Genome-wide discovery of epistatic loci affecting antibiotic resistance using evolutionary couplings

Benjamin Schubert[1,2,3,*], Rohan Maddamsetti[1,3,*], Jackson Nyman[1], Debora S. Marks[1,4,†]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[2]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

[3]cBio Center, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[4]Associated Member, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[*] Joint first author

[†]debbie@hms.harvard.edu

Keywords: *Neisseria gonorrhoeae*, antibiotic resistance, epistasis, genome-wide association study, evolutionary couplings, undirected graphical models

ABSTRACT

In theory, whole genome sequencing should allow the discovery of interdependent loci that cause antibiotic resistance. In practice, however, identifying this epistasis remains a challenge as the vast number of possible interactions erodes statistical power. To solve this problem, we infer the most strongly coupled loci using a method that has been successfully used to identify epistatic residues in proteins and here applied to whole genomes. Our method reduces the number of tests required for an epistatic genome-wide association study and increases the likelihood of identifying causal epistasis. We discover 45 loci and 291 epistatic pairs that influence the dose needed to inhibit growth for five different antibiotics on 1102 isolates of *Neisseria gonorrhoeae*, confirmed in an independent dataset of 495 isolates. Many of the epistatic pairs contain at least one locus or gene previously associated with antibiotic resistance, including interactions between *gyrA* and *parC* associated with ciprofloxacin resistance and between *penA* and *murE* with cefixime resistance. Our work provides a systematic identification of epistasis pairs in *N. gonorrhoeae* resistance and a generalizable method for epistatic genome-wide association studies.

Neisseria gonorrhoeae is rapidly evolving to be untreatable. It already exhibits resistance to extended-spectrum cephalosporins such as cefixime and ceftriaxone[1-4] and almost all other available mono-therapeutically used antimicrobials[5]. More alarmingly, a new strain of *Neisseria gonorrhoeae* resistant to the currently recommended therapy — a combination of azithromycin and ceftriaxone — was reported in the United Kingdom in March 2018[6].

Epistasis can both enable and prevent the evolution of antibiotic-resistant bacteria such as *N. gonorrhoeae*. On the one hand, variants that compensate for costly resistance variants can enable the evolution of multi-resistant bacteria[7]. On the other hand, epistasis can constrain the spread of antibiotic resistance if the fitness cost of resistance prohibits its spread by horizontal gene transfer or recombination. Statistically, it is difficult to discover epistatic interactions that affect antibiotic resistance because the number of possible interactions to test is prohibitively large. To circumvent this problem, researchers have used prior knowledge such as known protein-protein interactions or network information to find candidate pairs for epistatic testing[8]. Alternatively, researchers have restricted the statistical analysis on loci that individually are strongly associated with the trait of interest[9,10]. In the end, neither of these approaches allow a systematic search for epistatic interactions.

To conduct an epistatic genome-wide association study (GWAS), yet deal with the combinatorial explosion of statistical tests, we exploit coevolutionary information by using evolutionary couplings[11] to identify epistatic interactions. In recent years, evolutionary couplings methods have made breakthroughs in *ab initio* protein and RNA 3D structure[11-13], protein complex[14], and mutation effects prediction[15]. Now, these methods are being adopted for bacterial genome analysis[16]. Evolutionary couplings are able to separate causal interactions from indirect correlations[17] including, to a large extent, global correlations caused by population structure and phylogeny[18,19] or under-sampling. We hypothesized that evolutionary couplings, inferred from populations of bacterial pathogens, might represent functional or mechanistic dependencies between loci that affect bacterial fitness. If true, evolutionary couplings would be able to find epistatic interactions affecting phenotypes, and thus could be used to filter pairs for epistatic GWAS in a principled manner.

In this work, we report single nucleotide polymorphisms (SNPs) and epistatic interactions that were significantly associated with changes in minimal inhibitory concentration (MIC) in an exploratory dataset consisting of 1102 *N. gonorrhoeae* strains collected in the USA[20], and subsequently confirmed in a separate dataset consisting of 495 *N. gonorrhoeae* strains collected in Canada and England[21]. The structure of the epistatic associations reveals the genetic architecture that underlies antimicrobial sensitivity and resistance in *N. gonorrhoeae*. All epistatic interactions associated with azithromycin MICs involved a single position in 23S rRNA that is known to cause high-level resistance. Epistatic interactions associated with ciprofloxacin occurred across many genes, each containing few associated loci, in disparate cellular pathways. We also identified two epistatic interactions between *gyrA* and *parC* affecting ciprofloxacin that have experimental support in *Streptococcus pneumoniae*[22]. By contrast with ciprofloxacin, epistatic interactions associated with cefixime involved many associations in only a few genes. All epistatic interactions involved either the penicillin-binding protein *penA* or the peptidoglycan biosynthesis gene *murE*. Loci in

penA are known to cause cephalosporin resistance, while *murE* loci have not been reported before in the literature. Thus, our work provides a foundation for experimentalists and clinicians seeking to understand how epistatic interactions between genes affect the rapid evolution of antibiotic resistance in clinical pathogens such as *N. gonorrhoeae*.

RESULTS

Processing of N. gonorrhoeae *isolate genomes and resistance data*

We extracted MICs of five antibiotics (Penicillin: PEN, Tetracycline: TET, Cefixime: CFX, Ciprofloxacin: CIPRO, and Azithromycin: AZI) and whole genome sequencing data from Grad *et al.*[20] comprising 1102 strains collected in the USA and from Eyre *et al.*[21] comprising 495 strains collected in Canada (n = 246) and England (n = 249). The USA dataset was used as an exploratory test set and the other dataset was used as a confirmatory dataset. In the test set, 537 strains were resistant to PEN (resistance threshold > 1 mg/L, EUCAST); 661 were resistant to TET (> 1 mg/L); 266 were resistant to CFX (> 0.125 mg/L); 602 were resistant to CIPRO (> 0.06 mg/L); and 459 were resistant to AZI (> 0.5 mg/L). The confirmatory set contained 141 PEN, 400 TET, 17 CFX, 231 CIPRO, and 231 AZI resistant strains (Figure 2, Supplementary Table S1).

