STAP-AMRTime

(SNP Testing and Prediction)

Detecting rRNA and Protein Variants Conferring AMR in metagenomics data

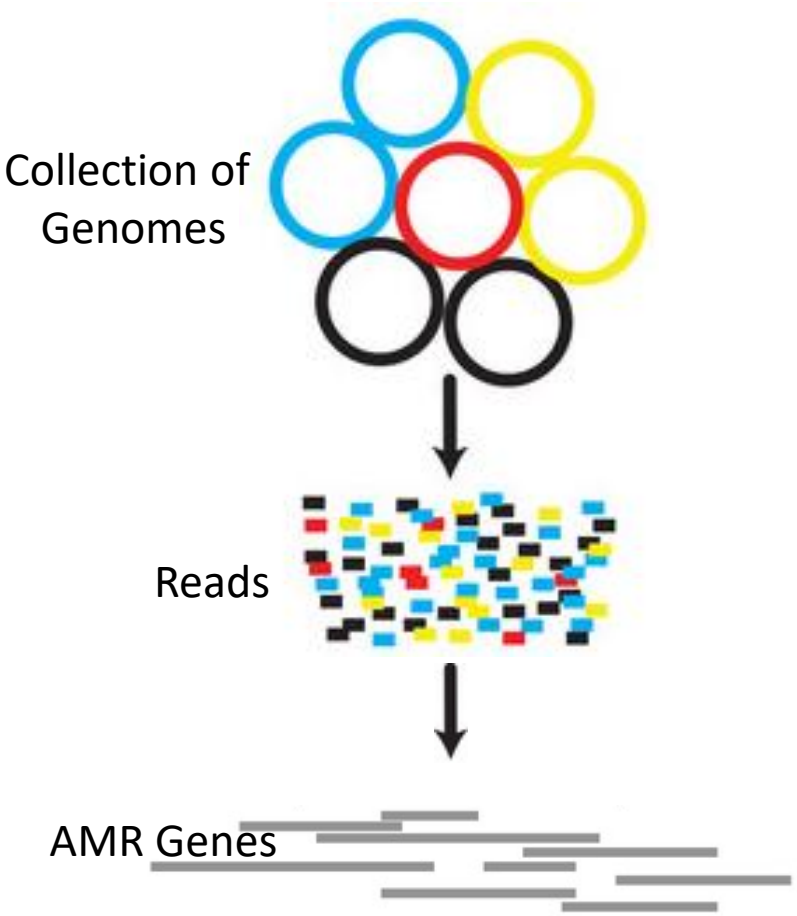
Jan 24th
Brinkman Lab Meeting

Antimicrobial Resistance (AMR) is a Problem Improved AMR Surveillance.

http://www.agriculture.gov.au/SiteCollectionDocuments/animal-plant/animal-health/amria.pdf

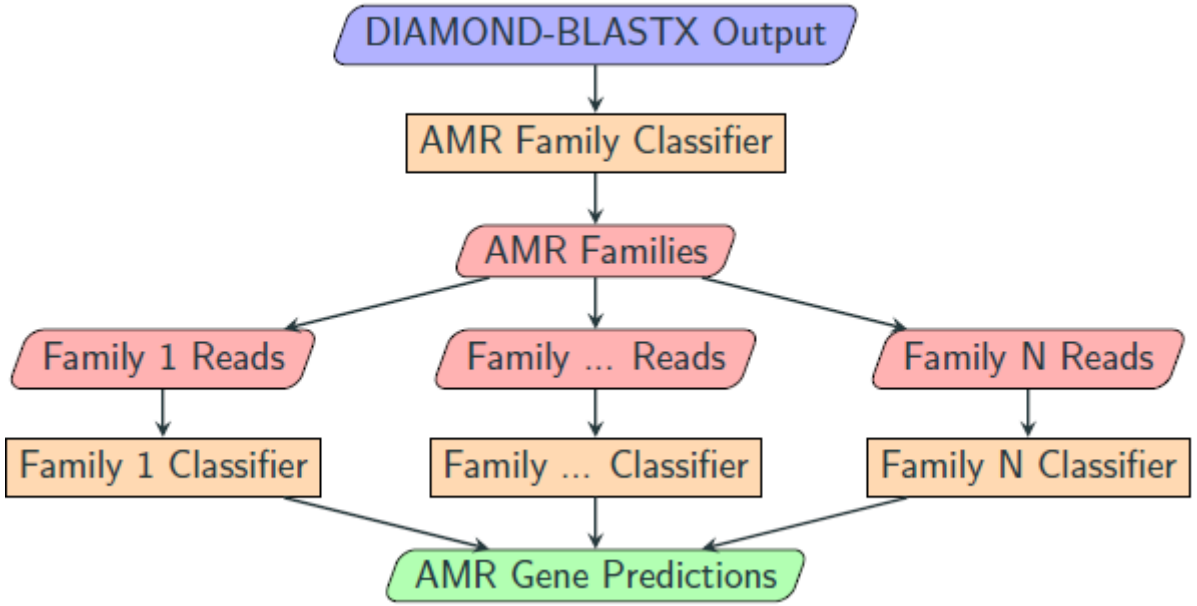
- Using AMR incidence, prevalence and epidemiological trends to prevent infectious disease outbreaks
- One-health
  - Approach to designing and implementing programmes, policies, legislation and research in which multiple sectors communicate and work together to achieve better public health outcomes.

