Using Machine Learning to Predict Antimicrobial Minimum Inhibitory Concentrations and Associated Genomic Features for Nontyphoidal *Salmonella*

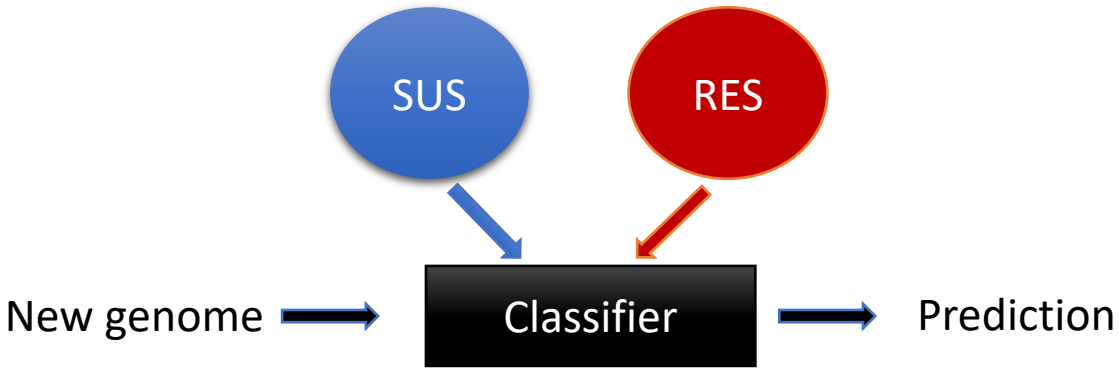
James J. Davis
Argonne National Laboratory
University of Chicago
jimdavis@uchicago.edu

Overall Goals

- Develop tools to predict Antimicrobial Resistance (AMR) phenotypes given a genome
- Predict genomic regions associated with AMR
- Use these tools for genome annotation (PATRIC and RAST)
- See if predictions can help guide research

- **Longer term:**
  - Could point of care sequencing with AMR prediction be a diagnostic?
  - Takes ~36-48hrs to lab-test faster growers like *E. coli* and *K. pneumoniae*

AMR prediction as a machine learning classification problem:



SCIENTIFIC REPORTS

OPEN

Antimicrobial Resistance Prediction in PATRIC and RAST

James J. Davis[1,2,†], Sébastien Boisvert[3], Thomas Brettin[1,2], Ronald W. Kenyon[4], Chunhong Mao[4], Robert Olson[1,2], Ross Overbeek[2,5], John Santerre[6], Maulik Shukla[1,2], Alice R. Wattam[4], Rebecca Will[4], Fangfang Xia[1,2] & Rick Stevens[1,2,6]

Received: 03 February 2016

OXFORD

Briefings in Bioinformatics, 2017, 1–9

doi: 10.1093/bib/bbx083
Paper

PATRIC as a unique resource for studying antimicrobial resistance

Dionysios A. Antonopoulos, Rida Assaf, Ramy Karam Aziz, Thomas Brettin, Christopher Bun, Neal Conrad, James J. Davis, Emily M. Dietrich, Terry Disz, Svetlana Gerdes, Ronald W. Kenyon, Dustin Machi, Chunhong Mao, Daniel E. Murphy-Olson, Eric K. Nordberg, Gary J. Olsen, Robert Olson,

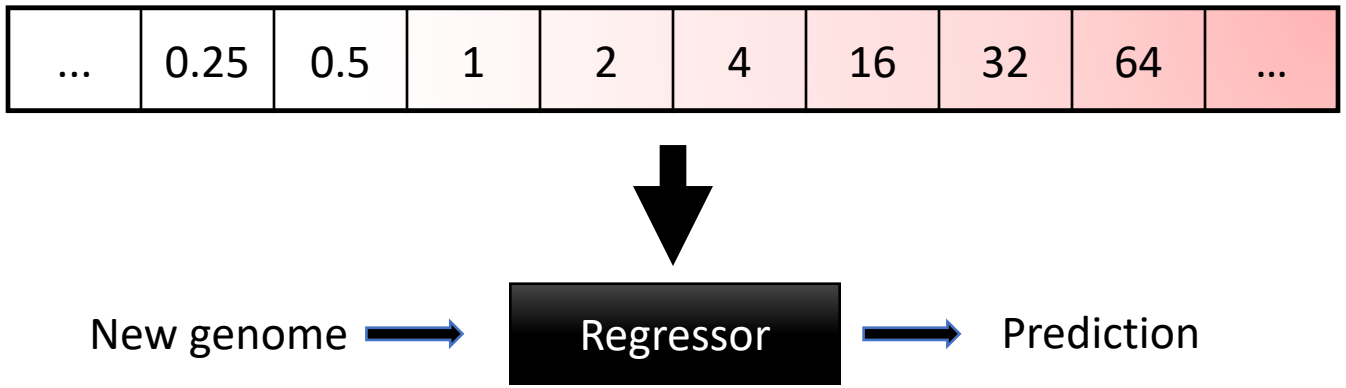
AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing *Klebsiella pneumoniae* Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307