We identified SNPs in each clinical isolate of the test and confirmatory sets using *N. gonorrhoeae* FA1090 as a reference strain. Only non-synonymous SNPs and SNPs in non-coding regions with experimentally determined functional annotation[23] that had a minor allele frequency ≥ 0.5% were considered, resulting in 10,087 loci (Methods). For epistatic GWAS, we selected 250,037 high confidence couplings involving 7119 loci in 1551 genes (see Methods). Of these 7119 selected positions, 5872 are within protein-coding regions, 1055 within transcription start sites, 181 within promoters, and 11 within ribosomal RNAs.

Probabilistic model to capture genome-wide interactions

In order to assess epistatic relationships between different loci, we formulated a generative probabilistic model of bacterial genomes. Our statistical model associates a genome sequence $\boldsymbol{\sigma}$ with a probability distribution $\boldsymbol{P}(\boldsymbol{\sigma})$ at equilibrium as

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\big(E(\boldsymbol{\sigma})\big),$$

where Z is a normalization constant. We define $E(\boldsymbol{\sigma})$ as the sum of coupling terms $\boldsymbol{J}_{ij}$ between every pair of loci in a sequence and a locus-wise bias terms h_i:

$$E(\boldsymbol{\sigma}) = \sum_i \boldsymbol{h}_i(\sigma_i) \quad + \sum_{i<j} \boldsymbol{J}_{ij}\left(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j\right)$$

To measure the evolutionary coupling (EC) strength between pairs of loci, the inferred parameters $\boldsymbol{J}_{ij}$ are summarized using the Frobenius norm[24]:

$$FN_{ij} \;=\; \sqrt{\sum\nolimits_{k,l} \boldsymbol{J}_{ij}(k,l)^2}$$

We use EC scores, after population-structure and under-sampling corrections, to stratify epistatic association testing. Thus, this is an agnostic alternative to pathway-based filtering approaches previously applied in epistatic GWAS[25]. Our working hypothesis is that pairs of loci with high EC scores are evolutionarily linked, thus are more likely to have an epistatic effect on some phenotype.

Genome-wide association study for single and paired loci

Next, we performed a GWAS to identify single positions and epistatic interactions that are associated with MIC change. To reduce computation and increase statistical power, we only tested high-scoring evolutionary couplings between distinct genes. We performed the GWAS on the exploratory dataset and confirmed the significant associations in an independent dataset. To measure the statistical association strength of individual loci and epistatic interactions, we used a linear mixed model[26,27] that accounts for population structure as well as cryptic relationships within the data (see Methods). To control type I errors and correct for multiple testing, we employed a Bonferroni correction and deemed all loci and interactions significantly associated at an adjusted $\alpha = 0.05$.

Single-site analysis recapitulates known associations and predicts new ones

We verified 45 associations in the confirmatory dataset (Tables 1 and 2) out of 98 significant associations in the exploratory dataset (Supplementary Table S2). Each association was specific to a single antibiotic. Of these 45 loci, 29 were associated with CIPRO MIC change (Table 1), 12 with CFX, and 2 with PEN (Table 2). One association was confirmed for TET and AZI each. Our GWAS recovered many known antibiotic resistance variants. The top CIPRO associations, *gyrA* p.S91 ($p = 8.25\text{x}10^{-101}$) and p.D95 ($p = 1.98\text{x}10^{-43}$) as well as p.S87 in *parC* ($p = 9.96\text{x}10^{-18}$), are known to cause CIPRO resistance[20]. The top AZI (23S rRNA g.C2617, previously described as C2611, $p = 2.19\text{x}10^{-31}$) and PEN (*porB* A121, $p = 2.45\text{x}10^{-4}$) associations[20] as well as a CFX-associated locus in *penA* (p.G546, $p = 5.08\text{x}10^{-12}$) have also been reported in the literature[4]. Our GWAS also uncovered 39 novel associations with antibiotic sensitivity and resistance.

Ciprofloxacin: Single-locus CIPRO associations occurred in 23 genes (Table 1). The two strongest novel associations, found in *dldH* and *gshB,* are involved in oxidative stress response[28], which is known to be affected by ciprofloxacin[29]. The *dldH* variant (p.A88, $p = 6.68\text{x}10^{-9}$) probably causes greater CIPRO sensitivity, since it is uncommon in resistant isolates and its regression coefficient (β) is negative. By contrast, the variant in *gshB* (p.E221, $p = 2.71\text{x}10^{-8}$) is associated with greater CIPRO resistance. The highest concentration of CIPRO associations was found in *lrgB* (3 loci), *lolB* (2 loci) and in the 5' untranslated region of *pyrG* (2 loci).

Cefixime: All twelve CFX associations occurred in the cell-wall biosynthesis genes *penA* and *murE* (Table 2). *penA* contained 8 associations, and *murE* contained 4 associations. While mosaic alleles of *penA* are known to be associated with cephalosporin resistance[5], to our knowledge *murE* has not been reported before.

Penicillin and Tetracycline: We found previously unreported loci in *porB* associated with greater antibiotic sensitivity. One was associated with PEN (p.V151, $p = 4.73\text{x}10^{-3}$) and another was associated with

TET (p.F131, $p = 1.33x10^{-3}$). Since variants in *porB* have also been associated with antibiotic resistance, this shows that *porB* variation can increase sensitivity as well as resistance to antibiotics.

Epistatic interactions associate with antibiotic resistance

We verified 291 epistatic associations in the confirmatory dataset (Figure 3), out of 822 significant epistatic associations in the exploratory dataset (Supplementary Table S3). 139 of the confirmed epistatic associations were associated with CIPRO MIC change; 97 with CFX; and 55 with AZI (Supplementary Table S4). We could not confirm any significant epistatic associations to TET or PEN MIC change.

All 291 epistatic interactions involved loci that were confirmed in our single locus GWAS. 24 epistatic associations connected two loci which were each significant in the single locus GWAS. The remaining 267 interactions connected an associated locus to a non-associated locus. Notably, 5 out of the 6 known resistance variants in our single locus GWAS occurred in the inferred epistatic pairs. These included 3 CIPRO resistance variants in *gyrA* and *parC* (Table 1 and Figure 3), and 1 CFX resistance variant in *penA* (Table 2). All epistatic associations to AZI involved a single, known antibiotic resistance variant C2617T (identical to C2611T in Grad *et al.* 2016) in the 23S rRNA (Table 2 and Figure 3A). In all, 123 epistatic associations involved known resistance variants. Table 3 contains a selection of the most significant epistatic interactions in our GWAS. Detailed results can be found in Supplementary Table 4.

