work
2019



5 … How did we get here?

B. Tech, Biotechnology (undergraduate)
Class of 2019

Biochemistry & Biomedical Sciences (grad. school)
2021

MSc, McArthur Lab
Thesis: *Predicting the Total Resistome*
Transferred to the **PhD program** in Fall 2022

QA & Manufacturing Intern
Coca-Cola (2019)

Desperately Applying to Grad. School
2019-2021



Post-grad work
2019



Repository of AMR information

6

The Comprehensive Antibiotic Resistance Database

A bioinformatic database of resistance genes, their products and associated phenotypes.

7170 Ontology Terms, 5194 Reference Sequences, 2008 SNPs, 3279 Publications, 5242 AMR Detection Models

Resistome predictions: 413 pathogens, 24291 chromosomes, 2662 genomic islands, 48212 plasmids, 172216 WGS assemblies, 276270 alleles

14 ontology terms | Hide

+ cell wall synthesis enzyme targeted by antibiotic
+ cell membrane component targeted by antibiotic
+ protein synthesis machinery targeted by antibiotic
+ nucleic acid synthesis machinery targeted by antibiotic
+ nucleic acid targeted by antibiotic
+ folate synthesis enzyme targeted by antibiotic
+ isoprenoid synthesis enzyme targeted by antibiotic

Antibiotic target

10 ontology terms | Hide

+ antibiotic target alteration [Resistance Mechanism]
+ antibiotic target replacement [Resistance Mechanism]
+ antibiotic target protection [Resistance Mechanism]
+ antibiotic inactivation [Resistance Mechanism]
+ antibiotic efflux [Resistance Mechanism]
+ reduced permeability to antibiotic [Resistance Mechanism]
+ resistance by absence [Resistance Mechanism]

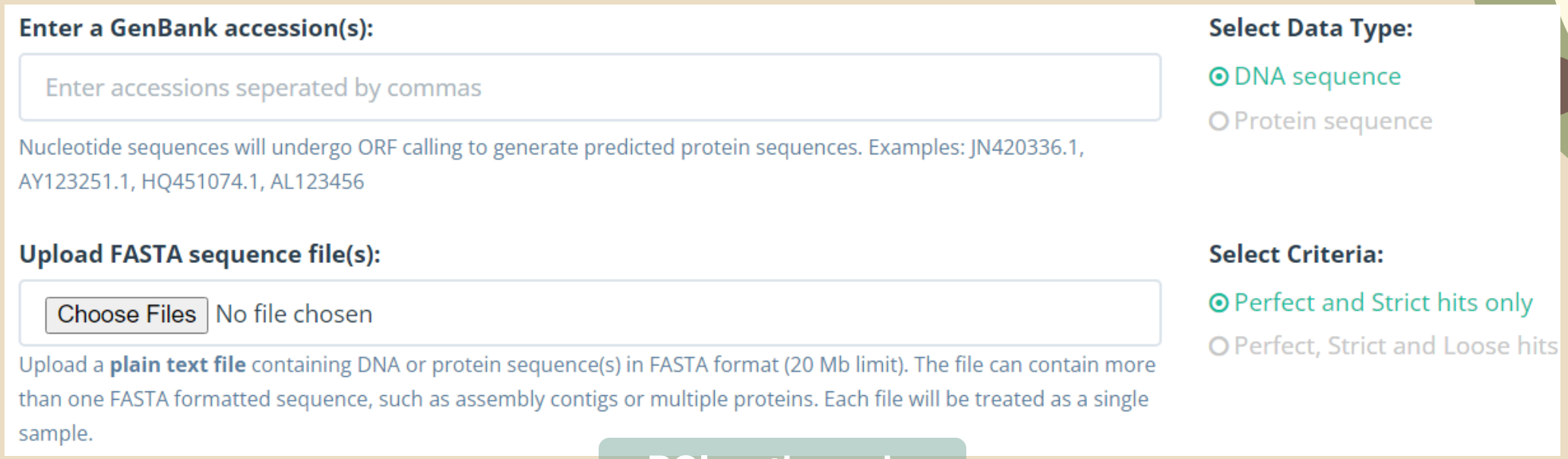
Mechanism of antibiotic resistance

7

Genome annotation tool

RGI Resistance Gene Identifier

RGI can be used to predict resistomes from protein or nucleotide data based on homology and SNP models. Analyses can be performed via this web portal (20 Mb limit), via the command line, or via use of a Galaxy wrapper. The command line version can be obtained from the Download section of the CARD website. You can additionally install RGI from Conda or run RGI from Docker.



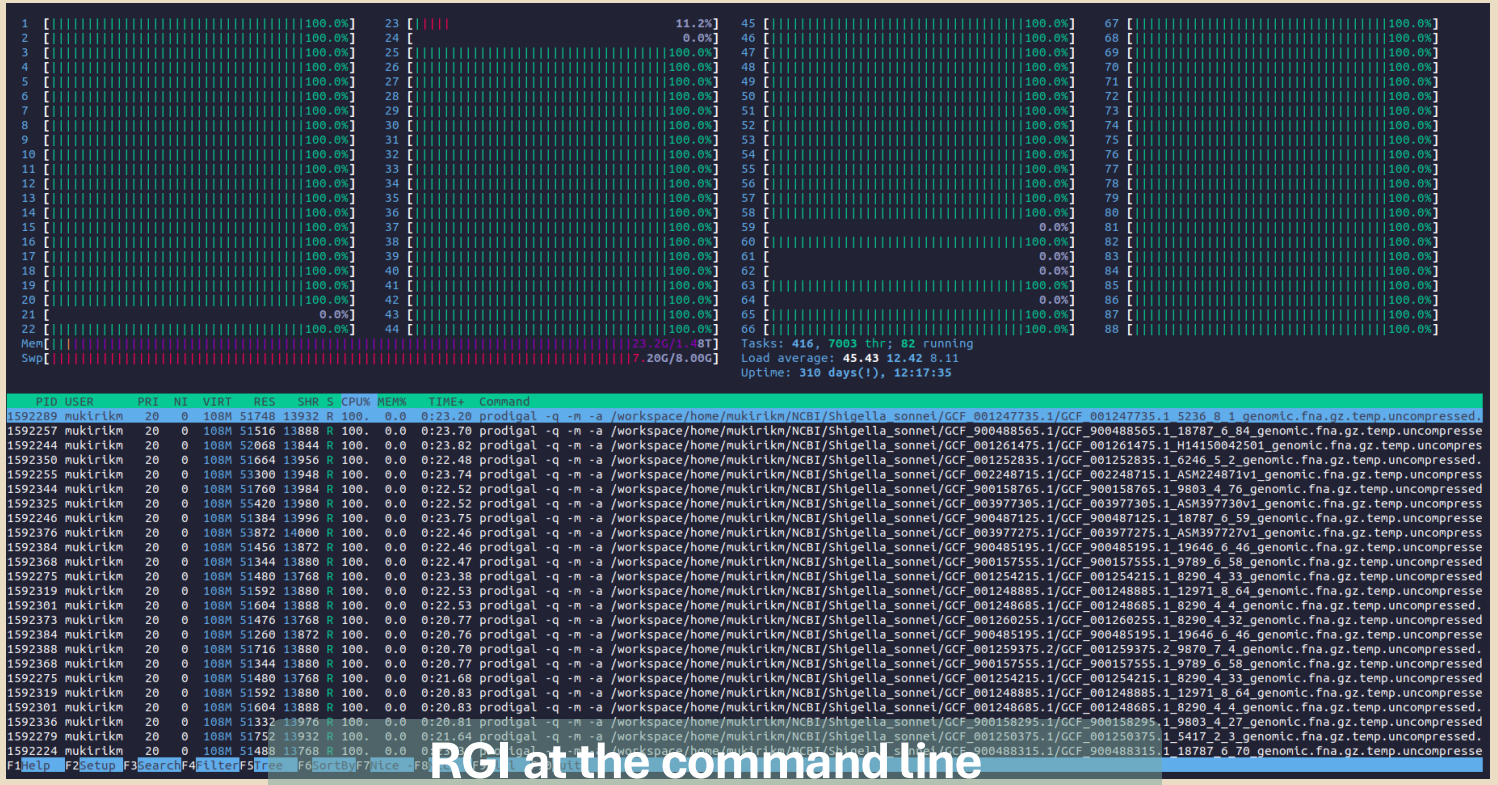
RGI on the web

7

RGI Resistance Gene Identifier

Genome annotation tool

RGI can be used to predict resistomes from protein or nucleotide data based on homology and SNP models. Analyses can be performed via this web portal (20 Mb limit), via the command line, or via use of a Galaxy wrapper. The command line version can be obtained from the Download section of the CARD website. You can additionally install RGI from Conda or run RGI from Docker.



RGI at the command line

Genome annotation tool

8

RGI Resistance Gene Identifier

RGI can be used to predict resistomes from protein or nucleotide data based on homology and SNP models. Analyses can be performed via this web portal (20 Mb limit), via the command line, or via use of a Galaxy wrapper. The command line version can be obtained from the Download section of the CARD website. You can additionally install RGI from Conda or run RGI from Docker.

