

# Genotype-to-phenotype prediction

*How well do machine learning methods capture causal  
mechanisms?*

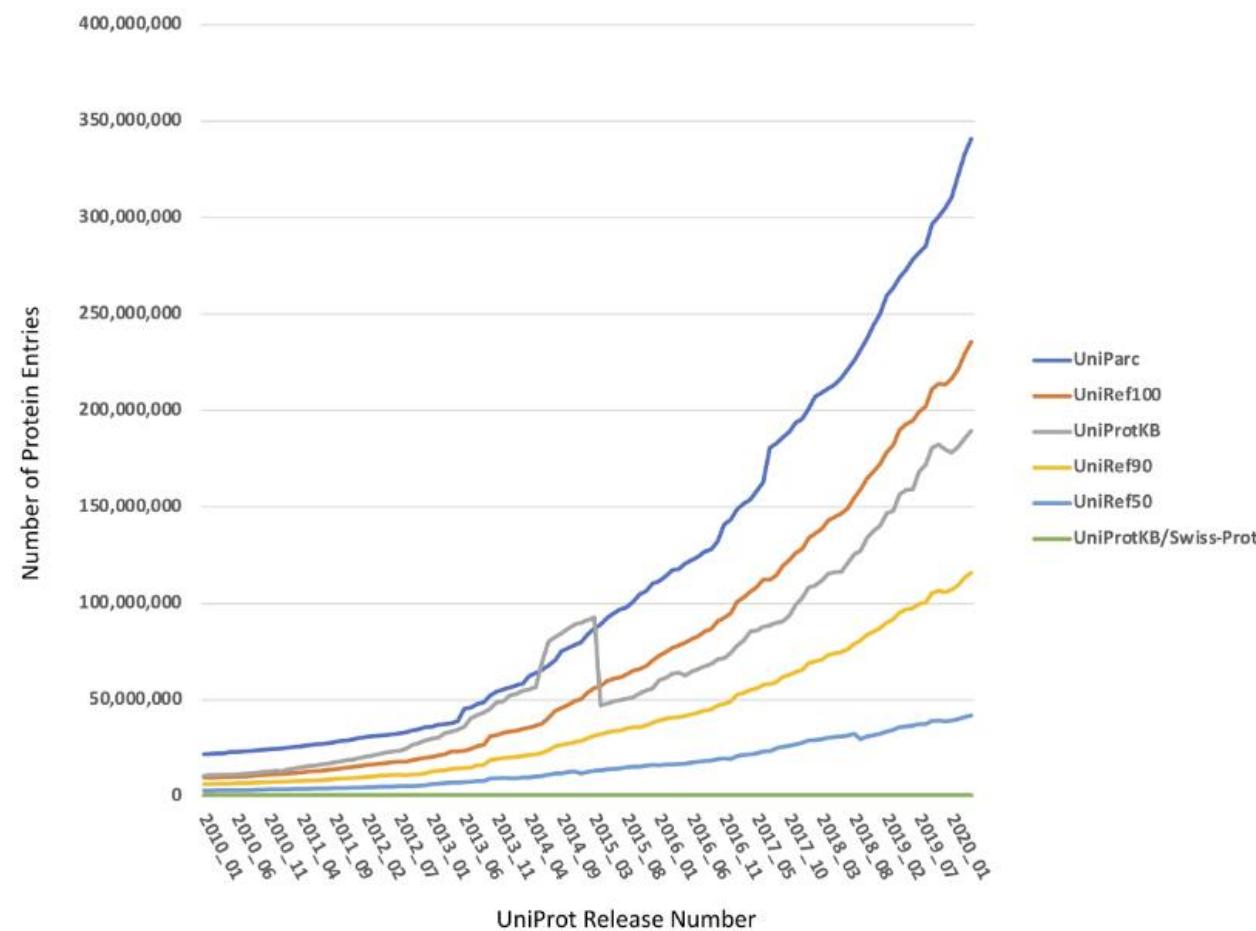
*Nicole Wheeler – University of Birmingham*

# Outline

- Performance of ML algorithms on new data
- Reasons for under-performance
- Diagnosing poor generalisability
- Characterising learning abilities

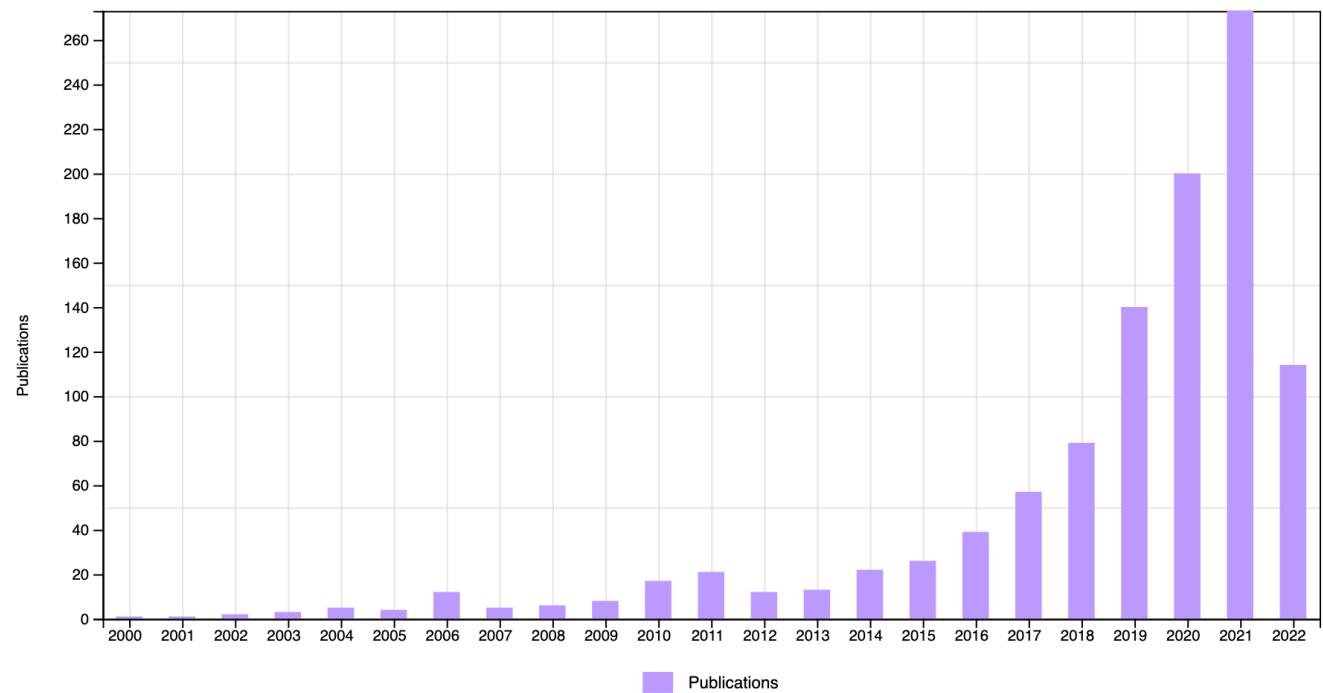
# Genomics and the 'big data' era

- Genome collections are growing exponentially
- Set to become the largest source of data in the world
- We are increasing power to link genotype to phenotype



# Machine learning in bacterial genomics

- Growing number of ML studies focused on bacterial pathogens
- Little consensus on best practices for training, testing and reporting on these models to date
  - Some guidelines now showing up in journals
- Most citations of these ML papers are other ML papers
  - Lack of integration with other research, clinical trials or diagnostics



Source: Web of Science search – “machine learning”, “bacteria”

# What we want these models to do

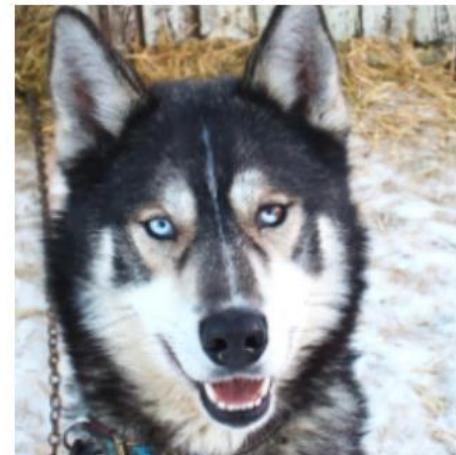
- Accurately predict phenotype
- Capture the biology of the trait (causal mechanisms)
- Learn unsupervised in a trustworthy manner
- Serve everyone equally – no subpopulation systematically disadvantaged

# What we want these models to do

- Accurately predict phenotype
- Capture the biology of the trait (causal mechanisms)
- Learn unsupervised in a trustworthy manner
- Serve everyone equally – no subpopulation systematically disadvantaged

The easiest and most accurate way to make predictions isn't necessarily the one we want

We need explainability to assess this



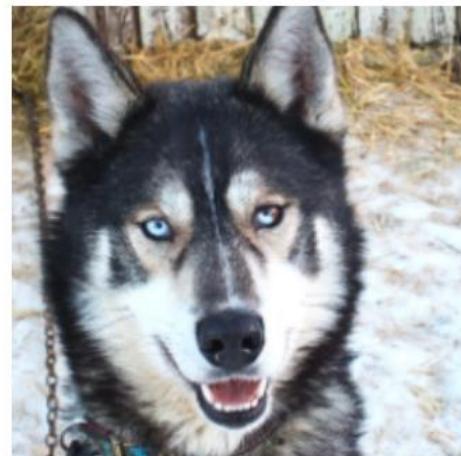
(a) Husky classified as wolf

# What we want these models to do

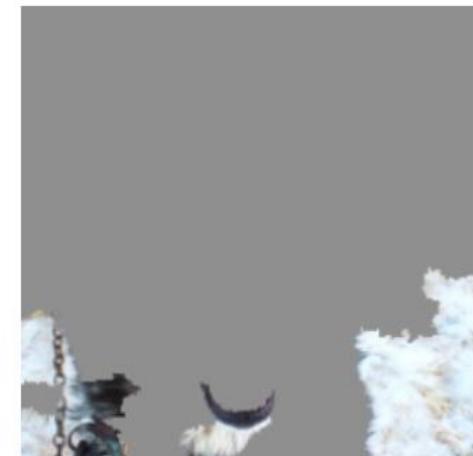
- Accurately predict phenotype
- Capture the biology of the trait (causal mechanisms)
- Learn unsupervised in a trustworthy manner
- Serve everyone equally – no subpopulation systematically disadvantaged

The easiest and most accurate way to make predictions isn't necessarily the one we want

We need explainability to assess this



(a) Husky classified as wolf



(b) Explanation

Performance of ML algorithms  
on new data

# Published algorithms falter on new data

[www.nature.com/scientificreports/](http://www.nature.com/scientificreports/)



