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

Benjamin Schubert[1,2,3,*], Rohan Maddamsetti[1,3,*], Jackson Nyman[1], Maha R. Farhat[4,5], Debora S. Marks[1,6,†]

[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]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

[5]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

[6]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

The analysis of whole genome sequencing data should, in theory, 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 extend a method that has been successfully used to identify epistatic residues in proteins to infer loci strongly coupled and associated with antibiotic resistance from 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 38 loci and 250 epistatic pairs that influence the dose needed to inhibit growth for five different antibiotics in 1102 isolates of *Neisseria gonorrhoeae*, that were confirmed in an independent dataset of 495 isolates. Many of the know resistance-affecting loci were recovered, and more sites within those genes, however the majority of loci occurred in unreported genes, including *murE* which was associated with cefixime. About half of the novel epistasis we report involves at least one locus previously associated with antibiotic resistance, including interactions between *gyrA* and *parC* associated with ciprofloxacin, leaving many combinations involving unreported loci and genes. 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. Clinical *N. gonorrhoeae* isolates already exhibit mono-resistance to extended-spectrum cephalosporins such as cefixime and ceftriaxone[1-4] and other relevant antimicrobials[5]. More alarmingly, a new strain of *Neisseria gonorrhoeae* that is multi-resistant to the currently recommended therapyeutic — 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[7] such as *N. gonorrhoeae*. On the one hand, variants that compensate for costly resistance variants[8] can enable the evolution of multi-resistant bacteria. On the other hand, epistasis can constrain the spread of antibiotic resistance if the fitness cost of resistance prohibits its transmission or the spread of the resistance elements by horizontal gene transfer. 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[9]. Alternatively, others have restricted the statistical analysis to loci that individually are strongly associated with the trait of interest[10,11]. In the end, neither of these approaches allow a systematic search for epistatic interactions.

To conduct a genome-wide association study[12] (GWAS) that tests for epistatic associations in face of the combinatorial explosion of statistical tests, we exploit sequence information by computing evolutionary couplings[13] to identify epistatic interactions. In recent years, evolutionary couplings methods have made breakthroughs in *ab initio* protein and RNA 3D structure[13-15], protein complex[16], and mutation effects prediction[17]. Now, these methods are being adopted for bacterial genome analysis[18]. Evolutionary couplings are able to separate causal interactions from indirect correlations[19] including, to a large extent, global correlations caused by population structure and phylogeny[20,21] 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 suggesting an agnostic alternative to pathway-based filtering approaches[22].

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 and subsequently confirmed in a separate dataset. The structure of the epistatic associations revealed the genetic architecture that underlies antimicrobial sensitivity and resistance in *N. gonorrhoeae*. 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

N. gonorrhoeae *isolate genomes and drug resistance*

We extracted the whole genome sequences and the MICs of five antibiotics (Penicillin: PEN, Tetracycline: TET, Cefixime: CFX, Ciprofloxacin: CIPRO, and Azithromycin: AZI) from clinical studies of *N. gonorrhoeae* infection, Supplementary Table 1. Genome wide association studies will often have some kind of validation test, though not strictly required to make the initial association. We therefore used one dataset of 1102 strains collected in the USA [23] for the initial exploration and a second dataset comprised of 495 strains [24] collected in Canada [25](n = 246) and England [26] (n = 249), which we used for independent confirmation (Supplementary Table 1. Whilst the phylogenetic analysis of all 1597 isolates does not show complete lineage separation, the cohorts are obviously not from the same lineages, supporting the use of the second dataset as confirmatory (Supplementary Figure 1, Methods). In a phenotypic analysis of susceptibility to the antibiotics, we found that 92 % of isolates from both cohorts were resistant to at least one drug, using clinically defined thresholds from the European Committee of Antimicrobial Susceptibility Testing[27] (Figure 2). Moving forward to the analysis of resistance association, using *N. gonorrhoeae* FA1090 as a reference, we identified non-synonymous SNPs and SNPs in non-coding regions within transcripts experimentally determined functional annotation[28] that had a minor allele frequency $\geq$ 0.5%, resulting in 8,686 loci (Methods).

Probabilistic model to capture genome-wide interactions

To test for epistatic contribution to the observed antibiotic resistance we would have to test over 37 million combinations of these 8,686 loci, resulting in low statistical power and a high probability of identifying spuriously correlated pairs. Since it seems reasonable to assume that only a small number of pairs of loci, if any, are causally related to the resistance phenotype, a logical approach would be to simply test the highest, most correlated pairs across the samples. However, many of the highest correlated pairs are likely to be non-causal that arise due to the inherent structure in the data that results in transitive correlations, a problem seen in many areas of biological data analysis - correlation does not mean causation, see [19,29] . Here we solve this by applied an agnostic approach using a maximum entropy model to identify which pairs of loci best explain all other observed pairs in the data. The model and inference approaches we develop are based on a method that identifies causal dependencies and epistasis between residues in proteins and RNAs and has led to successful folding 3D structure[13,19,30] and prediction of mutation effects[17] from sequences alone.