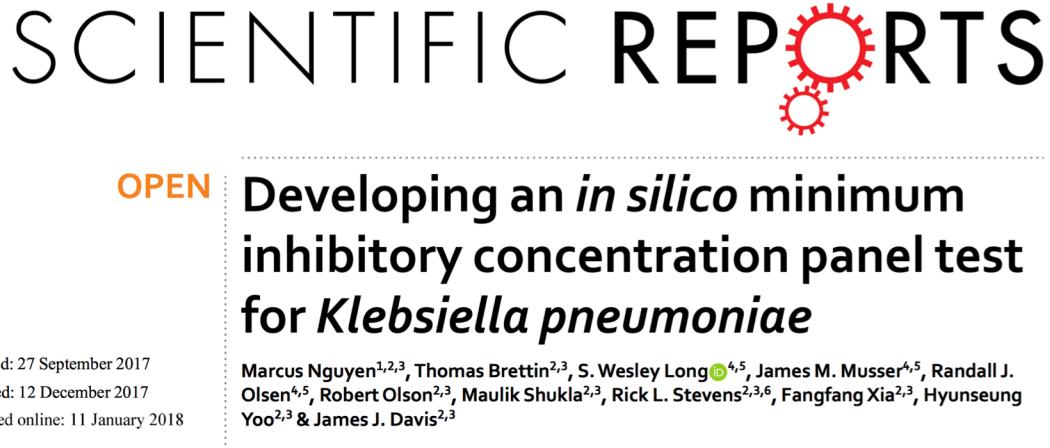
S. Wesley Long,[a,b] Randall J. Olsen,[a,b] Todd N. Eagar,[a] Stephen B. Beres,[a] Picheng Zhao,[a] James J. Davis,[c,d] Thomas Brettin,[c,d] Fangfang Xia,[c,d] James M. Musser[a,b]

Casting AMR prediction as a regression problem:

- **Predicting Minimum Inhibitory Concentrations (MICs)**
  - Training sets based on overlapping 10 or 15 nucleotide k-mers
    - *Not gene based*
  - Currently using XGBoost



Recent progress predicting MICs

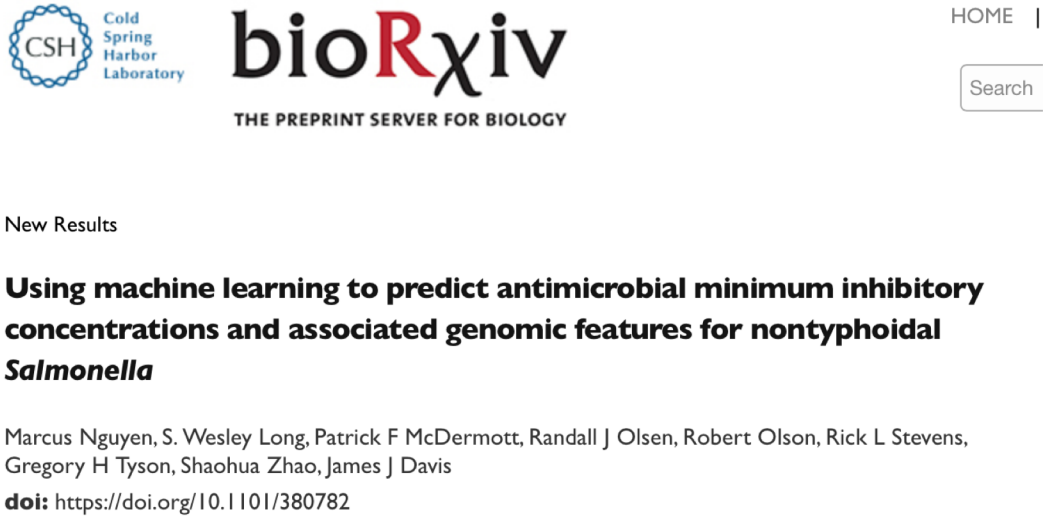

SCIENTIFIC REPORTS

OPEN

Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*

Received: 27 September 2017
Accepted: 12 December 2017
Published online: 11 January 2018