Shotgun Metagenomics is Central to Surveillance



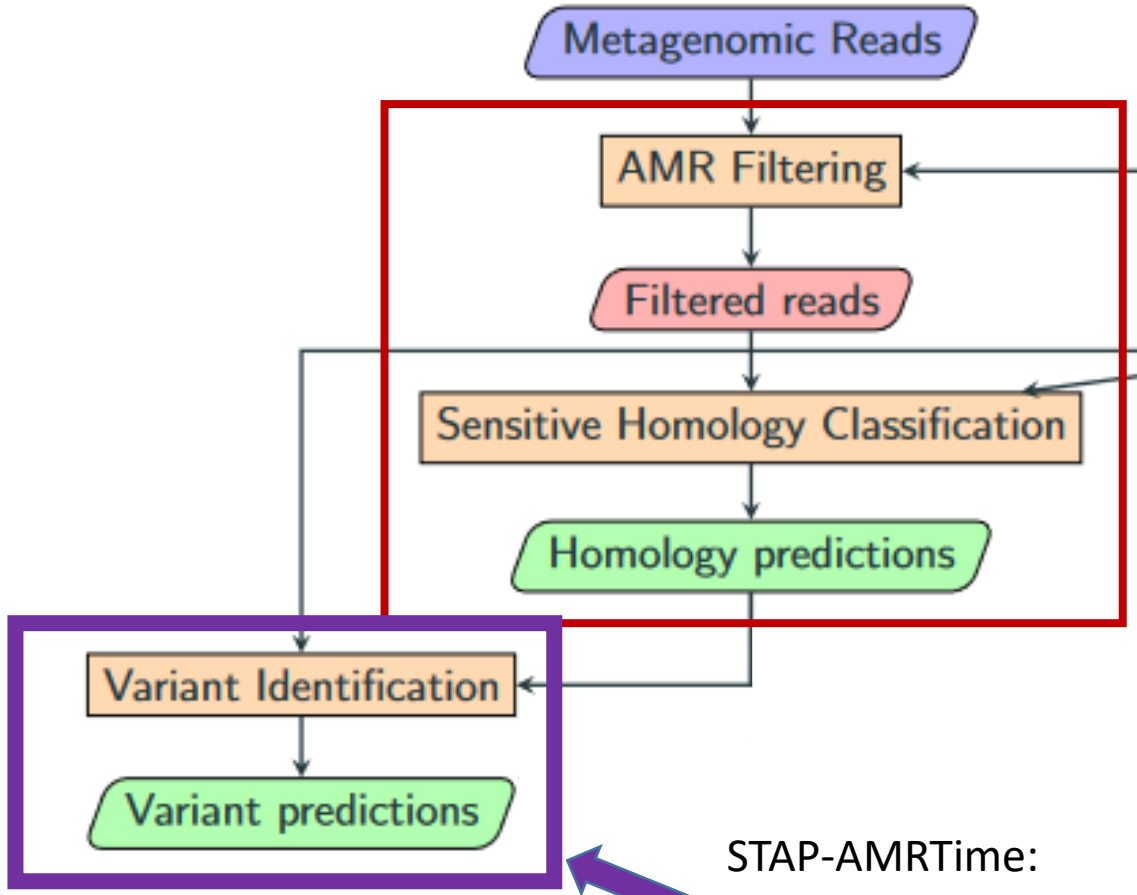
- Genomic techniques that profile microbes as a community (microbiome)
  - Culturing not necessary
- A priori information on pathogen not required
  - Identifying rare and novel pathogens
- Not just limited to humans - enables characterization of environment relevant to one-health
  - E.g. health care facilities, animal farms
- Problem: short reads and lots of noise

AMRTime:
Developing a metagenomic AMR predictor



*(Maguire et al. unpublished)

AMRTime: Tool for Detection of AMR Genes From Metagenomic Data.