We therefore computed the epistatic relationships between the 8,686 loci before exploring the association to antibiotic resistance phenotypes. Our statistical model associates the genome sequence σ , here consisting of an alignment of 8,686 loci with a probability distribution $P(\sigma)$ 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 term h_i:

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

We use a regularized pseudolikelihood maximization inference [31] to compute the parameters Jij and hi. . To measure the evolutionary coupling (EC) strength between pairs of loci, the inferred parameters $\boldsymbol{J}_{ij}$ are summarized using the Frobenius norm[31]:

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

To calculate the final EC score, we corrected the FN scores for population-structure and under-sampling using the average product correction [32]. From the full set of all pairs (~ 37 million) we selected the subset that had a more than 95% probability for being in the signal distribution when fitting the coupling strengths to a two-component mixture model (Methods). This resulted in a high-confidence set of 242,360 pairs involving 7,868 loci in 1,438 genes with the majority in protein–coding genes (7,437), 178 within transcription start sites or 5' untranslated regions, 224 within promoters, and 29 within ribosomal RNAs, Supplementary Table 2.

Genome-wide single-site analysis recapitulates known associations and predicts new ones

Before testing for epistatic pairs of loci identified we performed a single-site GWAS analysis. We used a linear mixed model to measure the association strength of individual loci with the measured change in minimum inhibitory concentrations, correcting for population structure, as in standard GWAS approaches [33,34] (Methods). We tested all 8,686 loci for single association, controlling for multiple hypothesis testing using a Bonferroni correction, such that all loci were then deemed significant at alpha set to 0.05 divided by the number of tests for each antibiotic. Nearly half (38/82) of the significant associations in the exploratory dataset were also significant in the confirmatory dataset (Tables 1 and 2, Supplementary Table 3).

Ciprofloxacin: Of these 38 loci, 22 were associated with CIPRO with known loci in *gyrA* (2 loci) and *parC* as the most significant (p= $8.25\text{x}10^{-101,}$ $1.98\text{x}10^{-43}$ and $9.96\text{x}10^{-18}$ respectively) (Table 1). We also saw single significant single associated loci in 20 genes and four non-coding regions (Table 1). The two strongest novel associations, found in *dldH (*dihydrolipoyl

dehydrogenase (p.A88V, $p = 6.68x10^{-9}$) and *gshB* (glutathione synthetase (p.E221K, $p = 2.71x10^{-8}$))are involved in oxidative stress response[35] which can be induced by fluoroquinolones[36] (e.g. ciprofloxacin). The *dldH* variant A88V is on the homodimer interface of the multimeric enzyme suggesting the multimerization stability is part of the mechanism (Figure 3). Our analysis suggests that *dldH* variant causes greater CIPRO sensitivity, since it is uncommon in resistant isolates (46/833) and its regression coefficient (β) is negative. By contrast, the variant E221K in *gshB* is associated with a higher CIPRO MIC and a positive regression coefficient (β), suggesting that the lysine mutation increases resistance. The highest density of CIPRO associations was found in *gyrA*, the AsnC family transcriptional regulator NGO1407 (three loci), *gshB*, and in the 5' untranslated region of *pyrG* (two loci each).

Cefixime: Eight of our CFX associated loci are to the known target *penA* (PBP2, penicillin binding protein 2), which is mosaic in the test isolates[23] with only one of the sites in a previously confirmed locus (p.G546S, $p = 5.08x10^{-12}$)[4]. The remaining four significantly associated loci are in the cell-wall biosynthesis gene, *murE,* mutations in which have increased β-lactam resistance in *S. pneumoniae*[37] but, which to our knowledge, have not been reported previously in *N. gonorrhoeae* (Figure 3).

Penicillin and Tetracycline: We recover one of the known *porB* sites associated with increased resistance (A121D/G/V , $p = 2.45x10^{-4}$) [23] in our confirmation dataset. We also identified previously unreported loci in *porB* that are likely to be associated with increased antibiotic sensitivity (its regression coefficient (β) is negative). One was associated with PEN (p.V151A, $p = 4.73x10^{-3}$) and another was associated with TET (p.F131Y, $p = 1.33x10^{-3}$). 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.
The only single association we see for AZI is the previously known locus in 23S rRNA g.C2617, previously described as C2611, $p = 2.19x10^{-31}$

Genome-wide association identifies epistatic pairs associated with antibiotic resistance

Next, we tested all 242,360 pairs of loci with high-scoring evolutionary couplings for epistatic association. To assess the contribution of each epistatic effect, we compared the full interaction model against a model only containing the individual sites as additive effects. As before, we applied a Bonferroni correction for multiple testing at an adjusted $\alpha = 0.05$.

We verified 250 epistatic associations in the confirmatory dataset (Figure 4, Table 3), out of 729 significant epistatic associations in the exploratory dataset (Supplementary Tables 4 and 5). We confirmed epistatic interactions to CIPRO< CFX and AZI but not to TET or PEN.
All 250 epistatic interactions involved loci that were confirmed in our single locus GWAS. Ten epistatic associations connected two loci which were each significant in the single locus GWAS. The remaining 240 interactions connected an associated locus to a non-associated locus. In addition, most of the known resistance variants that we identified in our single locus GWAS (5/6) also occurred in our epistatic GWAS analysis (141 occurrences). One hundred and eighteen of the confirmed epistatic associations were associated with CIPRO MIC change; 69 with CFX; and 63 with AZI. Although no interactions had both of the pair shared across antibiotics, four genes had epistatic associations with multiple antibiotics including, *rmsE*, a 16S rRNA (uracil(1498)-N(3))-methyltransferase between shared between AZI and CIPRO