OPEN

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

Received: 27 September 2017

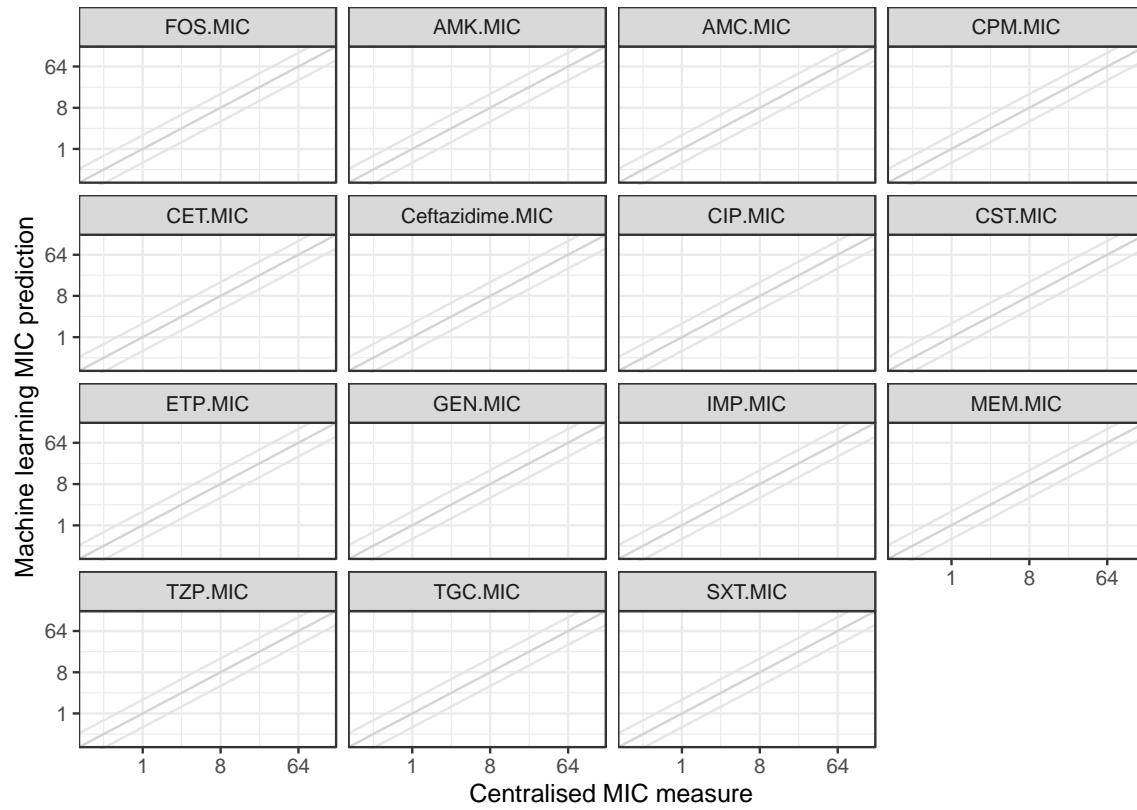
Accepted: 12 December 2017

Published online: 11 January 2018

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

# Published algorithms falter on new data

Predicted MIC



www.nature.com/scientificreports/

## 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<sup>1,2,3</sup>, Thomas Brettin<sup>2,3</sup>, S. Wesley Long<sup>4,5</sup>, James M. Musser<sup>4,5</sup>, Randall J. Olsen<sup>4,5</sup>, Robert Olson<sup>2,3</sup>, Maulik Shukla<sup>4,3</sup>, Rick L. Stevens<sup>2,3,6</sup>, Fangfang Xia<sup>2,3</sup>, Hyunseung Yoo<sup>2,3</sup> & James J. Davis<sup>2,3</sup>

Antibiotic	Reported accuracy (US samples)	Accuracy on European samples
Amikacin	97%	18%
Cefepime	61%	47%
Ciprofloxacin	98%	78%
Gentamicin	95%	51%
Imipenem	94%	74%
Piperacillin/tazobactam	78%	26%
Trimethoprim-sulphamethoxazole	95%	77%

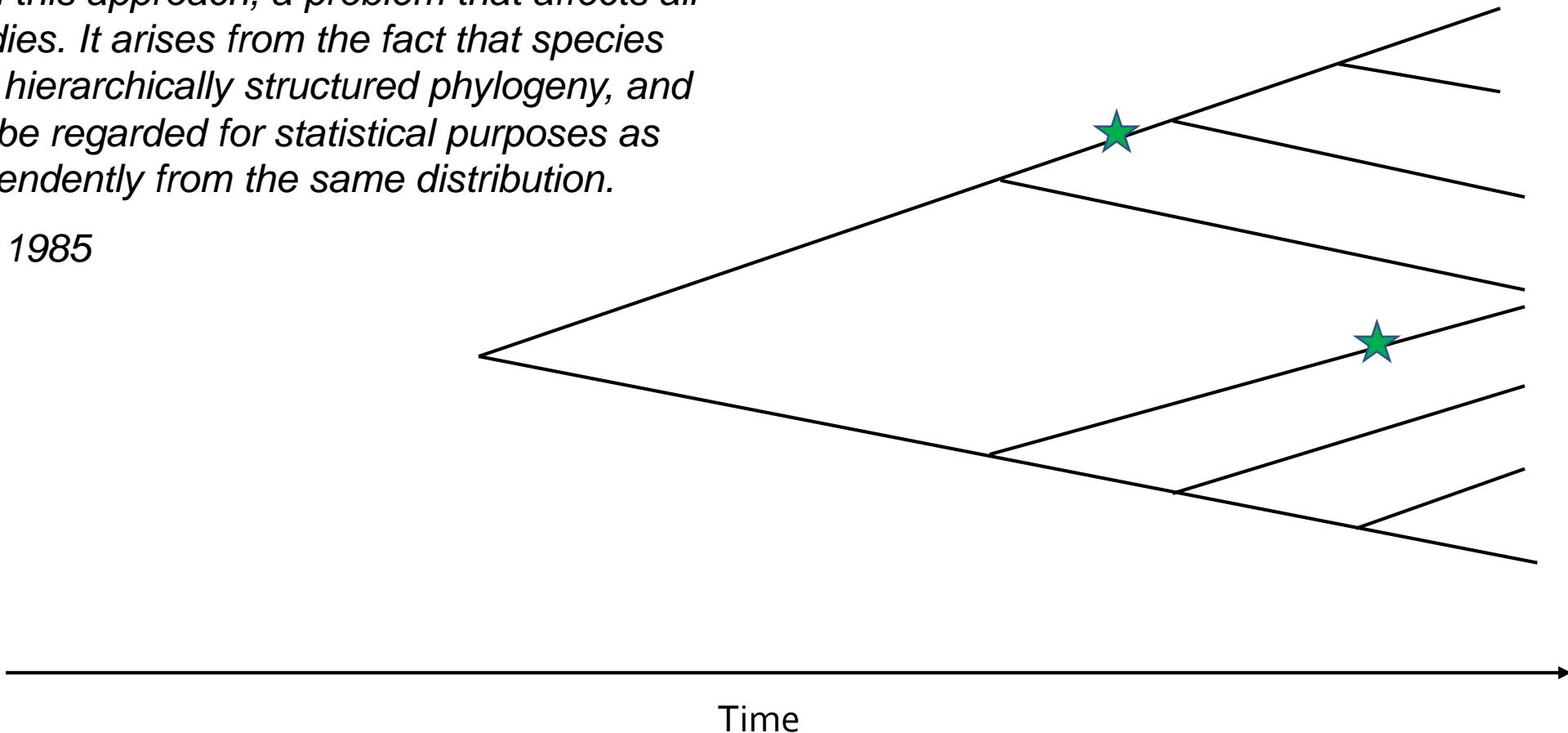
# Multiple possible reasons for poor performance

- Different mechanisms in different populations
- Failure to learn causal mechanisms
- Differences in phenotyping across labs

# Genomic data are autocorrelated

*My intention is to point out a serious statistical problem with this approach, a problem that affects all of these studies. It arises from the fact that species are part of a hierarchically structured phylogeny, and thus cannot be regarded for statistical purposes as drawn independently from the same distribution.*

*Felsenstein, 1985*



# Genome-wide association studies

- Test each variant for an association with a trait
- Have to correct for correlation structure in dataset (population structure)

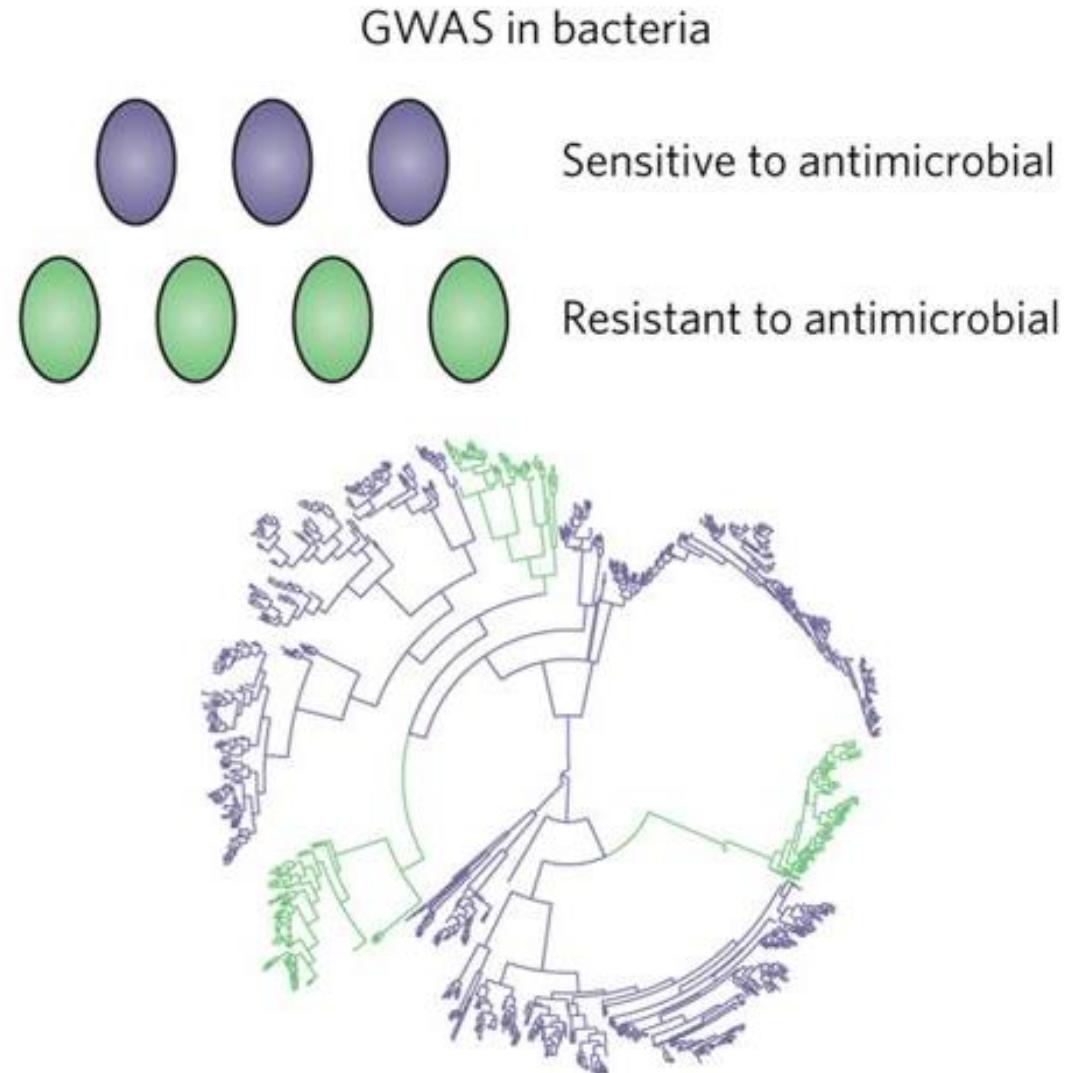
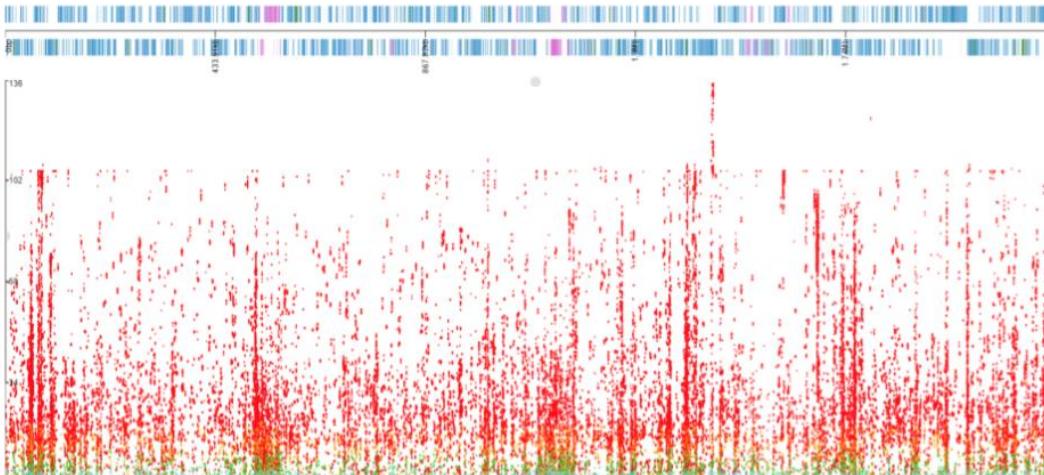


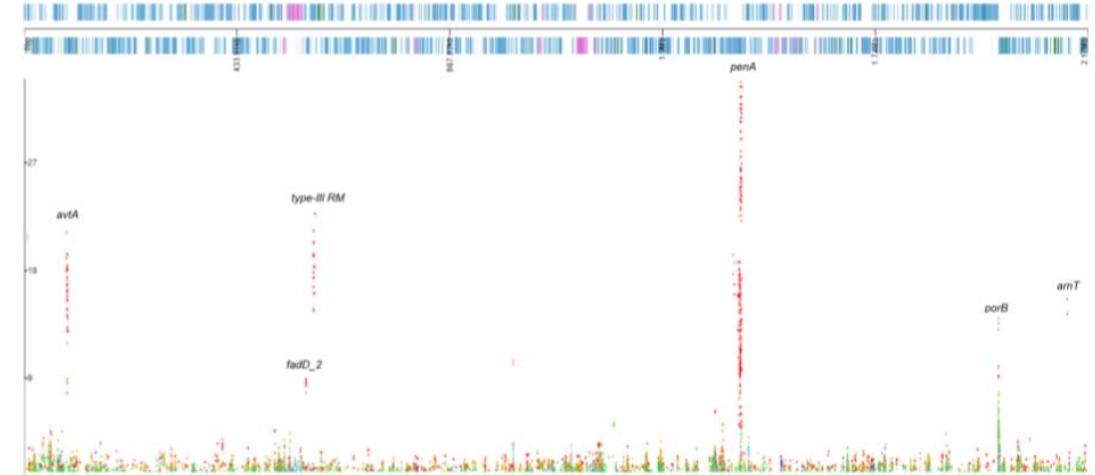
Image: Falush, D. (2016). Bacterial genomics: Microbial GWAS coming of age. *Nature Microbiology*, 1, 16059.