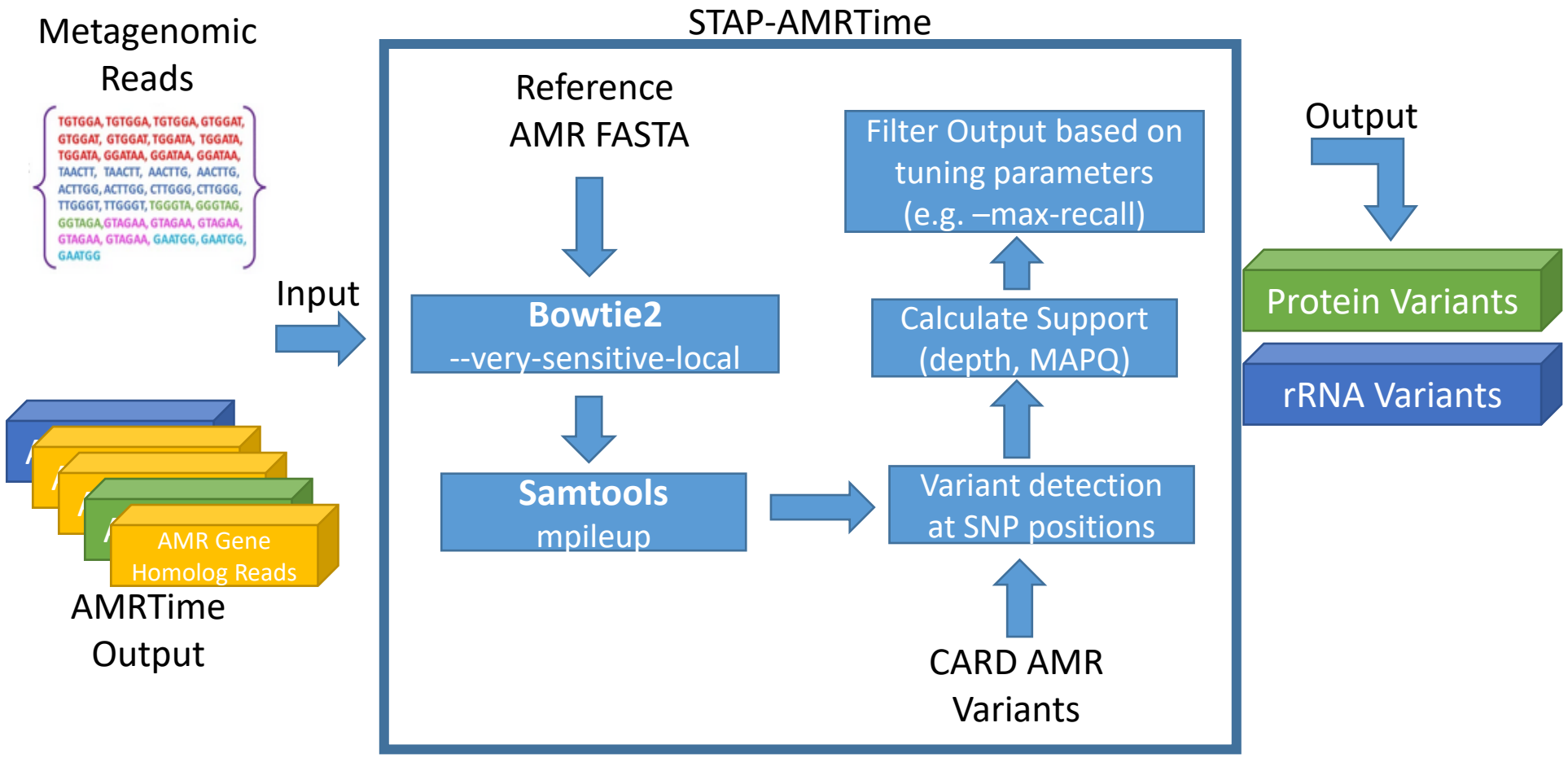
AMRTime:
- Diamond BLASTX
- AMR Homologs

CARD

STAP-AMRTime:
- SNP Testing and Prediction
- AMR Variant Detection

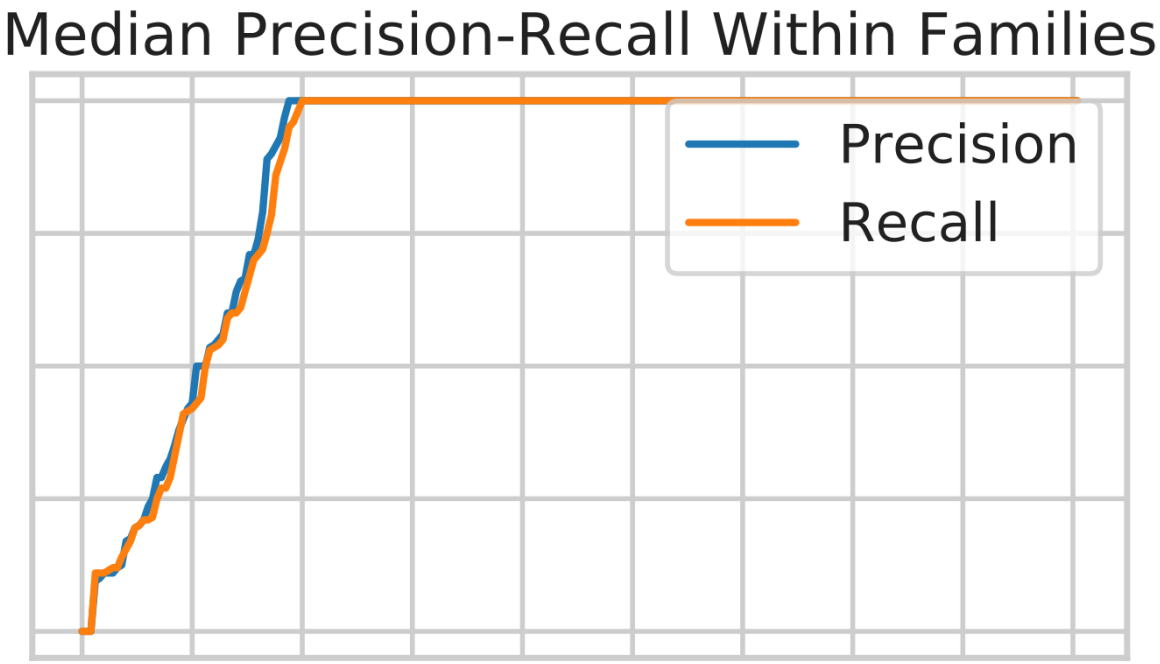
(Jia et al. In Preparation)
*Collaboration with Dr. Finlay McGuire @ Dalhousie