Azithromycin: We found an interaction ($p = 5.77x10^{-29}$) between the C2617T variant in 23S rRNA and F876V in *mtrD,* an efflux pump that is known to affect azithromycin resistance[30] (recent work on the same clinical isolates has identified epistasis within the *mtr* operon affecting AZI resistance[31]). Of the 8 genes containing two or more associated loci, 7 were either membrane proteins (NGO1512, NGO1861a, *porB*), cell wall biosynthesis proteins (*lgtE*, *lpxC*) or pilus proteins (*pilY1*, *pilA*). We speculate that variants in these membrane and cell wall biosynthesis proteins affect *N. gonorrhoeae's* permeability to AZI and other antibiotics[32-34].

Ciprofloxacin: As in the single locus analysis, epistatic interactions associated with CIPRO involved many genes (94 in all), each containing relatively few interactions in comparison to AZI and CFX (Figure 3B). The highest concentration of significant interactions occurred in two known antibiotic resistance genes: *parC* and *gyrA*. We discovered interactions between two known resistance-causing loci in *gyrA* and a resistance-causing locus in *parC* (*gyrA* p.S91 to *parC* p.S87, $p = 1.73x10^{-45}$; *gyrA* p.D95 to *parC* p.S87, $p = 1.26x10^{-30}$). We found experimental evidence that corroborates our discovery: precisely homologous epistatic interactions were reported to affect both fitness and CIPRO resistance in *S. pneumoniae*[22]. Apart from *gyrA* and *parC*, 16 other genes contained three or more epistatic associations. These genes have diverse annotations, including outer membrane proteins (*doxX*, *macA, cirA*), cell wall biosynthesis proteins (*lrgB, mltA*), pilus proteins (*pilA*), and proteins affecting the oxidative stress response[28] (*gshB, petC*).

Cefixime: By contrast with CIPRO, epistatic associations with CFX involved only 16 genes, each with many associated loci. All CFX-associated pairs involved either *penA* or *murE* (Figure 3C). Furthermore, epistatic associations with *penA* or *murE* often involved the same genes. The genes containing epistatic

associations to both *penA* and *murE* occur in diverse cellular pathways, including outer membrane proteins (*efeO, fetA, kefB*, *yebE*), cell wall biosynthesis proteins (*mdoB*), and cell division proteins (*ftsL*, *gidA*).

DISCUSSION

Here, we report the results of the first systematic GWAS of loci and epistatic interactions affecting antimicrobial sensitivity and resistance in *N. gonorrhoeae*. As expected, we recover many known resistance-causing variants, and many of the epistatic interactions that we discovered involve known resistance-causing variants.

In our study, reported loci can cause either heightened *sensitivity* or greater *resistance* to antimicrobials. Examples for heightened sensitivity include loci in *dldH, comF,* and *lolB* that are associated with CIPRO resistance (Table 1). We also found several unreported variants that associate strongly with greater antibiotic resistance, including several variants in *murE* linked to CFX resistance. Each antibiotic showed distinct epistatic association patterns. While epistatic associations with CIPRO occurred in many genes, all epistatic associations with AZI involved a single resistance locus in 23S rRNA, and all epistatic associations with CFX involve loci in a few genes, mostly concentrated in the peptidoglycan synthesis genes *penA* and *murE*. These observations suggest that the mechanism of CIPRO-induced death involves many genes across diverse cellular pathways, including oxidative stress, while the mechanism of CFX and AZI killing is focused on disabling peptidoglycan synthesis and 23S rRNA, respectively.

In principle, GWAS is able to identify causal variation. However, our study focuses only on point mutations found by comparison to the *N. gonorrhoeae* FA1090 reference genome. Thus, the effects of horizontally-transmitted genes and more complex variants such as gene duplications, indels, and mobile element transpositions will be missed. Some of these missing variants may be hidden yet causal variables that affect antibiotic resistance. If so, the effects of these missing variants will either be absent or will be associated to non-causative markers in our data that closely track the presence of the true, but missing, causal variants. We find some evidence of missing causal variables in our analysis as many epistatic associations are linked to genes that are highly variable in the genome alignment, such as transferrin-binding proteins[35]. Thus, it is possible that some associations involving highly variable genes are actually markers that represent the closest possible association to hidden causal variants. Nevertheless, our work demonstrates that evolutionary couplings analysis combined with standard GWAS methods is a promising new approach for discovering epistatic interactions affecting antibiotic resistance in a multi-resistant pathogen, *N. gonorrhoeae*. We anticipate that our findings as well as our method will be valuable in unraveling the genetics of complex traits in general.

ACKNOWLEDGEMENTS

We thank members of the Marks lab for helpful comments and Chris Sander for his support during this research project.

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

D.S.M. conceived the project and supervised research; D.S.M, B.S and R.M designed and planned research; B.S. implemented the analysis methods; B.S. and R.M. analyzed data with the help of J.N. and D.S.M. B.S., R.M, and D.S.M. wrote the paper.

REFERENCES

1 Cámara, J. *et al.* Molecular characterization of two high-level ceftriaxone-resistant Neisseria gonorrhoeae isolates detected in Catalonia, Spain. *Journal of antimicrobial chemotherapy* **67**, 1858-1860 (2012).

2 Ohnishi, M. *et al.* Is Neisseria gonorrhoeae initiating a future era of untreatable gonorrhea?: detailed characterization of the first strain with high-level resistance to ceftriaxone. *Antimicrobial agents and chemotherapy* **55**, 3538-3545 (2011).

3 Unemo, M. *et al.* High-level cefixime-and ceftriaxone-resistant Neisseria gonorrhoeae in France: novel penA mosaic allele in a successful international clone causes treatment failure. *Antimicrobial agents and chemotherapy* **56**, 1273-1280 (2012).

4 Unemo, M. & Nicholas, R. A. Emergence of multidrug-resistant, extensively drug-resistant and untreatable gonorrhea. *Future microbiology* **7**, 1401-1422 (2012).

5 Unemo, M. & Shafer, W. M. Antimicrobial resistance in Neisseria gonorrhoeae in the 21st century: past, evolution, and future. *Clinical microbiology reviews* **27**, 587-613 (2014).