# Confounding by population structure in GWAS

Before



After

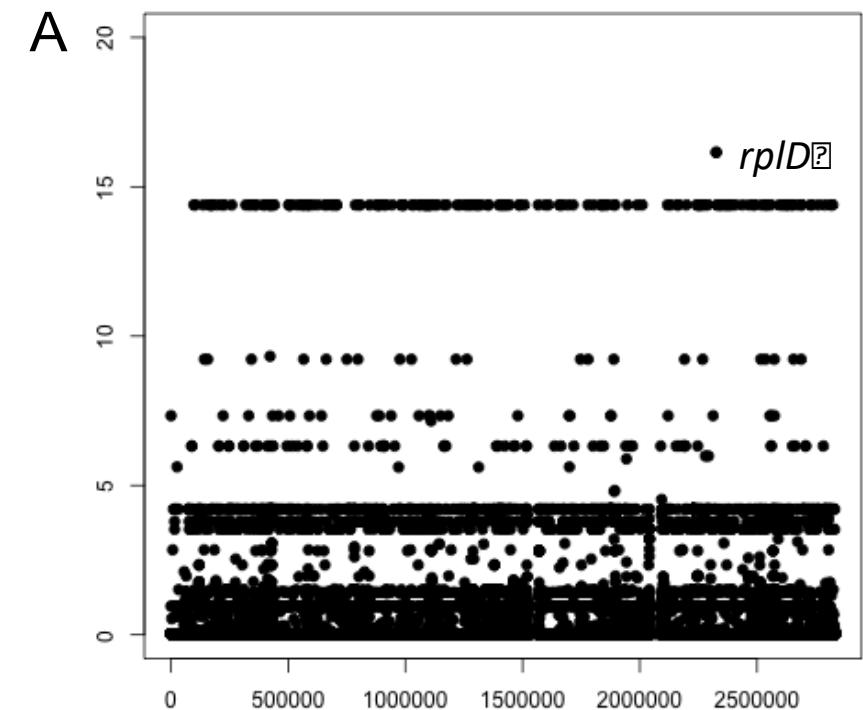


N.B. Y-axes between Manhattan plots are not directly comparable

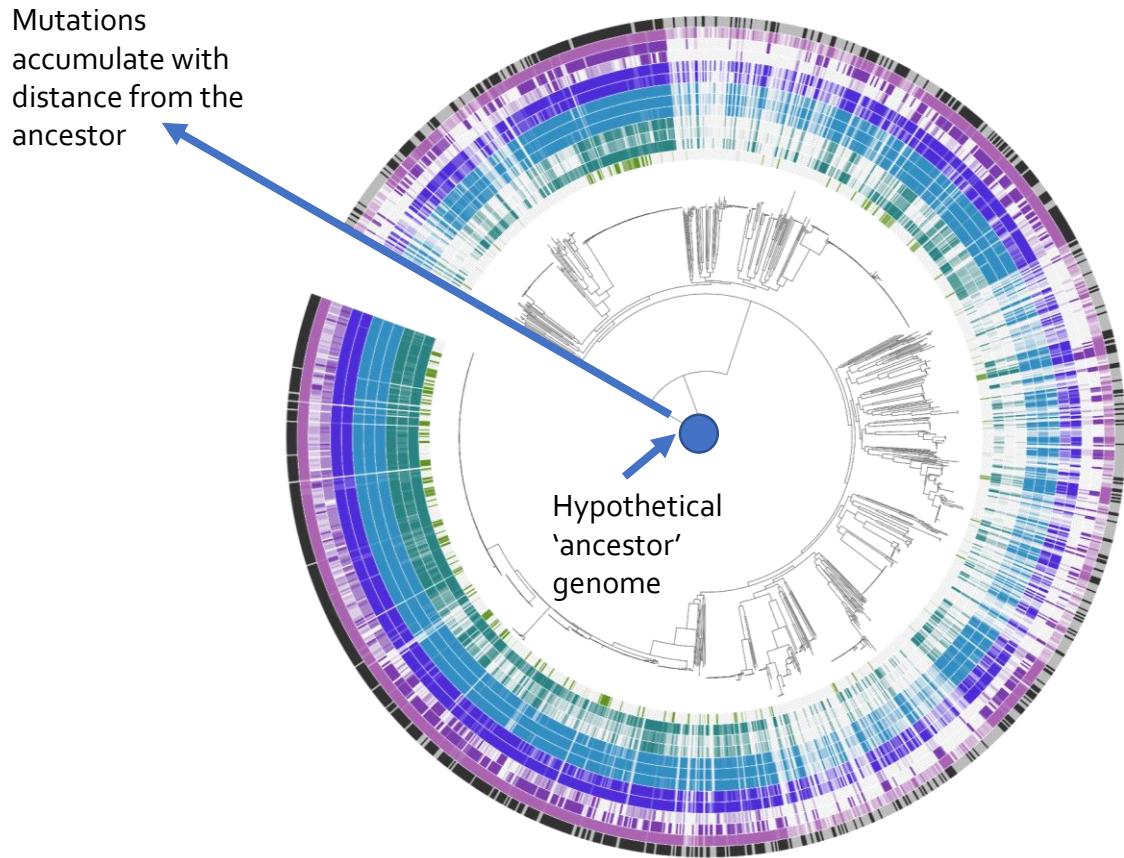
Image: Kevin Ma,  
Harvard Medical School

# Phylogeny vs biology– what is the model learning?

- Genetic markers linked to clones are likely to be incorporated with causal variants
- Correlated variables tend to be given lower individual “importance”, but may still have a high joint impact



# Pathogen populations make learning resistance mechanisms challenging



- An easy way to predict resistance is to ID successful clones
- An easy way to predict resistance to an antibiotic with a complex genetic mechanism is to predict based on an antibiotic with a simpler genetic mechanism

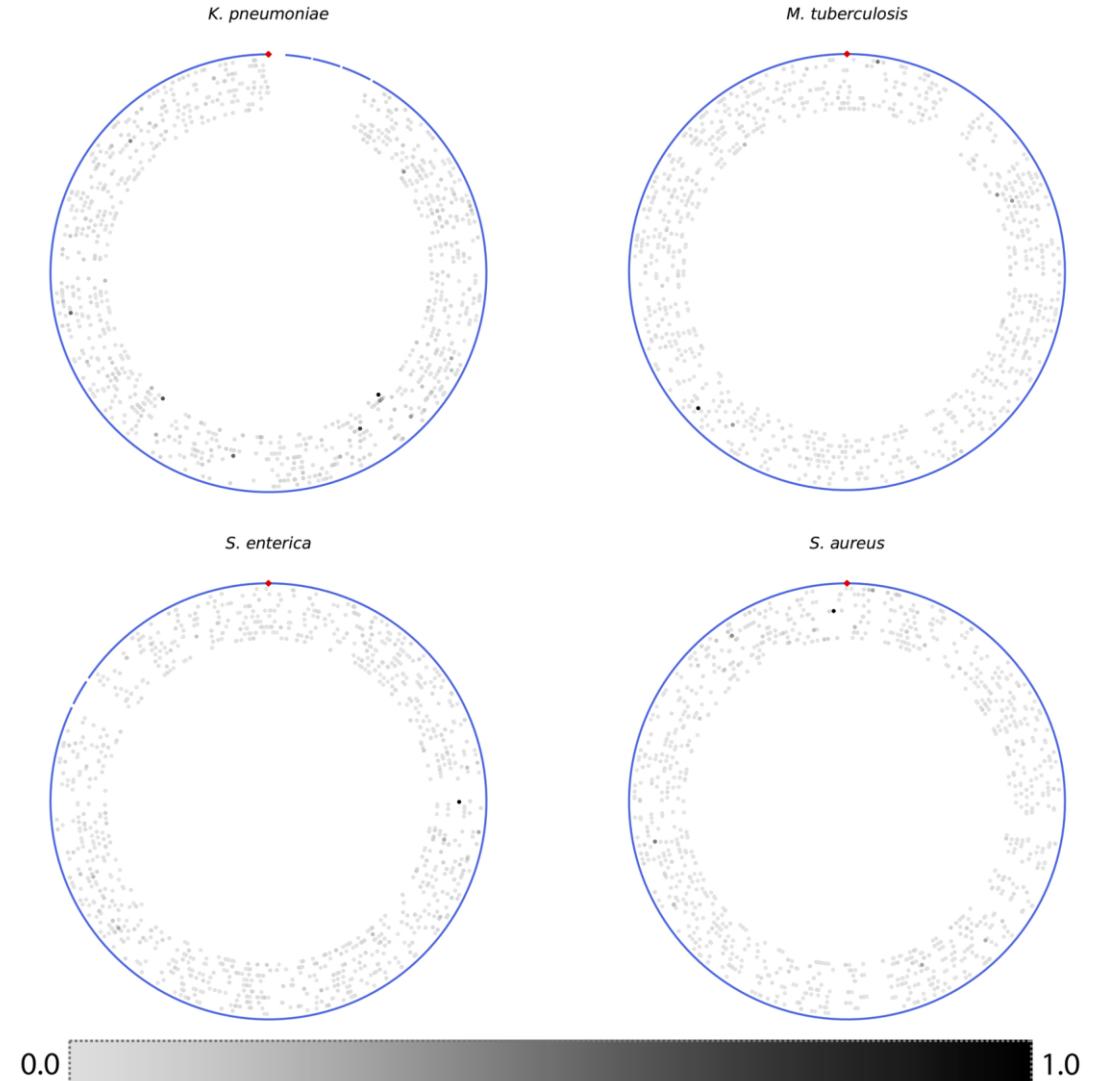
Can we trust ML algorithms to  
learn causal mechanisms?

# If the models are learning causal mechanisms

- Important variants should be more focused in certain parts of the genome
- The model should perform well on new populations

# Regions with high feature importance distributed across the genome

- Complete removal of known causal mechanisms doesn't decrease prediction accuracy
- Different subsamples of 100 core genes can produce models with high accuracy
- Predictive regions are spread across the genome rather than focused in particular locations



Nguyen, M. et al. (2020) 'Predicting antimicrobial resistance using conserved genes', *PLoS computational biology*, 16(10), p. e1008319.  
Aytan-Aktug, D. et al. (2021) 'Predicting Antimicrobial Resistance Using Partial Genome Alignments', *mSystems*, 6(3), p. e0018521.