STAP-AMRTime



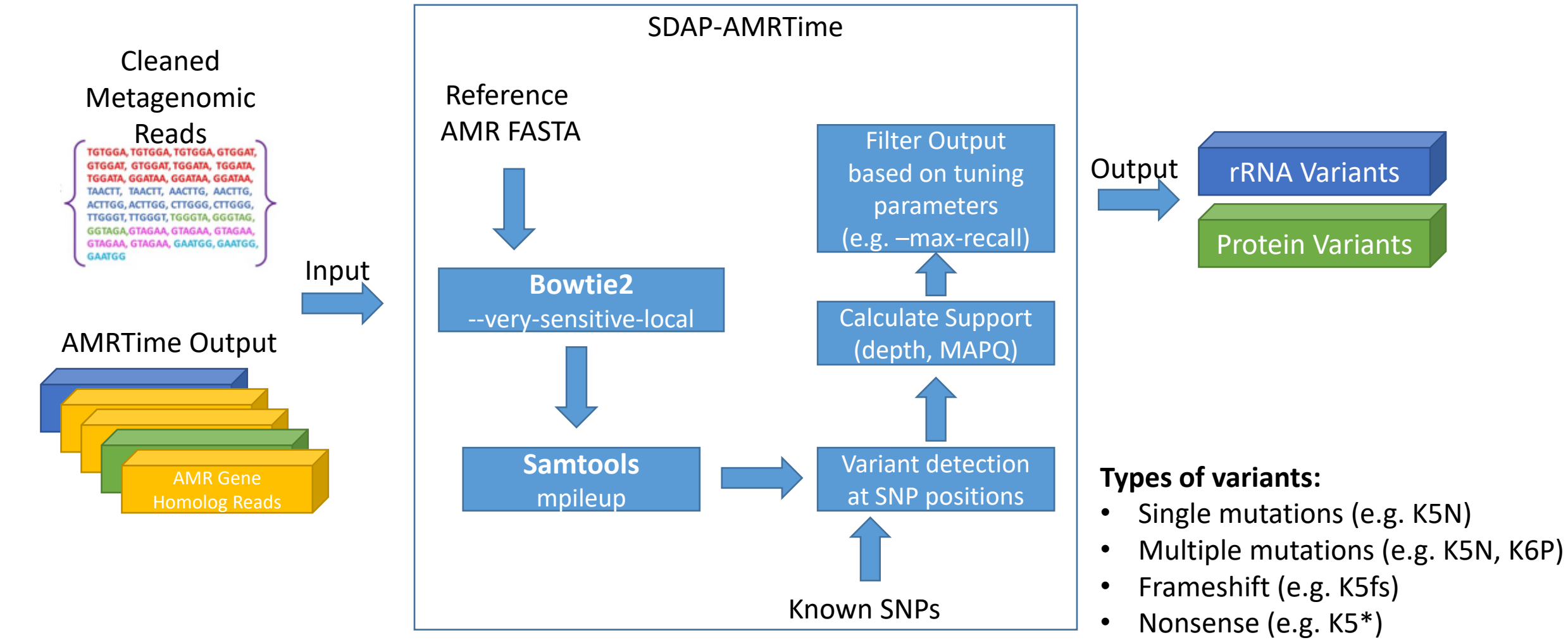
(https://github.com/bfjia/ProjectAMRVariant)

AMRTime is Not Perfect



- Suite of other tools available that does AMR prediction in metagenomic data.
- **No tools available to predict protein variants conferring resistance in metagenomic data**.

STAP-AMRTime Workflow



Types of variants:

- Single mutations (e.g. K5N)
- Multiple mutations (e.g. K5N, K6P)
- Frameshift (e.g. K5fs)
- Nonsense (e.g. K5*)
- Indels (e.g. K5+/-N)

https://github.com/imasianxd/ProjectAMRVariant

STAP-AMRTime Example Output (.tsv)

ARO	VariantClass	VariantType	ResistantVariant	SNP	Depth	AbsSupport	RelativeSupport	INFO
3004133	rna variant	Single	TRUE	A2058T	3	3	100.00%	T:3
3003285	protein variant	Single	TRUE	A473T	49	38	77.55%	A:19;A:8;T:30;T:8
3003735	protein variant	Single	TRUE	V90I	128	32	25.00%	V:75;V:51;I:16;I:16
3003902	protein variant	Nonsense	TRUE	W228*	79	1	1.27%	W:77;*:1;L:1

Evaluation

- 2 x Synthetic – *in silico* generated
- 2 x Mock Community – spiked in communities sequenced on MiSeq

Accuracy Metrics:

Synthetic Data (Ground Truth):

1. Resistant Variants (Q3K, Q3STOP, Q3Frameshift)
2. Non-Resistant Variants(Q3Q, Q3C…etc.)

Predictions:

1. Resistant Variants (K,*,fs)
2. Non-Resistant Variants (Q, C…etc)

T\P	Resistant	Non-Resistant
Resistant	X	X
Non-Resistant	X	X

Synthetic – Optimization Set

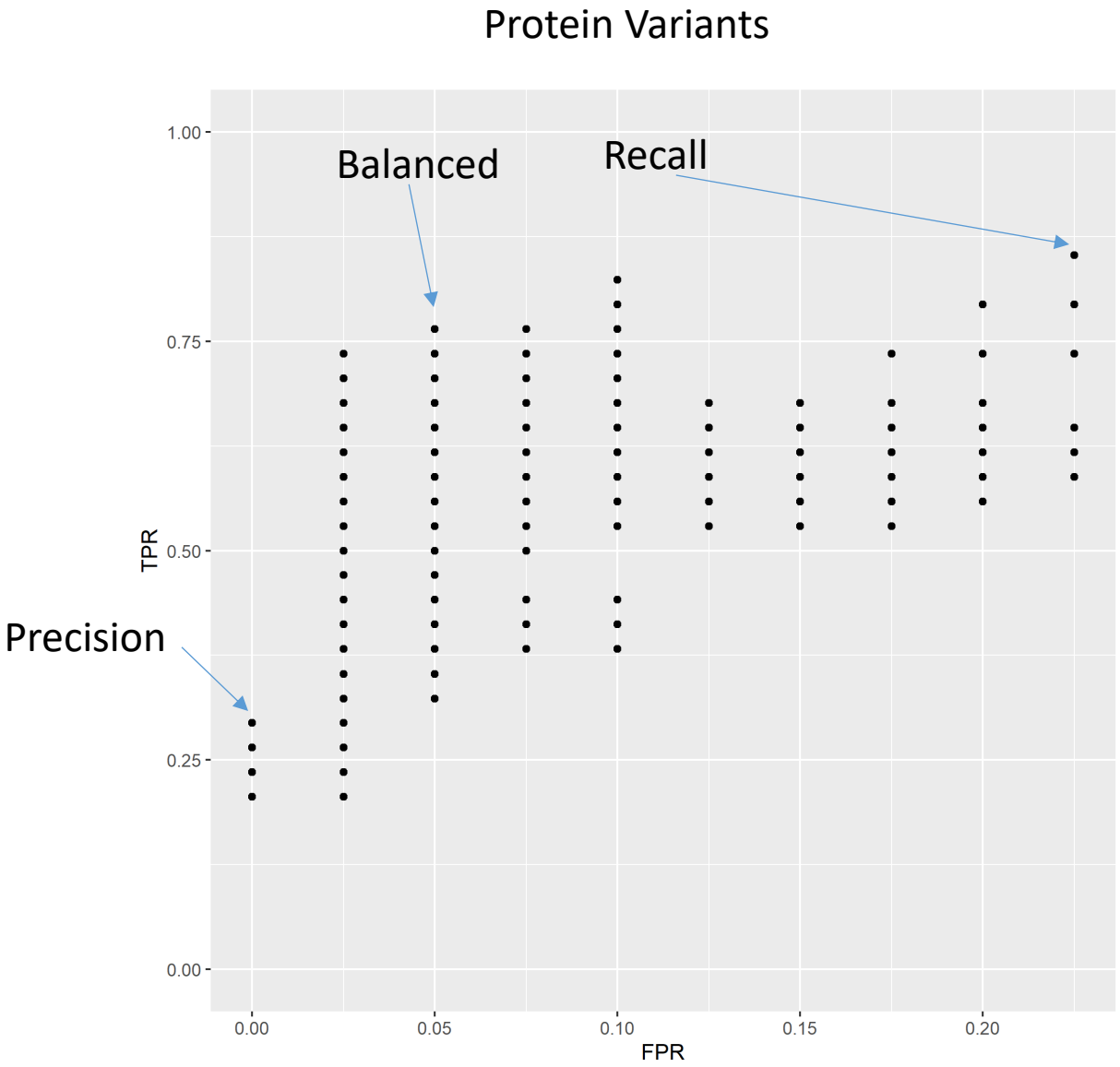
- 30 bacterial species from MAG Assessment Manuscript[1]