Ciprofloxacin: Epistatic interactions associated with CIPRO involved 79 genes (Figure 3B), and the top 15 most significant (all at $p = 1.73x10^{-45}$) involved the well-known S91F resistance variant in *gyrA*[5]. Two of the strongest epistatic pairs included interactions between the two known resistance-causing loci in *gyrA* and a resistance-causing locus in *parC* (*gyrA* p.S91F to *parC* p.S87F, $p = 1.73x10^{-45}$; *gyrA* p.D95A/G to *parC* p.S87, $p = 1.26x10^{-30}$). The gene products of *parC* and *gyrA* are both topoisomerases type II proteins and these well-known loci, S91(*gyrA*) and S87 (*parC*), are in homologous positions close to the DNA and ciprofloxacin binding sites where the mutations make the enzyme unable to religate the DNA[38] .To our knowledge this is the first report of epistatic effects between these loci in *N. gonorrhoeae*; notably, homologous epistatic interactions (affecting both ciprofloxacin resistance and fitness) have been seen in experiments with *S. pneumoniae*[39].

We also identified interactions between *gyrA* S91F and *birA (*E261K, $p = 1.73x10^{-45}$*)* and *glmU (*T425A*,* $p = 1.73x10^{-45}$) respectively. *birA* is known to increase sensitivity to ciprofloxacin when underexpressed in *Mycobacterium smegmatis*[40], while *glmU*'s protein abundance increases under ciprofloxacin stress in *Salmonella typhimurium,* possibly modulating cell wall permeability[41]. In total 15 genes contained three or more epistatic associations, dominated by inner and outer-membrane proteins (*doxX*, *lrgB*, *cirA*, *petC*, NGO1958, *tamB*), some of which are associated with oxidative stress[35]. **Azithromycin:** All epistatic associations to AZI involved a single, known antibiotic resistance variant C2617T (identical to C2611T in Grad *et al.* 2016[23]) in the 23S rRNA (Table 2 and Figure 3A). Of the eleven genes containing two or more associated loci, seven were either membrane proteins (NGO1496, NGO1861a, NGO1463a, *porB*), cell wall biosynthesis proteins (*lgtE*) or pilus proteins (NGO1911a, *pilA*). **Cefixime:** By contrast with CIPRO, epistatic associations with CFX involved only 15 genes, each with multiple associated loci. Eleven epistatic interactions involved the known single resistance locus (p.G546S) in *penA*. The fifteen genes were dominated by cell membrane, cell wall, and cell wall biosynthesis proteins (*penA*, *fetA, efeO, yebE, mdoB, ffh, arnT, oafA*), but also included cell division proteins (*ftsL*, *gidA*), and DNA repair (*dinG*).

Taken together, the epistatic interactions often involved known antibiotic resistance loci. Of the interactions that did not involve known antibiotic resistance loci, some were found in genes associated with cell division, and oxidative stress. The majority of novel epistatic loci, however, were found in genes encoding inner- and outer- membrane proteins, or cell wall biosynthesis proteins that potentially affect *N. gonorrhoeae's* permeability to antibiotics[42-44].

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 and comF* that are associated with CIPRO (Table 1). We also found several unreported variants and interactions that associate strongly with greater antibiotic resistance, including several variants in *murE* linked to CFX resistance and two interactions between *gyrA* and *parC* linked to CIPRO 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 gene *penA*. 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[45]. 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 combined with standard GWAS methods is a promising new approach for discovering epistatic interactions affecting antibiotic resistance in a multi-resistant pathogen, *N. gonorrhoeae*. The loci and interactions captured may support efforts in predicting phenotypic resistance from whole genome sequence data. 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 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

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}**	**2.727**	**239/833**
620905*	***gyrA* (NGO0629)**	**D95(A/G)**	**1.98×10^{-43}**	**2.091**	**270/833**
1210781*	***parC* (NGO1259)**	**S87R**	**9.96×10^{-18}**	**1.929**	**702/833**
894348*	*dldH* (NGO0915)	A88V	6.68×10^{-9}	–1.265	46/833
1167114*	*gshB* (NGO1217)	E221K	2.71×10^{-8}	0.920	242/833
1687733*	*comF* (NGO1726)	F43Y	1.86×10^{-7}	–1.004	55/833
93476	NGO0086	D363(D/E)	1.66×10^{-5}	–0.885	27/833
1167527	*gshB* (NGO1217)	E83A	1.91×10^{-5}	0.647	241/833
89733	*pglD* (NGO0083)	A360T	2.07×10^{-5}	0.728	652/833
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
447535	*fimT* (NGO0452)	G100D	9.02×10^{-5}	–0.553	67/833
1162600	*pyrG* (NGO1212)	–12	0.000176	0.647	299/833
1162602	*pyrG* (NGO1212)	–14	0.000176	0.647	299/833
964051*	NGO0993	K161E	0.000389	-0.526	11/833
1378462	*nqrB* (NGO1414)	A25V	0.000661	-0.578	22/833
565114	*uvrC* (NGO0578)	N258D	0.000696	0.664	228/833
1747633	*gapA-B* NGO1776	V267A	0.000882	–0.568	35/833
1373003	NGO1407	L66M	0.00116	–0.666	3/833
1372748	NGO1407	–61	0.00116	–0.666	3/833
1373005	NGO1407	L66M	0.00116	–0.666	3/833

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 (–) from the closest gene. 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 S5).