# How do we tell if an algorithm has learned causal mechanisms?

TABLE 1

Known AMR genes identified by the k-mer-based AMR classifiers<sup>a</sup>

'correct' predictors



Antibiotic	Drug class	Known AMR gene(s) to the antibiotic <sup>b</sup>
Ampicillin	Beta-lactam	TEM-1**, CTX-M-15, <i>yicJ</i> *
Aztreonam	Beta-lactam	CTX-M-55*
Cefepime	Beta-lactam	CTX-M-1**, CTX-M-15, CTX-M-55
Cefoxitin	Beta-lactam	CMY-2*, <i>ybiW</i> *, <i>betT</i> , <i>chiP</i> , <i>cra</i> , <i>envZ</i> , <i>htrE</i> , <i>lyxK</i> , <i>mdlA</i> , <i>yeeJ</i> , <i>yghA</i>
Ciprofloxacin	Fluoroquinolone	<i>gyrA</i> **
Gentamicin	Aminoglycoside	AAC(3)-IId**, AAC(6')-Ib7**, <i>aadA13</i> *, AAC(3)-IIe*, AAC(6')-Ib9*, <i>aadA7</i> , ANT(2')-Ia
Levofloxacin	Fluoroquinolone	<i>gyrA</i> **
Tetracycline	Tetracycline	<i>tet(A)</i> **, <i>tet(B)</i> **, <i>mdfA</i>
Tobramycin	Aminoglycoside	AAC(3)-IId**, AAC(6')-Ib-cr**, AAC(3)-IIe, AAC(6')-Ib7
Trimethoprim	Diaminopyrimidine	

# ML learns the wrong resistance mechanisms

TABLE 1

Known AMR genes identified by the k-mer-based AMR classifiers<sup>a</sup>

'correct' predictors  
↓  
*Cause resistance to a different drug*  
'wrong' predictors

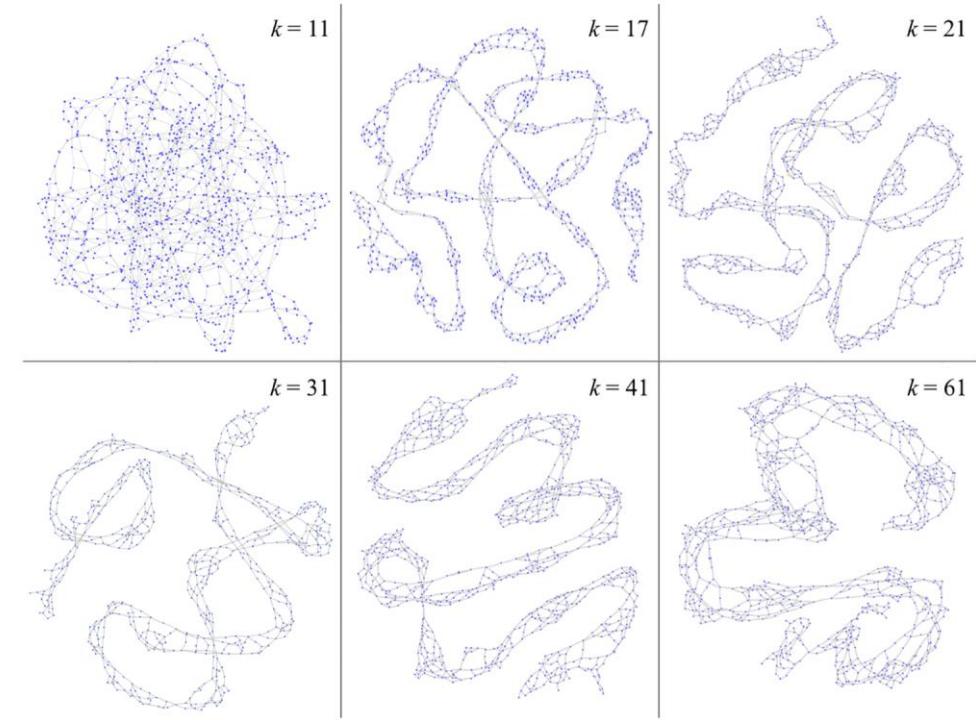
Antibiotic	Drug class	Known AMR gene(s) to the antibiotic <sup>b</sup>	Known AMR genes associated with other antibiotics <sup>b</sup>
Ampicillin	Beta-lactam	TEM-1**, CTX-M-15, <i>yicJ</i> *	<i>sulI</i> **, <i>folP</i> **, APH(3")-Ib, <i>katE</i> *, <i>yadV</i> *, <i>arnC</i> , <i>fsr</i> , <i>nmpC</i> , <i>pepT</i> , <i>yeeJ</i> , <i>yhdJ</i>
Aztreonam	Beta-lactam	CTX-M-55*	AAC(6')-Ib-cr, <i>acrD</i> , <i>catIII</i> , <i>nmpC</i> , <i>pitA</i> , <i>yicI</i> , <i>cpdB</i> , <i>yoaE</i> , <i>rapA</i> , <i>dinG</i> , <i>yeeJ</i> , <i>oppA</i> , <i>arnC</i>
Cefepime	Beta-lactam	CTX-M-1**, CTX-M-15, CTX-M-55	<i>dfrA25</i> *, AAC(6')-Ib10*, AAC(3)-IId, <i>catB3</i> , AAC(6')-Ib-cr, <i>folA</i> *, <i>yadV</i> *, <i>citF</i> , <i>yeeJ</i> , <i>ftsI</i>
Cefoxitin	Beta-lactam	CMY-2*, <i>ybiW</i> *, <i>betT</i> , <i>chiP</i> , <i>cra</i> , <i>envZ</i> , <i>htrE</i> , <i>lyxK</i> , <i>mdlA</i> , <i>yeeJ</i> , <i>yghA</i>	<i>dfrA25</i> , AAC(3)-IId, <i>catIII</i> , <i>blc</i> , <i>yaiY</i> , <i>folA</i> , <i>putA</i> , <i>lpoA</i>
Ciprofloxacin	Fluoroquinolone	<i>gyrA</i> **	OXA-1*, CTX-M-15*, <i>arnC</i> , <i>nmpC</i> , <i>htrE</i> , <i>cpdB</i> , <i>arcA</i> , <i>flu</i>
Gentamicin	Aminoglycoside	AAC(3)-IId**, AAC(6')-Ib7**, <i>aadA13</i> *, AAC(3)-Ile*, AAC(6')-Ib9*, <i>aadA7</i> , ANT(2')-Ia	<i>floR</i> , CTX-M-15, <i>dfrA17</i> , <i>mphA</i> , <i>intS</i> *, <i>fliC</i> *, <i>arnC</i> , <i>yicJ</i>
Levofloxacin	Fluoroquinolone	<i>gyrA</i> **	<i>lacI</i> *, <i>yqiK</i> , <i>flu</i> , <i>arcA</i> , <i>fimC</i> , <i>phoE</i> , <i>ybiH</i> , <i>dadA</i>
Tetracycline	Tetracycline	<i>tet(A)</i> **, <i>tet(B)</i> **, <i>mdfA</i>	APH(6)-IId, <i>sul2</i> , <i>yeeJ</i> , <i>folP</i> , <i>csiD</i>
Tobramycin	Aminoglycoside	AAC(3)-IId**, AAC(6')-Ib-cr**, AAC(3)-Ile, AAC(6')-Ib7	<i>catB3</i> *, CTX-M-55, <i>dfrA17</i> , OXA-1, <i>fliC</i> *, <i>pinR</i> , <i>ydfU</i> , <i>dnaQ</i>
Trimethoprim	Diaminopyrimidine		ANT(2")-Ia**, <i>sul2</i> *, <i>aadA16</i> *, <i>aadA25</i> *, APH(3")-Ib*, TEM-1, <i>tet(A)</i> , APH(6)-Id, <i>mphA</i> , TEM-150, <i>sul1</i> , <i>folP</i> *, <i>dosP</i> , <i>valS</i> , <i>nmpC</i> , <i>htrE</i> , <i>groL</i> , <i>putP</i>

ML algorithms learn more 'wrong' predictors than right ones

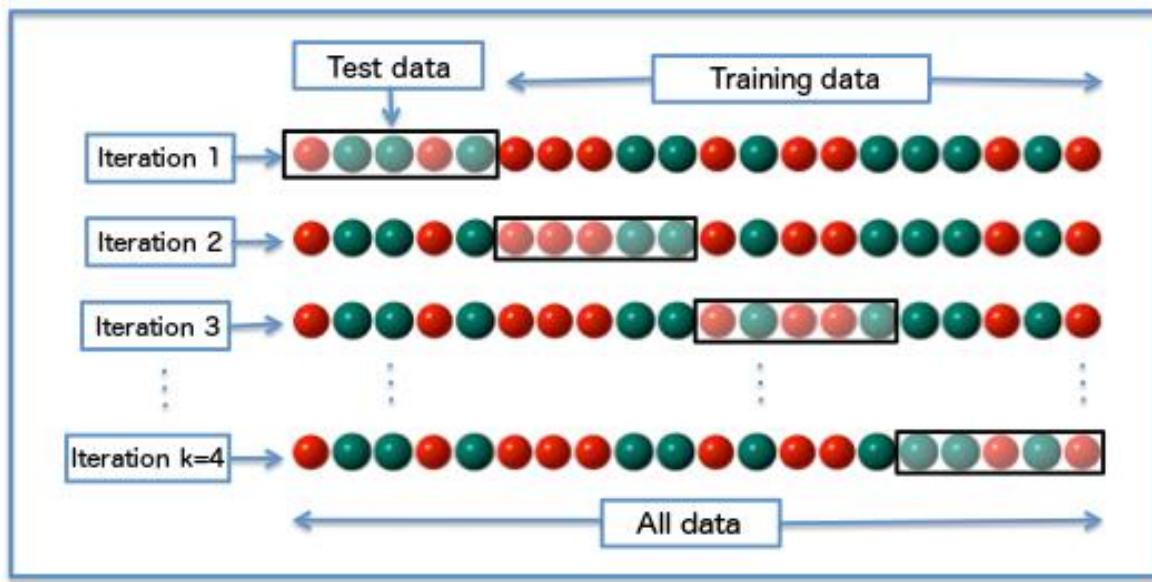
# Diagnosing this problem

# Test dataset

- 3970 *Neisseria gonorrhoeae* genomes
- Encoded as a unitig graph
  - Efficient, flexible representation of genomic diversity
  - Dataset usually 5% size of kmer representation
- MIC data
- Models trained with grid search of hyperparameters



# Measuring performance

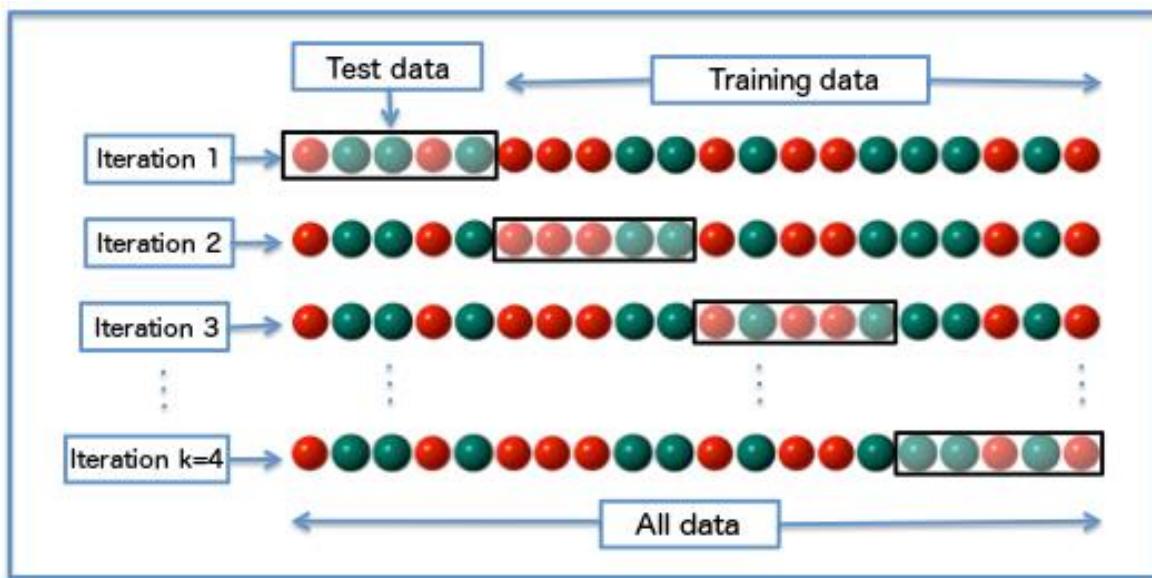


k-fold cross-validation  
data are randomly partitioned  
into k subsamples, then k  
models are built, with most of  
the data used for training and  
the subsample used for testing