Marcus Nguyen[1,2,3], Thomas Brettin[2,3], S. Wesley Long[4,5], James M. Musser[4,5], Randall J. Olsen[4,5], Robert Olson[2,3], Maulik Shukla[2,3], Rick L. Stevens[2,3,6], Fangfang Xia[2,3], Hyunseung Yoo[2,3] & James J. Davis[2,3]


Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME |

Search

New Results

Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal *Salmonella*

Marcus Nguyen, S. Wesley Long, Patrick F McDermott, Randall J Olsen, Robert Olson, Rick L Stevens, Gregory H Tyson, Shaohua Zhao, James J Davis

doi: https://doi.org/10.1101/380782

Poster #88, Wesley Long, HMHS Tuesday 2PM

Poster #90, Andrew Warren, PATRIC

https://github.com/PATRIC3/mic_prediction

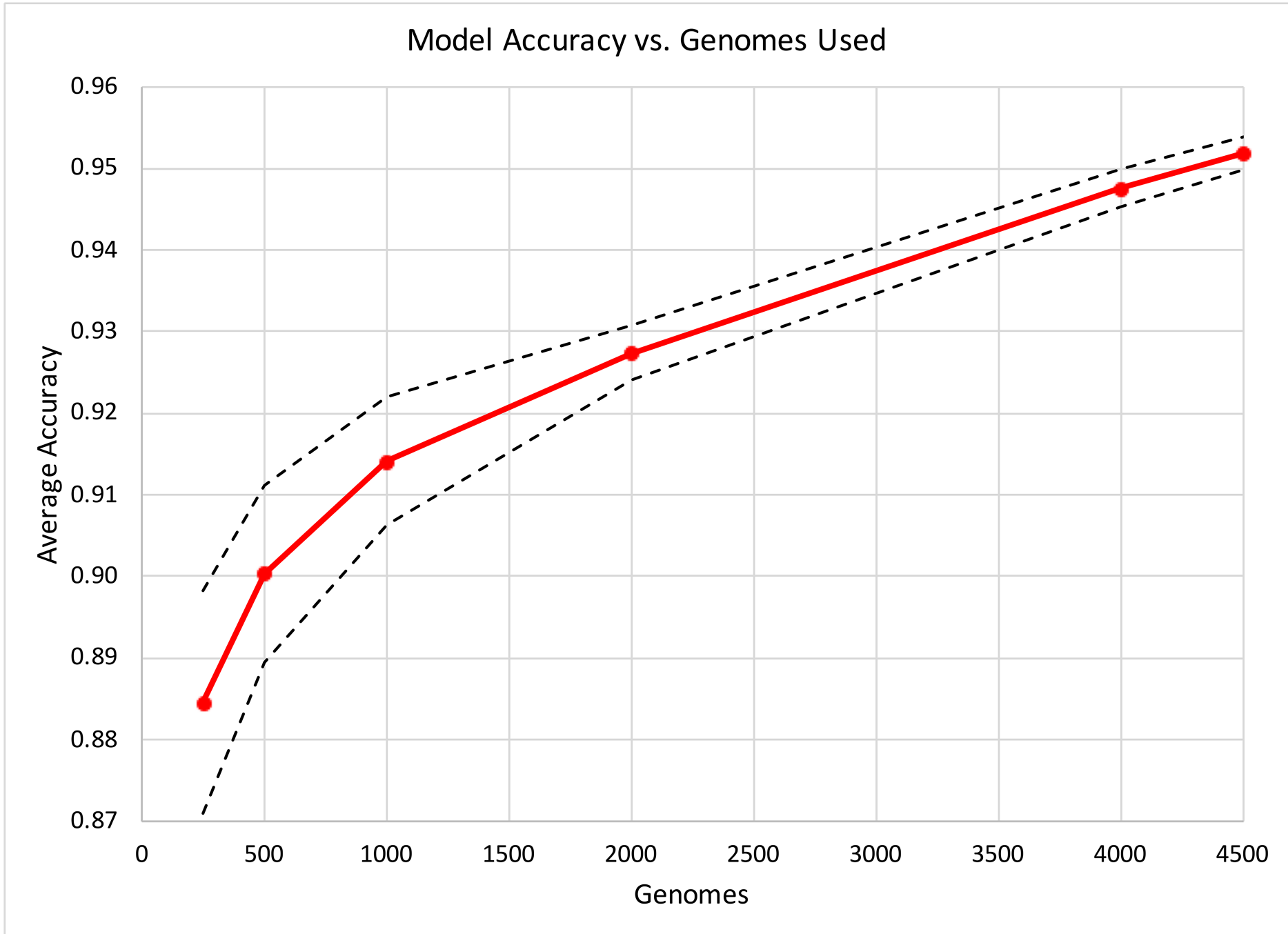
FDA Collaboration

- Patrick McDermott and colleagues
  - NARMS (National Antimicrobial Resistance Monitoring System)
- 5278 *Salmonella enterica* genomes with MIC data
- Collected from 2002-2016
  - Some are taken randomly
- Most are food contamination isolates
- Some animal isolates
- 91 distinct serotypes
- 15 antibiotics

