BACTERIAL ASSOCIATION STUDIES FOR DRUG RESISTANCE & VIRULENCE

Danny Wilson
Associate Professor and Sir Henry Dale Fellow
Nuffield Department of Medicine, University of Oxford

BACKGROUND GENOME-WIDE ASSOCIATION STUDIES IN BACTERIA

GWAS in bacteria

Example questions

- *Do Staphylococcus aureus genomes vary in their propensity to cause invasive disease vs asymptomatic carriage?*
- *Are there genetic differences in the Clostridium difficile population that explain differential mortality among infections?*
- *Can we pinpoint the mutations conferring antibiotic resistance in Mycobacterium tuberculosis?*

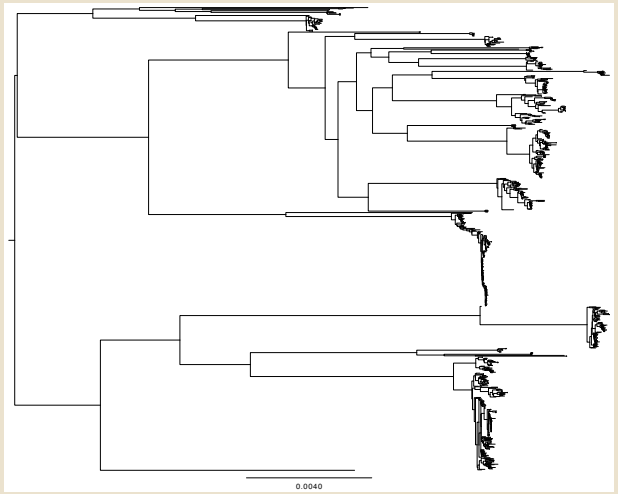
GWAS in bacteria: methods

Can we adapt Genome Wide Association Studies (GWAS) to address challenges particular to bacteria?

- *Strong structuring of populations into distinct clusters of highly-related strains*
- *Weak decay of linkage disequilibrium within the core genome*
- *Highly mobile accessory genomes*

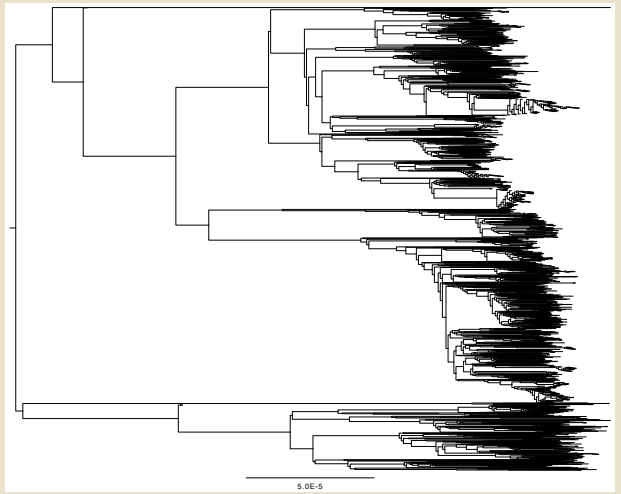
Will Bacterial GWAS have power to fine map the genetic basis of key traits?