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	1230597	23S rRNA	NGO1277	C2617T	315/690	D173G	358/690	0.877	5.77E-29
AZI	1116541	1427960	23S rRNA	*opA54* (NGO1463a)	C2617T	315/690	V123L	32/690	0.877	5.77E-29
AZI	1116541	1530961	23S rRNA	*putA* (NGO1552a)	C2617T	315/690	E1169G	377/690	0.877	5.77E-29
AZI	1116541	1789096	23S rRNA	*porB* (NGO1812)	C2617T	315/690	N134D/H	352/690	0.877	5.77E-29
CFX	1516646	1517334	*penA* (NGO1542)	*ftsL* (NGO1543)	N173S	277/283	L52V	0/283	0.956	1.63E-10
CFX	1516646	2025976	*penA* (NGO1542)	*efeO* (NGO2050)	N173S	277/283	S360G	0/283	0.959	1.63E-10
CFX	1516685	2070085	*penA* (NGO1542)	*fetA* (NGO2093)	V160A	277/283	A638V	132/283	0.959	1.63E-10
CFX	1516685	1665722	*penA* (NGO1542)	*yebE* (NGO1709)	V160A	277/283	K90E	258/283	0.959	1.63E-10
CIPRO	620905	1210781	*gyrA* (NGO0629)	*parC* (NGO1259	D95A/G	270/833	S87R	702/833	1.984	1.26E-30
CIPRO	620917	1210781	*gyrA* (NGO0629)	*parC* (NGO1259)	S91F	239/833	S87R	702/833	2.413	1.73E-45
CIPRO	620917	2028211	*gyrA* (NGO0629)	*glmU* (NGO2053)	S91F	239/833	T425A	191/833	2.413	1.73E-45
CIPRO	620917	1974020	*gyrA* (NGO0629)	*birA* (NGO2001)	S91F	239/833	E261K	56/833	2.413	1.73E-45
CIPRO	620917	693931	*gyrA* (NGO0629)	*hsdS* (NGO0699)	S91F	239/833	A173T	47/833	2.413	1.73E-45
CIPRO	620917	894348	*gyrA* (NGO0629)	*dldH* (NGO0915)	S91F	239/833	A88V	46/833	2.413	1.73E-45
CIPRO	620917	1167114	*gyrA* (NGO0629)	*gshB* (NGO1217)	S91F	239/833	E221K	242/833	2.413	1.73E-45
CIPRO	620917	1522835	*gyrA* (NGO0629)	*ftsN* (NGO1549)	S91F	239/833	Q142R	244/833	2.413	1.73E-45

NOTE– Loci in bold are known resistance variants. Non-coding variants are labeled with the distance upstream (–) from the closest gene. 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