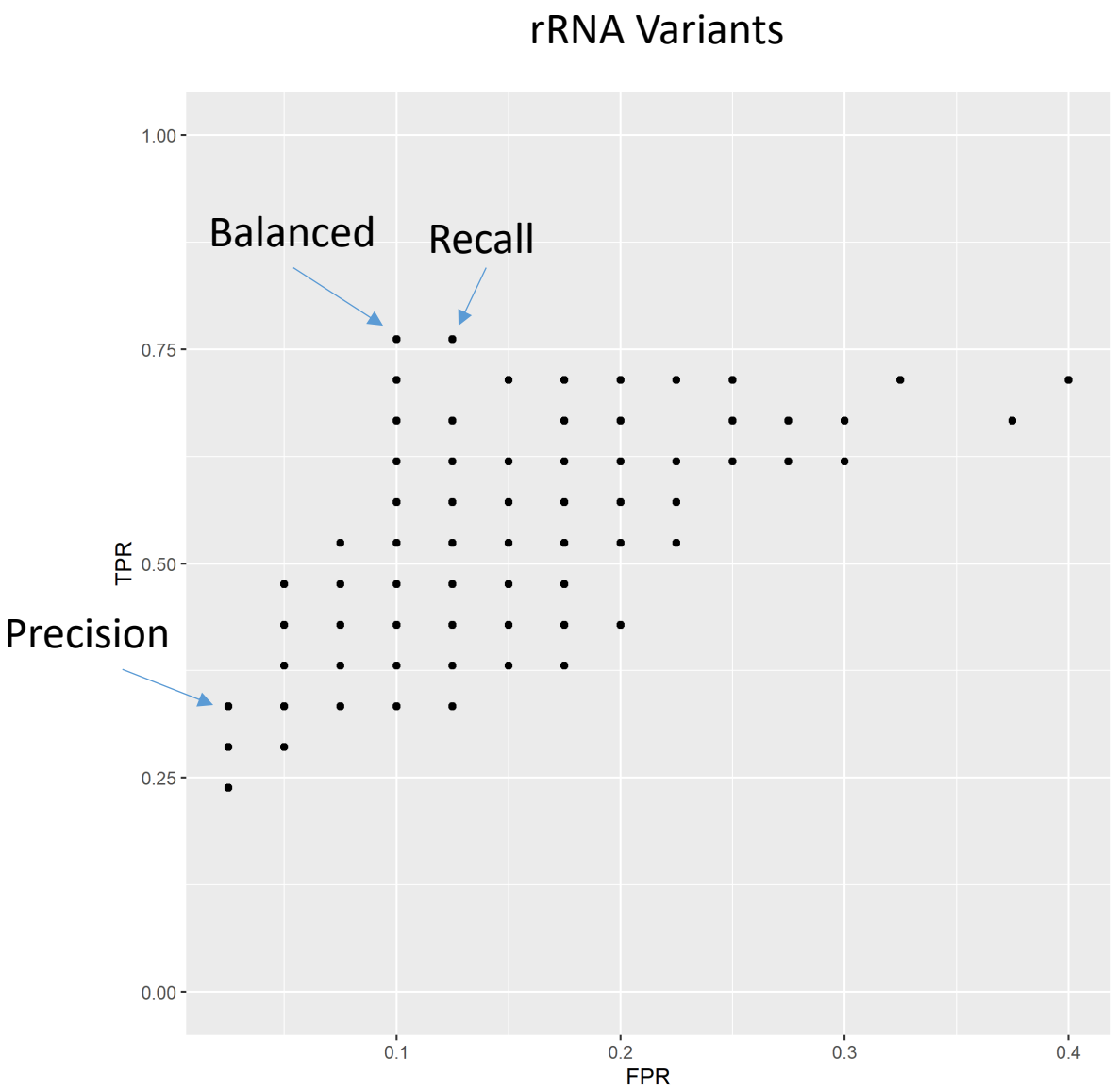
- Known resistant and non-resistant variant inserted at random position of reference genomes.
  - 34 resistant + 40 non-resistant protein variants
  - 21 resistant + 40 non-resistant rRNA variants

- Simulate Metagenome[2]
  - Genomes: log normal abundance distribution ($\mu = 1$, $\sigma = 2$)
  - Plasmids: randomly assigned copy number (low, medium, high) and scaled with gamma distribution ($\alpha = 4$, $\beta = 1$)

1. (Maguire and Jia et al. PMID: 33001022)
2. (https://github.com/fmaguire/AMR_Metagenome_Simulator)

ROC curves


Protein Variants
Balanced
Recall
Precision
TPR
FPR
1.00
0.75
0.50
0.25
0.00
0.00
0.05
0.10
0.15
0.20


rRNA Variants
Balanced
Recall
Precision
TPR
FPR
1.00
0.75
0.50
0.25
0.00
0.1
0.2
0.3
0.4

Synthetic – Optimization Set

Protein Variants

Balanced

T\P	Resistant	Non-Resistant
Resistant	27	7
Non-Resistant	2	38

TPR=0.82
PPV=0.86
MCC=80.88

Max Recall

T\P	Resistant	Non-Resistant
Resistant	29	5
Non-Resistant	9	31

TPR=0.86
PPV=0.76
MCC=70.20

Max Precision

T\P	Resistant	Non-Resistant
Resistant	10	24
Non-Resistant	0	40

TPR=0.31
PPV=1
MCC=32.88

Synthetic – Optimization Set

rRNA Variants

Balanced

T\P	Resistant	Non-Resistant
Resistant	16	5
Non-Resistant	4	36

TPR=0.76
PPV=0.8
MCC=50.33

Max Recall

T\P	Resistant	Non-Resistant
Resistant	16	5
Non-Resistant	4	36

TPR=0.76
PPV=0.8
MCC=50.33

Max Precision

T\P	Resistant	Non-Resistant
Resistant	7	14
Non-Resistant	0	40

TPR=0.33
PPV=1
MCC=33.45

Synthetic 2:
CAMI2 Human Gastrointestinal Tract Toy Sets[1]

- Simulated Illumina HiSeq metagenome data (2 x 150bp)
- 2 randomly chosen samples (sample 1 and sample 11)