>my_query

MRAHFRRKYQLSQNAKDSAQLRAVGGKVFLSGDLARMVRPDAPEKDDGFNLIRAFQGLKF
SGPMFTGSKHSRADKTTNEEPCRVIRRTDAQDTLLFSSSQCSVCEISIGSMFSSCAGSVC
ARPHLCVRASVCVRPLHQSSRSAGAARLFQEIPDTGSNGWMNLLRFWRDGRFSRMHKHME
ESFRRLGPIYREHLGSQSSVNIMLPMDTGELFRSEGLHPRRMTLQPWATHRETRRHSKGV
FLKNGTEWRADRLLLNREVMVSSSVHRFLPLLDEVAQDFCRSLRRRVQADGFEKAGQHTL
TLDPSPDLFRFALEASCHVLYGERIGLFSSCPSDESERFISAVERMLATTPPLLYLPPRL
LLRLRASLWTTHATAWDDIFSHAEQRIQRSYQRLQARPSAAPDCSFPGVLGKLMEAGQLS
LELIRANITELMAGGVDTTAVPLQFALFELARNPDVQECVRAQVLSSWQQASGDPLKALQ
GAPLLKGTIKETLRLYPVGITVQRYPVRDIVLQNYHVPAGTLVQVCLYPLGRSAEVFSRP
ECFDPSRWSADADAGSAGGFRSLAFGFGSRQCVGRRIAENEMQLLLMHILRTFKLTVSST
EELSTKYTLILQPECPPRITFSTLTHQH

RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
Loose	tmrB		protein homolog model	tunicamycin resistance protein	nucleoside antibiotic	reduced permeability to antibiotic	34.25	91.37
Loose	DHA-7		protein homolog model	DHA beta-lactamase	cephalosporin, cephamycin	antibiotic inactivation	100.0	89.18
Loose	NmcR		protein homolog model	NmcA beta-lactamase	carbapenem, cephalosporin, cephamycin, penam	antibiotic inactivation	46.69	98.64
Perfect	TEM-1		protein homolog model	TEM beta-lactamase	monobactam, cephalosporin, penam, penem	antibiotic inactivation	100.0	100.00
Perfect	AAC(3)-IId		protein homolog model	AAC(3)	aminoglycoside antibiotic	antibiotic inactivation	100.0	100.00
Perfect	NDM-1		protein homolog model	NDM beta-lactamase	carbapenem, cephalosporin, cephamycin, penam	antibiotic inactivation	100.0	100.00

Sequenced-based detection models

protein homolog model	4,805
BLASTP bit-score	4,805

protein variant model	532
single resistance variant	188
BLASTP bit-score	188
multiple resistance variants	22
no association with resistance TB	17
deletion mutation from nucleotide sequence	17
nonsense mutation	16
high confidence TB	14
insertion mutation from nucleotide sequence	14
minimal confidence TB	9
moderate confidence TB	8
insertion mutation from peptide sequence	7
deletion mutation from peptide sequence	7
indeterminate confidence TB	7
co-dependent single resistance variant	6
frameshift mutation	5
snp in promoter region	3
co-dependent nonsense SNP	2
disruptive mutation in regulatory element	1
insertion mutation	1

protein knockout model	18
BLASTP bit-score	18

protein overexpression model	36
single resistance variant	15
BLASTP bit-score	15
multiple resistance variants	1
deletion mutation from peptide sequence	1
nonsense mutation	1
deletion mutation from nucleotide sequence	1
insertion mutation from nucleotide sequence	1
frameshift mutation	1

rRNA gene variant model	187
BLASTN bit-score	87
single resistance variant	86
high confidence TB	4
no association with resistance TB	3
moderate confidence TB	3
multiple resistance variants	2
disruptive mutation in regulatory element	1
minimal confidence TB	1

Meta-model detection models

efflux pump system meta-model	33
efflux pump components	33

gene cluster meta-model	14
gene order	14

protein domain meta-model	1
domain order	1

One big problem: **RGI can't detect every CARD gene or mutation because it's missing code**

The full suite of AMR genes and mutations in bacteria

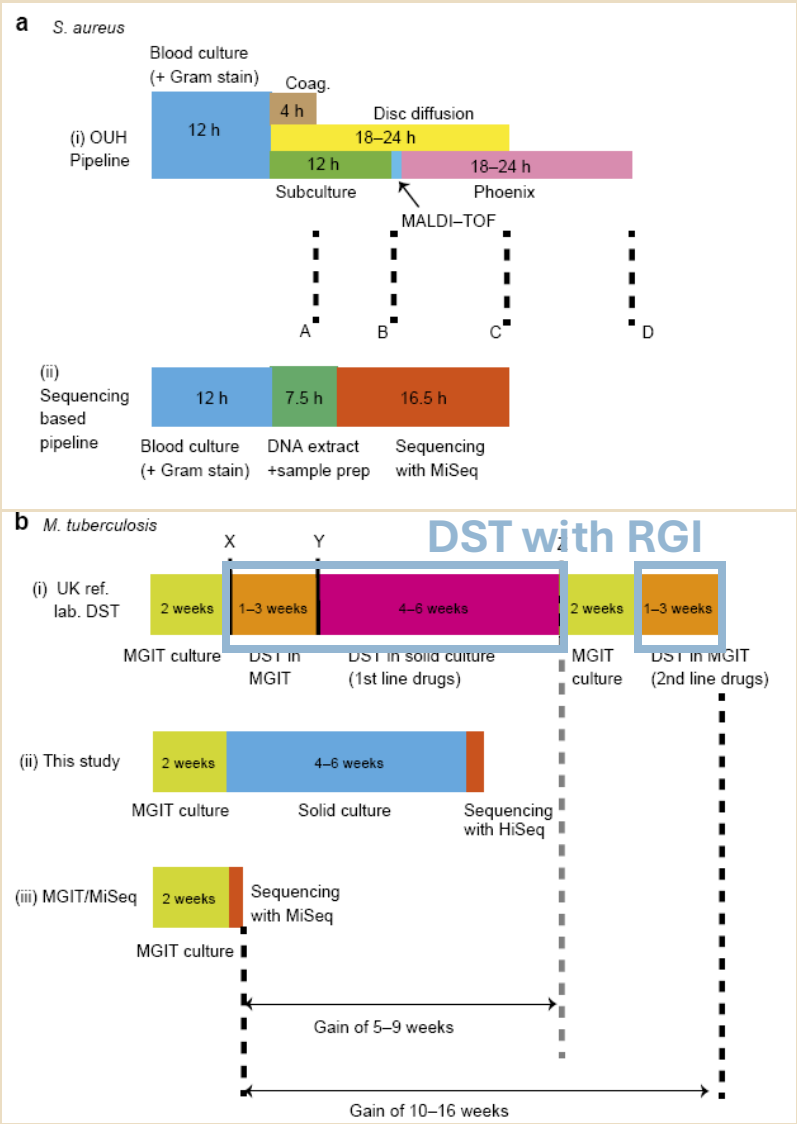
10 Why *Predict the Total Resistome*?

Being able to detect all AMR genes and mutations computationally gets us closer to phenotype prediction from genotype

A more complete resistome prediction can speed up diagnostic pipelines

Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*

Phelim Bradley, N. Claire Gordon, Timothy M. Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J. Pankhurst, Luke Anson, Mariateresa de Cesare, Paolo Piazza, Antonina A. Votintseva, Tanya Golubchik, Daniel J. Wilson, David H. Wyllie, Roland Diel, Stefan Niemann, Silke Feuerriegel, Thomas A. Kohl, Nazir Ismail, Shaheed V. Omar, E. Grace Smith, David Buck, Gil McVean, ... Zamin Iqbal



10

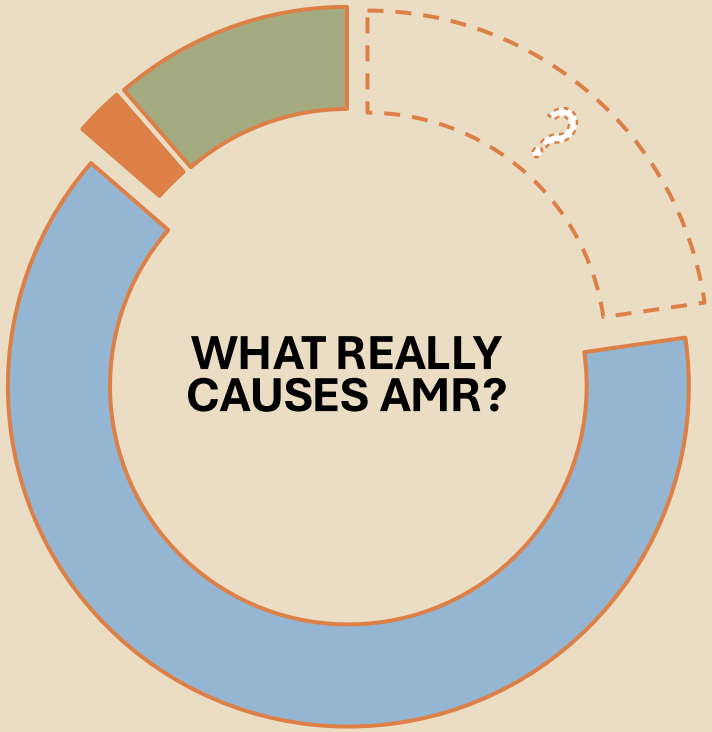
Why *Predict the Total Resistome*?