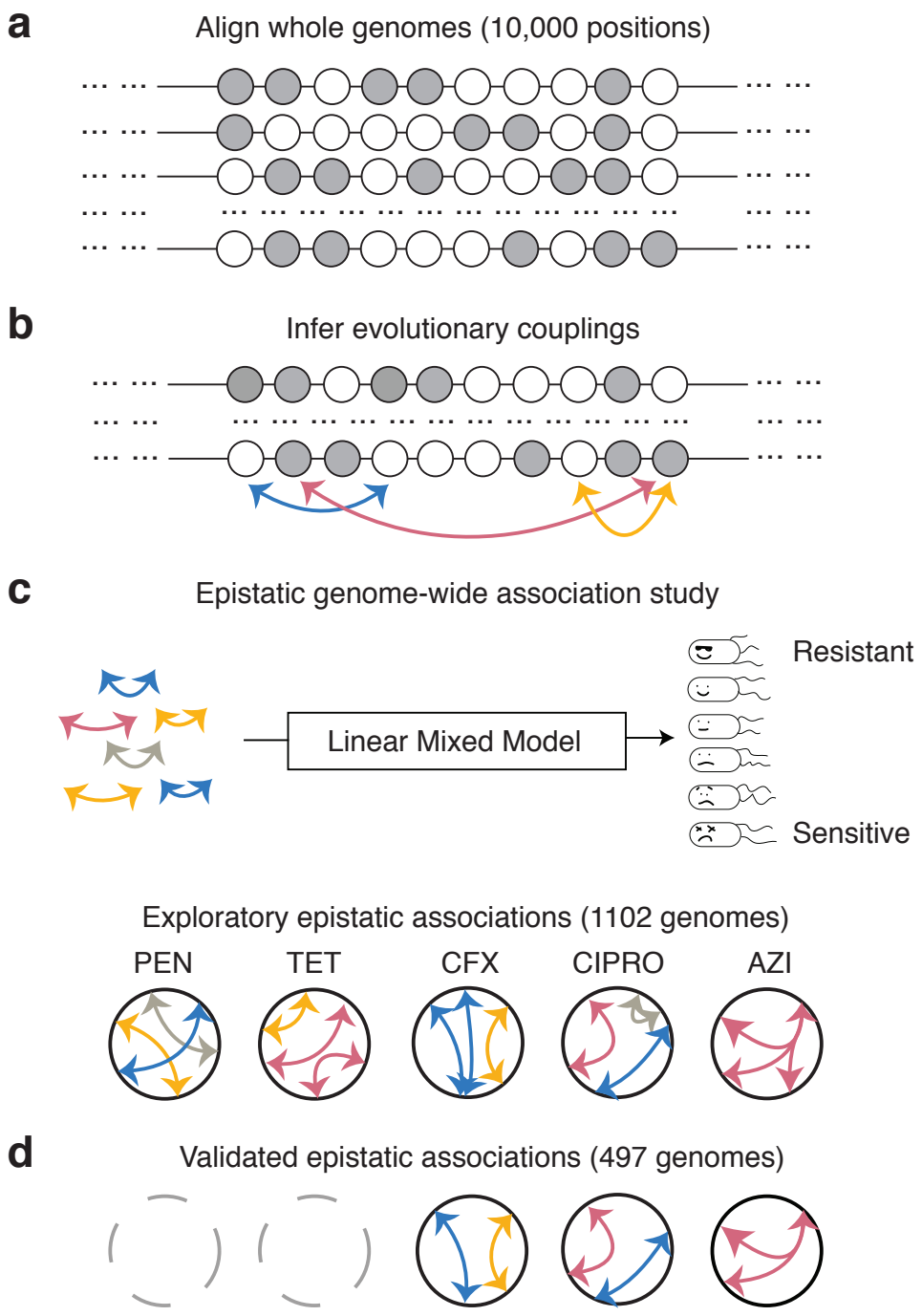


Figure 1: Evolutionary couplings power epistatic GWAS. a) First, non-synonymous SNPs and non-coding variants with experimentally-derived functional annotation are identified and used to generate a whole-genome alignment. b) Highly evolutionarily coupled loci are identified using an undirected graphical 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.

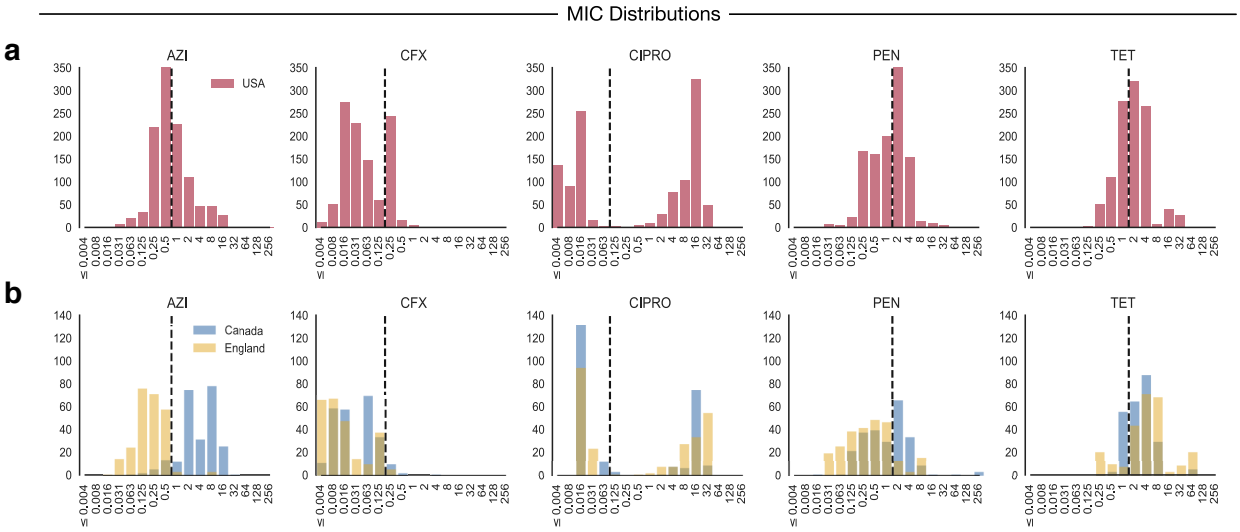


Figure 2: Distributions of minimum inhibitory concentration (MIC), in milligrams per liter, for the *N. gonorrhoeae* strains studied. The dotted lines indicate clinical breakpoints for *N. gonorrhoeae* as defined by the European committee on antimicrobial susceptibility testing (EUCAST). The MIC distribution of the 1102 strains in the exploratory test dataset. In the test set, 537 strains were resistant to penicillin (PEN; resistance threshold > 1 mg/L), 661 were resistant to tetracycline (TET; > 1 mg/L); 266 were resistant to cefixime (CFX; > 0.125 mg/L); 602 were resistant to ciprofloxacin (CIPRO; > 0.06 mg/L); and 459 were resistant to azithromycin (AZI; > 0.5 mg/L). b) The MIC distribution of the 495 strains in the confirmatory dataset. The confirmatory set contained 141 PEN, 400 TET, 17 CFX, 231 CIPRO, and 231 AZI resistant.