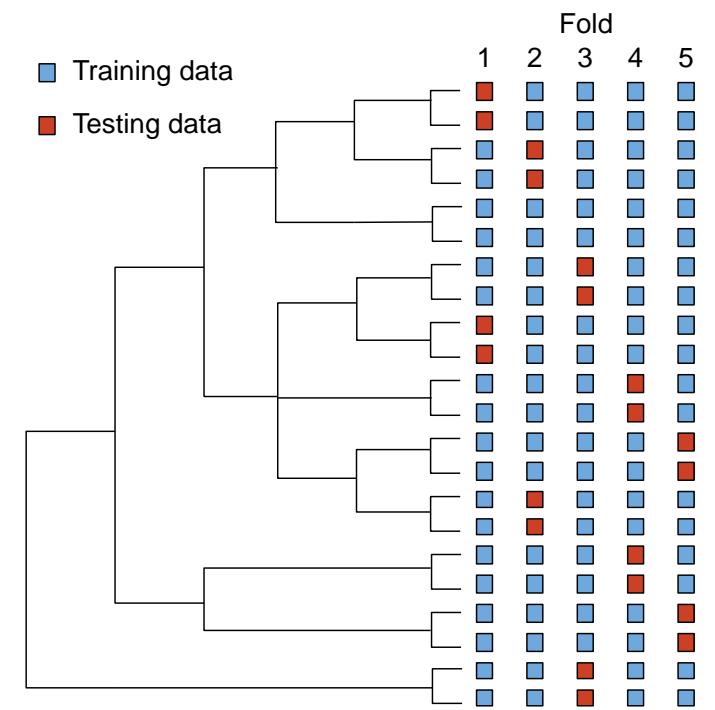
Traditional cross-validation is typically reported in the literature, but this can overestimate performance

# Measuring the performance of ML algorithms

Traditional cross-validation

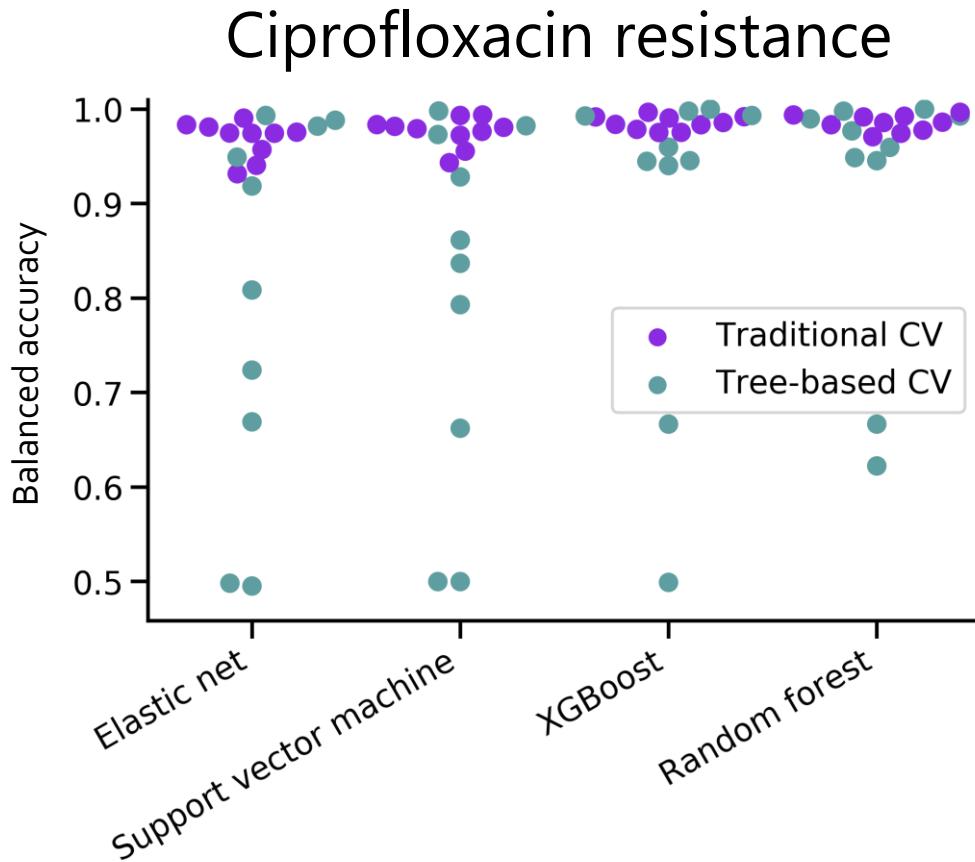


Tree-based cross-validation

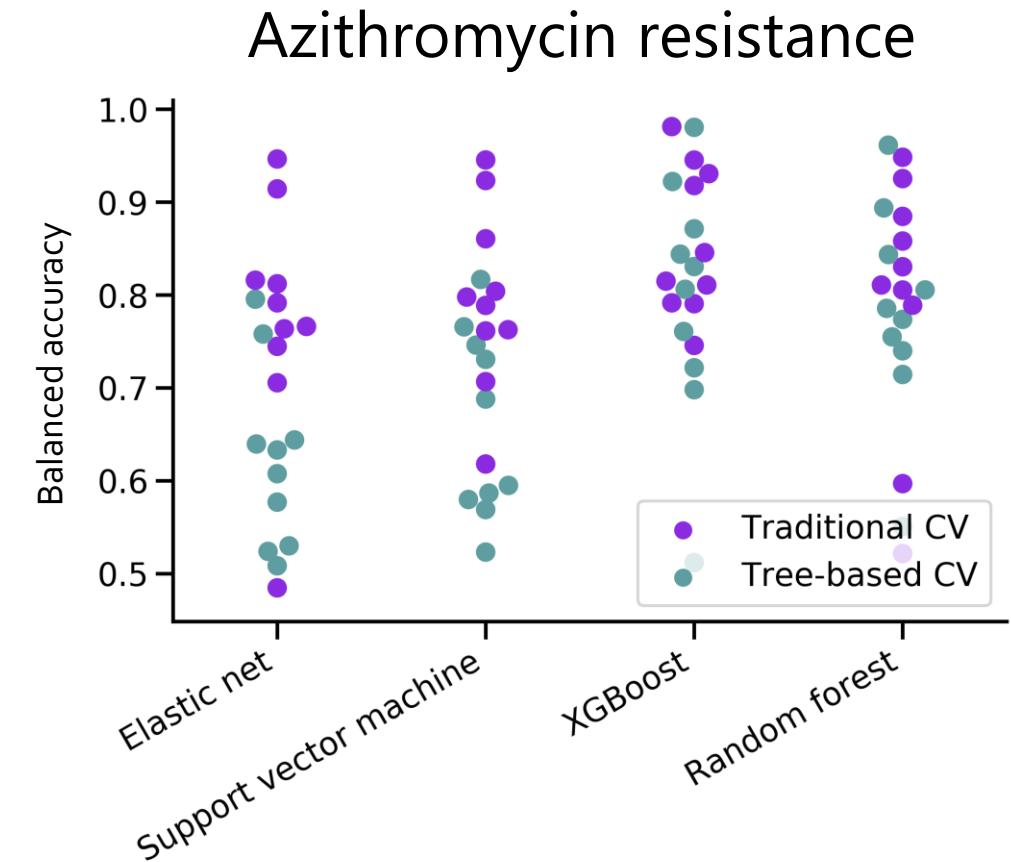


Traditional cross-validation is typically reported in the literature, but this can overestimate performance if the same strains appear in training and testing data

# Traditional cross-validation underestimates overfitting

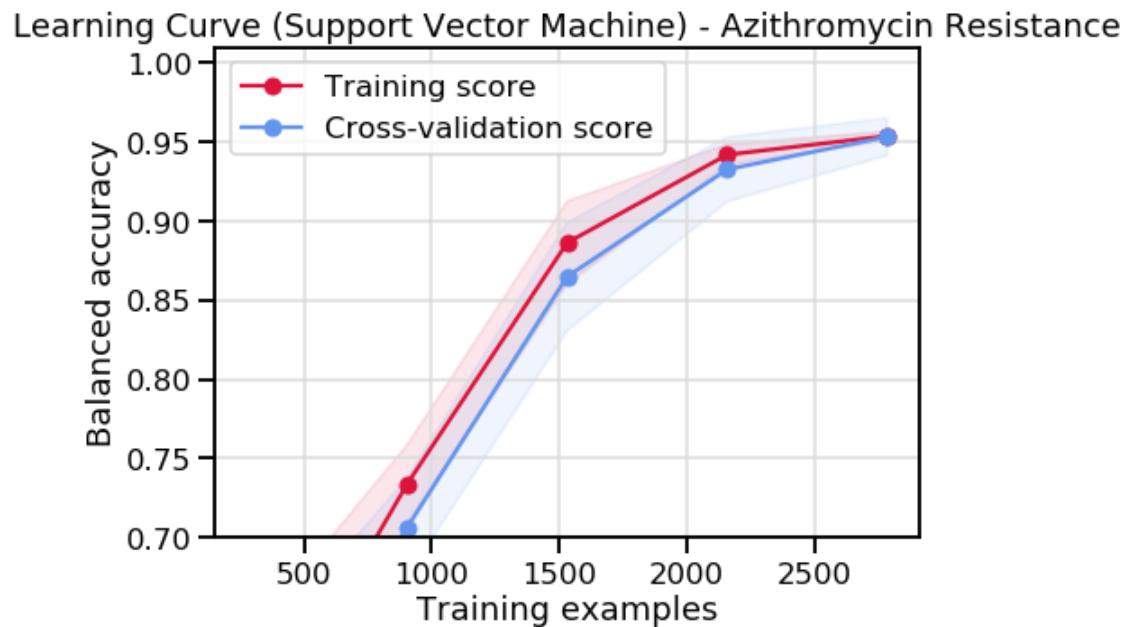
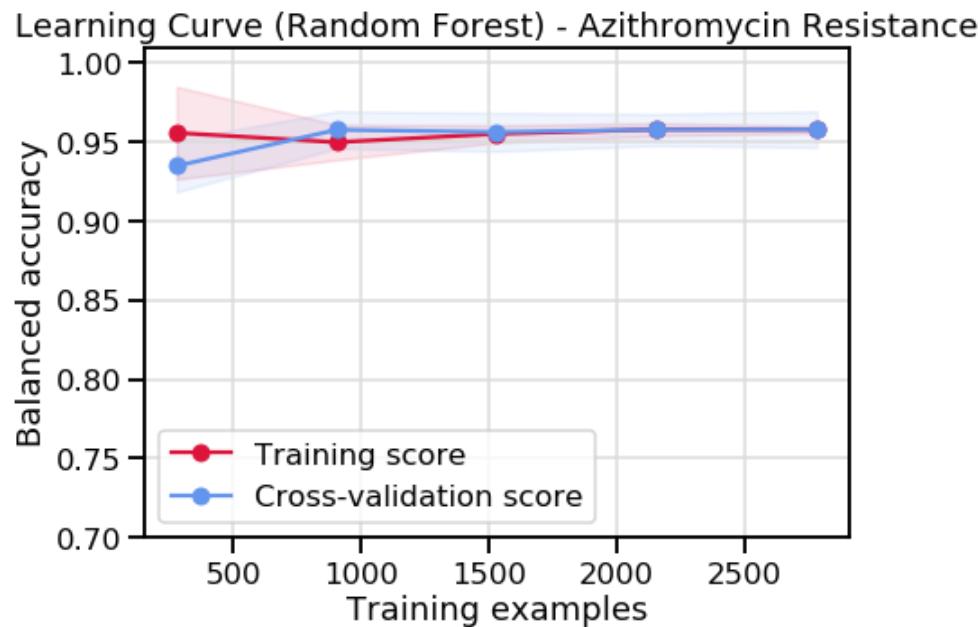


Top features don't include known mechanisms



Top features include known mechanisms

# An aside



Different algorithms can reach the same accuracy, but with very different numbers of samples