The full suite of AMR genes and mutations in bacteria

Being able to detect all AMR genes and mutations computationally gets us closer to phenotype prediction from genotype

A more complete resistome prediction can speed up diagnostic pipelines

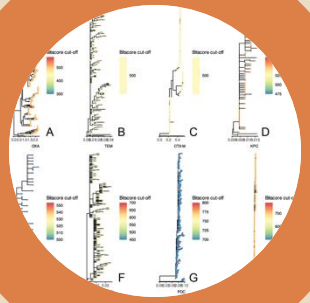
We'll be able to figure out which AMR genes are most prevalent epidemiologically



So why isn’t the code done, Karyn?????

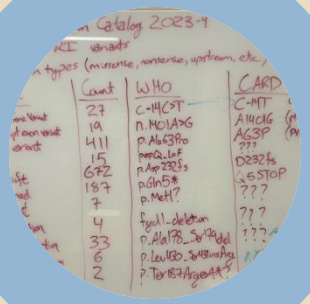
Research isn't [that] linear

Step 1: Figure out how bad our curation is



- Bad curation leads to bad AMR predictions

Step 2: Fix bad curation and add more genes/mutations



- More data == a better look at the "resistome"

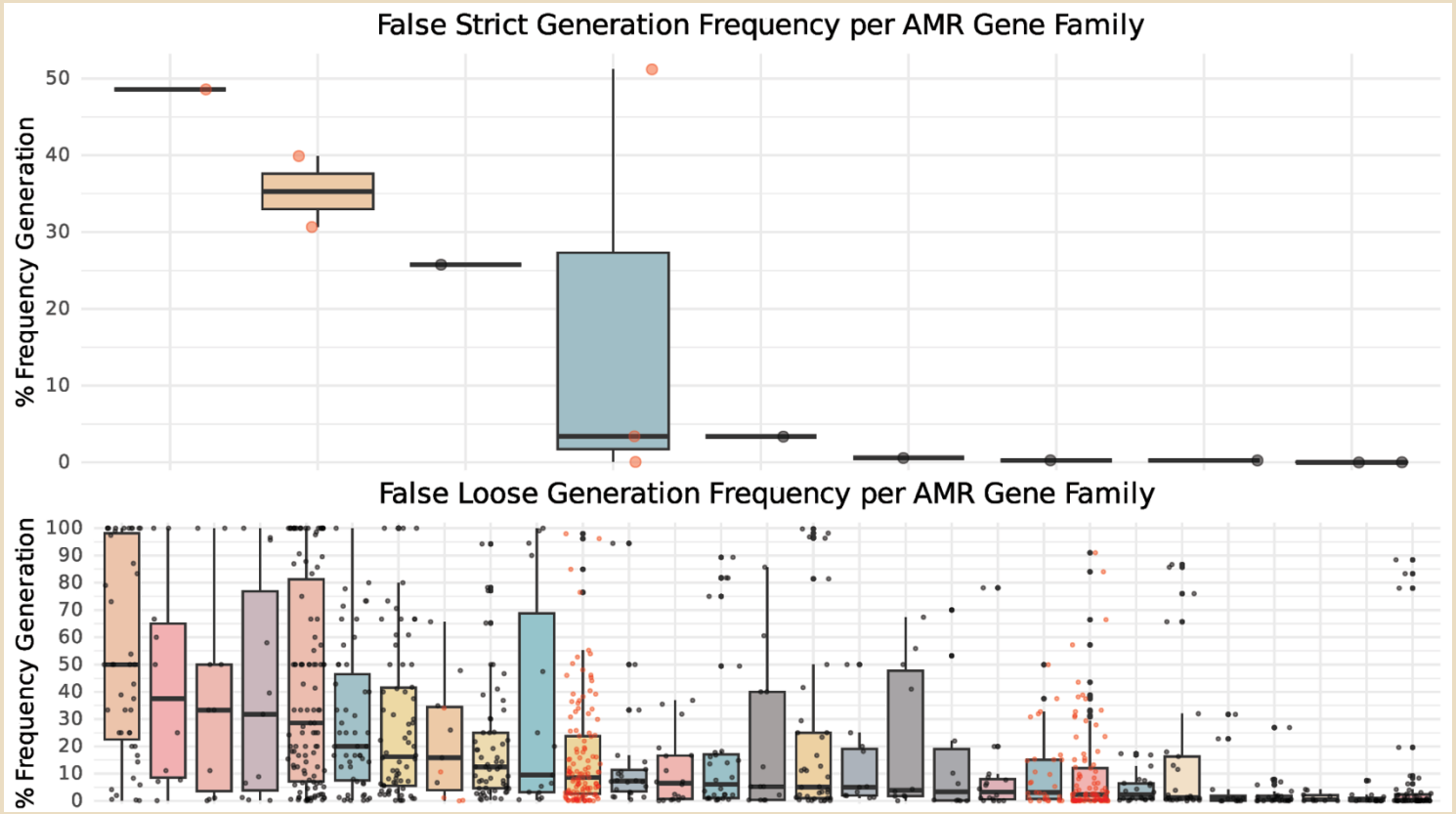
Step 3: Write the code

- Without the code, RGI is blind

Step 4: Validate the code

- Did we get it right? How accurate are we? Did I break it?

13 Step 1: Figure out how bad curation is



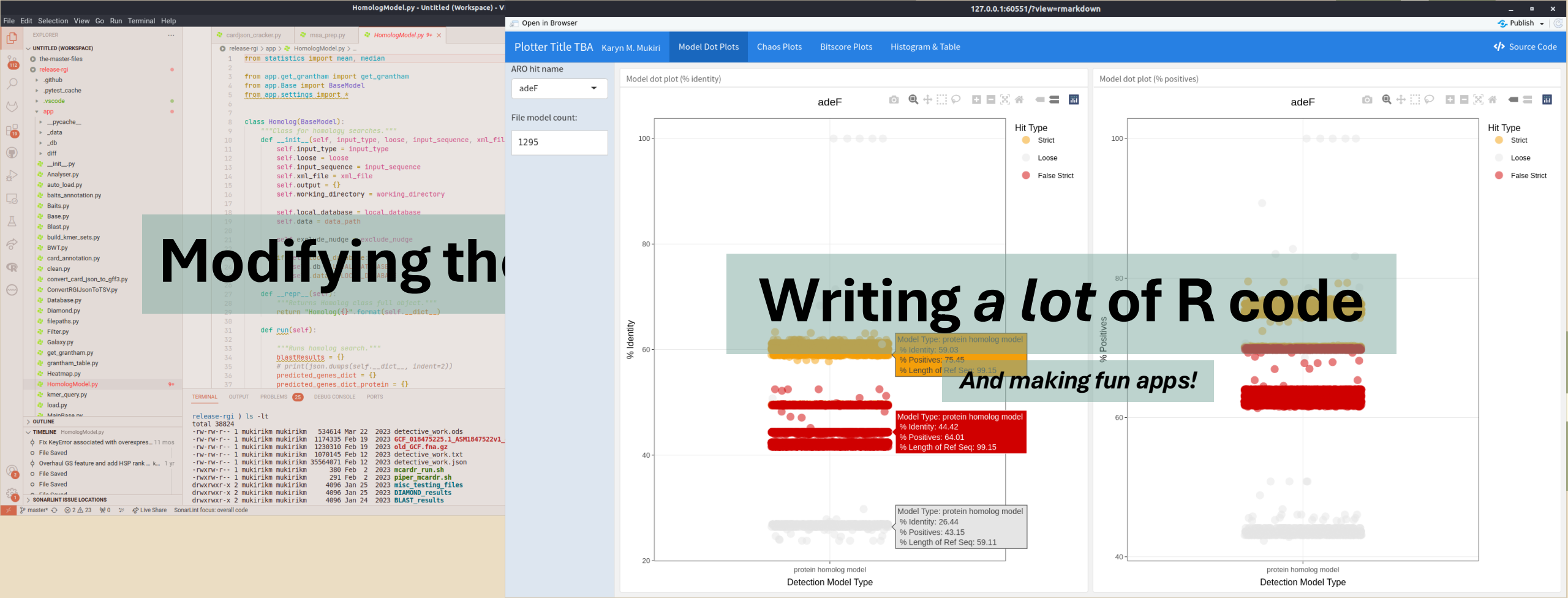
The answer: Not terrible!