1. The Challenge of Bacterial Population Structure



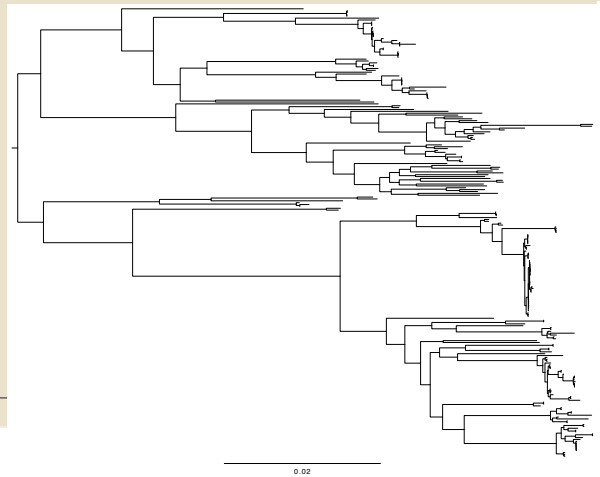
Staphylococcus aureus

Clostridium difficile



Mycobacterium tuberculosis

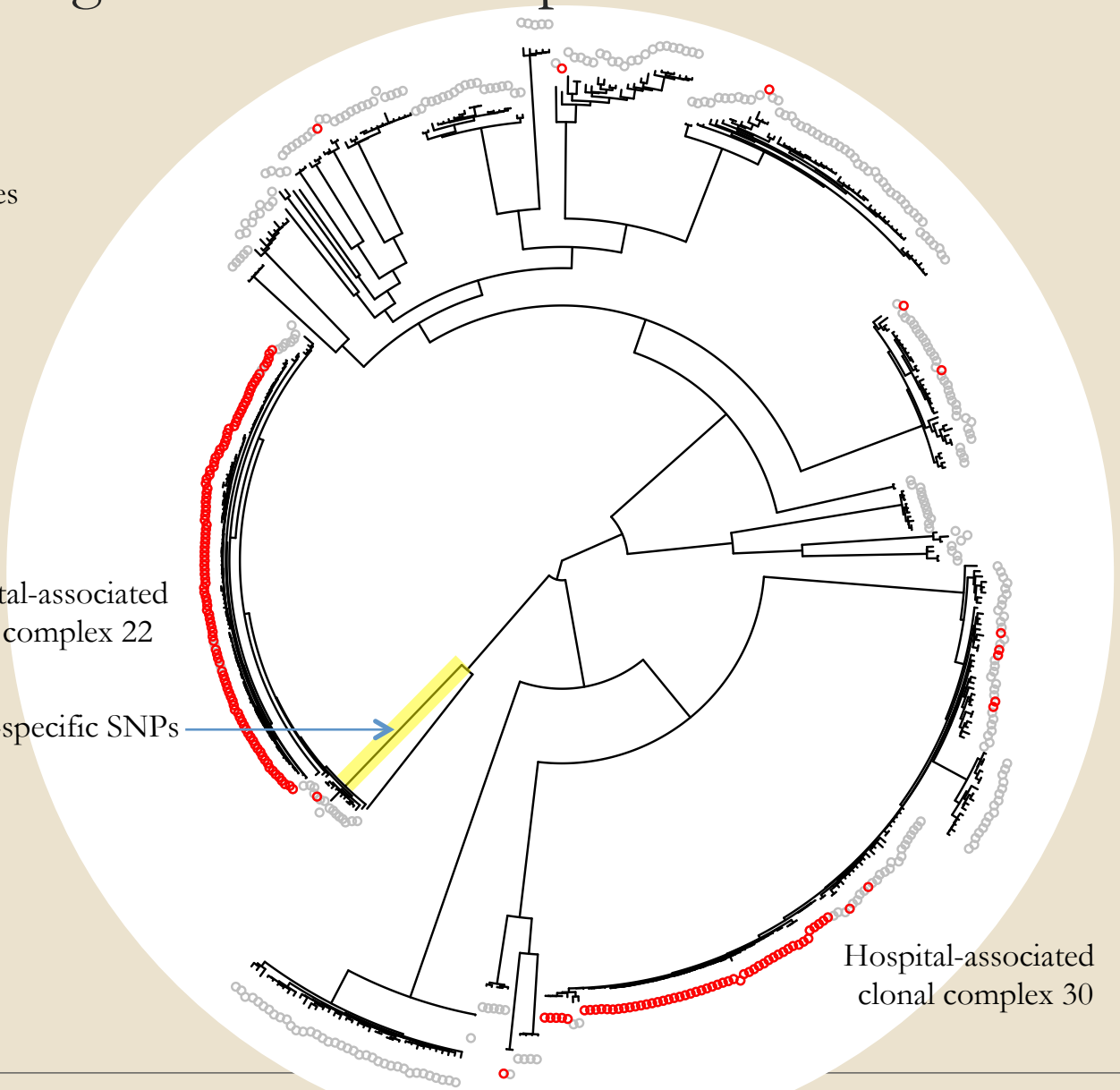
Escherichia coli



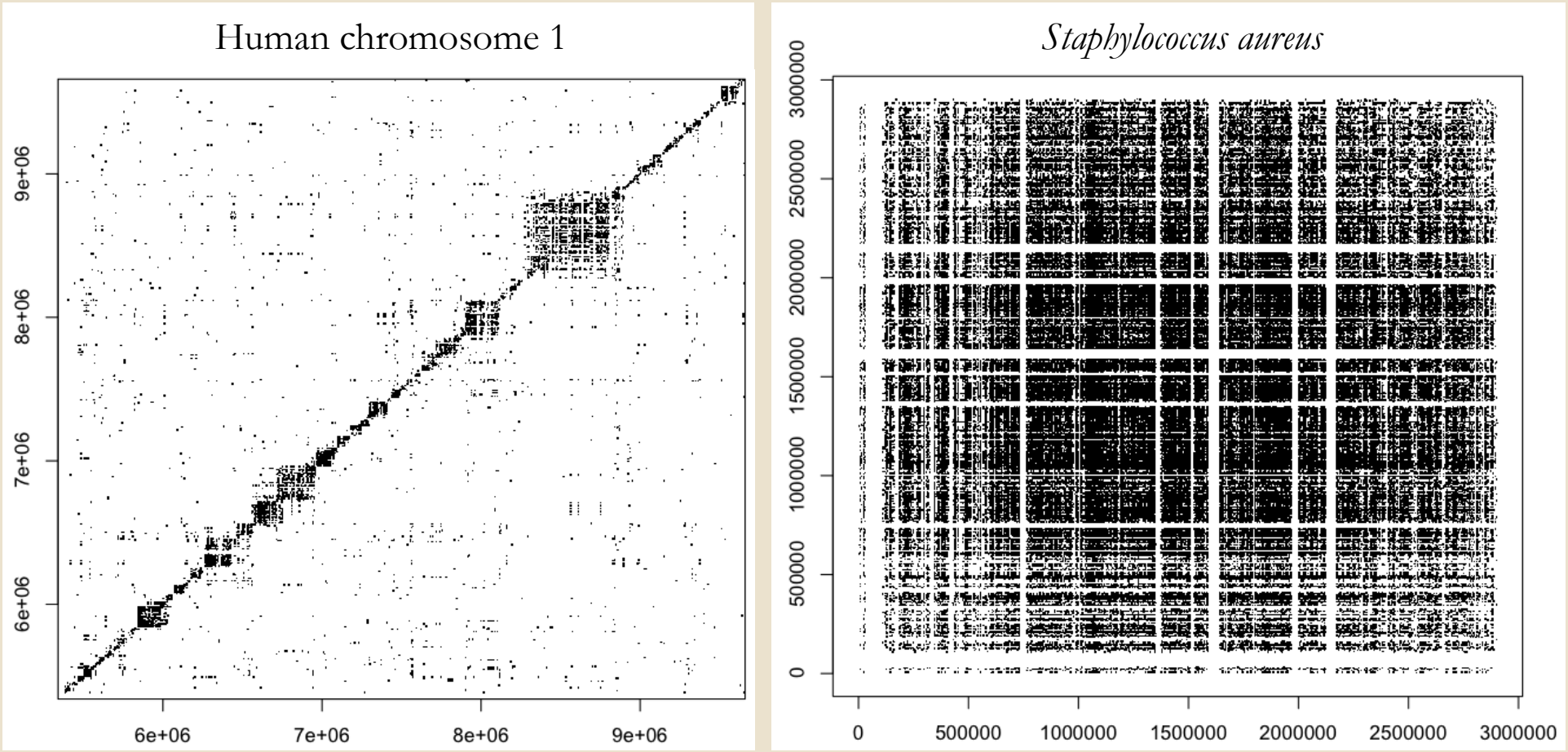
Klebsiella pneumoniae

1. The Challenge of Bacterial Population Structure

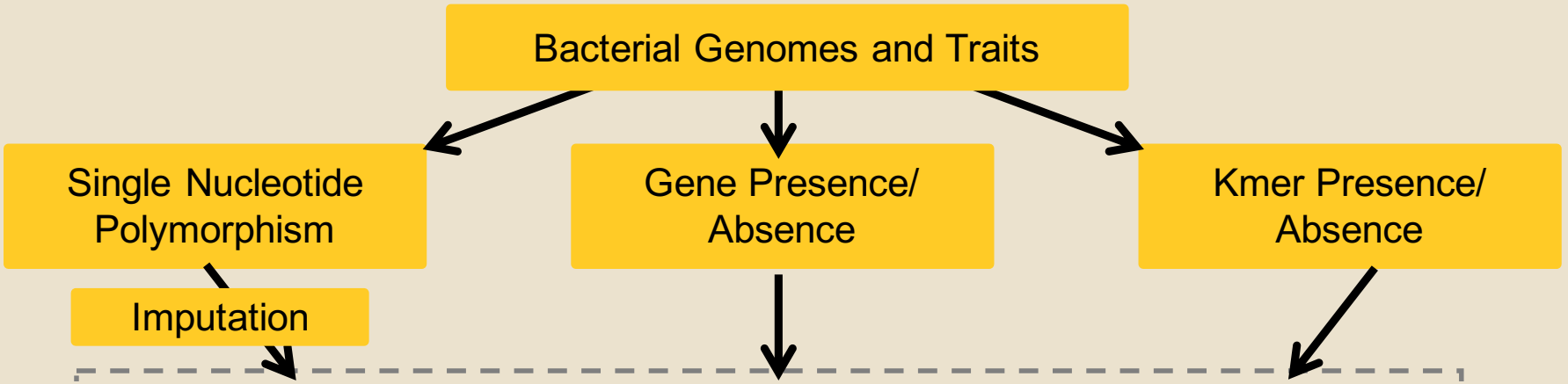
- Ciprofloxacin Resistant isolates
- Ciprofloxacin Susceptible isolates



2. The Challenge of Bacterial Linkage Structure



Bacterial GWAS pipeline



- To capture diverse forms of bacterial genetic diversity (SNPs, indels, repeats, gene presence/ absence, mobile elements) we adopt a multipronged approach

Null hypothesis: *no genetic variant is associated with case vs control status*

gle nucleotide polymorphisms (SNPs)

	Controls (Susceptible)	Cases (Resistant)
C /Ser)	326	10
T /Leu)	3	161

Resistance to ciprofloxacin conferred by S84L substitution in the *gyrA* gene
$p = 10^{-114.9}$

Presence vs absence of entire genes

	Controls (Susceptible)	Cases (Resistant)
Absent ($blaZ^-$)	59	3
Present ($blaZ^+$)	0	439

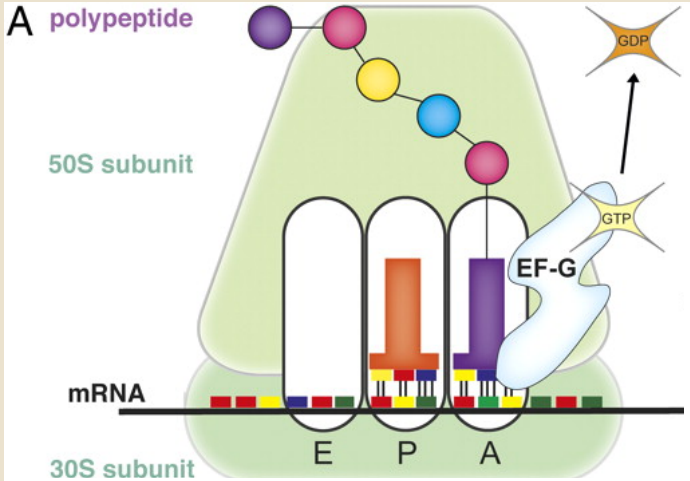
e.g. Resistance to penicillin conferred by the *blaZ* gene
$p = 10^{-71.3}$

Presence vs absence of short haplotypes (kmers)

	Controls (Susceptible)	Ca (Resi
Absent ($kmer^-$)	342	
Present ($kmer^+$)	0	1

e.g. Resistance to methicillin conferred by the *mecA* gene that contains th
AAAACAAGTTATAAAATCGATGGTAAAGGTT
$p = 10^{-109.2}$

Fusidic acid resistance in *Staphylococcus aureus*



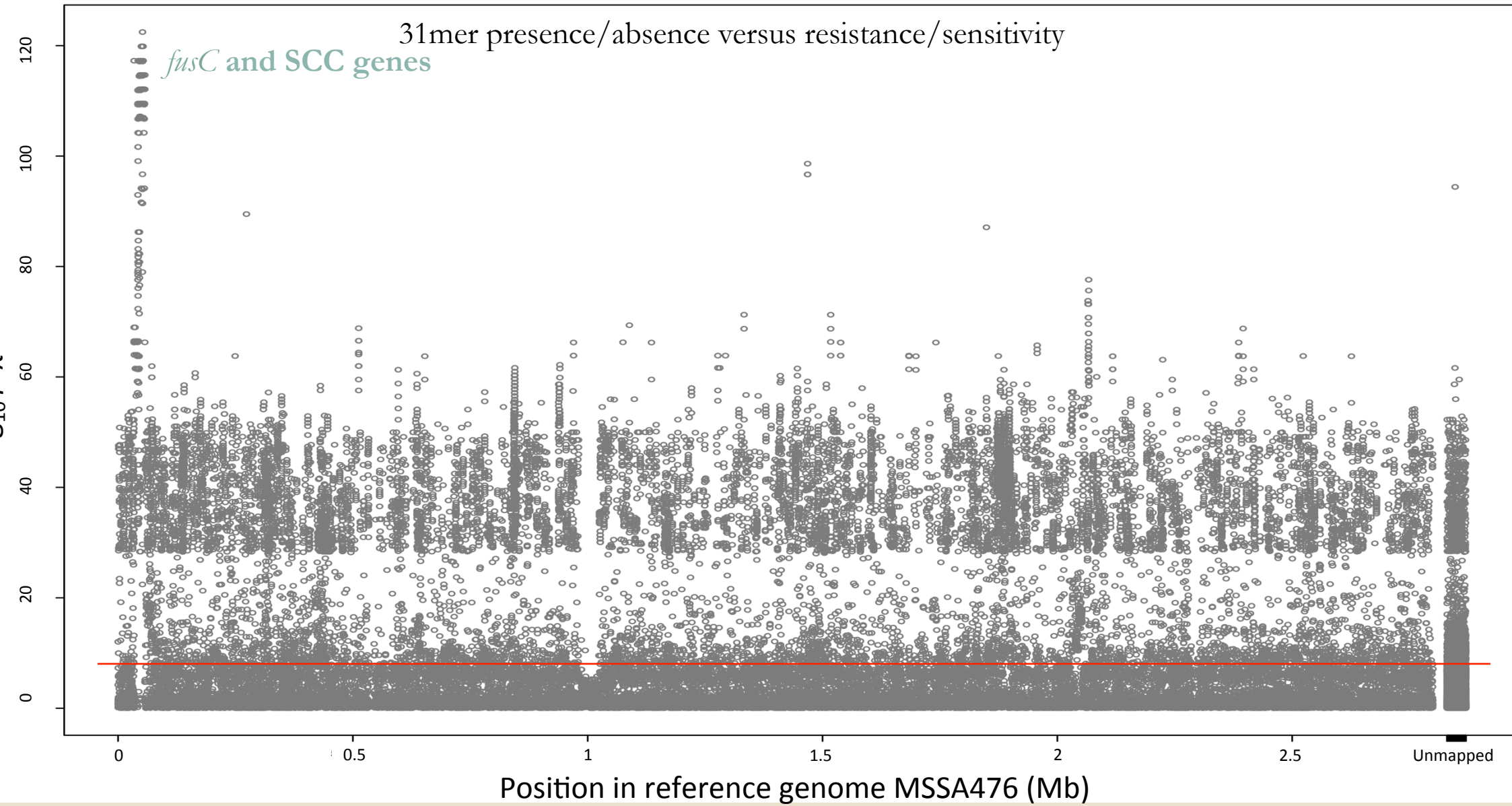
Protein synthesis requires elongation factor G (EF-G) to proceed normally

Accessory genes *fusB* and *fusC* encode chaperones that bind to EF-G and destablize binding of FA. This promotes dissociation of FA and allows protein synthesis to proceed.

Fusidic acid (FA) binds to EF-G, locking it to the ribosome and stalling protein synthesis. Mutations in *fusA,* the gene encoding EF-G, prevent FA binding, allowing protein synthesis to proceed.

Credit: Georgina Cox

Manhattan plot: Fusidic acid resistance χ^2 test



ADDRESSING THE CHALLENGES OF BACTERIAL GWAS

Controlling for population stratification

- This phrase covers the following artefacts

  - Linkage disequilibrium with genuine causal variants that are population-stratified
  - Uncontrolled environmental variables that are population-stratified
  - Population-stratified differences in sampling
  - Sensitivity to over-sampling close relatives (i.e. same genotypes and environments)

What can go wrong in GWAS?