# Characterising learning abilities

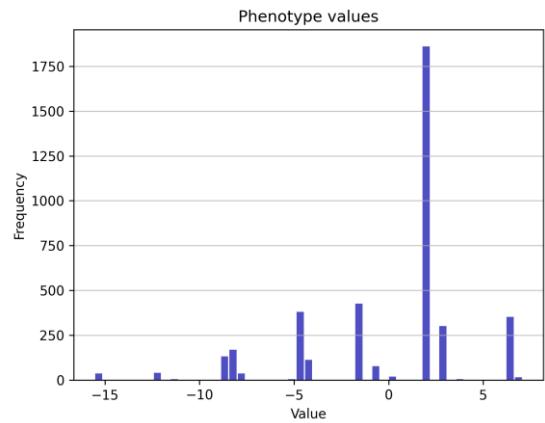
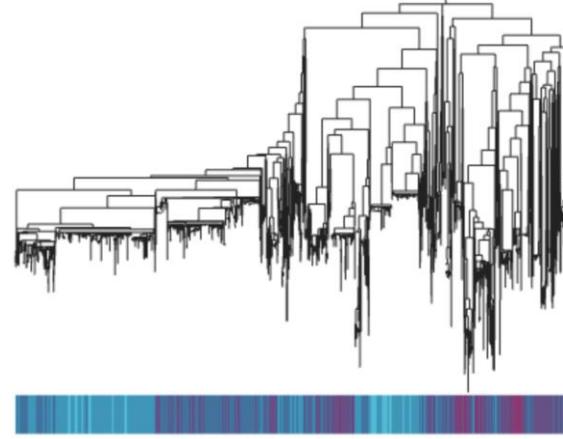
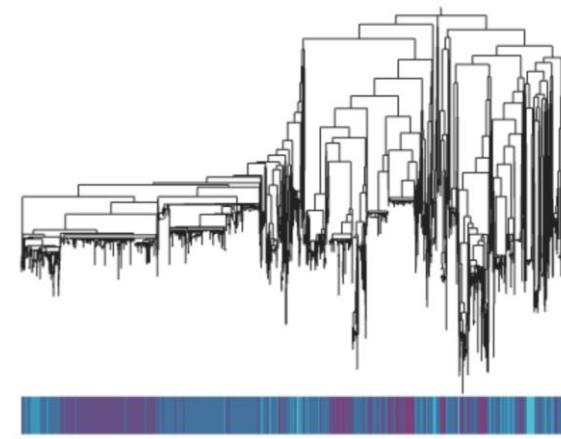
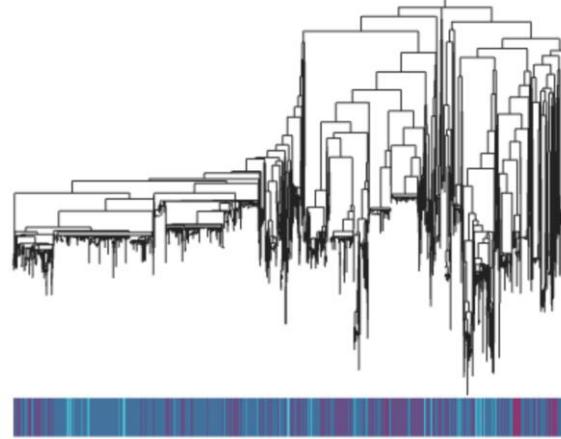
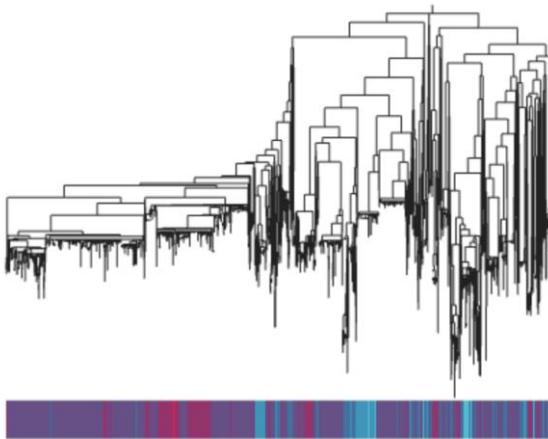
# Simulating phenotypes

- Same *N. gonorrhoeae* genomes
- Pick out causal genes from the core-ish (80%) genome
- ID unitigs that map to those genes
- Filter by unitig frequency --maf 0.05
- Set N unitigs per gene as causal
- Simulate phenotypes with GCTA
  - Heritability = 1
  - Quantitative trait

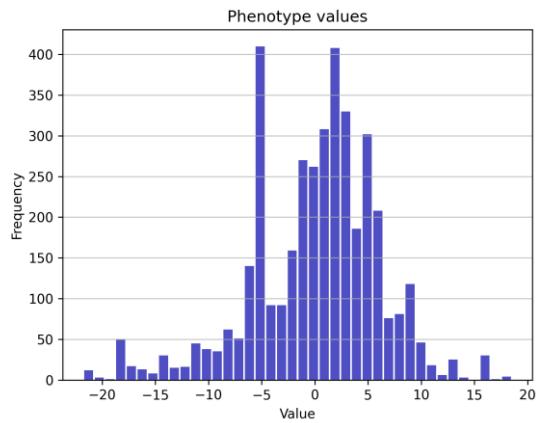
# Scenarios

- 5 causal unitigs sampled from 1 causal gene
  - 25 causal unitigs sampled from 5 causal genes
  - 100 causal unitigs sampled from 5 causal genes
  - 250 causal unitigs sampled from 50 causal genes
- 
- 5 repeats of phenotype generation each
  - 5 repeats of ML training each – different train/test split each time
  - Elastic net and random forest

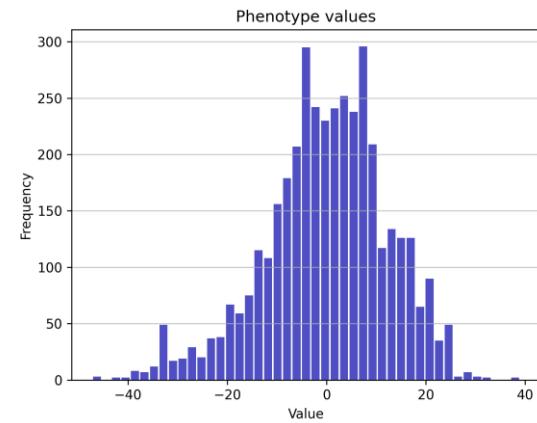
# Phenotype data



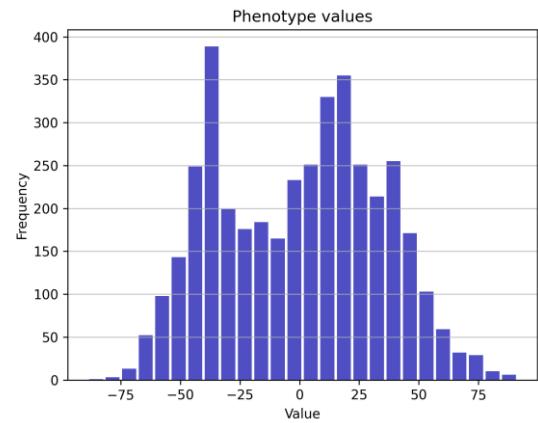
5 unitigs from 1  
gene



25 unitigs from 5  
genes

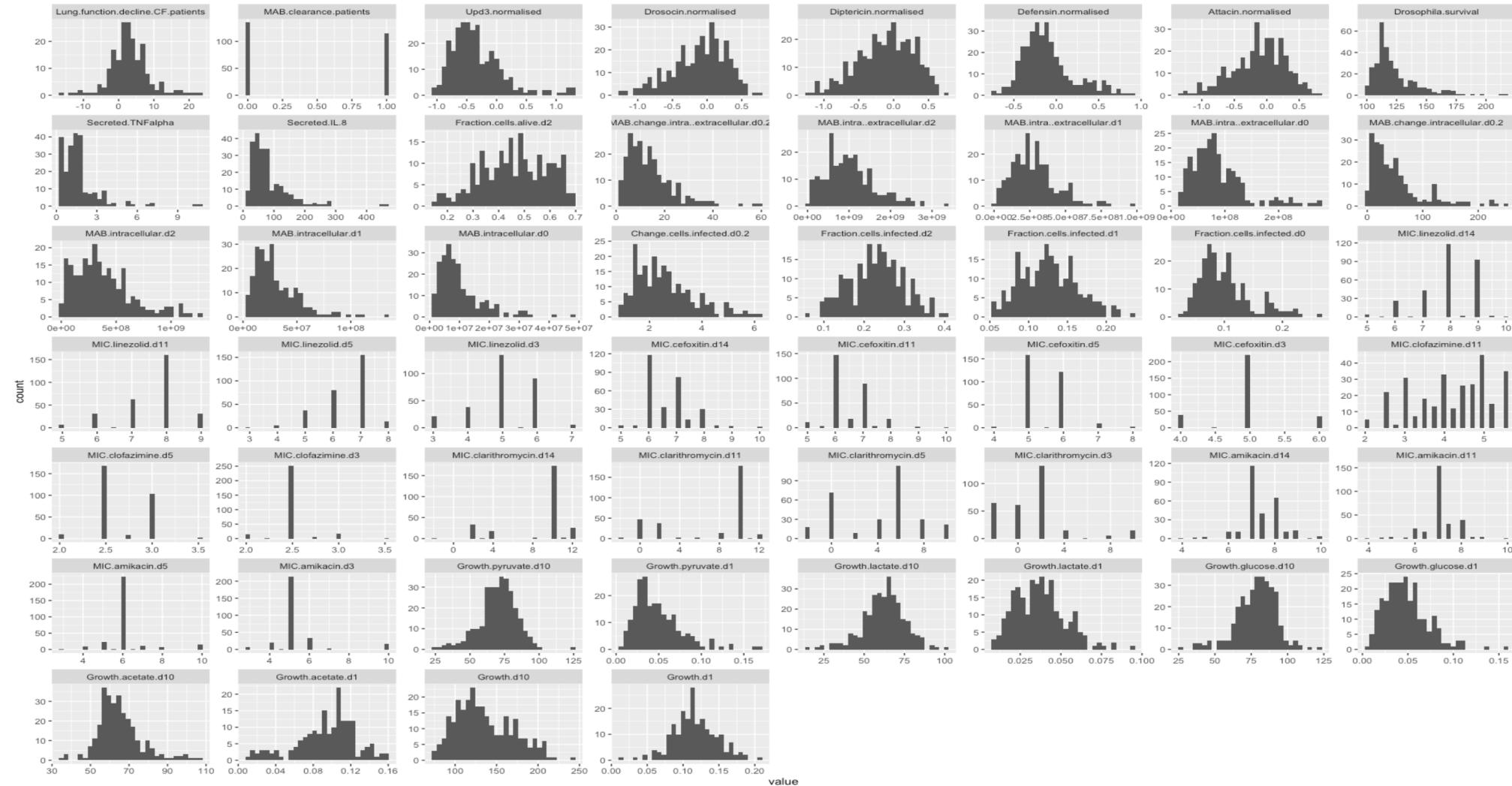


100 unitigs from 5  
genes



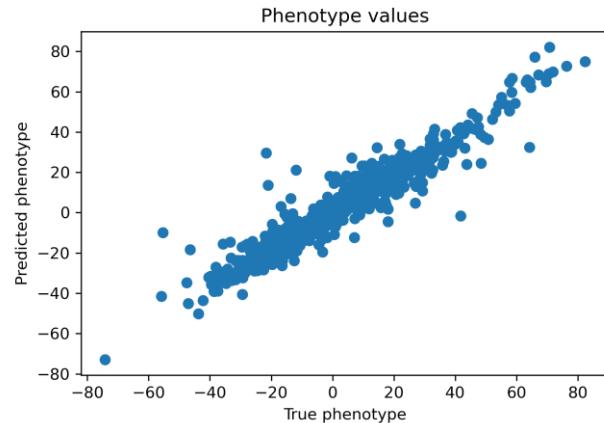
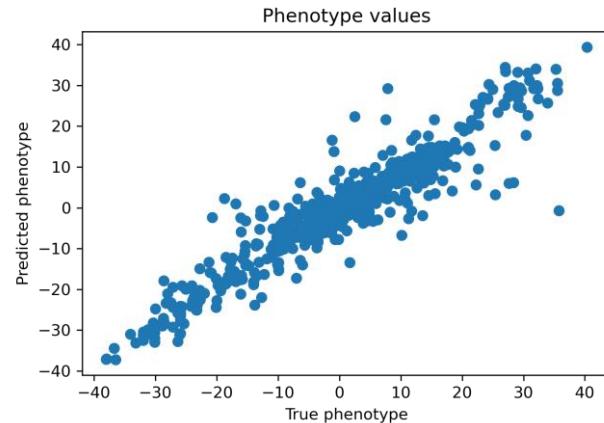
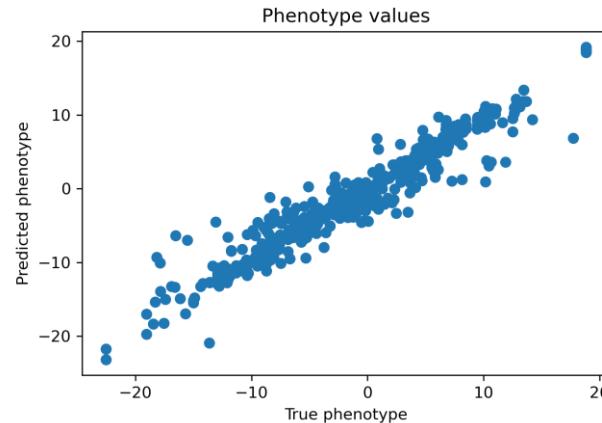
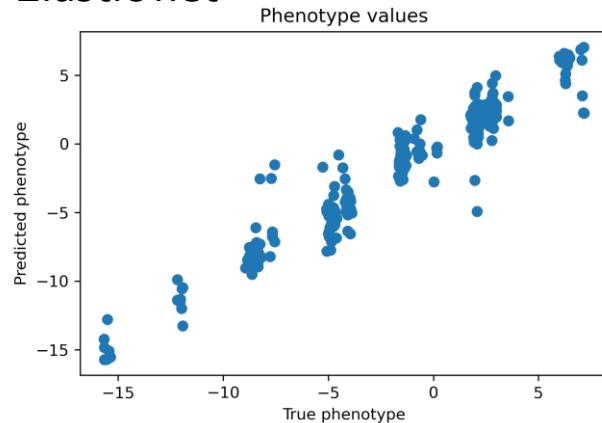
250 unitigs from 50  
genes