1. (Meyer et al. https://www.biorxiv.org/content/10.1101/2021.07.12.451567v1)

Synthetic – CAMI Gut Sample 1

Truth

3003378	protein overexpression	Y137H, G103S
3004562	protein variant model	R377G, I139R
3003369	protein variant model	R234F
3004562	protein variant model	I139R
3003889	protein variant model	E448K
3004446	protein variant model	D350N, S357N

Found in:
Precision Mode
Balanced Mode
Recall Mode

Balanced

T\P	Resistant	Non-Resistant
Resistant	7	1
Non-Resistant	1	-

88% Recall
88% Precision

Max Recall

T\P	Resistant	Non-Resistant
Resistant	7	1
Non-Resistant	6	-

Max Precision

T\P	Resistant	Non-Resistant
Resistant	5	3
Non-Resistant	1	-

62% Recall
83% Precision

Synthetic – CAMI Gut Sample 11

Truth

3003369	protein variant model	R234F
3003283	protein variant model	L511R
3004562	protein variant model	I139R
3003685	protein variant model	A473V
3004562	protein variant model	R377G, I139R
3004446	protein variant model	D350N, S357N
3004334	protein variant model	S83F
3003889	protein variant model	E448K
3003684	protein variant model	D87N
3003890	protein variant model	E350Q
3003582	protein variant model	L71R
3003378	protein overexpression	Y137H, G103S, S3N
3000818	protein overexpression	S209R, G71E

Found in:

Precision Mode

Balanced Mode

Recall Mode

Balanced

T\P	Resistant	Non-Resistant
Resistant	14	3
Non-Resistant	4	-

83% Recall
78% Precision

Max Recall

T\P	Resistant	Non-Resistant
Resistant	14	3
Non-Resistant	8	-

Max Precision

T\P	Resistant	Non-Resistant
Resistant	3	14
Non-Resistant	0	-

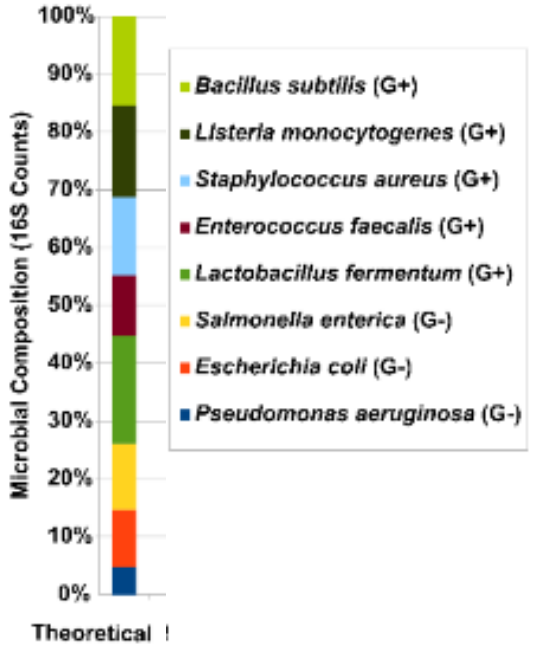
18% Recall
100% Precision

Stats:

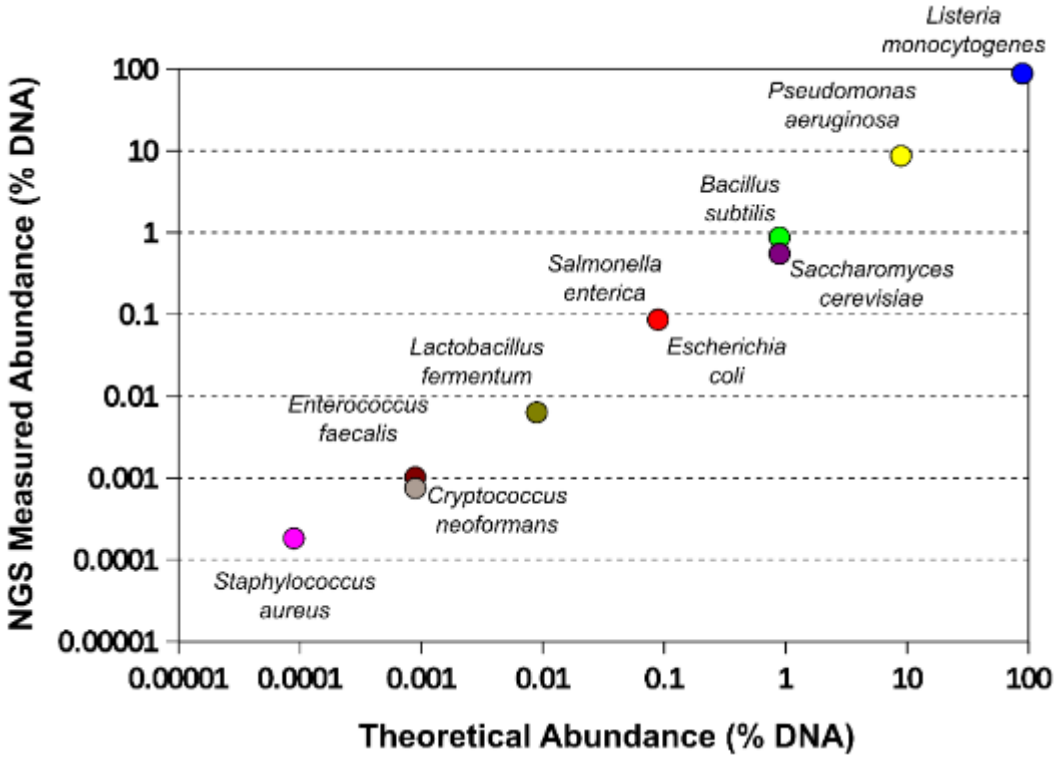
- 16 Threads
- 209s tun time
- Peak memory: 2G