	Type I errors	**Type II errors**
0. Grand null No association	Find any variant	n/a
1. Simple alternative One causal variant	Find the wrong variant	Miss the right variant
2. Complex null Confounding	Find any variant	n/a
3. Complex alternative Confounding plus one interesting variant	Find the wrong variant	Miss the right variant

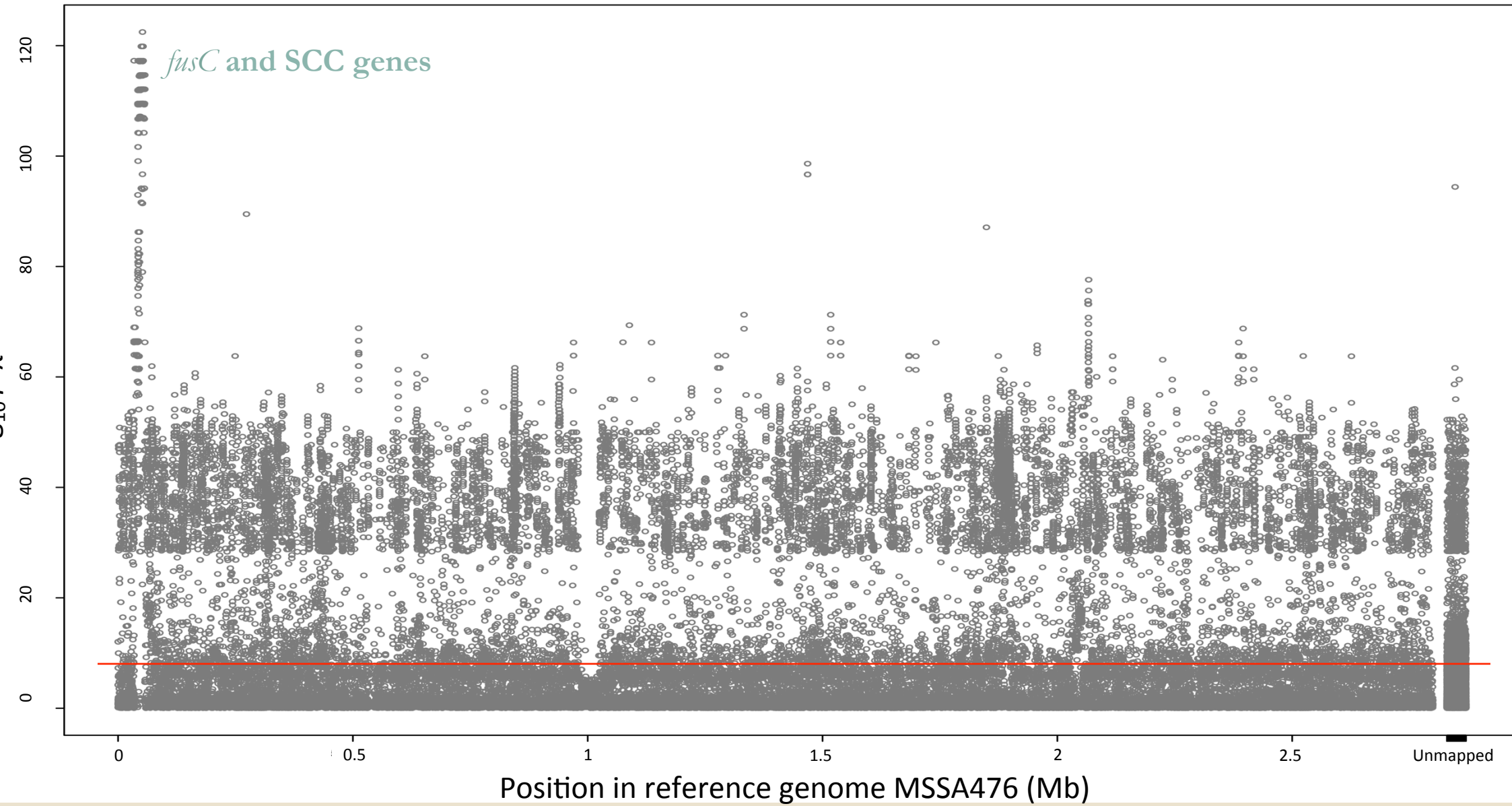
In the eukaryotic setting, where LD is localized and block-like, determination of whether a GWAS hit is right or wrong is often considered at the level of the LD block, rather than individual variants.
For bacteria, where LD is not block-like, this does not work.

Control of population structure

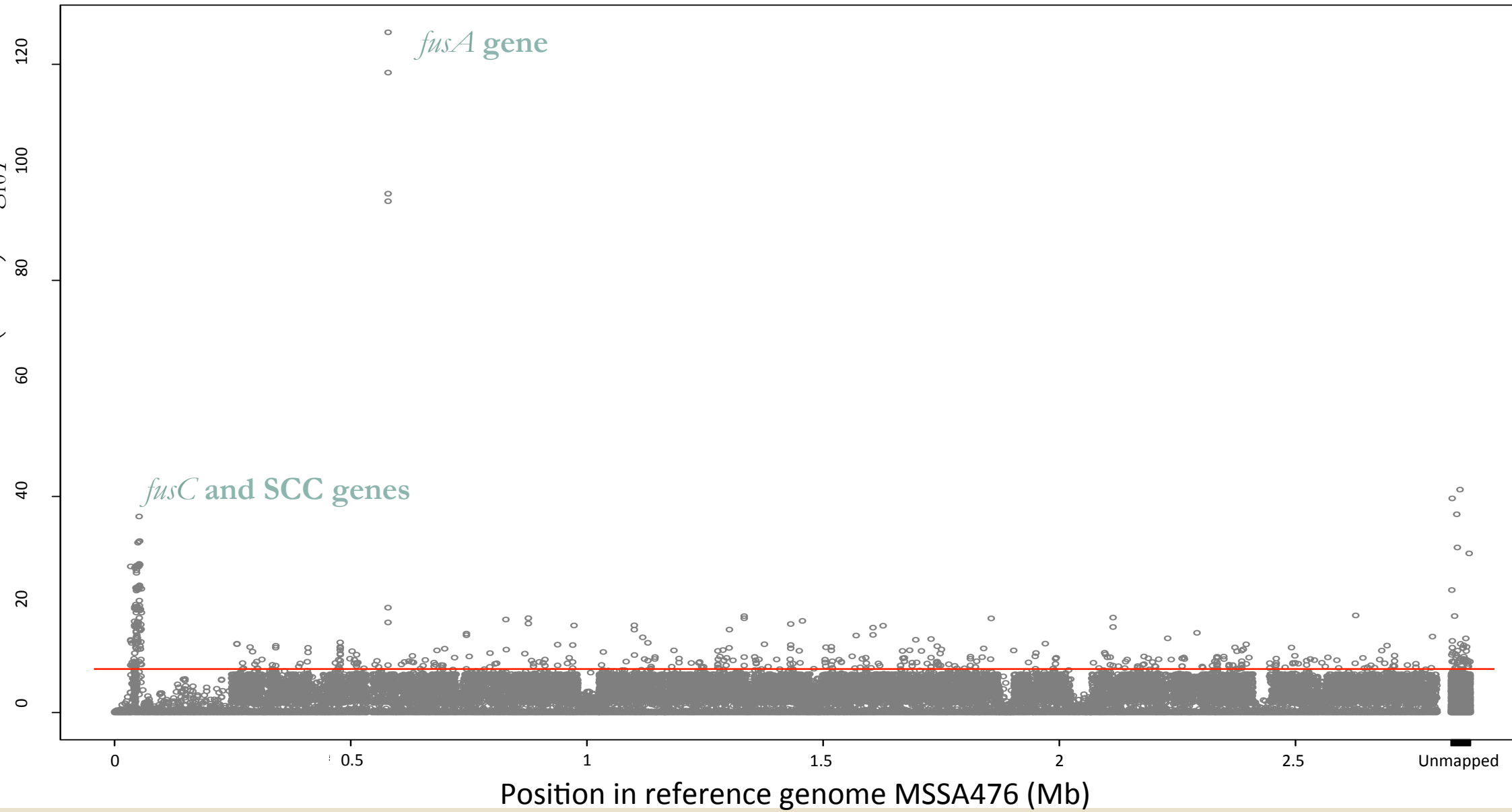
- Genomic control
  - Directly adjust the p-values so the vast majority are non-significant
- Subpopulations
  - Expect individuals in the same subpopulation to have phenotypes that are more similar
- Principal components analysis (PCA)
  - Expect individuals with similar principal components to have phenotypes that are more similar
- Linear mixed models
  - Expect individuals that are closely related to have phenotypes that are more similar