Antibiotic	Susceptible genomes	Intermediate genomes	Resistant genomes
AMP	3682	2	1593
AUG	4145	355	778
AXO	4508	1	769
AZI	2409	0	7
CHL	5026	87	164
CIP	5217	53	7
COT	5219	0	58
FIS	3356	0	1573
FOX	4501	98	679
GEN	4577	68	633
KAN	837	3	84
NAL	5233	0	45
STR	872	0	1919
TET	2364	28	2885
TIO	4517	8	753

How many genomes do you need for an accurate model?

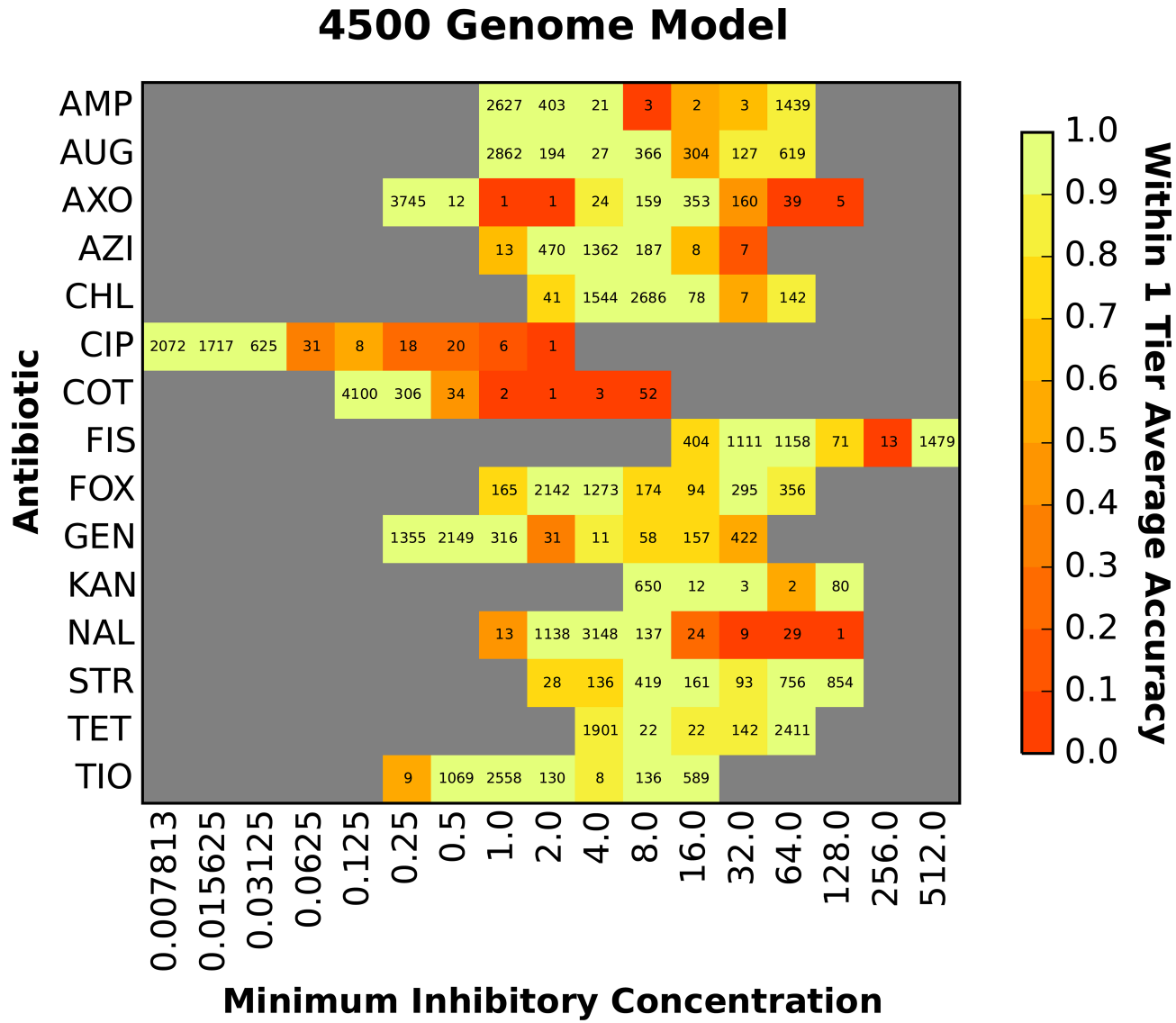
- Systematically reduce the number of genomes while maintaining diversity
  - Manhattan metric distance between genomes
  - Most diverse set at 250,500 1000… genomes
  - Measured accuracy (± 1 2-fold dilution step)