6 England, P. H. UK case of Neisseria gonorrhoeae with high-level resistance to azithromycin and resistance to ceftriaxone acquired abroad. *Health Protection Report* **12** (2018).

7 de Sousa, J. M., Balbontín, R., Durão, P. & Gordo, I. Multidrug-resistant bacteria compensate for the epistasis between resistances. *PLoS biology* **15**, e2001741 (2017).

8 Moore, J. H., Asselbergs, F. W. & Williams, S. M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **26**, 445-455, doi:10.1093/bioinformatics/btp713 (2010).

9 Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics* **18**, 41, doi:10.1038/nrg.2016.132 (2016).

10 Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nature Genetics* **50**, 307-316, doi:10.1038/s41588-017-0029-0 (2018).

11 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).

12 Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072-1080, doi:10.1038/nbt.2419 (2012).

13 Weinreb, C. *et al.* 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963-975 (2016).

14 Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).

15 Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature Biotechnology* (2017).

16 Skwark, M. J. *et al.* Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genetics* **13**, e1006508 (2017).

17 Lapedes, A. S., Giraud, B. G., Liu, L. & Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lecture Notes-Monograph Series*, 236-256 (1999).

18 Qin, C. & Colwell, L. J. Power law tails in phylogenetic systems. *Proceedings of the National Academy of Sciences*, 201711913 (2018).

19 Cocco, S., Monasson, R. & Weigt, M. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS computational biology* **9**, e1003176 (2013).

20 Grad, Y. H. *et al.* Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000–2013. *Journal of Infectious Diseases*, jiw420 (2016).

21 Eyre, D. W. *et al.* WGS to predict antibiotic MICs for Neisseria gonorrhoeae. *J Antimicrob Chemother* **72**, 1937-1947, doi:10.1093/jac/dkx067 (2017).

22 Rozen, D. E., McGee, L., Levin, B. R. & Klugman, K. P. Fitness costs of fluoroquinolone resistance in Streptococcus pneumoniae. *Antimicrobial agents and chemotherapy* **51**, 412-416 (2007).

23 Remmele, C. W. *et al.* Transcriptional landscape and essential genes of Neisseria gonorrhoeae. *Nucleic acids research* **42**, 10579-10595 (2014).

24 Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E* **87**, 012707 (2013).

25 Emily, M., Mailund, T., Hein, J., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics* **17**, 1231 (2009).

26 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354 (2010).

27 Sul, J. H. *et al.* Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS genetics* **12**, e1005849 (2016).

28 Seib, K. L. *et al.* Defenses against oxidative stress in Neisseria gonorrhoeae: a system tailored for a challenging environment. *Microbiology and molecular biology reviews* **70**, 344-361 (2006).

29 Kohanski, M. A., Dwyer, D. J. & Collins, J. J. How antibiotics kill bacteria: from targets to networks. *Nature Reviews Microbiology* **8**, 423 (2010).

30 Folster, J. P., Dhulipala, V., Nicholas, R. A. & Shafer, W. M. Differential regulation of ponA and pilMNOPQ expression by the MtrR transcriptional regulatory protein in Neisseria gonorrhoeae. *Journal of bacteriology* **189**, 4569-4577 (2007).

31 Vaara, M. Outer membrane permeability barrier to azithromycin, clarithromycin, and roxithromycin in gram-negative enteric bacteria. *Antimicrobial agents and chemotherapy* **37**, 354-356 (1993).

32 Zgurskaya, H. I., López, C. A. & Gnanakaran, S. Permeability barrier of Gram-negative cell envelopes and approaches to bypass it. *ACS infectious diseases* **1**, 512-522 (2015).

33 Delcour, A. H. Outer membrane permeability and antibiotic resistance. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1794**, 808-816 (2009).
34 Barber, M. F. & Elde, N. C. Escape from bacterial iron piracy through rapid evolution of transferrin. *Science* **346**, 1362-1366 (2014).
35 Toth-Petroczy, A. *et al.* Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158-170. e112 (2016).
36 Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607-1621, doi:10.1016/j.cell.2012.04.012 (2012).
37 Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Engineering and Analyzing Multicellular Systems: Methods and Protocols*, 165-188 (2014).
38 Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333-340 (2008).

TABLES

Table 1. Single loci associated with change in sensitivity or resistance to Ciprofloxacin (CIPRO).

Locus	Gene	Variant	*p*-value	β	Frequency in resistant isolates
620917*	***gyrA* (NGO0629)**	**S91F**	**8.25×10^{-101}**	**2.73**	**239/833**
620905*	***gyrA* (NGO0629)**	**D95(A/G)**	**1.98×10^{-43}**	**2.09**	**270/833**
1210781*	***parC* (NGO1259)**	**S87R**	**9.96×10^{-18}**	**1.93**	**702/833**
894348	*dldH* (NGO0915)	A88V	6.68×10^{-9}	–1.26	46/833
1167114*	*gshB* (NGO1217)	E221K	2.71×10^{-8}	0.920	242/833
1687733*	*comF* (NGO1726)	F43Y	1.86×10^{-7}	–1.00	55/833
433930	*lolB* (NGO0439)	S74G	5.04×10^{-7}	–0.790	218/833
433932	*lolB* (NGO0439)	S74G	5.04×10^{-7}	–0.790	218/833
2024713*	*efeO* (NGO2050)	–187	1.55×10^{-6}	–0.633	37/833
621987	*gyrA* (NGO0629)	–798	4.32×10^{-6}	–0.832	99/833
93476	*rfaB* (NGO0086)	D363(D/E)	1.66×10^{-5}	–0.885	27/833
1167527	*gshB* (NGO1217)	E83A	1.91×10^{-5}	–0.647	241/833
1135241*	*lrgB* (NGO1193)	S9G	1.98×10^{-5}	–1.05	90/833
1135244*	*lrgB* (NGO1193)	V10I	1.98×10^{-5}	–1.05	90/833
1135246*	*lrgB* (NGO1193)	V10I	1.98×10^{-5}	–1.05	90/833
89733	*pglD* (NGO0083)	A360T	2.07×10^{-5}	0.728	652/833
1372483*	*gcvT* (NGO1406)	–186	2.14×10^{-5}	–0.961	4/833
1393066*	*cirA* (NGO1430)	R122P	2.83×10^{-5}	0.947	625/833
1921503*	*doxX* (NGO1948)	L139R	4.09×10^{-5}	0.837	242/833
1923171*	*prfB* (NGO1951)	–84	4.93×10^{-5}	–0.930	32/833
1370385	*gcvH* (NGO1404)	T25I	5.25×10^{-5}	–0.708	44/833
777611	NGO0783	L63P	5.57×10^{-5}	0.565	638/833
447535	*fimT* (NGO0452)	G100D	9.02×10^{-5}	–0.553	67/833
1162600	*pyrG* (NGO1212)	–12 promoter	1.76×10^{-4}	0.647	299/833
1162602	*pyrG* (NGO1212)	–14 promoter	1.76×10^{-4}	0.647	299/833
964051*	NGO0993	K161E	3.89×10^{-4}	–0.526	11/833
1378462	*nqrB* (NGO1414)	A25V	6.61×10^{-4}	–0.578	22/833
565114	*uvrC* (NGO0578)	N258D	6.96×10^{-4}	0.664	228/833
1747633	*gapA-B* (NGO1776)	V267A	8.82×10^{-4}	–0.569	35/833