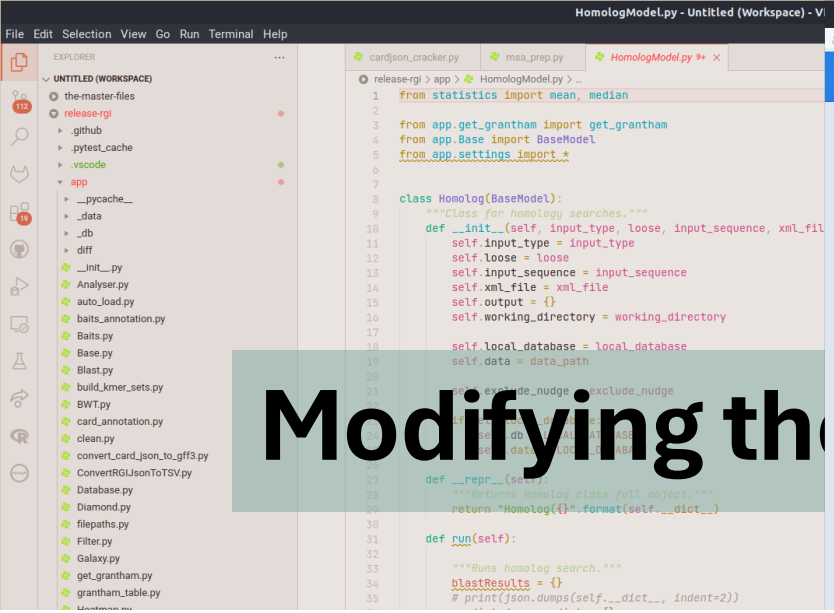
Two findings:

1. RGI has a **very strict prediction algorithm**, so we rarely overpredict resistance
   - **You can trust RGI-Perfect and RGI-Strict annotations**
2. RGI may be **underpredicting resistance**
   - Mainly beta-lactamases! Terrible nomenclature…

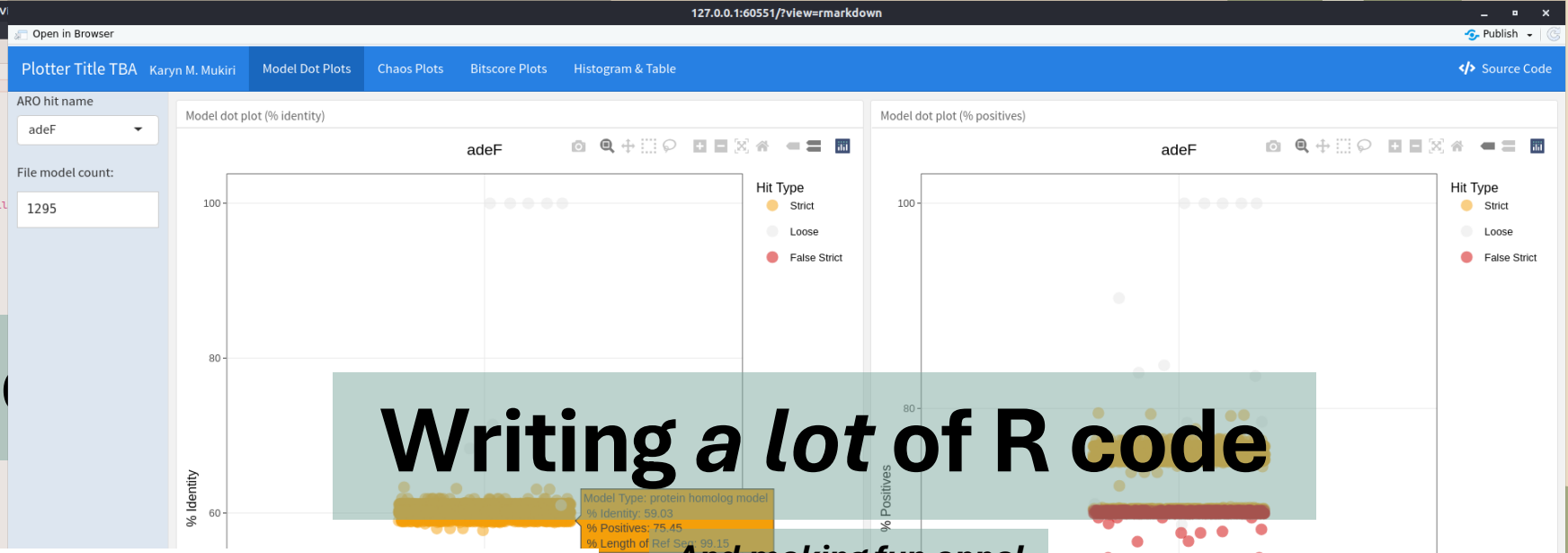
The last slide condensed 3 years of work into one figure… Painful.

Modifying the RGI code


HomologModel.py - Untitled (Workspace) - V
File Edit Selection View Go Run Terminal Help
Modifying th
127.0.0.1:60551/?view=rmarkdown
Open in Browser
Publish
Plotter Title TBA
Karyn M. Mukiri
Model Dot Plots
Chaos Plots
Bitscore Plots
Histogram & Table
Source Code
ARO hit name
adeF
File model count:
1295
Model dot plot (% identity)
Model dot plot (% positives)
Hit Type
Strict
Loose
False Strict
% Identity
% Positives
Model Type: protein homolog model
% Identity: 59.03
% Positives: 75.45
% Length of Ref Seq: 99.15
Model Type: protein homolog model
% Identity: 44.42
% Positives: 64.01
% Length of Ref Seq: 99.15
Model Type: protein homolog model
% Identity: 26.44
% Positives: 43.15
% Length of Ref Seq: 59.11
protein homolog model
Detection Model Type
Writing a lot of R code
And making fun apps!