Figure 3

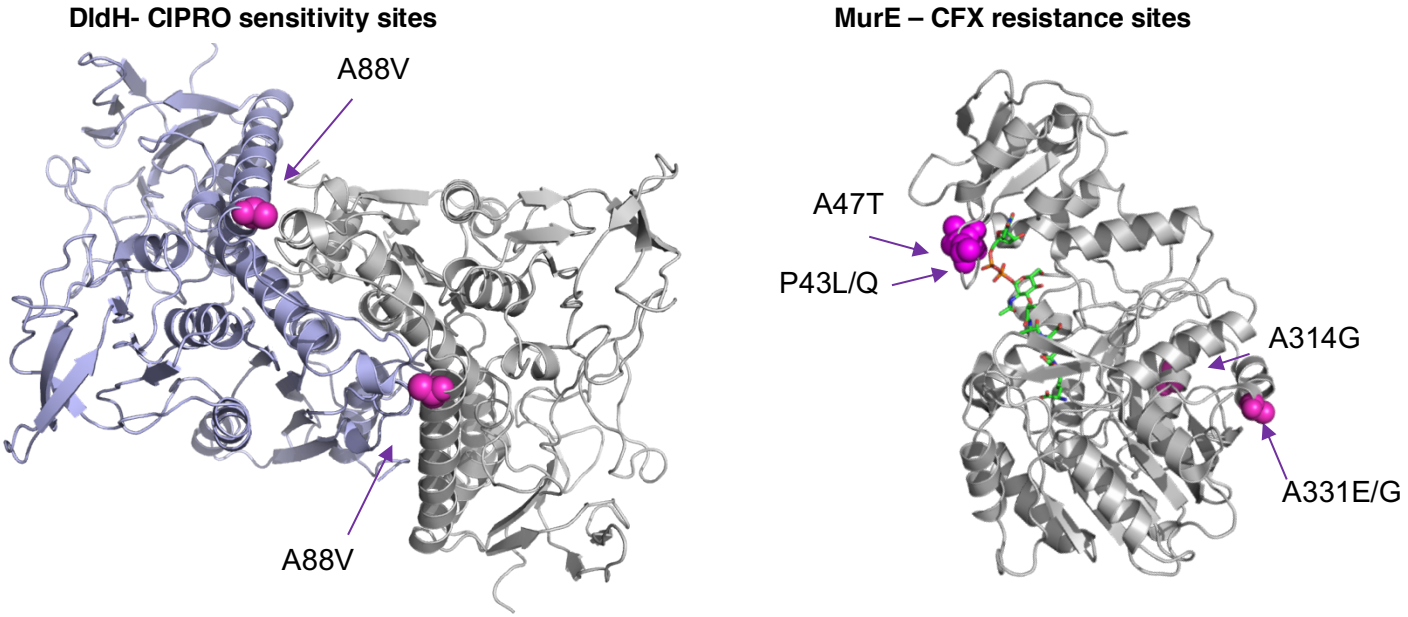


Figure 3 Novel associations that occur in structural binding sites

Left: *dldH* A88 (magenta), associated with CIPRO resistance, maps on the interface of the homodimer mapped onto the 3D structure of the homologous protein in *Pisum sativum*, 1dxl[46]. **Right:** Four variants associated with CFX resistance in *murE,* mapped onto the 3D structure of the homologous protein in *E.coli*, 1e8c[47]. See Tables 1 and 2 for details.