NOTE– Loci in bold are known resistance variants. Asterisked loci are associated with significant epistatic interactions shown in Figure 3. Non-coding variants are labeled with the distance upstream (–) or downstream (+) from the closest gene, and are labeled with 'promoter' or 'UTR' if occurring in a promoter or untranslated region of a known RNA transcript. β is the regression coefficient associated with each locus in the linear mixed model. Reported *p*-values follow a Bonferroni correction for multiple tests. The count of each variant in all *N. gonorrhoeae* isolates (both test and confirmatory datasets) is reported, as is the fraction of resistant isolates (both test and confirmatory datasets) with that variant.

Table 2. Single loci associated with change in sensitivity or resistance to Azithromycin (AZI), Penicillin (PEN), Tetracycline (TET), and Cefixime (CFX).

Antibiotic	Locus	Gene	Variant	*p*-value	β	Frequency in resistant isolates
AZI	**1116541***	**23S rRNA (NGOr02)**	**C2617T (C2611T)**	**2.19×10^{-31}**	**0.889**	**315/690**
PEN	**1789058**	***porB* (NGO1812)**	**A121(D/G/V)**	**2.45×10^{-4}**	**0.348**	**317/678**
PEN	1789148	*porB* (NGO1812)	V151A	4.78×10^{-3}	–0.249	47/678
TET	1789088	*porB* (NGO1812)	F131Y	1.33×10^{-3}	–0.174	139/1061
CFX	1516685*	*penA* (NGO1542)	V160A	2.14×10^{-15}	0.958	277/283
CFX	1516646*	*penA* (NGO1542)	N173S	2.14×10^{-15}	0.958	277/283
CFX	**1515528***	***penA* (NGO1542)**	**G546S**	**5.08×10^{-12}**	**0.859**	**274/283**
CFX	1516861*	*penA* (NGO1542)	D101E	1.54×10^{-10}	0.954	277/283
CFX	1516798*	*penA* (NGO1542)	V122(F/V)	1.54×10^{-10}	0.954	277/283
CFX	1514399	*murE* (NGO1541)	A331(E/G)	1.41×10^{-9}	0.835	271/283
CFX	1514450*	*murE* (NGO1541)	A314G	1.41×10^{-9}	0.835	271/283
CFX	1516552*	*penA* (NGO1542)	D204E	7.54×10^{-9}	0.802	277/283
CFX	1516556*	*penA* (NGO1542)	E203G	7.54×10^{-9}	0.802	277/283
CFX	1516563*	*penA* (NGO1542)	Y201H	7.54×10^{-9}	0.802	277/283
CFX	1515252	*murE* (NGO1541)	A47T	2.83×10^{-4}	0.465	277/283
CFX	1515263	*murE* (NGO1541)	P43(L/Q)	2.83×10^{-4}	0.465	277/283

NOTE– See Table 1 Note for explanation.

Table 3. Selected set of epistatic pairs of loci associated with change in antibiotic sensitivity or resistance (Detailed results in Supplementary Table S4).

Antibiotic	Locus i	Locus j	Gene i	Gene j	Variant i	Frequency in resistant isolates	Variant j	Frequency in resistant isolates	β	*p*-value
AZI	**1116541**	**1789096**	**23S rRNA**	***porB* (NGO1812)**	**C2617T (C2611T)**	**315/690**	**N134D/H**	**352/690**	**0.88**	**5.77×10^{-29}**
AZI	**1116541**	**1323670**	**23S rRNA**	***mtrD* (NGO1364)**	**C2617T (C2611T)**	**315/690**	**L876F**	**131/690**	**0.88**	**5.77×10^{-29}**
AZI	**1116541**	**2136352**	**23S rRNA**	**NGO2159**	**C2617T (C2611T)**	**315/690**	**R262S**	**229/690**	**0.88**	**5.77×10^{-29}**
AZI	**1116541**	**1888880**	**23S rRNA**	***pilY1 (NGO1911a)***	**C2617T (C2611T)**	**315/690**	**R462Q**	**315/690**	**0.88**	**5.77×10^{-29}**
AZI	**1116541**	**1830645**	**23S rRNA**	**NGO1861a**	**C2617T (C2611T)**	**315/690**	**S90G/R**	**194/690**	**0.88**	**5.77×10^{-29}**
CFX	1516646	1517334	*penA* (NGO1542)	*ftsL* (NGO1543)	N173S	277/283	L52V	0/283	0.96	1.63×10^{-10}
CFX	1516685	2070085	*penA* (NGO1542)	*fetA* (NGO2093)	V160A	277/283	A638V	132/283	0.96	1.63×10^{-10}
CFX	1516685	1665722	*penA* (NGO1542)	*yebE* (NGO1709)	V160A	277/283	K90E	258/283	0.96	1.63×10^{-10}
CFX	1516685	2025976	*penA* (NGO1542)	*efeO* (NGO2050)	V160A	277/283	S360G	0/283	0.96	1.63×10^{-10}
CFX	1516646	1657081	*penA* (NGO1542)	*arnT* (NGO2128)	N173S	277/283	D558Y	0/283	0.96	1.63×10^{-10}
CIPRO	**620917**	**964051**	***gyrA* (NGO0629)**	***lrgB* (NGO1193)**	**S91F**	**239/833**	**V10I**	**90/833**	**2.41**	**1.73×10^{-45}**
CIPRO	**620917**	**1135244**	***gyrA* (NGO0629)**	***gshB* (NGO1217)**	**S91F**	**239/833**	**E221K**	**242/833**	**2.41**	**1.73×10^{-45}**
CIPRO	**620917**	**1165188**	***gyrA* (NGO0629)**	***parC* (NGO1259)**	**S91F**	**239/833**	**S87R**	**702/833**	**2.41**	**1.73×10^{-45}**
CIPRO	**620917**	**1167114**	***gyrA* (NGO0629)**	***ftsN* (NGO1549)**	**S91F**	**239/833**	**E221K**	**242/833**	**2.41**	**1.73×10^{-45}**
CIPRO	**620917**	**693931**	***gyrA* (NGO0629)**	***hsdS (NGO0699)***	**S91F**	**239/833**	**A173T**	**47/833**	**2.41**	**1.73×10^{-45}**