Model Accuracy vs. Genomes Used
Average Accuracy
0.96
0.95
0.94
0.93
0.92
0.91
0.90
0.89
0.88
0.87
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Genomes

Model accuracy by antibiotic

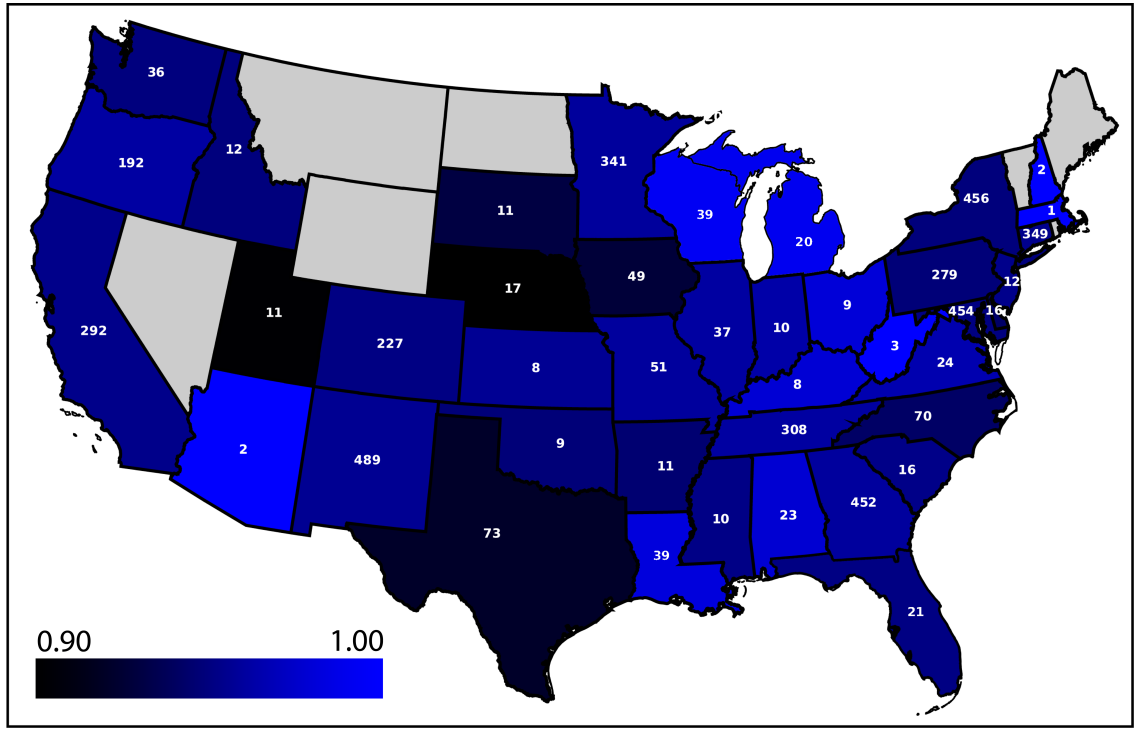
Accuracies within ±1 two-fold dilution step of the actual MIC value for the 4500 genome model.		
Antibiotic	**Accuracy within ±1 2-fold dilution step**	**Within ±1 2-fold dilution step 95% CI**
All	0.95	[0.95-0.95]
AMP	0.92	[0.90-0.93]
AUG	0.93	[0.93-0.94]
AXO	0.95	[0.95-0.96]
AZI	0.97	[0.96-0.98]
CHL	0.99	[0.98-0.99]
CIP	0.97	[0.97-0.98]
COT	0.98	[0.97-0.98]
FIS	0.95	[0.95-0.96]
FOX	0.96	[0.96-0.97]
GEN	0.91	[0.90-0.92]
KAN	0.98	[0.97-1.00]
NAL	0.96	[0.95-0.97]
STR	0.93	[0.92-0.94]
TET	0.90	[0.90-0.91]
TIO	0.99	[0.99-0.99]