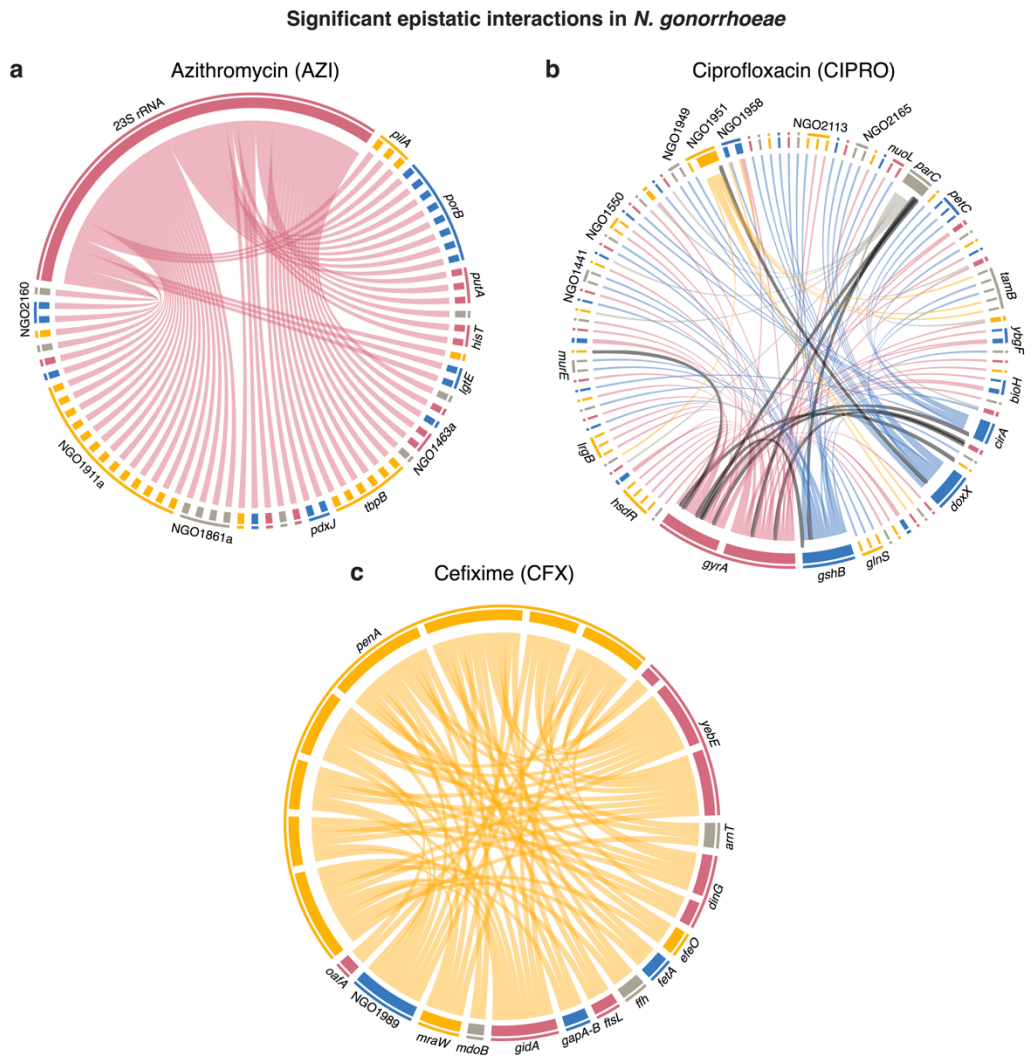


Figure 4 : Significant epistatic interactions affecting antibiotic resistance and sensitivity in *Neisseria gonorrhoeae.* Edges between pairs of positions that are both individually associated with changes in minimum inhibitory concentration (MIC) are outlined in black. a) Epistatic interactions affecting 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).

Alignment construction and annotation

We used *breseq*[1] 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. Promoters and 5' untranslated regions associated with experimentally determined RNA transcripts[2] were annotated as follows. Regions bounded by the transcriptional start site and the location of the first gene on the transcript were annotated as 5' untranslated regions, as long as the annotated region was no more than 150 nucleotides (nt) long. This decision rule was based on experimental results on the distribution of 5' untranslated region lengths in *Bacillus subtilis*, *Escherichia coli*, and *Pseudomonas aeruginosa*[3]. Sequences up to 70 nt upstream of the transcript (or, the end of the upstream gene on the same strand if within 70 nt) were annotated as promoters, following Remmele *et al.*[2]. 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.

Population structure analysis

We used FastTree[4] to make an approximately maximum-likelihood phylogeny using a genome alignment of all 1597 strains. We did not filter the alignment for loci with evidence of recombination, for two reasons. First, we were more interested in seeing how strains cluster (i.e. visualizing population structure), than in inferring the evolutionary history of *N. gonorrhoeae* based on a single clonal frame[5]. Second, Hedge and Wilson (2014)[6] have shown that bacterial phylogenetic reconstruction from whole genomes is robust to recombination; furthermore, filtering out loci with evidence of recombination could actually worsen our estimate of *N. gonorrhoeae* phylogeny by removing many informative sites.

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[7-11]. 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)[7]:

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

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)[12]. Following Toth-Petroczy *et al.*[11], 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[13,14] 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[14]. 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 Johns, N. I. *et al.* Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nature Methods* **15**, 323 (2018).

4 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).

5 Milkman, R. & Bridges, M. M. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. *Genetics* **126**, 505-517 (1990).

6 Hedge, J. & Wilson, D. J. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* **5**, e02158-02114 (2014).

7 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).

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

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

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

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

12 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).

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

14 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).

SUPPLEMENT

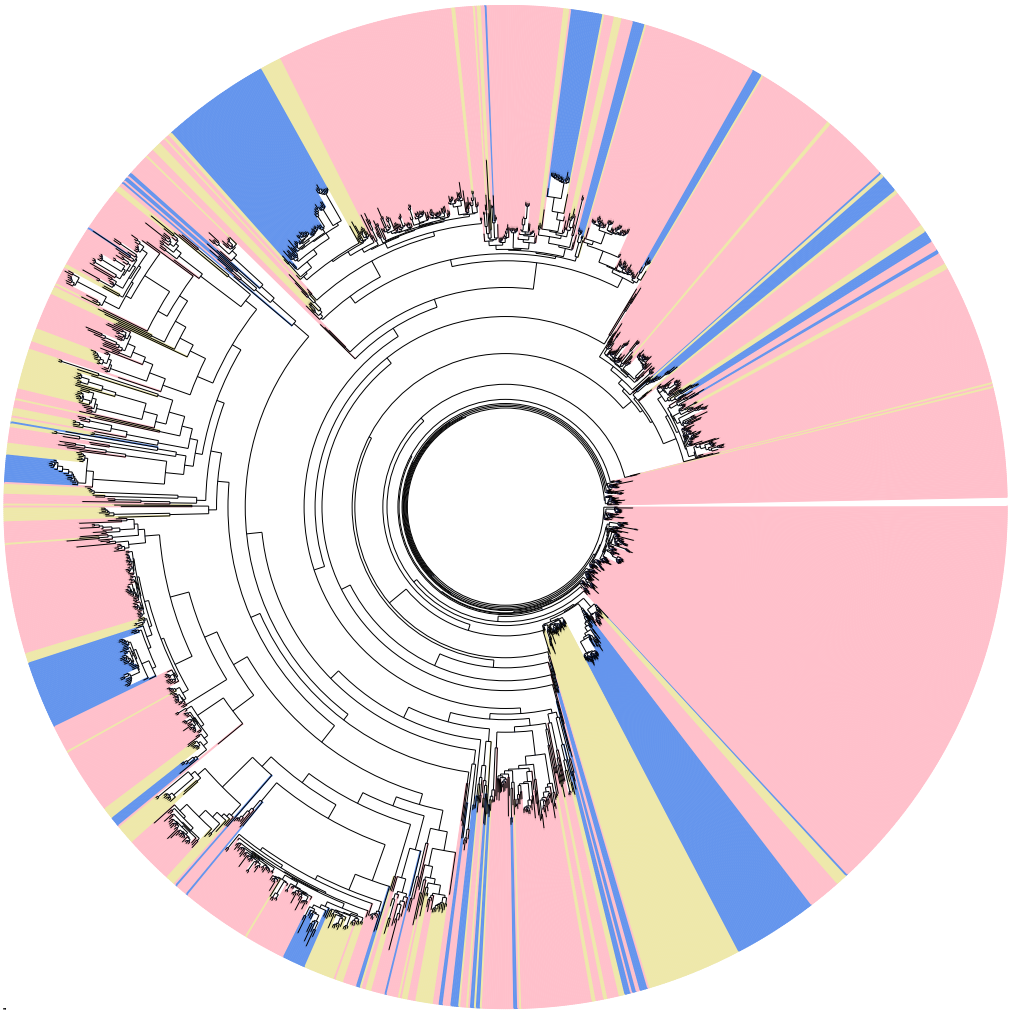
Supplementary Table 1: Minimum inhibitory concentration of exploratory and confirmatory strains and their NCBI SRA identifiers.