# Compared to real phenotypes

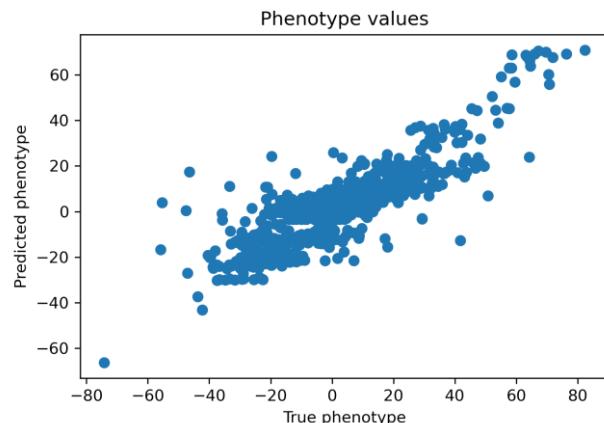
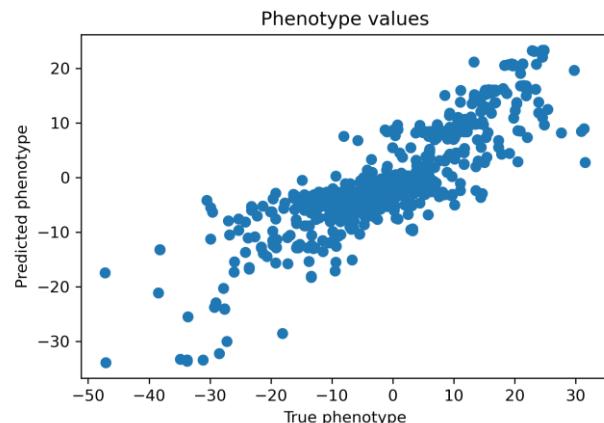
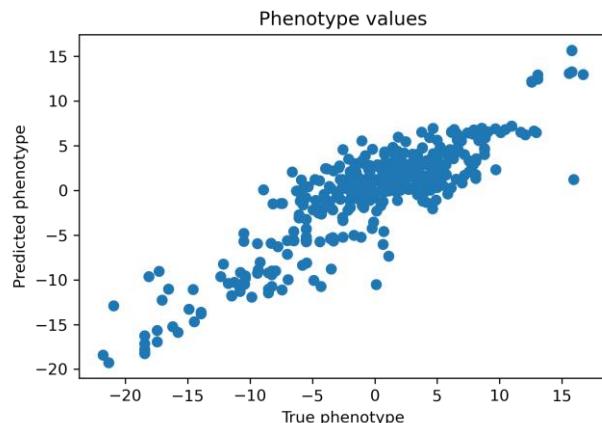
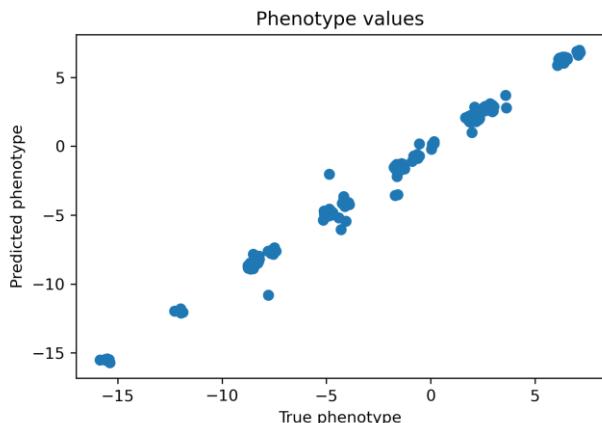


# Prediction of simulated values

Elastic net



Random forest



\*values here are jittered for visibility

5 unitigs from 1 gene

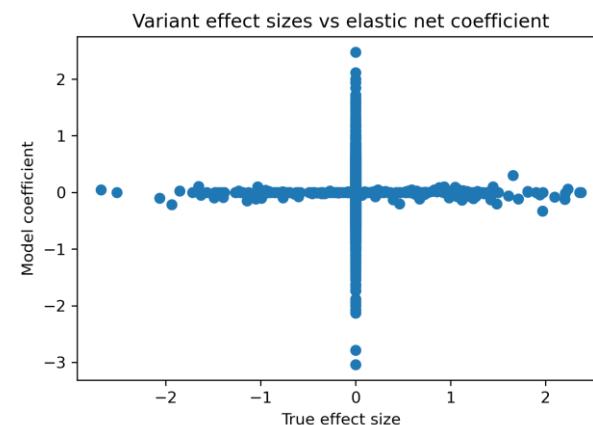
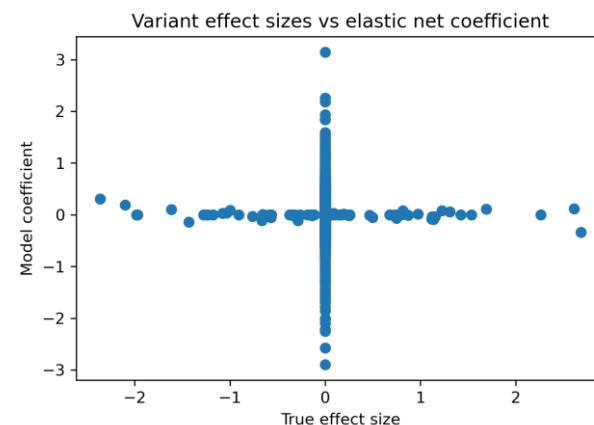
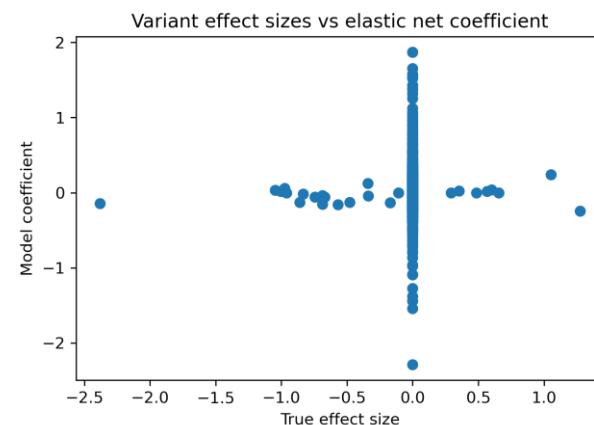
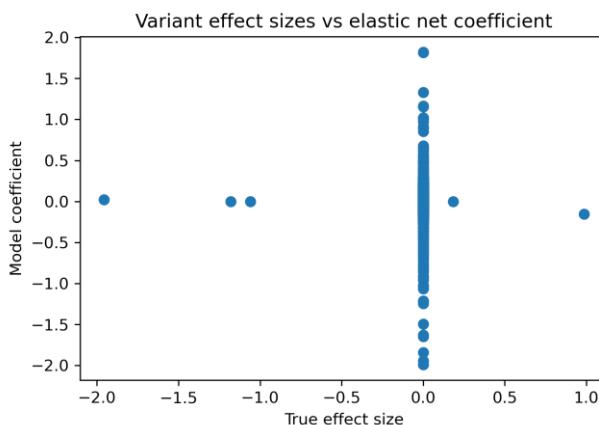
25 unitigs from 5 genes

100 unitigs from 5 genes

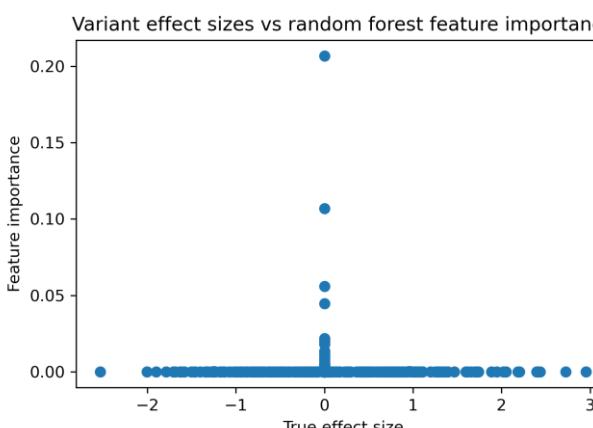
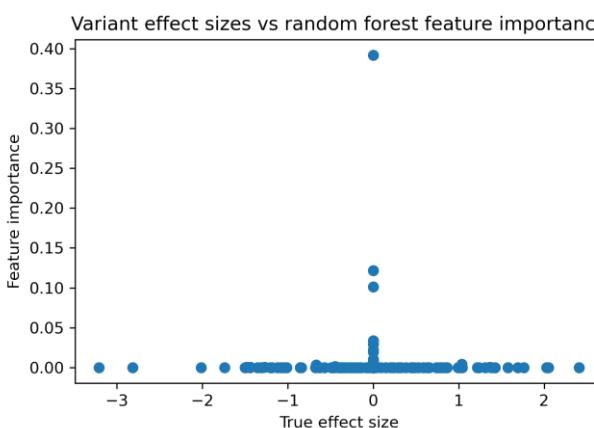
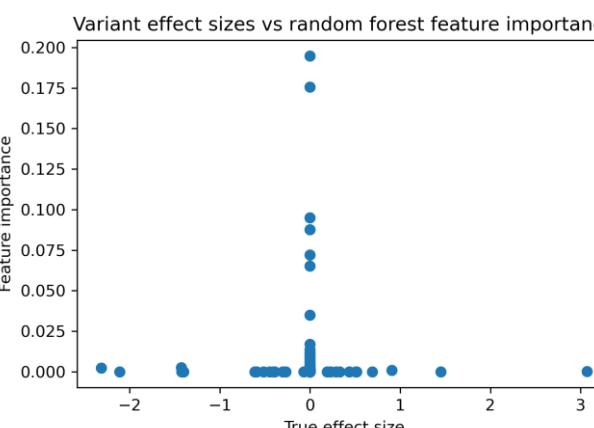
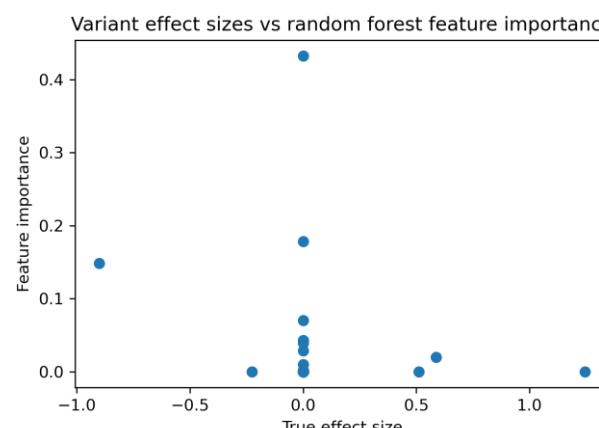
250 unitigs from 50 genes