NOTE– Loci in bold are known resistance variants. Non-coding variants are labeled with the distance upstream (–) or downstream (+) from the closest gene, and labeled with 'promoter' or 'UTR' if occurring in a promoter or untranslated region of a known RNA transcript. β is the regression coefficient associated with each locus in the linear mixed model. Reported *p*-values follow a Bonferroni correction for multiple tests. The count of each variant in all *N. gonorrhoeae* isolates (both test and confirmatory datasets) is reported, as is the fraction of resistant isolates (both test and confirmatory datasets) with that variant.

FIGURES

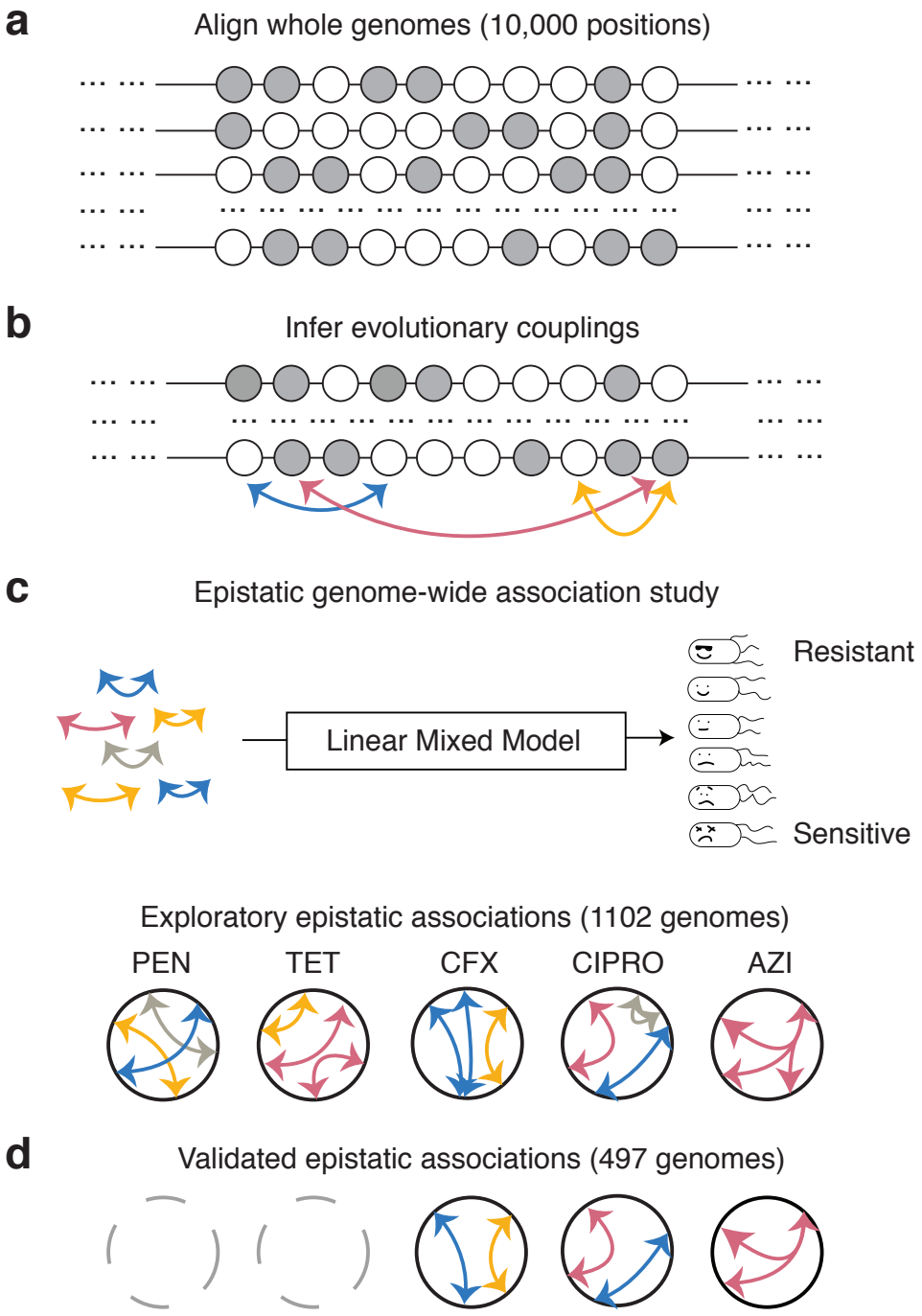

Figure 1: Evolutionary couplings power epistatic GWAS. a) First, non-synonymous SNPs and non-coding variants with experimentally-derived functional annotation are identified and used to generate a whole-genome alignment. b) Highly evolutionarily coupled loci are identified using an undirected graphical model given the multiple sequence alignment. c) The inferred evolutionary couplings are then tested in an exploratory set of 1102 *Neisseria gonorrhoeae* strains, using a linear mixed model, for their epistatic association to minimum inhibitory concentrations for five antibiotics: penicillin (PEN), tetracycline (TET), cefixime (CFX), ciprofloxacin (CIPRO), and azithromycin (AZI). d) Significant epistatic associations are then validated in an independent and geographically distinct set of 497 *N. gonorrhoeae* genomes.