BACTERIAL GWAS IN PRACTICE

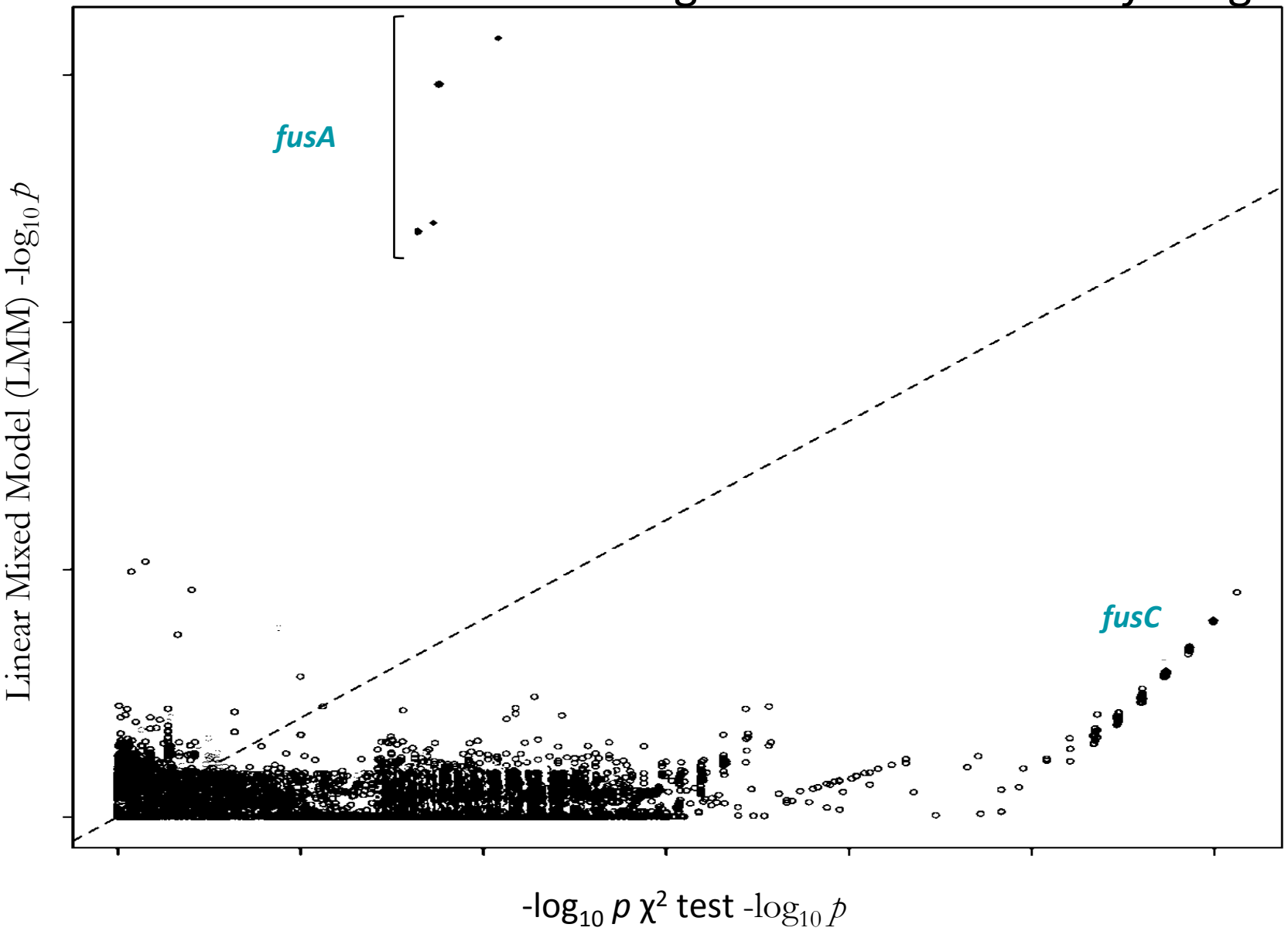
Fusidic acid resistance χ^2 test



Fusidic acid resistance χ^2 test



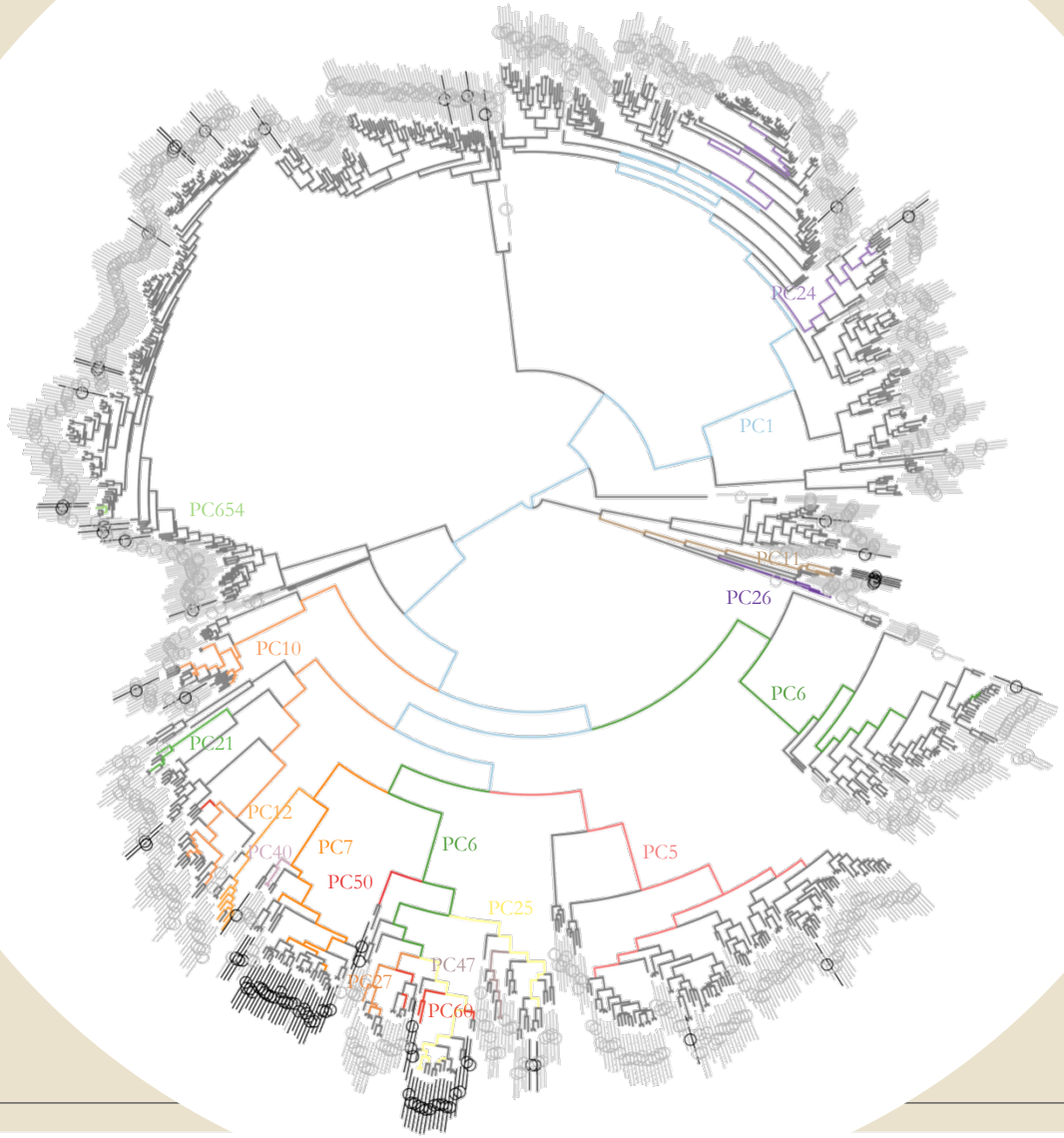
Where did the signal of association at *fusC* go?



Controlling for population structure reduced significance of *fusC* kmers from $p < 10^{-122}$ to $p < 10^{-39}$

Kmers capturing *fusA* increased from $p < 10^{-122}$ to $p < 10^{-39}$

Leading principal components correspond to major lineages

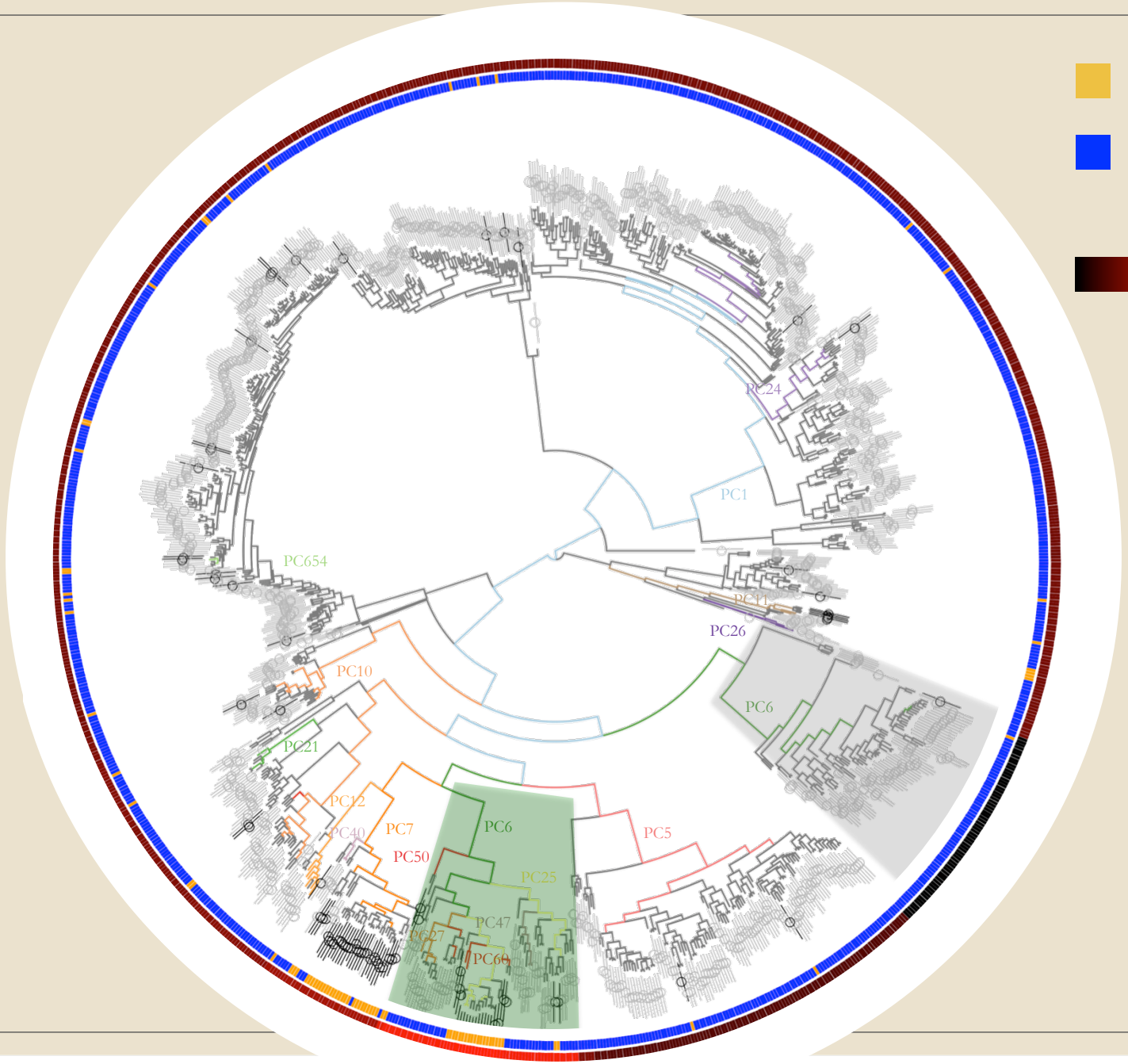


True phenotype:

- Sensitive
- Resistant

LMM predicted phenotype:

- Sensitive
- Resistant



Resistant

Sensitive

Projection of individual onto PC 6

True phenotype:

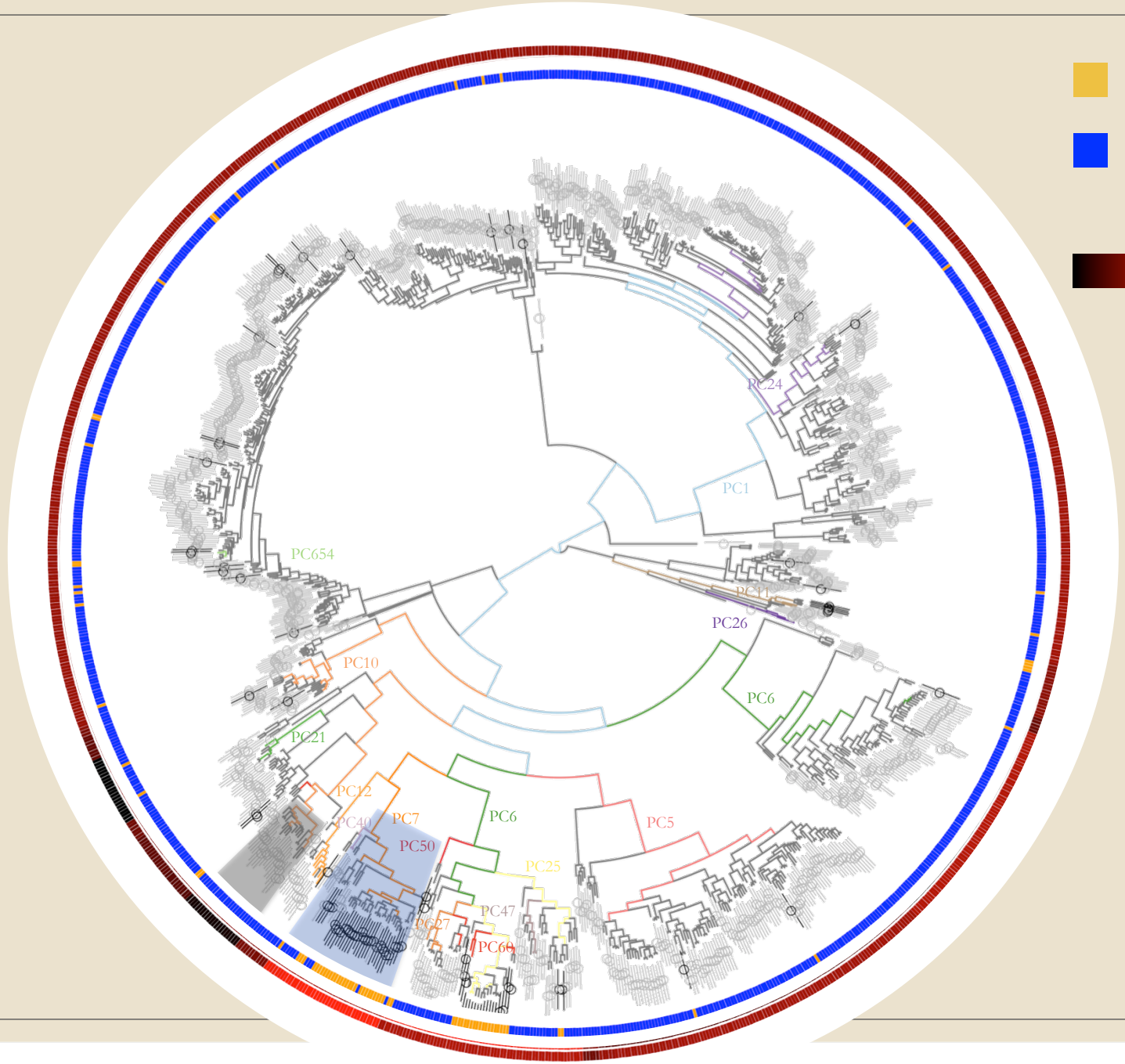
Sensitive

Resistant

LMM predicted phenotype:

Sensitive

Resistant



Resistant

Sensitive

Projection of individual onto PC 9

True phenotype:

Sensitive

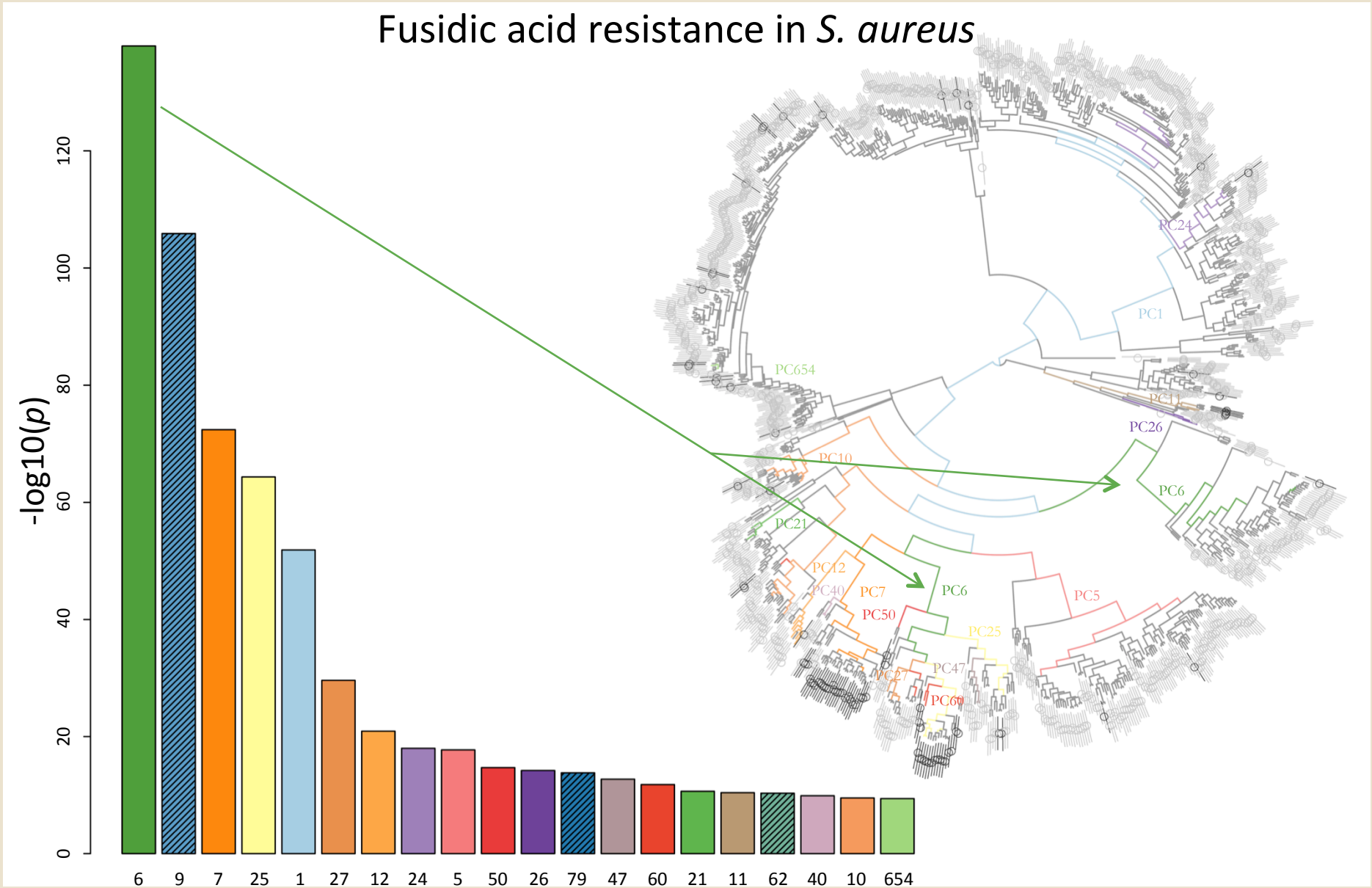
Resistant

LMM predicted phenotype:

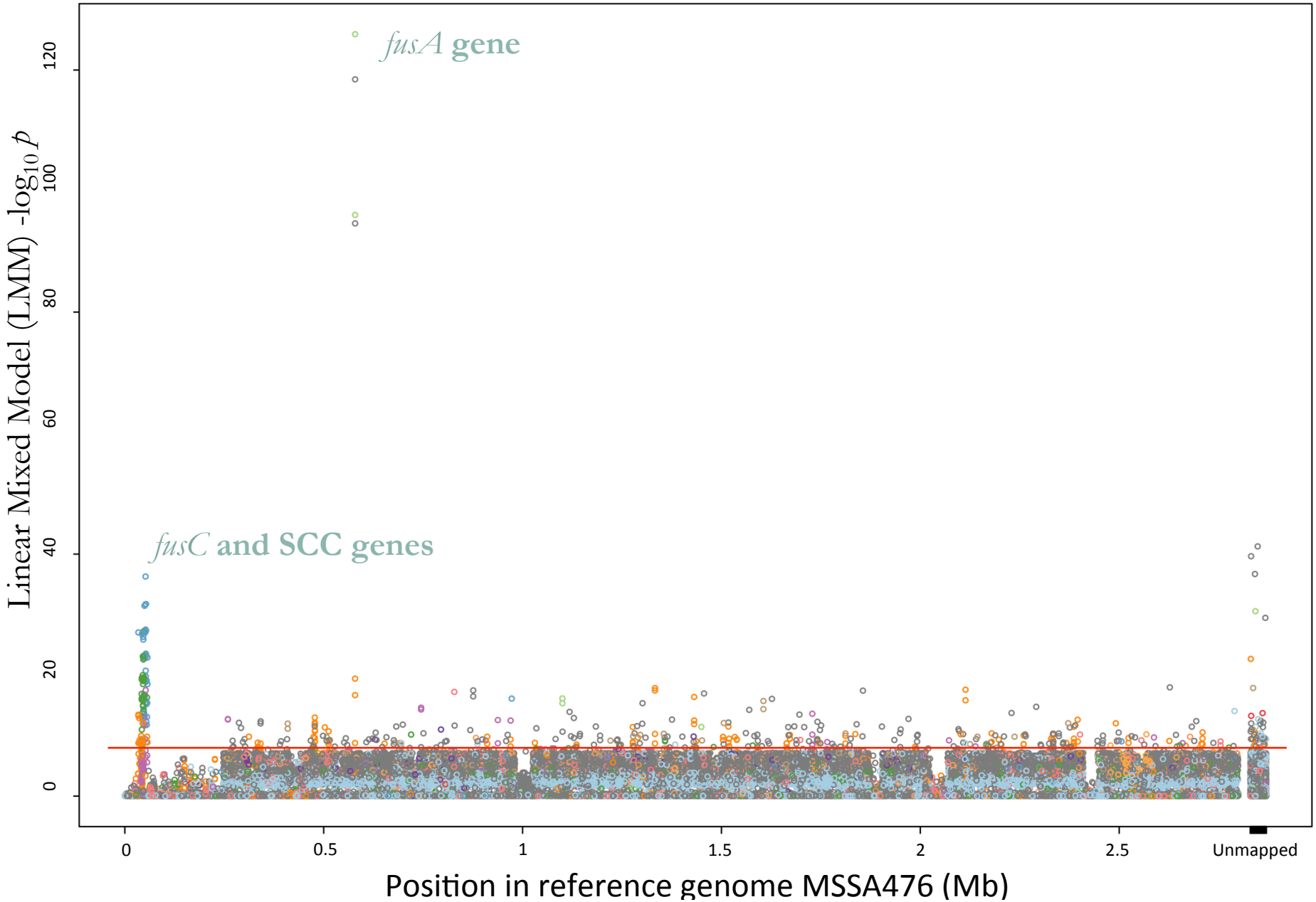
Sensitive

Resistant

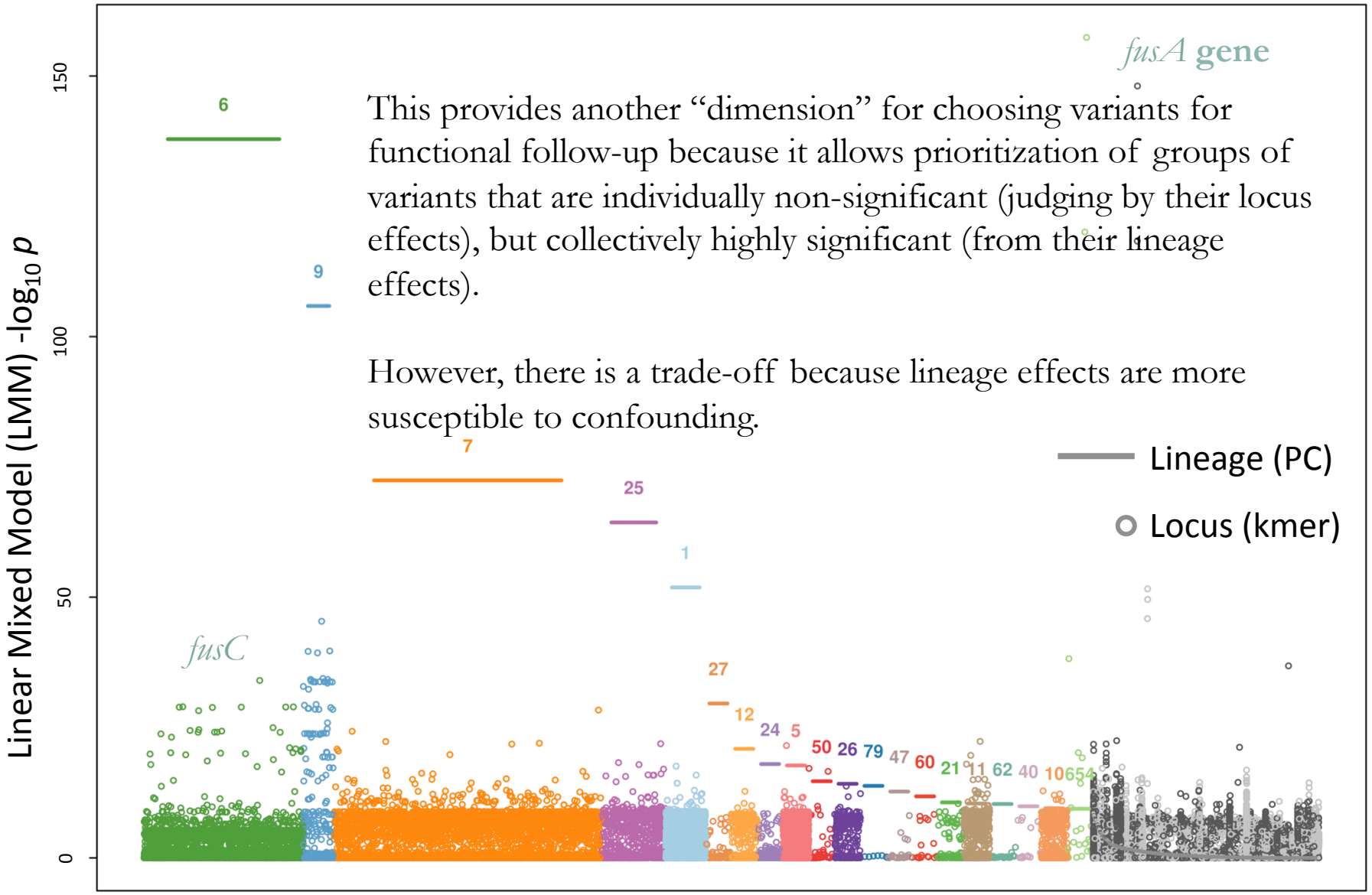
bugwas Wald Test for Lineage Effects



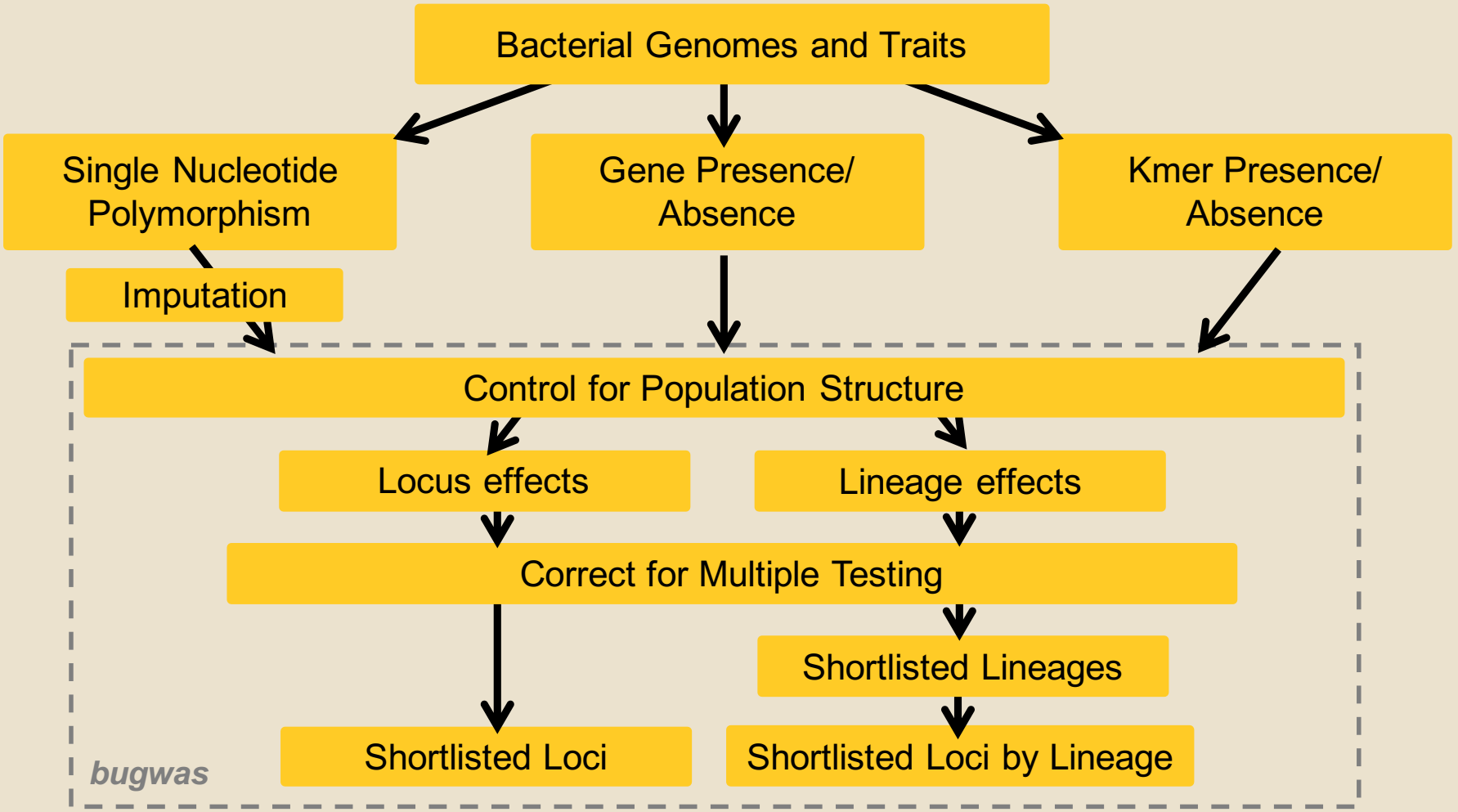
Fusidic acid resistanc: Manhattan plot



Fusidic acid resistance: *bugwas* plot



Bacterial GWAS pipeline



Antibiotic	# R	# S	Resistance mechanism	SNP / gene LMM	Kmer LMM
oli					
Ampicillin	189	52	β-lactamase genes *blaTEM*		
Cefazolin	62	179	β-lactamase genes *blaCTX-M*		
Cefuroxime	81	160	β-lactamase genes *blaCTX-M*		
Ceftriaxone	55	186	β-lactamase genes *blaCTX-M*		
Ciprofloxacin	91	150	SNPs in *gyrA*[a], *gyrB*, *parC*[b] or *parE* or presence of PMQR		
Gentamicin	48	193	AAC (*aac(3)-II*), ANT, APH or rRNA methylase		
Tobramycin	67	174	AAC (*aac(3)-II*), ANT, APH or rRNA methylase		
neumoniae					
Cefazolin	38	138	β-lactamase genes *blaCTX-M*		
Cefuroxime	46	130	β-lactamase genes *blaCTX-M*		
Ceftriaxone	35	141	β-lactamase genes *blaCTX-M*		
Ciprofloxacin	34	142	SNPs in *gyrA*, *gyrB*, *parC* or *parE* or presence of PMQR (*qnr-B1*[a], *qnr-B19*[b])		
Gentamicin	31	145	AAC (*acc(3)-II*), ANT, APH or rRNA methylase		
Tobramycin	36	140	AAC (*acc(3)-II*), ANT, APH or rRNA methylase		
uberculosis					
Ethambutol	41	1589	*embB*		
Isoniazid	239	1470	*katG*, *mabA* or *fabG1*		
Pyrazinamide	45	1662	*pncA*		
Rifampin	86	1487	*rpoB*		
ureus					
Ciprofloxacin	242	750	*grlA* or *gyrA*		
Erythromycin	216	776	*ermA*, *ermC*, *ermT* or *msrA*		
Fusidic acid	84	908	SNPs in *fusA*[a] or presence of *fusB* or *far*[b]		
Gentamicin	11	981	*aacA/aphD*		
Penicillin	824	168	*blaZ*		
Methicillin	216	776	*mecA*		
Tetracycline	46	946	*tetK*, *tetL* or *tetM*		
Trimethoprim	15	308	*SNPs in dfrB, presence of dfrG or dfrA*		
Rifampicin	8	984	*rpoB*		

Earle et al 2016 *Nature Microbiology*

Resistance determined by gene presence

Resistance determined by SNPs

Resistance determined by gene presence or SNPs or both

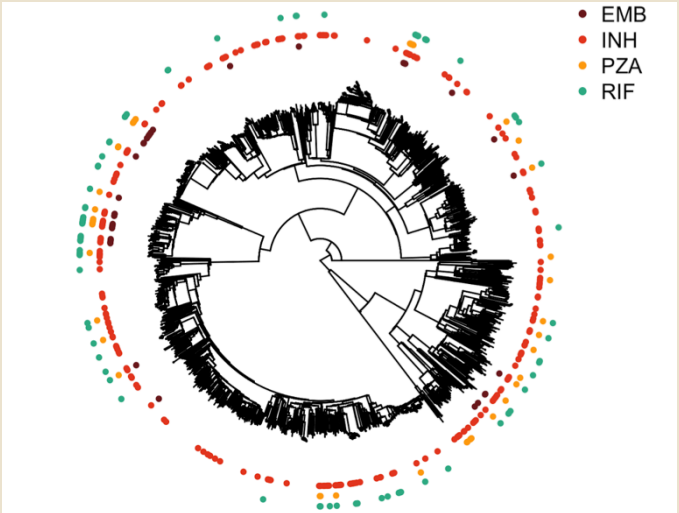
Most significant variant was the expected mechanism

Most significant variant was in physical linkage (PL) with the expected mechanism