Model accuracy by MIC



Model accuracy by metadata category

- No obvious biases by:
  - Serotype
  - Phylogeny
  - State of isolation
  - Collection year
  - Contamination source (chicken, turkey, beef, pork)



Accuracy of models built from previous years

The ability of models trained on genomes from prior years to predict MICs for genomes collected in later years							
Training set years	**Test set years**	**Accuracy**	**95% CI**	**Training Bins**	**Testing Bins**	**Training Genomes**	**Testing Genomes**
2002-2008	2009-2016	0.88	[0.88-0.89]	36563	22412	1819	2681
2002-2009	2010-2016	0.88	[0.88-0.89]	31196	27779	2255	2245
2002-2010	2011-2016	0.88	[0.88-0.88]	28376	30599	2485	2015
2002-2011	2012-2016	0.88	[0.88-0.89]	25408	33567	2699	1801
2002-2012	2013-2016	0.88	[0.87-0.88]	21714	37261	2956	1544
2002-2013	2014-2016	0.86	[0.86-0.87]	17921	41054	3221	1279
2002-2014	2015-2016	0.92	[0.92-0.92]	10807	48168	3728	772

Could this ever be a diagnostic?

- VME
  - Very major error
  - R is called S
- ME
  - Major error
  - S is called R

Antibiotic	VME Avg	VME 95% CI	ME Avg	ME 95% CI	Resistant Samples	Susceptible Samples
All	0.027	[0.024-0.030]	0.001	[0.001-0.002]	10979	47366
AMP	0.028	[0.022-0.033]	0.000	[0.000-0.001]	1442	3054
AUG	0.012	[0.000-0.025]	0.000	[0.000-0.000]	746	3449
AXO	0.022	[0.011-0.032]	0.000	[0.000-0.001]	740	3758
AZI	0.857	[0.508-1.207]	0.000	[0.000-0.000]	7	2040
CHL	0.000	[0.000-0.000]	0.000	[0.000-0.001]	149	4271
CIP	0.417	[-0.099-0.933]	0.000	[0.000-0.000]	7	4445
COT	0.670	[0.515-0.825]	0.000	[0.000-0.001]	55	4443
FIS	0.039	[0.026-0.053]	0.000	[0.000-0.000]	1479	2757
FOX	0.009	[-0.001-0.020]	0.000	[0.000-0.000]	651	3754
GEN	0.090	[0.066-0.113]	0.000	[0.000-0.000]	579	3862
KAN	0.074	[0.012-0.136]	0.000	[0.000-0.000]	82	662
NAL	0.917	[0.819-1.014]	0.000	[0.000-0.001]	39	4460
STR	0.014	[0.008-0.020]	0.027	[0.013-0.040]	1703	744
TET	0.000	[0.000-0.000]	0.018	[0.012-0.025]	2575	1901
TIO	0.004	[-0.001-0.009]	0.000	[0.000-0.000]	725	3766

Could this ever be a diagnostic?

- VME
  - Very major error
  - R is called S
- ME
  - Major error
  - S is called R

Antibiotic	VME Avg	VME 95% CI	ME Avg	ME 95% CI	Resistant Samples	Susceptible Samples
All	0.027	[0.024-0.030]	0.001	[0.001-0.002]		
AMP	0.028	[0.022-0.033]	0.000	[0.000-0.001]	1442	3054
AUG	0.012	[0.000-0.025]	0.000	[0.000-0.000]	746	3449
AXO	0.022	[0.011-0.032]	0.000	[0.000-0.001]	740	3758
AZI	0.857	[0.508-1.207]	0.000	[0.000-0.000]	7	2040
CHL	0.000	[0.000-0.000]	0.000	[0.000-0.001]	149	4271
CIP	0.417	[-0.099-0.933]	0.000	[0.000-0.000]	7	4445
COT	0.670	[0.515-0.825]	0.000	[0.000-0.001]	55	4443
FIS	0.039	[0.026-0.053]	0.000	[0.000-0.000]	1479	2757
FOX	0.009	[-0.001-0.020]	0.000	[0.000-0.000]	651	3754
GEN	0.090	[0.066-0.113]	0.000	[0.000-0.000]	579	3862
KAN	0.074	[0.012-0.136]	0.000	[0.000-0.000]	82	662
NAL	0.917	[0.819-1.014]	0.000	[0.000-0.001]	39	4460
STR	0.014	[0.008-0.020]	0.027	[0.013-0.040]	1703	744
TET	0.000	[0.000-0.000]	0.018	[0.012-0.025]	2575	1901
TIO	0.004	[-0.001-0.009]	0.000	[0.000-0.000]	725	3766

Could this ever be a diagnostic?