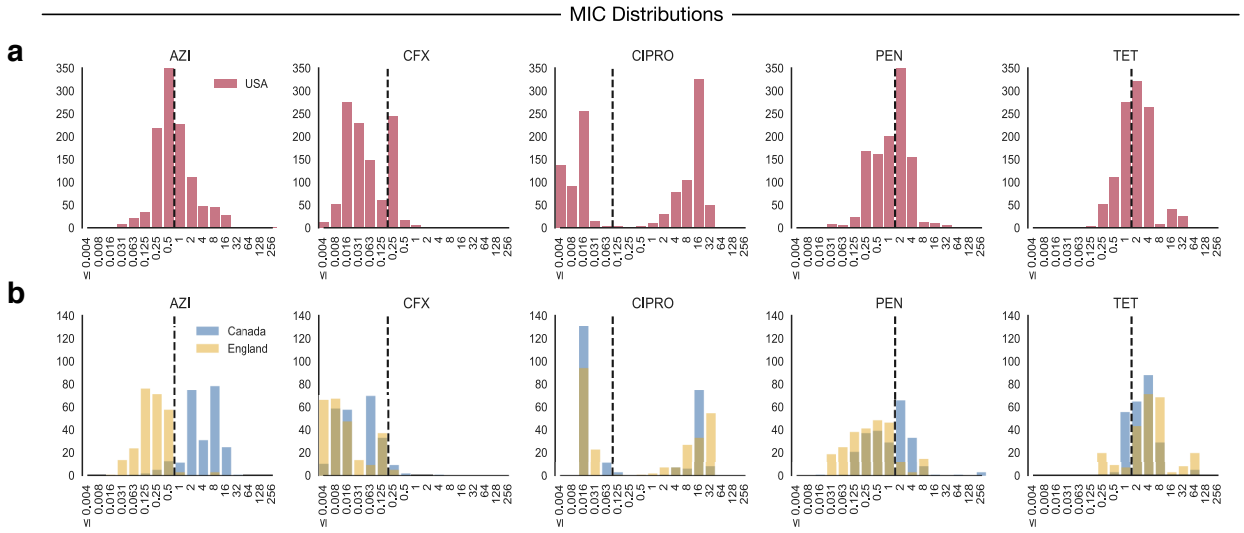

Figure 2: Distributions of minimum inhibitory concentration (MIC), in milligrams per liter, for the *N. gonorrhoeae* strains studied. The dotted lines indicate clinical breakpoints for *N. gonorrhoeae* as defined by the European committee on antimicrobial susceptibility testing (EUCAST). The five antibiotics in this study are azithromycin (AZI), cefixime (CFX), ciprofloxacin (CIPRO), penicillin (PEN), and tetracycline (TET). a) The MIC distribution of the 1102 strains in the exploratory test dataset. b). The MIC distribution of the 495 strains in the confirmatory dataset.