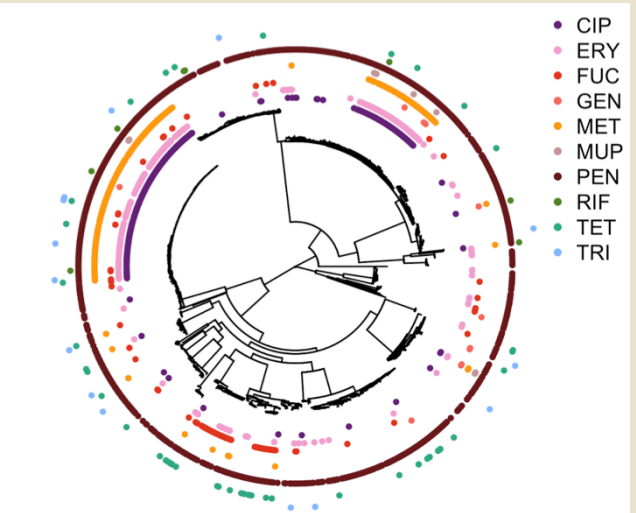
Most significant variant was not the expected mechanism or in PL with the expected mechanism

Genuine resistance-conferring variants were detected in all but one study

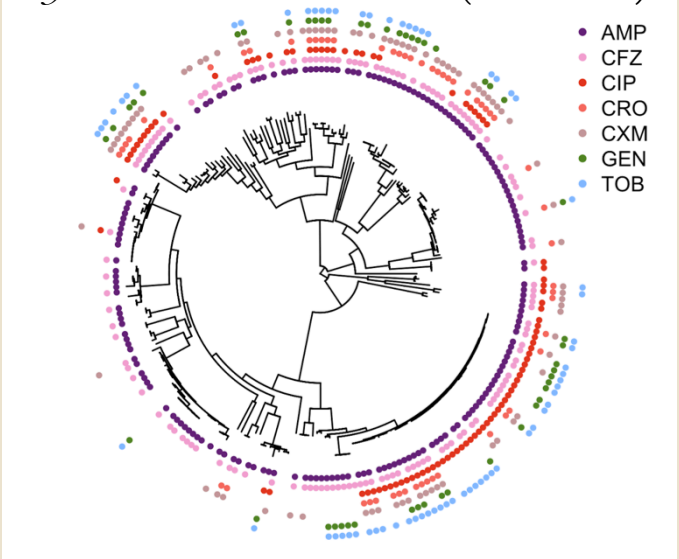
Testing the ability of GWAS to detect genes and genetic variants underlying antimicrobial resistance



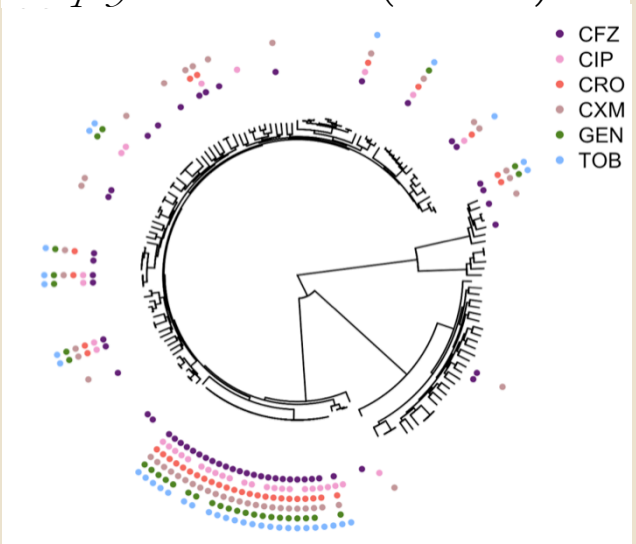
Mycobacterium tuberculosis ($n \leq 1954$)



Staphylococcus aureus ($n \leq 992$)

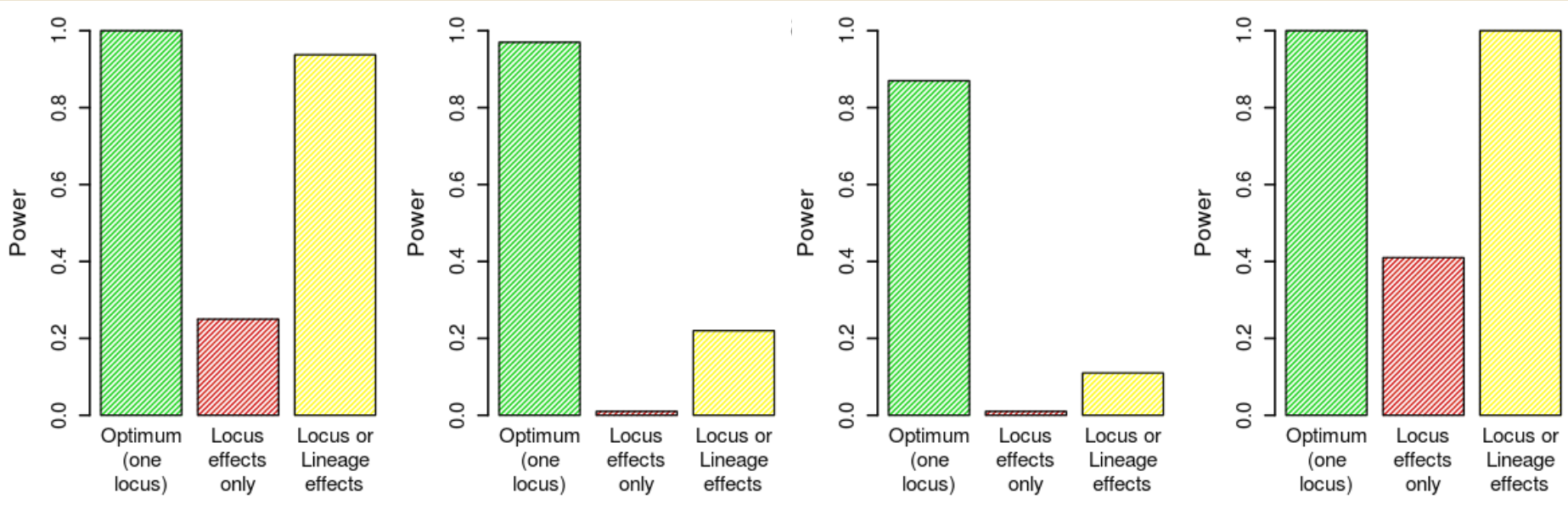


Escherichia coli ($n = 241$)



Klebsiella pneumoniae ($n = 176$)

GWAS simulations



Staphylococcus aureus ($n = 992$)

Escherichia coli ($n = 241$)

Klebsiella pneumoniae ($n = 176$)

Mycobacterium tuberculosis ($n \leq 1954$)

FURTHER GWAS EXAMPLES

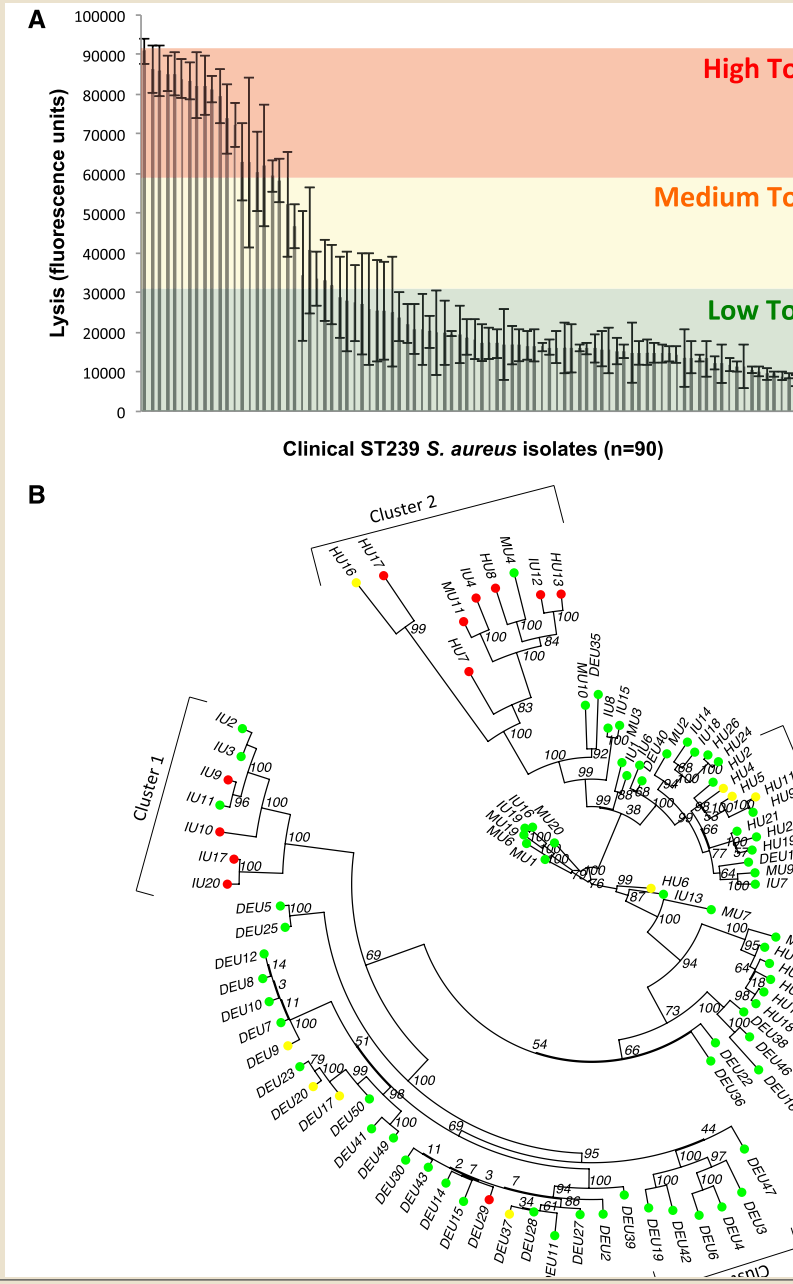
GWAS for toxicity

earch

edicting the virulence of MRSA from its genome quence

sem Laabei,[1,11] Mario Recker,[2,11] Justine K. Rudkin,[1] Mona Aldeljawi,[1]
nep Gulay,[3] Tim J. Sloan,[4] Paul Williams,[4] Jennifer L. Endres,[5] Kenneth W. Bayles,[5]
D. Fey,[5] Vijaya Kumar Yajjala,[5] Todd Widhelm,[5] Erica Hawkins,[1] Katie Lewis,[1]
Parfett,[1] Lucy Scowen,[1] Sharon J. Peacock,[6] Matthew Holden,[7] Daniel Wilson,[8]
othy D. Read,[9] Jean van den Elsen,[1] Nicholas K. Priest,[1] Edward J. Feil,[1]
ence D. Hurst,[1] Elisabet Josefsson,[10] and Ruth C. Massey[1,12]

genes, and in intergenic regions. Two genes previously shown to affect the expression of toxins contained significantly associated SNPs: *mecA* (Rudkin et al. 2012) and *agrC* (Ji et al. 1995; Novick and Geisinger 2008), which provided some proof of principle for the validity of our approach. Mobile genetic elements, such as the *S. aureus* pathogenicity Island I (*Sa*PI1) (Ruzin et al. 2001) and the beta-haemolytic converting phage (Bae et al. 2006), also contained several associated genetic changes, implying that variability in many diverse regions of the genome contributes to the toxicity of a given isolate. Some of the polymorphisms appeared