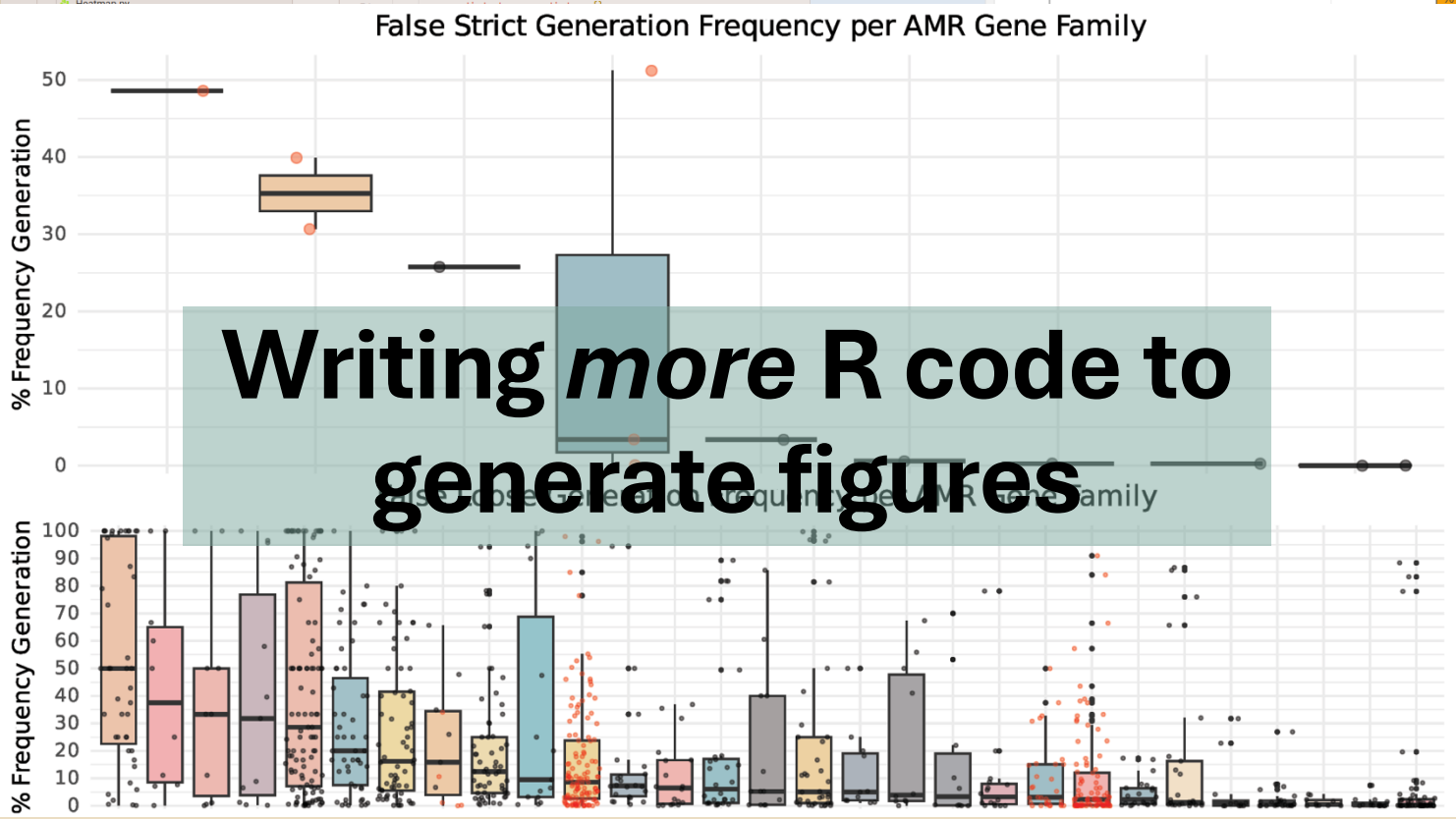


Modifying th

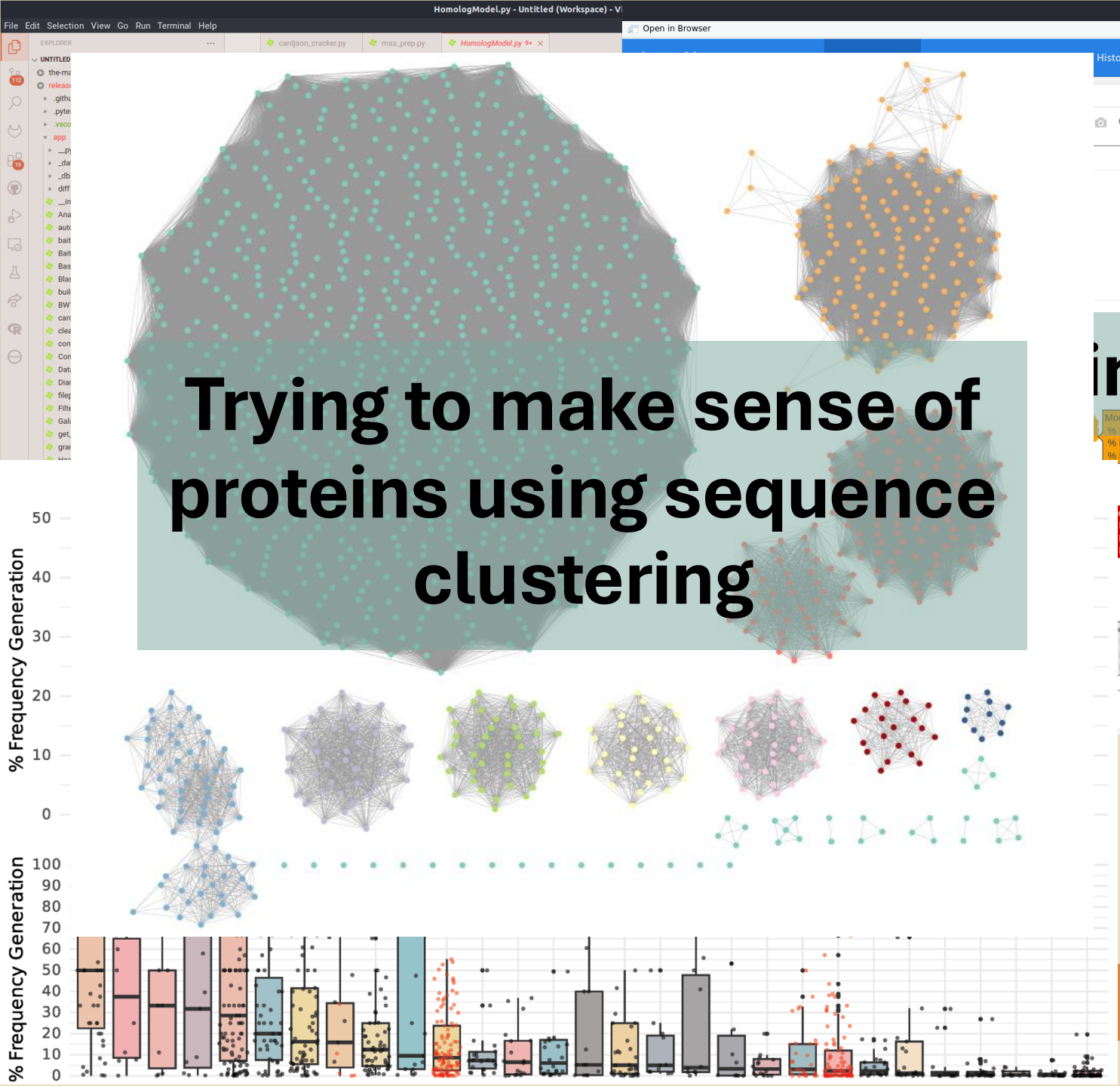


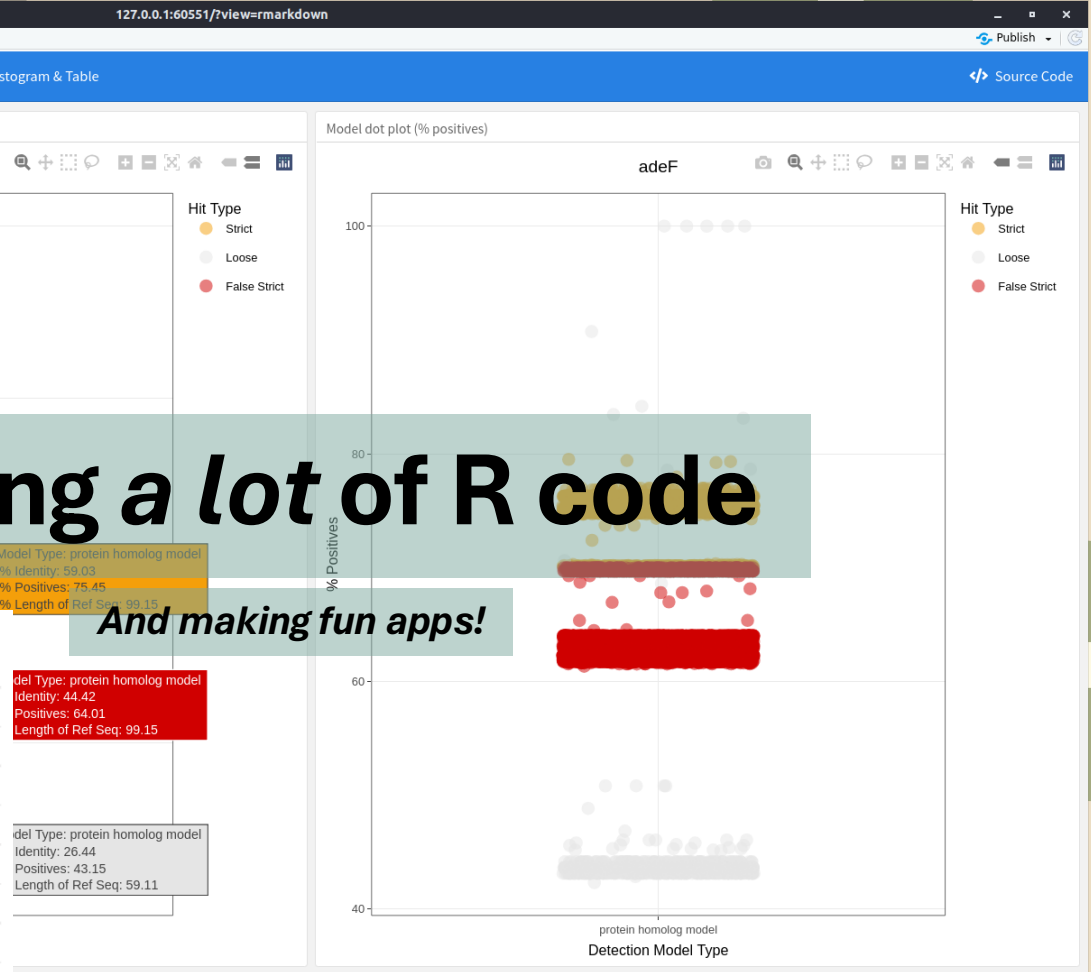
Writing *a lot* of R code

And making fun apps!