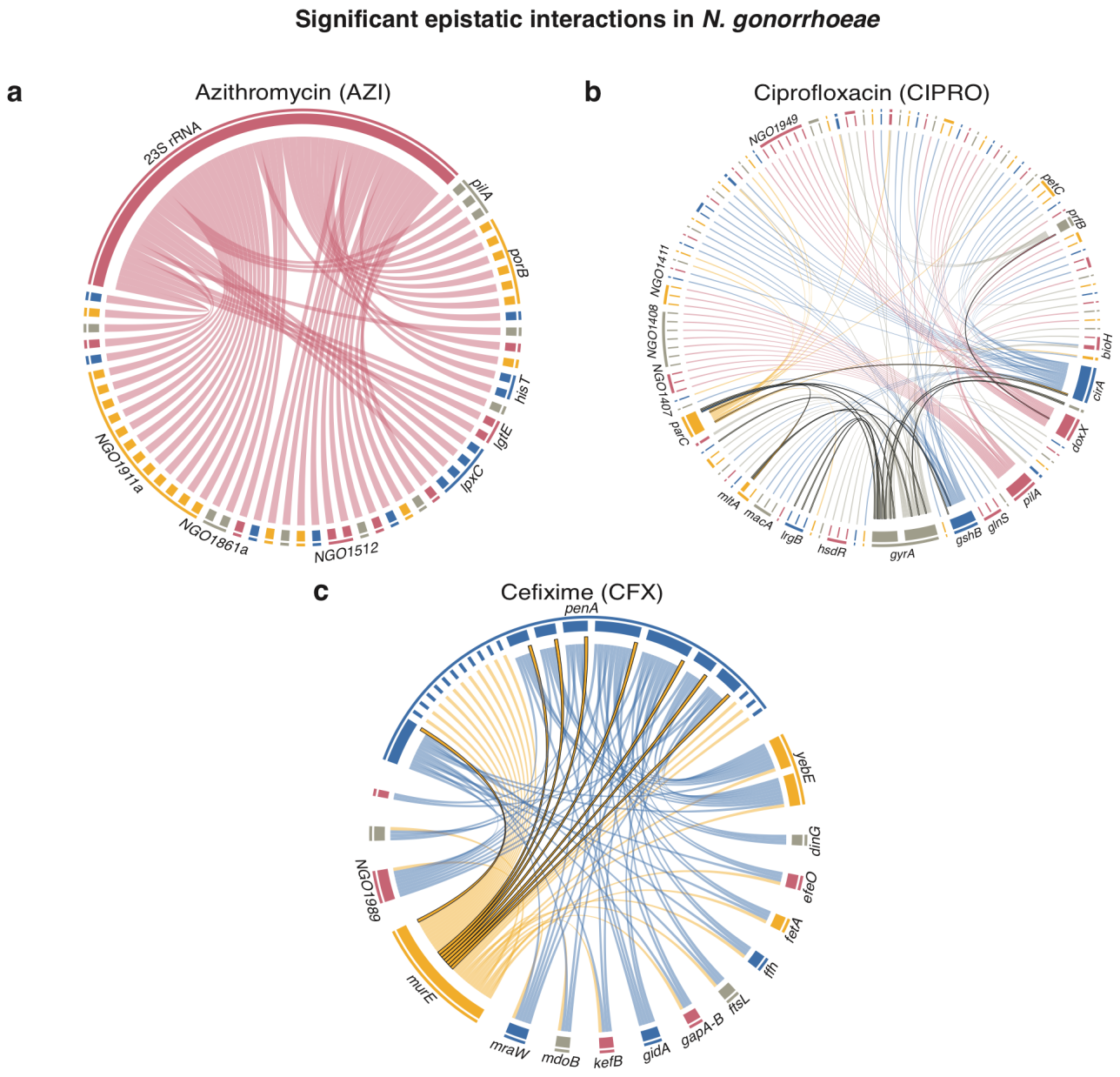

Figure 3: Significant epistatic interactions affecting antibiotic resistance and sensitivity in *Neisseria gonorrhoeae.* Edges between pairs of positions that are both individually associated with changes in minimum inhibitory concentration (MIC) are outlined in black. a) Epistatic interactions affecting resistance to azithromycin (AZI). All interactions are connected to a single C2617T variant in 23S rRNA, which is known to cause significant resistance to AZI. b). Epistatic interactions affecting resistance to ciprofloxacin (CIPRO). c) Epistatic interactions affecting resistance to cefixime (CFX).

SUPPLEMENT

S1 Table: Minimum inhibitory concentration of test and confirmatory strains and their NCBI SRA identifiers.

S2 Table. Significant single-locus associations in the exploratory dataset.

S3 Table. Significant epistatic associations in the exploratory dataset.

S4 Table. Significant epistatic associations confirmed in the confirmatory dataset.

ONLINE METHODS

Alignment construction and annotation

We used *breseq*[36] to call variants in each isolate with reference to the *Neisseria gonorrhoeae* FA1090 genome. To identify variants in ribosomal rRNAs, we masked all except the first copy of the rRNA operon in the reference genome. We ran *breseq* in consensus-mode using the masked reference genome. We filtered variants in the consensus-mode runs for single nucleotide polymorphisms (SNPs), then mapped the SNPs onto the reference sequence using *gdtools* (part of *breseq*) to generate a genome alignment. We ran *breseq* in polymorphism-mode to infer the copy number of rRNA variants in each isolate. The FA1090 reference genome was used to annotate the location of coding regions (CDS) and other genomic features. Transcription start (TSS) and end sites (TES), as well as promoter annotations were extracted from Remmele *et al.*[23]. The genome alignment was filtered for variants with a known functional annotation that occurred in at least 5 strains (0.5% minor allele frequency). Synonymous variants within CDS were excluded from the alignment.

Genome-wide evolutionary couplings inference

To identify strongly coupled positions within a genome of length L, we fit a $q = 2$ state undirected graphical model of the form:

$$P(\delta\,)_b = \frac{1}{Z}\exp\left(\sum_{i=1}^{L} h_i(\,\delta_i) + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} J_{ij}(\delta_i, \delta_j)\right),$$

using a multiple sequence alignment of *N* bacterial genomes. Minor alleles were encoded as '1' and major alleles were encoded as '0'. To prevent overfitting, we used L2-regularized pseudo-likelihood estimation, similar to previous work[13-15,24,37]. The regularization parameters for the site-specific term was set to $\lambda_h = 0.01$ and the pair-wise term was set to $\lambda_e = \lambda_h(L-1)q$. To summarize the evolutionary couplings strength of each pair of loci, we use the Frobenius norm (FN)[24]:

$$FN_{ij} = \left\|J_{ij}(\delta_i, \delta_j)\right\|_2 = \sqrt{\sum_{k,l=1}^{q} J_{ij}(k,l)^2}\ .$$

Identification of highly evolutionarily coupled loci

To correct for undersampling and phylogenetic bias in the alignment, we adjusted the summarized FN-scores with an average-product correction (APC)[38]. Following Toth-Petroczy *et al.*[37], a two-component mixture model was fitted against the corrected FN-score distribution to identify strongly evolutionarily coupled genomic loci:

$$FN_{ij}{}^{APC} \sim (1-\pi)SN(0, \omega^2, \xi) + \pi LN(\mu, \sigma^2).$$

Here π represents the mixing parameter; SN is a skew-normal distribution with location at 0 and unknown scale ω and shape ξ parameters; and LN represents a log-normal distribution with unknown mean μ and variance σ. All evolutionary couplings (ECs) whose posterior probability of membership in the log-normal component p_{ij} is larger than 0.95 are deemed significant and used for later analysis.

Detection of antibiotic-resistant loci and co-evolutionarily coupled pairs

To test single-locus and epistatic associations to a specific phenotype, we employ a linear mixed model[26,27] of the following form:

$$\boldsymbol{y} \sim \mathrm{N}\left(\boldsymbol{S\beta}, \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \mathbf{I}\right),$$

with $\boldsymbol{S} := \left[\mathbf{1}, \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{X}_i\boldsymbol{X}_j\right]$ or $\boldsymbol{S} := [\mathbf{1}, \boldsymbol{X}_i]$ denoting the design matrix with the locus or epistatic interaction of interest. The bacterial sequences are first binary-encoded in $\boldsymbol{X}$ with $\boldsymbol{X}_i = 0$ as major and $\boldsymbol{X}_i = 1$ as minor allele, and then standardized for the purpose of association testing. $\boldsymbol{K} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$ represents the kernel matrix that is used to correct for population structure and other cryptic cofounders. Single-locus associations are tested for statistical significance using a Wald's-test, while epistatic associations are tested using a likelihood-ratio test comparing the full interaction model to a null model disregarding the interaction effect. Bonferroni correction was used for multiple-testing correction at $\alpha = 0.05$.

Implementation

The undirected graphical model was inferred using PLMC (https://github.com/debbiemarkslab/plmc). The linear mixed model was implemented in Python and is based on Pylmm[27]. The remaining data analysis was performed in Python and Jupyter notebooks.

REFERENCES

1 Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Engineering and Analyzing Multicellular Systems: Methods and Protocols*, 165-188 (2014).

2 Remmele, C. W. *et al.* Transcriptional landscape and essential genes of Neisseria gonorrhoeae. *Nucleic acids research* **42**, 10579-10595 (2014).

3 Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E* **87**, 012707 (2013).

4 Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature Biotechnology* (2017).

5 Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).

6 Weinreb, C. *et al.* 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963-975 (2016).

7 Toth-Petroczy, A. *et al.* Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158-170. e112 (2016).

8 Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333-340 (2008).

9 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354 (2010).

10 Sul, J. H. *et al.* Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS genetics* **12**, e1005849 (2016).