- VME
  - Very major error
  - R is called S
- ME
  - Major error
  - S is called R

Antibiotic	VME Avg	VME 95% CI	ME Avg	ME 95% CI	Resistant Samples	Susceptible Samples
All	0.027	[0.024-0.030]	0.001	[0.001-0.002]		
AMP	0.028	[0.022-0.033]	0.000	[0.000-0.001]	1442	3054
AUG	0.012	[0.000-0.025]	0.000	[0.000-0.000]	746	3449
AXO	0.022	[0.011-0.032]	0.000	[0.000-0.001]	740	3758
AZI	0.857	[0.508-1.207]	0.000	[0.000-0.000]	7	2040
CHL	0.000	[0.000-0.000]	0.000	[0.000-0.001]	149	4271
CIP	0.417	[-0.099-0.933]	0.000	[0.000-0.000]	7	4445
COT	0.670	[0.515-0.825]	0.000	[0.000-0.001]	55	4443
FIS	0.039	[0.026-0.053]	0.000	[0.000-0.000]	1479	2757
FOX	0.009	[-0.001-0.020]	0.000	[0.000-0.000]	651	3754
GEN	0.090	[0.066-0.113]	0.000	[0.000-0.000]	579	3862
KAN	0.074	[0.012-0.136]	0.000	[0.000-0.000]	82	662
NAL	0.917	[0.819-1.014]	0.000	[0.000-0.001]	39	4460
STR	0.014	[0.008-0.020]	0.027	[0.013-0.040]	1703	744
TET	0.000	[0.000-0.000]	0.018	[0.012-0.025]	2575	1901
TIO	0.004	[-0.001-0.009]	0.000	[0.000-0.000]	725	3766

Could this ever be a diagnostic?

- VME
  - Very major error
  - R is called S
- ME
  - Major error
  - S is called R

Antibiotic	VME Avg	VME 95% CI	ME Avg	ME 95% CI	Resistant Genomes	Susceptible Genomes
All	0.027	[0.024-0.030]	0.001	[0.001-0.002]		
AMP	0.028	[0.022-0.033]	0.000	[0.000-0.001]	1442	3054
AUG	0.012	[0.000-0.025]	0.000	[0.000-0.000]	746	3449
AXO	0.022	[0.011-0.032]	0.000	[0.000-0.001]	740	3758
AZI	0.857	[0.508-1.207]	0.000	[0.000-0.000]	7	2040
CHL	0.000	[0.000-0.000]	0.000	[0.000-0.001]	149	4271
CIP	0.417	[-0.099-0.933]	0.000	[0.000-0.000]	7	4445
COT	0.670	[0.515-0.825]	0.000	[0.000-0.001]	55	4443
FIS	0.039	[0.026-0.053]	0.000	[0.000-0.000]	1479	2757
FOX	0.009	[-0.001-0.020]	0.000	[0.000-0.000]	651	3754
GEN	0.090	[0.066-0.113]	0.000	[0.000-0.000]	579	3862
KAN	0.074	[0.012-0.136]	0.000	[0.000-0.000]	82	662
NAL	0.917	[0.819-1.014]	0.000	[0.000-0.001]	39	4460
STR	0.014	[0.008-0.020]	0.027	[0.013-0.040]	1703	744
TET	0.000	[0.000-0.000]	0.018	[0.012-0.025]	2575	1901
TIO	0.004	[-0.001-0.009]	0.000	[0.000-0.000]	725	3766

Important resistance k-mers

The highest-ranking AMR-related protein function with a matching k-mer from the XGBoost models