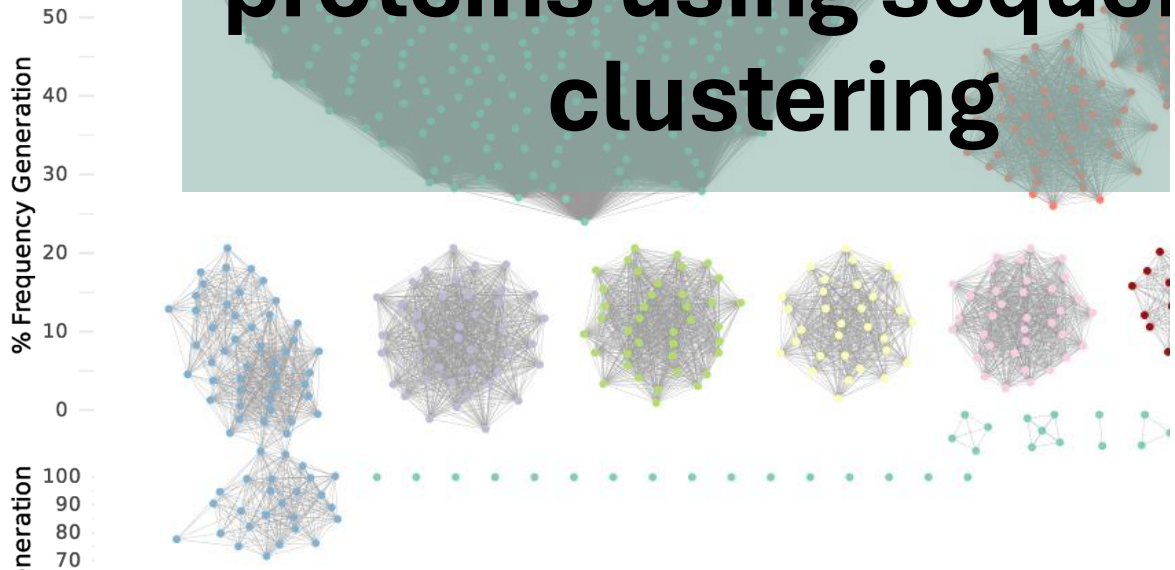


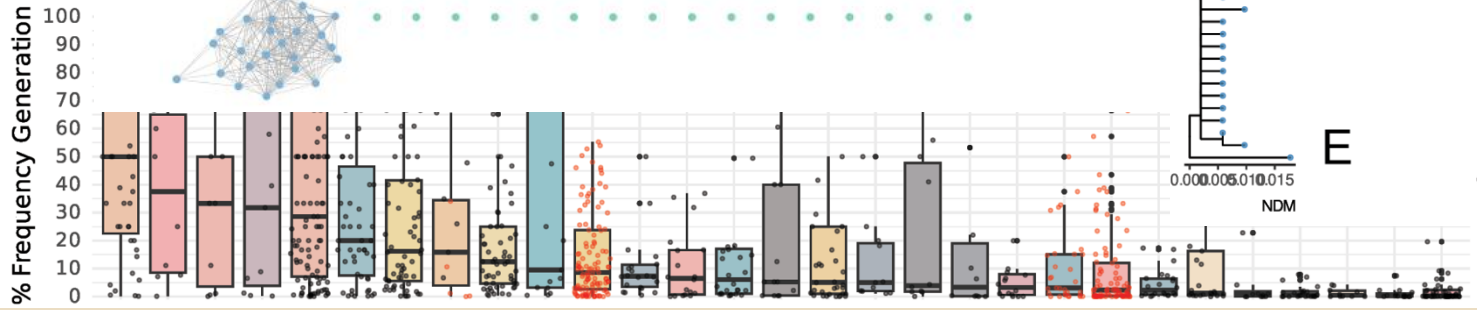
Writing *more* R code to generate figures


HomologModel.py - Untitled (Workspace) - V
File Edit Selection View Go Run Terminal Help
cardjson_cracker.py
msa_prep.py
HomologModel.py 9+
Trying to make sense of proteins using sequence clustering
% Frequency Generation
50
40
30
20
10
0
% Frequency Generation
100
90
80
70
60
50
40
30
20
10
0

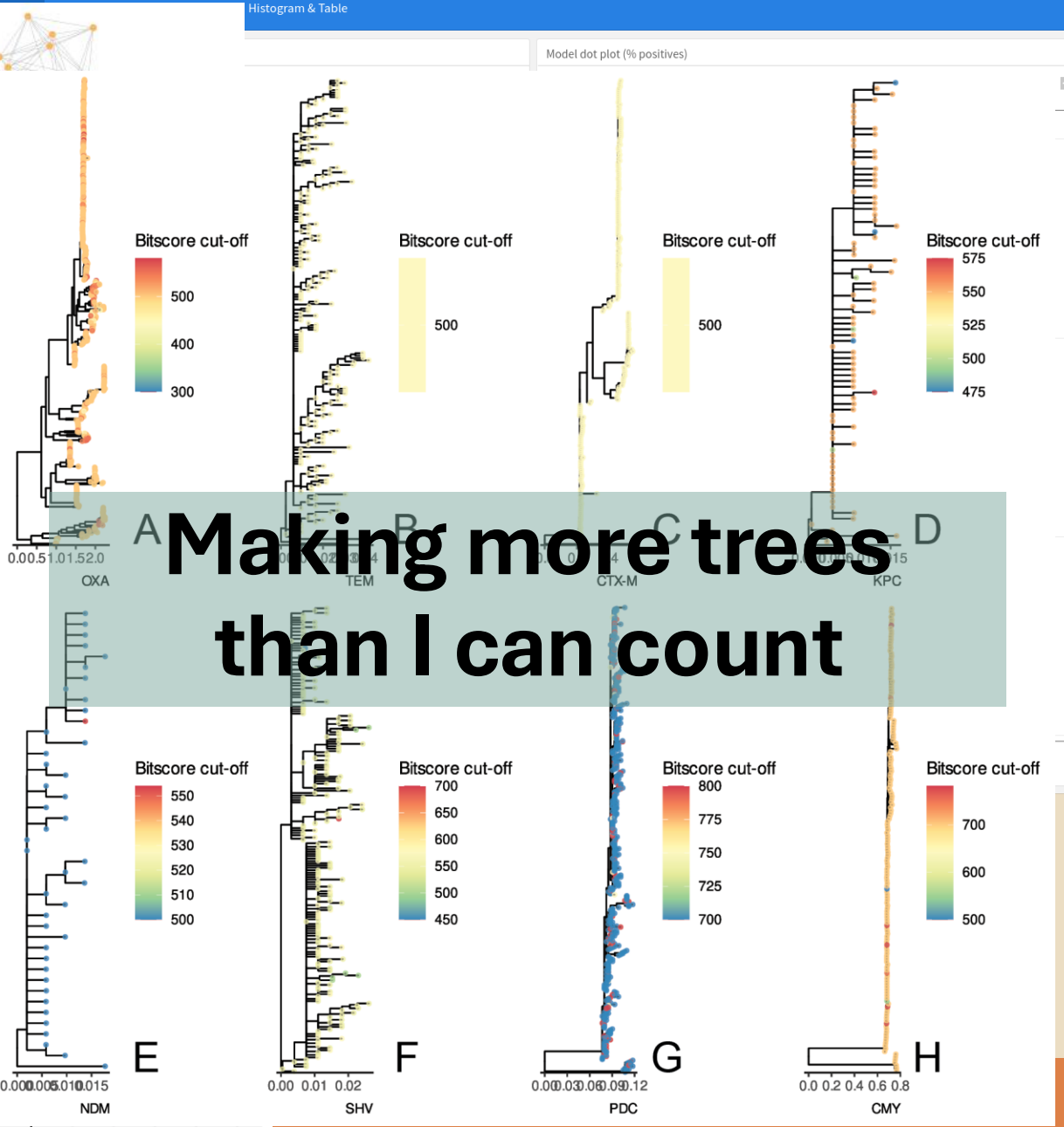

127.0.0.1:60551/?view=rmarkdown
Open in Browser
Publish
Histogram & Table
Source Code
Model dot plot (% positives)
adeF
Hit Type
Strict
Loose
False Strict
ing a lot of R code
And making fun apps!
Model Type: protein homolog model
% Identity: 59.03
% Positives: 75.45
% Length of Ref Seq: 99.15
del Type: protein homolog model
Identity: 44.42
Positives: 64.01
Length of Ref Seq: 99.15
del Type: protein homolog model
Identity: 26.44
Positives: 43.15
Length of Ref Seq: 59.11
% Positives
100
80
60
40
protein homolog model
Detection Model Type

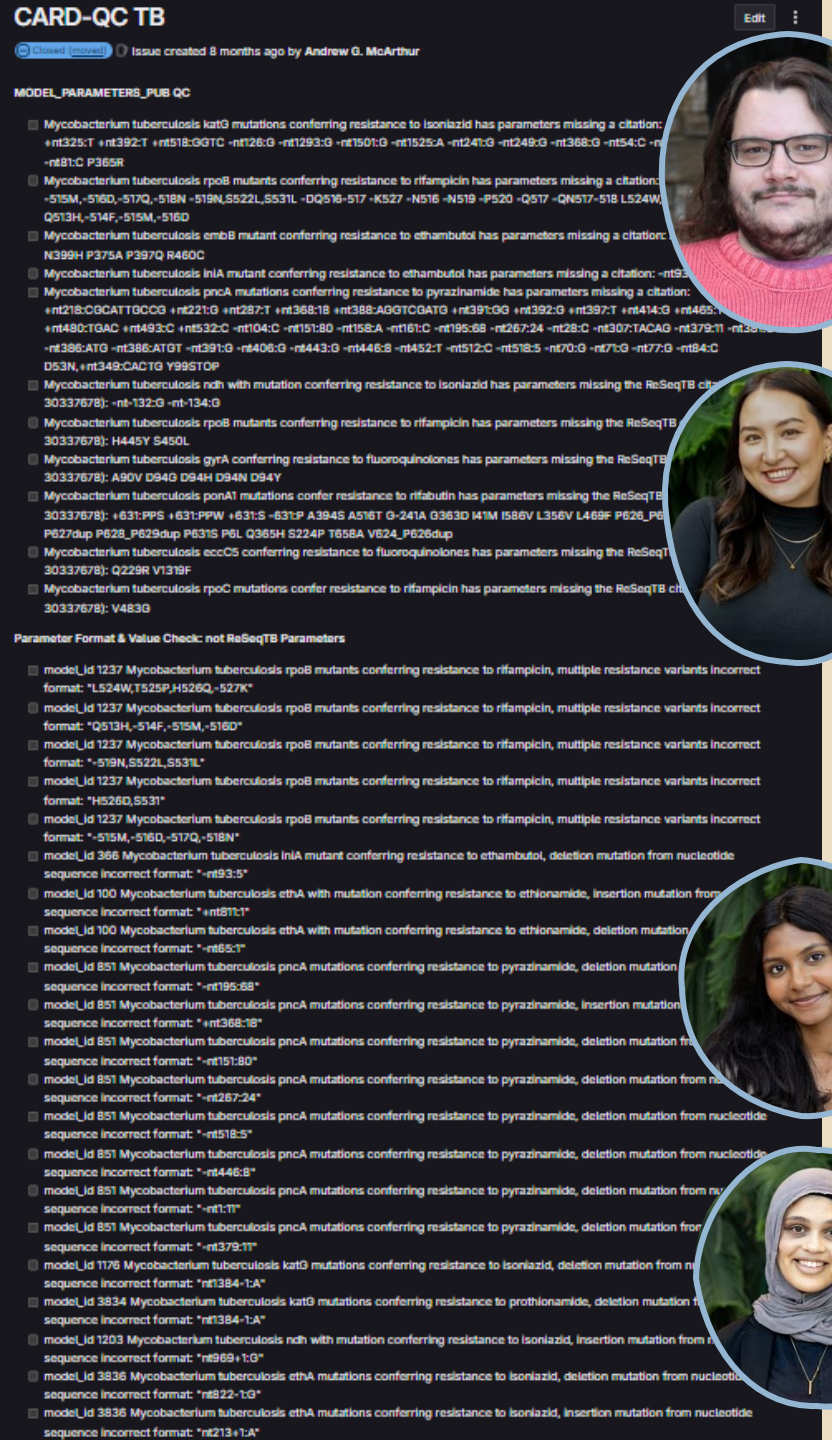
Trying to make sense proteins using sequen clustering