GWAS for host association

Genome-wide association study identifies vitamin B_5 biosynthesis as a host specificity factor in *Campylobacter*

Samuel K. Sheppard[a,b,1], Xavier Didelot[c], Guillaume Meric[b], Alicia Torralbo[d], Keith A. Jolley[a], David J. Kelly[e], Stephen D. Bentley[f,g], Martin C. J. Maiden[a], Julian Parkhill[f], and Daniel Falush[h]

[a]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom; [b]Institute of Life Scienc Swansea SA2 8PP, United Kingdom; [c]School of Public Health, St. Mary's Campus, Imperial College London, of Animal Health, University of Cordoba, 14071 Cordoba, Spain; [e]Department of Molecular Biology and B S10 2TN, United Kingdom; [f]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge of Medicine, University of Cambridge, Addenbrookes Hospital, Cambridge CB2 0SP, United Kingdom; and Anthropology, 04103 Leipzig, Germany

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved June 3, 2013 (received f

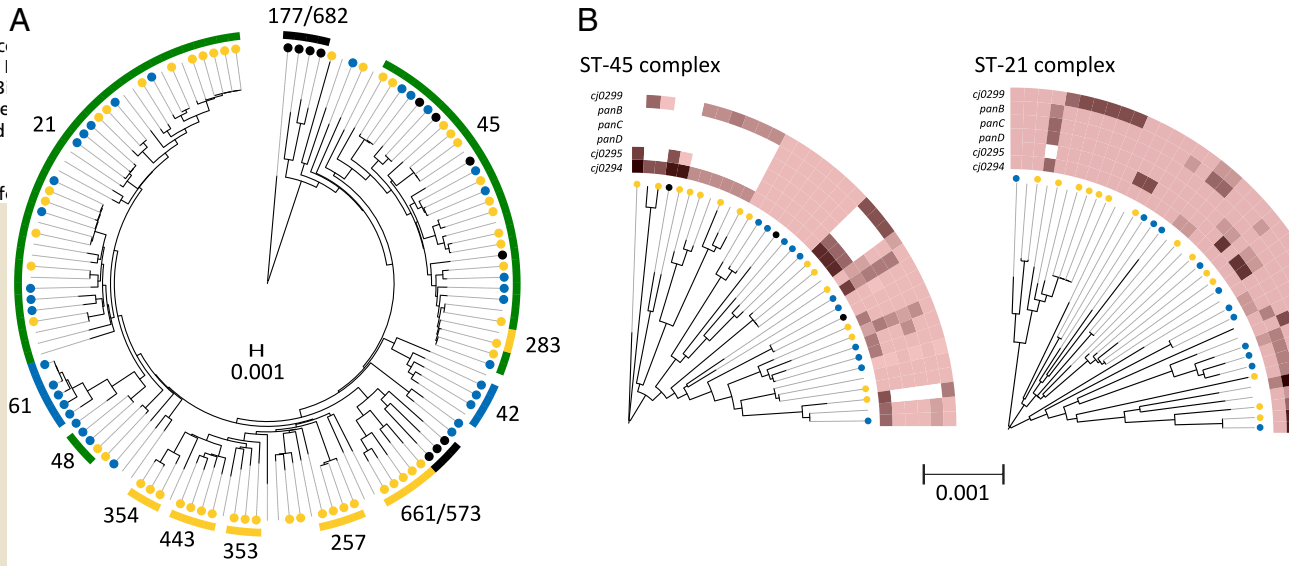


Fig. 1. Genetic structure of *C. jejuni isolates* from different hosts. (*A*) Neighbor-joining tree of all isolates based on 1.53-Mb concatenated sequences (variable sites) of 1,623 loci in the NCTC11168 *C. jejuni* isolate genome. Host origin is indicated for chickens (yellow), cattle (blue), and wild birds a environment (black). Other isolates are primarily from human infections. Clonal complex designations based on MLST are labeled around the tre associations, based on the 2,764 isolates in the pubMLST database (pubMLST.org), are indicated in the halo using the same color scheme, with ge lineages shown in green. (*B*) Tree of the ST-45 and ST-21 clonal complex isolates only, estimated using ClonalFrame. Allelic variation for six genes in th associated region is shown at the end of each branch. Each shade of red represents a unique allele at that locus. White denotes the absence or trunca the gene.

Using GWAS results for predicting resistance

ture
MMUNICATIONS

TICLE

ed 17 Apr 2015 | Accepted 28 Oct 2015 | Published 21 Dec 2015

DOI: 10.1038/ncomms10063 OPEN

pid antibiotic-resistance predictions from nome sequence data for *Staphylococcus aureus* d *Mycobacterium tuberculosis*

m Bradley[1], N. Claire Gordon[2], Timothy M. Walker[2], Laura Dunn[2], Simon Heys[1], Bill Huang[1], Sarah Earle[2], e J. Pankhurst[2], Luke Anson[2], Mariateresa de Cesare[1], Paolo Piazza[1], Antonina A. Votintseva[2], a Golubchik[2], Daniel J. Wilson[1,2], David H. Wyllie[2], Roland Diel[3], Stefan Niemann[4,5], Silke Feuerriegel[4,5], nas A. Kohl[4], Nazir Ismail[6,7], Shaheed V. Omar[6], E. Grace Smith[8], David Buck[1], Gil McVean[1], rah Walker[2,9], Tim E.A. Peto[2,9], Derrick W. Crook[2,9,10] & Zamin Iqbal[1]

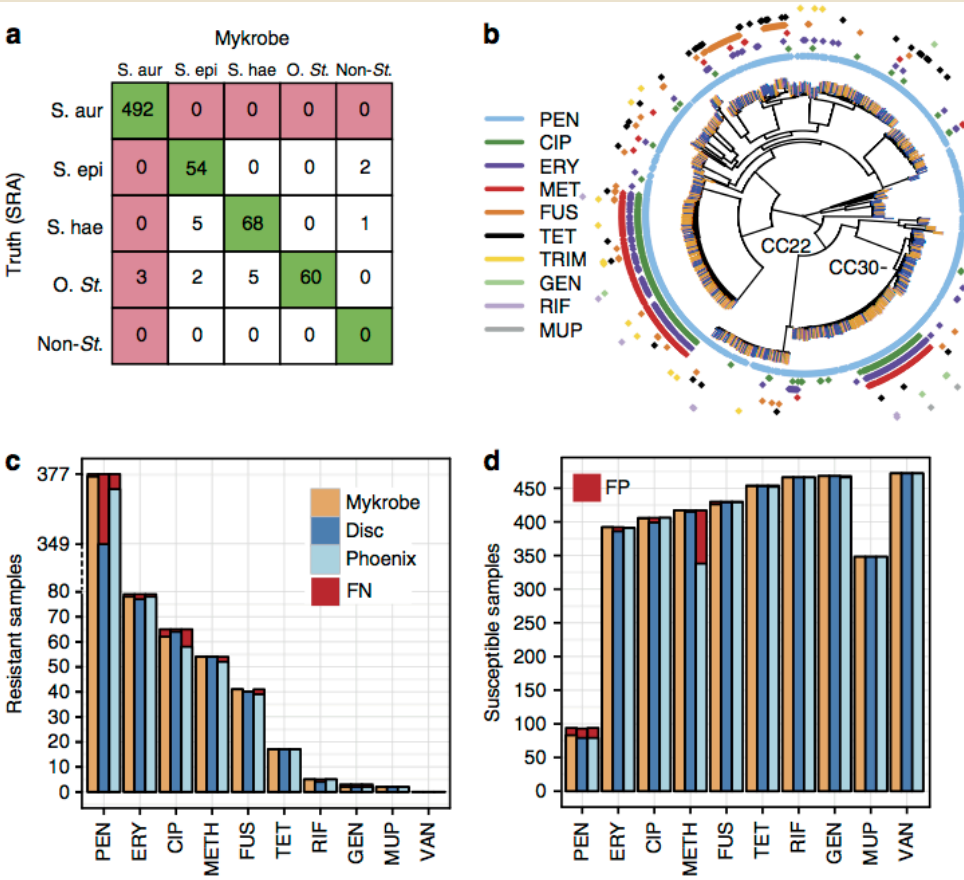


Figure 2 | Species and susceptibility predictions for *S. aureus*. (a) Species classification results on species validation set St

SUMMARY

Genome-wide association studies for bacterial pathogens

- Accessibility of genome sequencing paves the way to discover genetic basis of diverse traits
- Studies must overcome a range of pitfalls, notably
  - False positives caused by many tests
  - Artefactual associations generated by confounders, especially population structure
  - Limited power
- Proof of principle for bacterial GWAS established by studies of antimicrobial resistance
- New studies are shedding light on other phenotypes such as components of virulence
- Knowledge about genotype-phenotype relations will be used in future in the clinic for rapid prediction of genotypes including antimicrobial resistance

Acknowledgments

Analysis and development
Sarah Earle
Jessie Wu
Jane Charlesworth

Microbiology
Nicole Stoesser
Claire Gordon
Tim Walker
Katie Hopkins

Statistical Expertise
Chris Spencer
Zam Iqbal
David Clifton
Gil McVean
Sarah Walker

Clinical Collaborators
Neil Woodford
Grace Smith
Nazir Ismail
Martin Llewelyn
Tim Peto & Derrick Crook

Bacterial GWAS
Earle et al (2016) Nature Microbiology 1: 16041
Laabei et al (2014) Genome Research 24: 839

Evolution during infection
Laabei et al (2015) PLoS Biology 13: e1002229
Young et al (2012) PNAS 109: 4550-4555

Evolution of horizontal gene transfer
Everitt et al (2014) Nature Communications 5: 3956

Reviews
Didelot et al (2016) Nature Reviews Microbiology 14: 150
Wilson (2012) PLoS Pathogens 8: e1002874

Contact daniel.wilson@ndm.ox.ac.uk
Web www.daniel wilson.me.uk
www.virulogenomics.com

Funding

UKCRC Modernising Medical Microbiology
wellcome trust sanger institute
Health Protection Agency
University of Oxford
Department
Mérieux Research Grants
wellcome trust
The Royal Society
Oxford Biomedical Research Centre
Oxford University Hospitals NHS Trust
University of Oxford
NHS National Institute for Health Research