Antibiotic	K-mer Rank	Distance between k-mer and AMR gene	k-mer	PATRIC Annotation(s)
AMP	1	Direct match	CTTAATCAGTGAGGC	Class A beta-lactamase (EC 3.5.2.6) => TEM family
AUG	1	Direct match	AAACGTCTTACTAAC	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family
AXO[2]	1	566.0 ± 39.7	AAAGAGAAAGAAAGG	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family
AZI	8	Direct match	CCCATTTCCGCCGCC	Macrolide 2'-phosphotransferase => Mph(A) family
CHL[2]	1	611.8 ± 5.1	AGACAAGTAAGCCGC	Chloramphenicol/florfenicol resistance, MFS efflux pump => FloR family
CIP	1	313.5 ± 70.5	ACAGTCCATCCAGGA	Pentapeptide repeat protein QnrB family => Quinolone resistance protein QnrB10
COT[2]	1	248.4 ± 7.0	AAAAACGATAGCTGC	Dihydrofolate reductase (EC 1.5.1.3)
FIS[2]	1	408.9 ± 11.3	CGCAACGGCTCAAGC	Dihydropteroate synthase type-2 (EC 2.5.1.15) @ Sulfonamide resistance protein
FOX	1	Direct match	AAAAAAACCTTGGCA	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family
GEN	1	439.0 ± 70.4, 451.2 ± 7.0	AGTTAAGCCGCGCCG	Aminoglycoside 3''-nucleotidyltransferase (EC 2.7.7.-) => APH(3'')-Ia (AadA family); Aminoglycoside N(3)-acetyltransferase (EC 2.3.1.81) => AAC(3)-II,III,IV,VI,VIII,IX,X
KAN	1	Direct match	AAAAAGCCGTTTCTG	Aminoglycoside 3'-phosphotransferase (EC 2.7.1.95) => APH(3')-I
NAL	1	Direct match	ATTCCGCAGTGTATG	DNA gyrase subunit A (EC 5.99.1.3)
STR	1	Direct match	ATTTGTACGGCTCCG	Aminoglycoside 3''-nucleotidyltransferase (EC 2.7.7.-) => APH(3'')-Ia (AadA family)
TET	1	335.6 ± 12.6, 596.5 ± 22.5	CGTTCTGCCTTGCGC	Tetracycline resistance regulatory protein TetR; Tetracycline resistance, MFS efflux pump => Tet(A)
TIO[2]	1	566 ± 39.7	AAAGAGAAAGAAAGG	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family

Important susceptibility k-mers

Important k-mers used by the individual antibiotic models for predicting susceptible MICs

Antibiotic	k-mer	Sus	Res	Frac Sus	Frac Res	Genomic region	PATRIC annotation or genomic region
NAL	ATTCCGCAGTGTATG	5233	45	1.00	0.38	PEG	DNA gyrase subunit A (EC 5.99.1.3)
AXO	TGGTATTCGCATCAA	4508	769	0.78	0.48	PEG	Phosphoethanolamine transferase EptA
KAN	CTGGCTTTTTTTTTT	837	84	0.30	0.00	RNA	RyhB RNA
STR	CCCTTATCCAACACG	872	1919	0.85	0.55	PEG	Respiratory nitrate reductase delta chain (EC 1.7.99.4)
AXO	CAGAACCAGAATTTG	4508	769	0.74	0.46	PEGs	Formate-dependent nitrite reductase complex subunit NrfF, and Cytochrome c-type heme lyase subunit nrfE, nitrite reductase complex assembly
TIO	AGAGAAGCCTGCCGC	4517	753	0.68	0.40	PEG	Oxaloacetate decarboxylase alpha chain (EC 4.1.1.3)
AXO	ATCCCCGCCATTACA	4508	769	0.73	0.46	PEG	Tagatose-1,6-bisphosphate aldolase GatY (EC 4.1.2.40)
AXO	TGCTGCAAAACGCCA	4508	769	0.69	0.45	PEG	Protein AraJ precursor
AXO	GAAAACAGGGTGTAG	4508	769	0.47	0.23	INT	Upstream of IlvGMEDA operon leader peptide
FOX	GGATACCACGCCGGG	4501	679	0.58	0.35	PEGs	Glucose dehydrogenase, PQQ-dependent (EC 1.1.5.2), and IncF plasmid conjugative transfer protein TraP

Ongoing work

- Moving MIC models into PATRIC
- Improve AMR gene sampling
- Looking at global strains
- Classification of Oxford Nanopore data
- Understanding susceptible k-mers
- Matched patient studies

Acknowledgments

Marcus Nguyen
University of Chicago

FDA

- Maureen Davidson
- Chih-Hao Hsu
- Patrick McDermott
- Heather Tate
- Gregory Tyson
- Shaohua Zhao

Houston Methodist

- S. Wesley Long
- James Musser
- Randall Olsen

PATRIC Team Members Including:

University of Chicago

- Tom Brettin
- Emily Dietrich
- Marcus Nguyen
- Bob Olson
- John Santerre
- Maulik Shukla
- **Rick Stevens**
- Fangfang Xia
- Harry Yoo

Sébastien Boisvert

Biocomplexity Institute / Virginia Tech

- Chunhong Mao
- Rebecca Wattam
- Rebecca Will

Fellowship for Interpretation of Genomes

- Svetlana Gerdes
- Margo VanOeffelen

National Institute of Allergy and Infectious Diseases
National Institutes of Health, Contract No. HHSN272201400027C