Making more trees than I can count

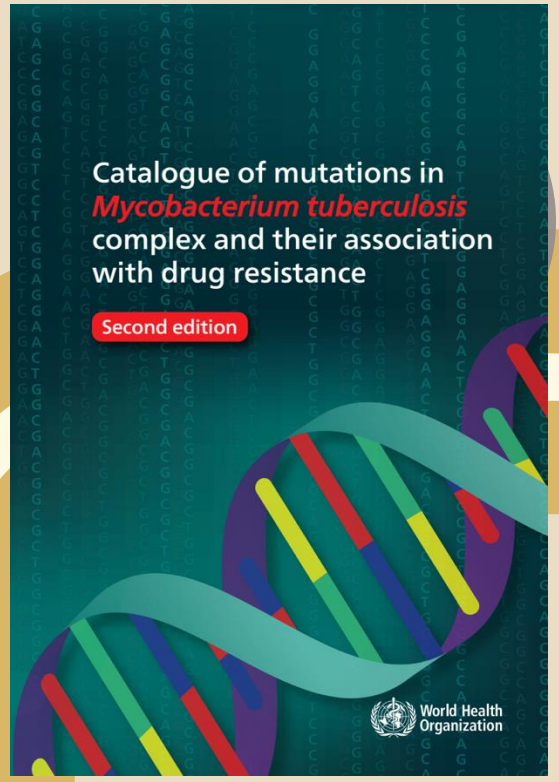




Step 2: Fix the old, add the new

+

CRyPTIC

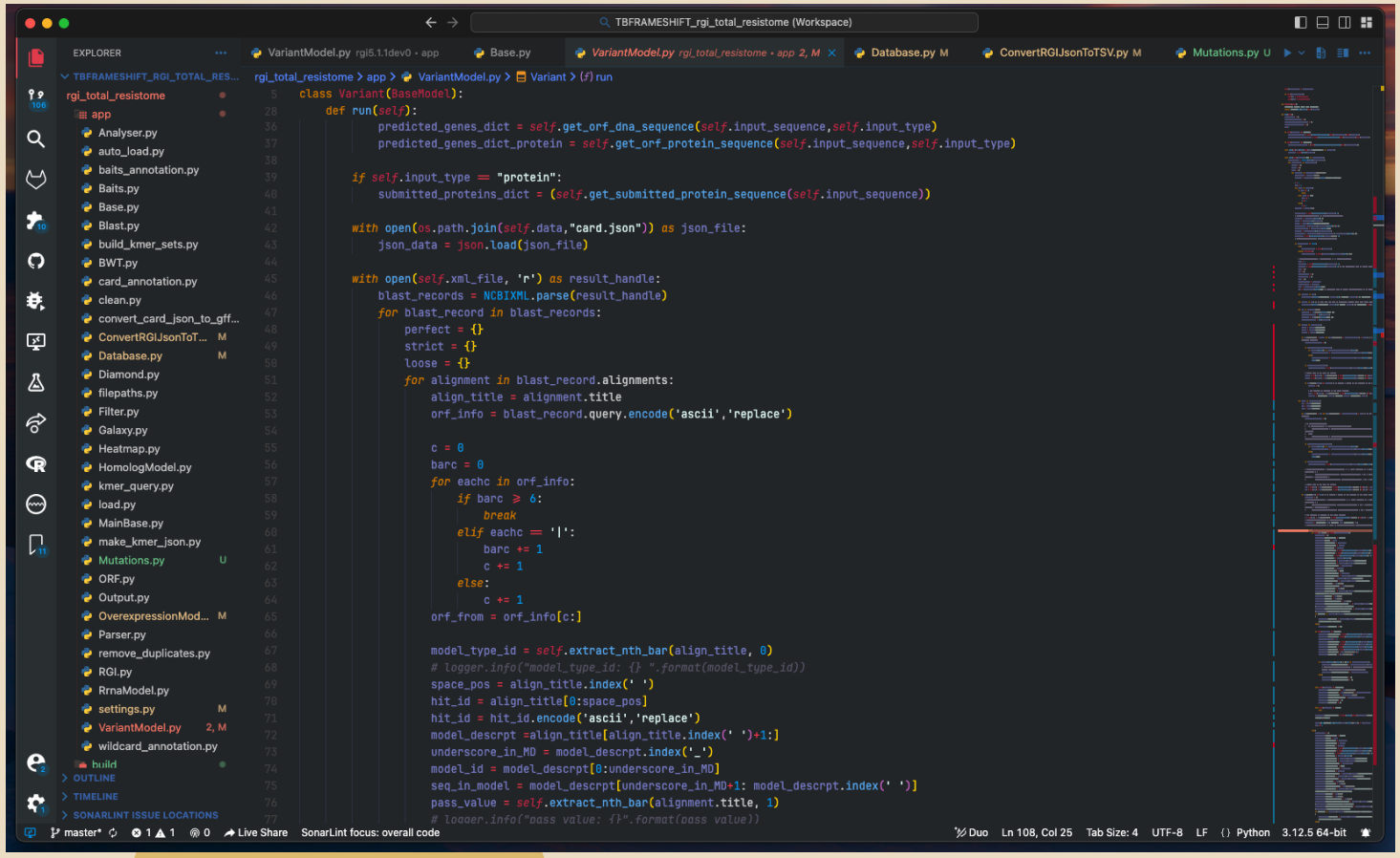


Data from >30,000 MTB patient samples

30,000 MTB mutations

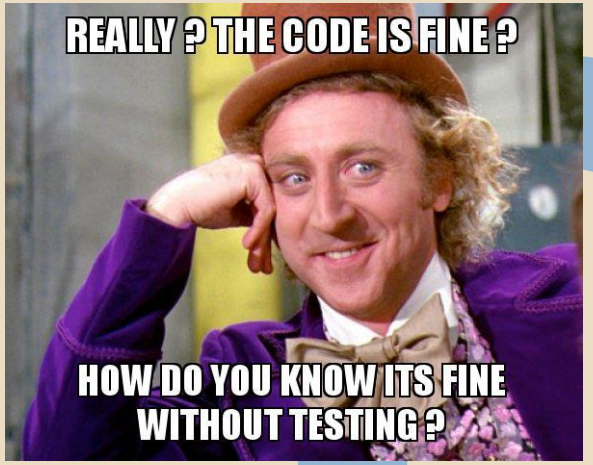
16

Step 3: Write the code (finally)



Almost there...