Mock Data 1: Zymo Microbial Community Standard

Zymo Even



Zymo Log



(Zymo Research; Catalog D6300/D6305 Data Sheet)

Mock – Zymo Even

Truth

ARO	Type	SNP
3003369	protein variant model	R234F
3003582	protein variant model	L71R
3000818	protein overexpression model	G71E
3003889	protein variant model	E448K
3003890	protein variant model	E350Q
3004446	protein variant model	D350N, S357N

Balanced

T\P	Resistant	Non-Resistant
Resistant	5	1
Non-Resistant	2	-

83% Recall
71% Precision

Max Recall

T\P	Resistant	Non-Resistant
Resistant	5	1
Non-Resistant	16	-

Max Precision

T\P	Resistant	Non-Resistant
Resistant	4	2
Non-Resistant	0	-

66.7% Recall
100% Precision

Found in:
Precision Mode
Balanced Mode
Recall Mode

Stats:
- 16 Threads
- 181s tun time
- Peak memory: 1.5G

(Nicholls et al. https://doi.org/10.1093/gigascience/giz043)

Mock – Zymo Log

Truth

ARO	Type	SNP
3003369	protein variant model	R234F
3003582	protein variant model	L71R
3000818	protein overexpression model	G71E
3003889	protein variant model	E448K
3003890	protein variant model	E350Q
3004446	protein variant model	D350N, S357N

Balanced

T\P	Resistant	Non-Resistant
Resistant	4	2
Non-Resistant	2	-

66.7% Recall
66.7% Precision

Max Recall

T\P	Resistant	Non-Resistant
Resistant	5	1
Non-Resistant	20	-

Max Precision

T\P	Resistant	Non-Resistant
Resistant	3	3
Non-Resistant	0	-

50% Recall
100% Precision

Found in:
Precision Mode
Balanced Mode
Recall Mode

(Nicholls et al. https://doi.org/10.1093/gigascience/giz043)

Mock Data 2: Peabody et al.

Genus	Species	Strain	
Bacillus	*amyloliquefaciens*	FZB42	
Bacillus	*cereus*	ATCC 14579	
Burkholderia	*cenocepacia*	J2315	
~~*Escherichia*~~	~~*coli*~~	~~K-12~~	*
~~*Frankia*~~	~~*sp.*~~	~~Cel3~~	*
Micrococcus	*luteus*	NCTC 2665	
~~*Pseudomonas*~~	~~*aeruginosa*~~	~~PAO1~~	*
Pseudomonas	*aeruginosa*	UCBPP-PA14	
~~*Pseudomonas*~~	~~*fluorescens*~~	~~Pf-5~~	*
Pseudomonas	*putida*	KT2440	
Rhodobacter	*capsulatus*	SB 1003	
Streptomyces	*coelicolor*	A3(2)	

*No reference genome specified in manuscript.

(Peabody et al.; https://doi.org/10.1186/s12859-015-0788-5)

Mock – Peabody et al.

Truth

ARO	Type	SNP
3000818	protein overexpression	S209R, G71E

Balanced

T\P	Resistant	Non-Resistant
Resistant	1	0
Non-Resistant	1	-

Max Recall

T\P	Resistant	Non-Resistant
Resistant	1	0
Non-Resistant	8	-

Max Precision

T\P	Resistant	Non-Resistant
Resistant	1	0
Non-Resistant	0	-

Found in:
Precision Mode
Balanced Mode
Recall Mode

(Nicholls et al. https://doi.org/10.1093/gigascience/giz043)

Availability:

- Currently: Conda environment + git repo
- Future: Conda Package

- Object oriented Python Program
  - If any new variant class or types gets added, its very easy to modify to include them.

Caveats, Limitations, and Future Directions

- Sequencing depth, sequencing depth, sequencing depth
  - Higher depth = better fine tuning = better accuracy
  - **Need for some user interpretation**

ARO	VariantCla	VariantTyp	ResistantV	SNP	Depth	AbsSuppo	RelativeSu	INFO	
3003995	protein va	Single	TRUE	L345I	5	1	20.00%	L:3;I:1	FP
3003889	protein va	Single	TRUE	E448K	8	7	87.50%	K:7	TP
3003285	protein va	Single	TRUE	A473T	2	1	50.00%	T:1	?

- 86% Recall cap, Why?
  - No alignment with reference gene?
  - Reads aligned but no variants found?

- Can precision be further improved?
  - Benefit of using report all mapped reads parameter in bowtie2 (-a)
  - Benefit of local versus end-to-end alignment mode in bowtie2 (--local)

- Simple ML based filtering model?