Supplementary Table 2. Significant evolutionarily coupled loci in the exploratory dataset.

Supplementary Table 3. Entire single-locus GWAS analysis of the exploratory and confirmatory dataset.

Supplementary Table 4. Confirmed and Bonferroni-corrected epistatic associations.

Supplementary Table 5. Entire epistatic GWAS analysis of the exploratory and confirmatory datasets.



Supplementary Figure. Population structure of the 1102 exploratory (USA) and 495 confirmatory (England and Canada) *N. gonorrhoeae* isolates. USA isolates are colored pink, England isolates are yellow, and Canada isolates are blue.

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 Palmer, A. C. & Kishony, R. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nature Reviews Genetics* **14**, 243 (2013).

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

9 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).

10 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).

11 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).

12 Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Current opinion in microbiology* **25**, 17-24 (2015).

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

14 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).

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

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

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

18 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).

19 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).

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

21 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).

22 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).

23 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).

24 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).

25 Demczuk, W. *et al.* Whole-genome phylogenomic heterogeneity of Neisseria gonorrhoeae isolates with decreased cephalosporin susceptibility collected in Canada between 1989 and 2013. *J Clin Microbiol* **53**, 191-200, doi:10.1128/JCM.02589-14 (2015).
26 De Silva, D. *et al.* Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae: an observational study. *Lancet Infect Dis* **16**, 1295-1303, doi:10.1016/S1473-3099(16)30157-8 (2016).
27 *The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters, version 7.1. Dec 2017*, <http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_7.1_Breakpoint_Tables.pdf>
28 Remmele, C. W. *et al.* Transcriptional landscape and essential genes of Neisseria gonorrhoeae. *Nucleic acids research* **42**, 10579-10595 (2014).
29 Giraud, B. G., Heumann, J. M. & Lapedes, A. S. Superadditive correlation. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **59**, 4983-4991 (1999).
30 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).
31 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).
32 Harrison, O. B. *et al.* Genomic analyses of Neisseria gonorrhoeae reveal an association of the gonococcal genetic island with antimicrobial resistance. *J Infect* **73**, 578-587, doi:10.1016/j.jinf.2016.08.010 (2016).
33 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354 (2010).
34 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).
35 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).
36 Kohanski, M. A., Dwyer, D. J. & Collins, J. J. How antibiotics kill bacteria: from targets to networks. *Nature Reviews Microbiology* **8**, 423 (2010).
37 Todorova, K. *et al.* Transfer of penicillin resistance from Streptococcus oralis to Streptococcus pneumoniae identifies murE as resistance determinant. *Molecular microbiology* **97**, 866-880 (2015).
38 Redgrave, L. S., Sutton, S. B., Webber, M. A. & Piddock, L. J. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends Microbiol* **22**, 438-445, doi:10.1016/j.tim.2014.04.007 (2014).
39 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).
40 Duckworth, B. P. *et al.* Bisubstrate adenylation inhibitors of biotin protein ligase from Mycobacterium tuberculosis. *Chemistry & biology* **18**, 1432-1441 (2011).
41 Correia, S. *et al.* Comparative subproteomic analysis of clinically acquired fluoroquinolone resistance and ciprofloxacin stress in Salmonella Typhimurium DT104B. *PROTEOMICS-Clinical Applications* **11** (2017).
42 Vaara, M. Outer membrane permeability barrier to azithromycin, clarithromycin, and roxithromycin in gram-negative enteric bacteria. *Antimicrobial agents and chemotherapy* **37**, 354-356 (1993).
43 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).
44 Delcour, A. H. Outer membrane permeability and antibiotic resistance. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1794**, 808-816 (2009).
45 Barber, M. F. & Elde, N. C. Escape from bacterial iron piracy through rapid evolution of transferrin. *Science* **346**, 1362-1366 (2014).
46 Faure, M. *et al.* Interaction between the lipoamide-containing H-protein and the lipoamide dehydrogenase (L-protein) of the glycine decarboxylase multienzyme system 2. Crystal structures of H- and L-proteins. *Eur J Biochem* **267**, 2890-2898 (2000).
47 Gordon, E. *et al.* Crystal structure of UDP-N-acetylmuramoyl-L-alanyl-D-glutamate: meso-diaminopimelate ligase from Escherichia coli. *J Biol Chem* **276**, 10999-11006, doi:10.1074/jbc.M009835200 (2001).