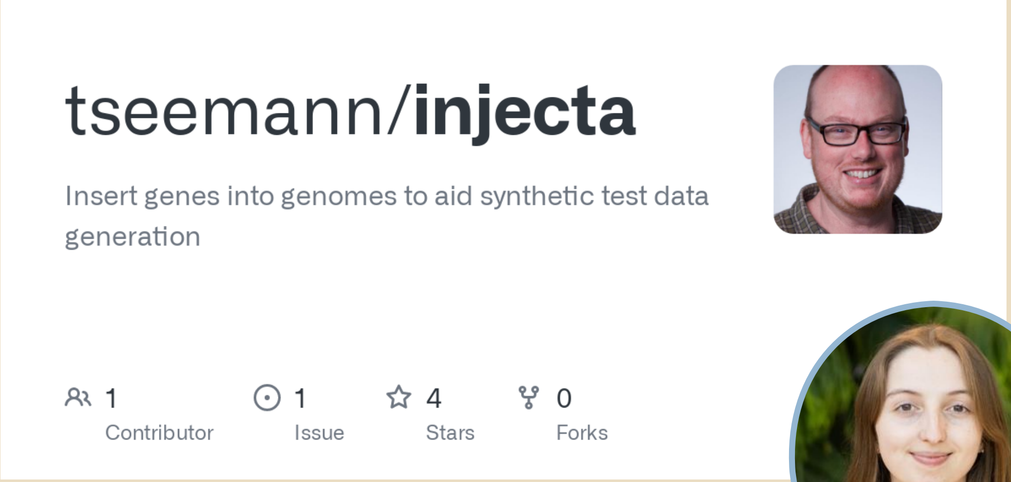
- RGI is **production-level code**
- All **object-oriented Python**
- New modules need accompanying **testing code**

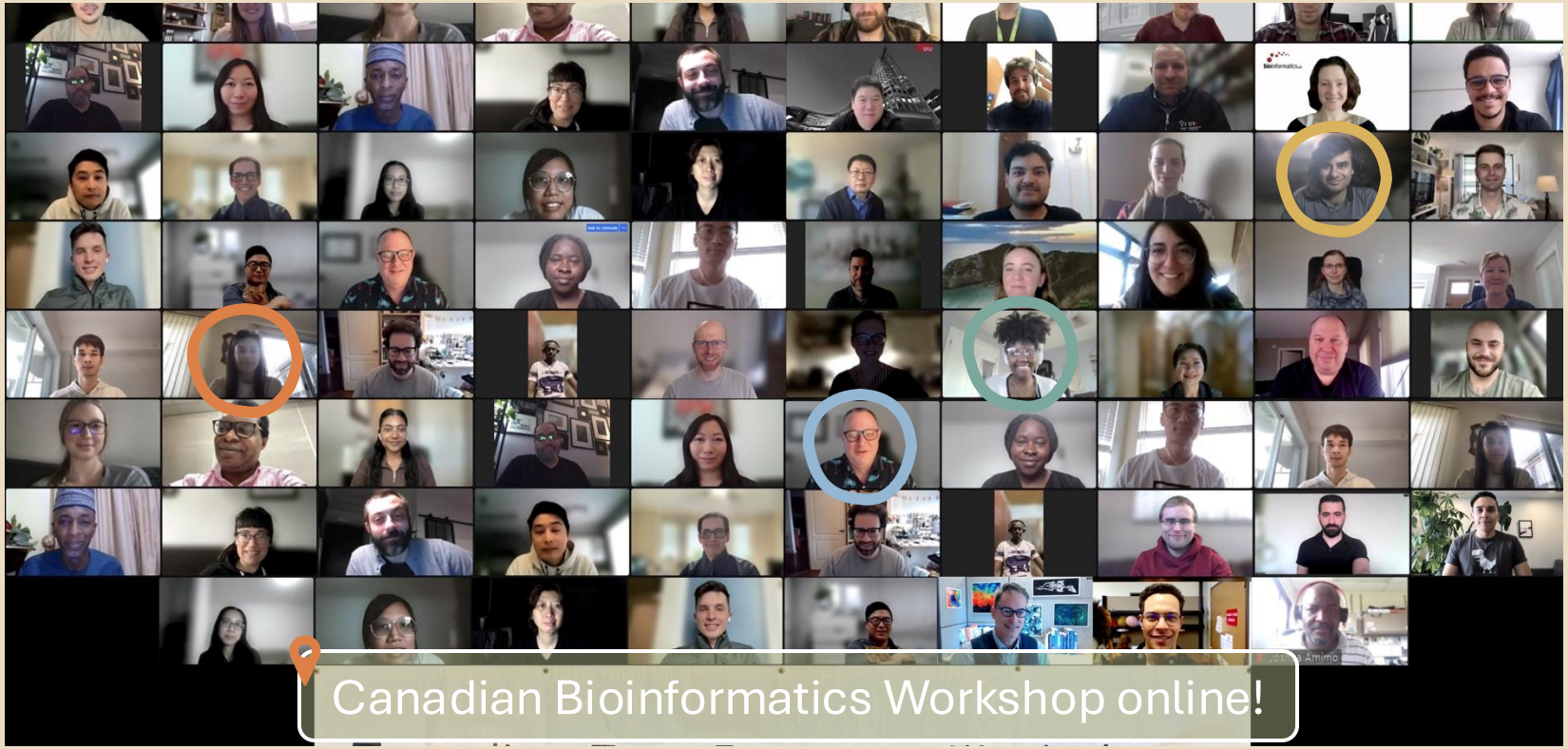


17 Step 4: Validate the code with data

Real datasets

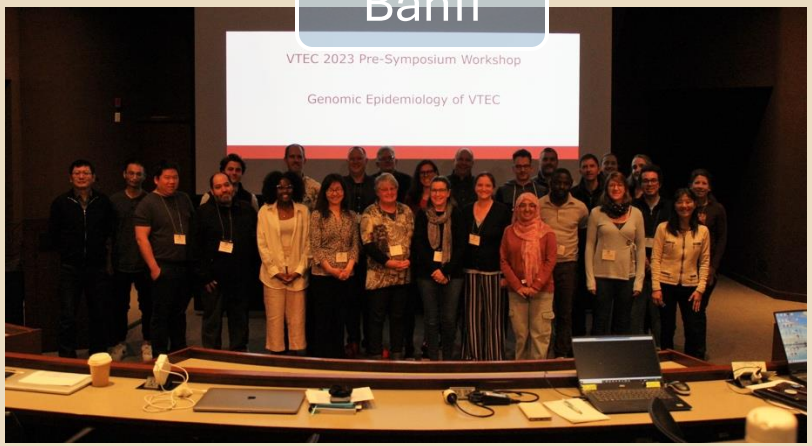
Synthetic datasets

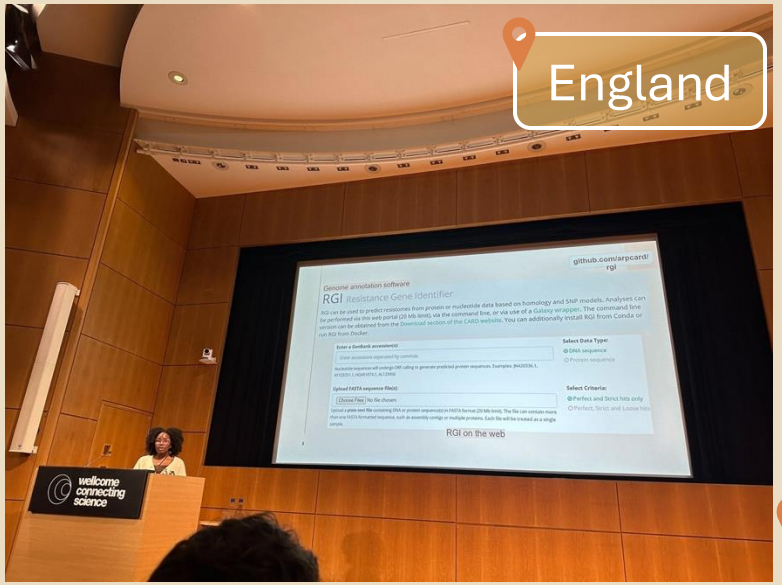




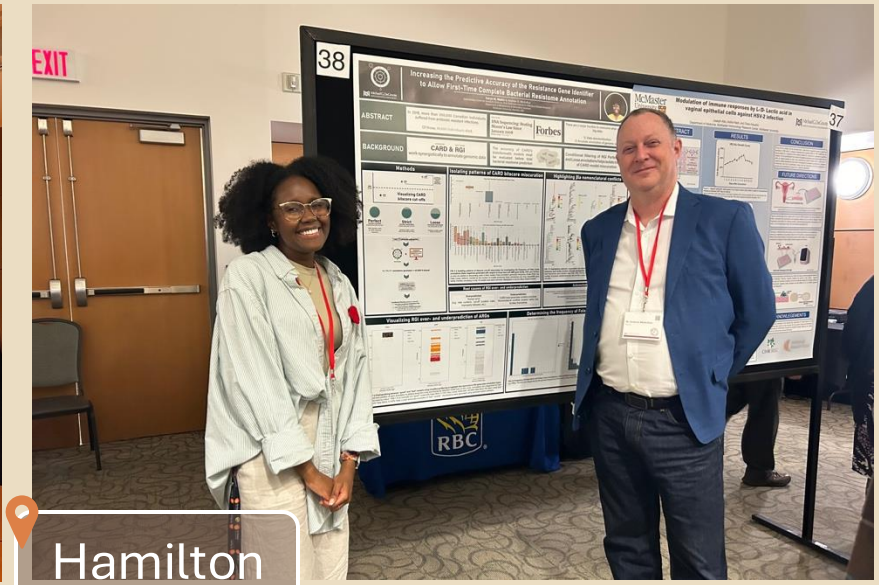
Canadian Bioinformatics Workshop online!

Banff





England



Hamilton

Halifax

England