# Capture of causal unitigs

## Elastic net



## Random forest



5 unitigs from 1 gene

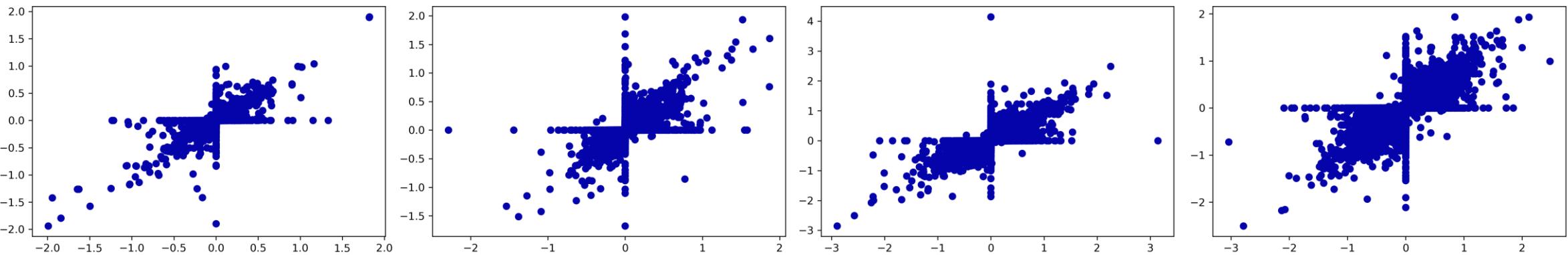
25 unitigs from 5 genes

100 unitigs from 5 genes

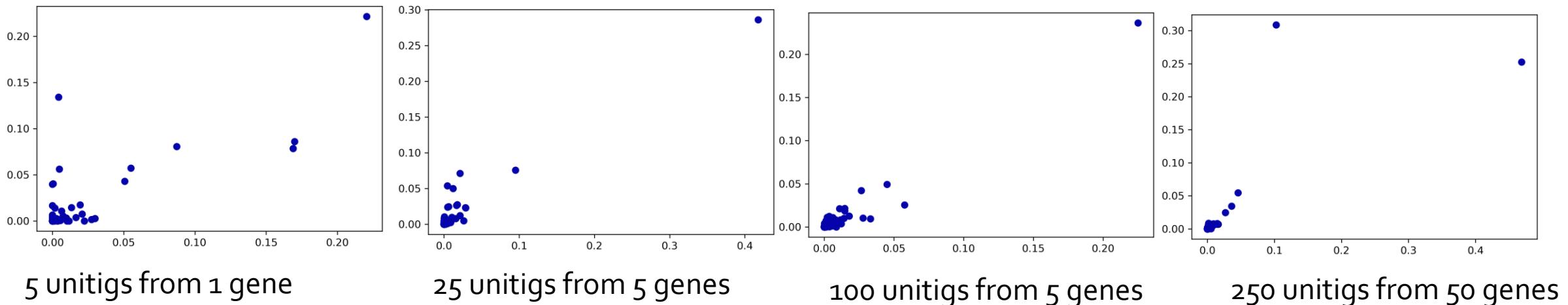
250 unitigs from 50 genes

# Are the same predictive unitigs chosen each time?

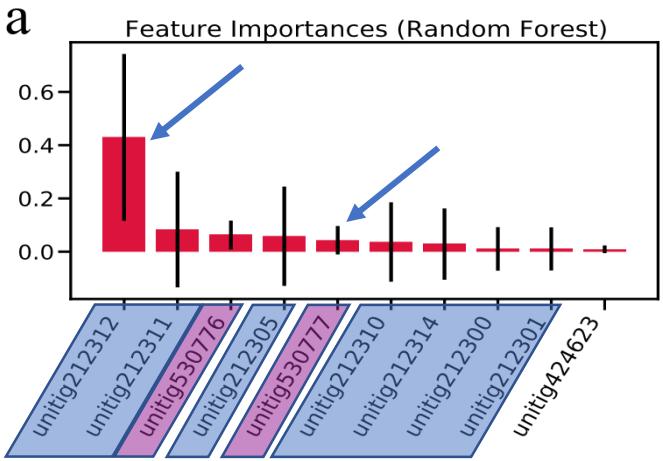
Elastic net



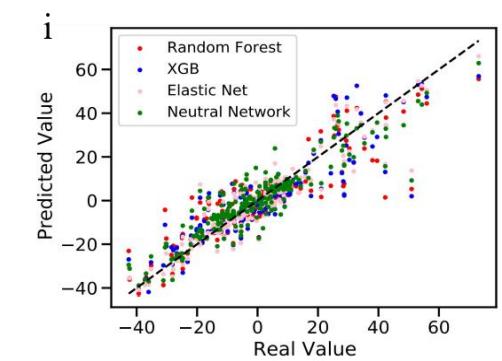
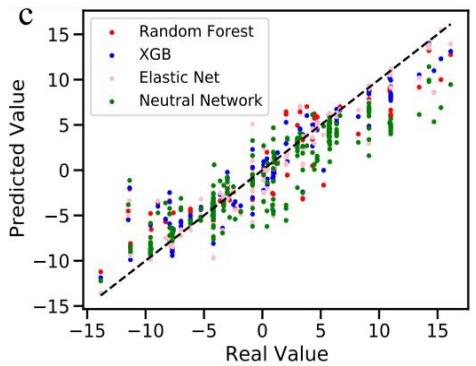
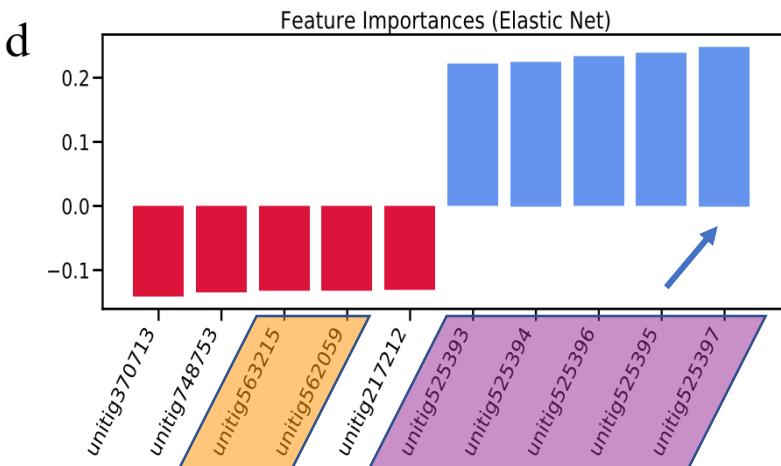
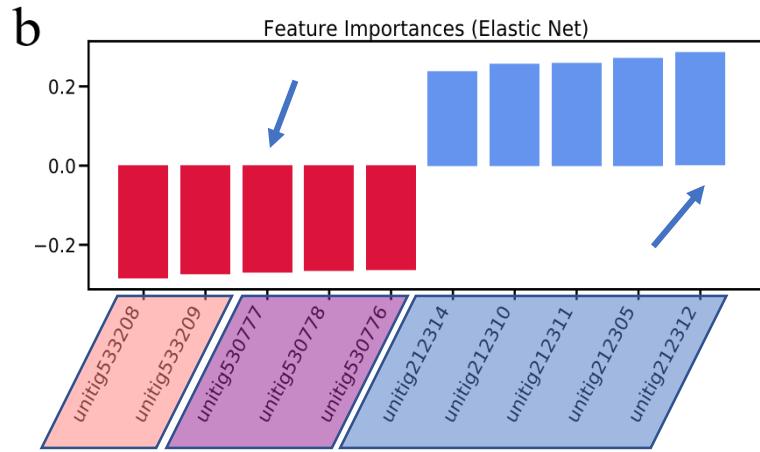
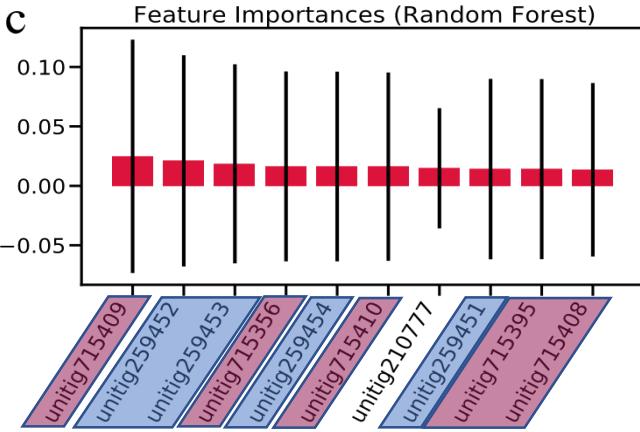
Random forest

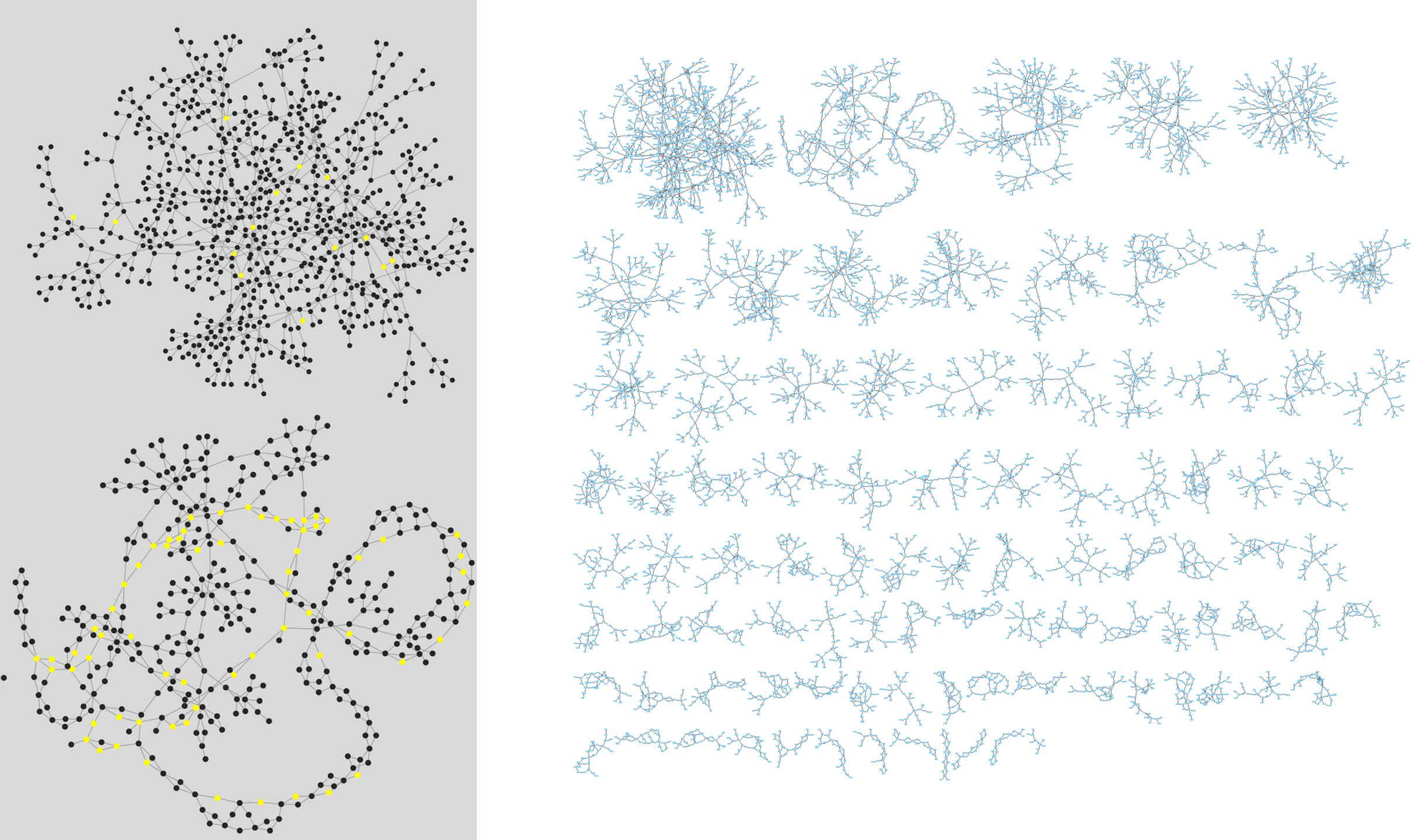


10 causal  
unitigs



200 causal  
unitigs







# Conclusions

- If phenotype can be measured perfectly, ML models can predict (quantitative) traits of varying complexity with high accuracy
- Accurate predictions of the trait can be made without learning the correct magnitude or direction of effect of causal unitigs
- Some *regions* of the unitig graph can be reliably identified as causal
  - Within these, there may be multiple good solutions for predicting phenotype from genotype within the training data

# Improvements/next steps

- Evaluating and reporting on ML algorithms
  - 4000 samples could be great or terrible – papers should report effective sample number
  - Show the mapping of the trait to a phylogenetic tree – how many independent evolutionary events have been captured?
  - Better measure of the generalizability of algorithms in publications
- Publishing
  - Make the model easy to run on new data
- Better communication of uncertainty
  - Communicate when a new sample falls outside the diversity of previously seen samples

# Thank you!

- Acknowledgements
  - Ge Zhou
  - John Lees
  - Jukka Corander
  - Gerry Tonkin Hill
  - Yonatan Grad
  - Lucas Boeck

Contact: N.Wheeler@bham.ac.uk

@nwheeler443

