NAR Genomics and Bioinformatics, 2024, **6**, lqae061
https://doi.org/10.1093/nargab/lqae061
Advance access publication date: 6 June 2024
Methods Article

Detecting co-selection through excess linkage disequilibrium in bacterial genomes

Sudaraka Mallawaarachchi[1,*], Gerry Tonkin-Hill[1], Anna K. Pöntinen[1,2], Jessica K. Calland[3], Rebecca A. Gladstone[1], Sergio Arredondo-Alonso[1], Neil MacAlasdair[1], Harry A. Thorpe[1], Janetta Top[4], Samuel K. Sheppard[5], David Balding[6], Nicholas J. Croucher[7,8,†] and Jukka Corander[1,9,10,*,†]

[1]Department of Biostatistics, University of Oslo, Oslo, Norway
[2]Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway
[3]Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway
[4]Department of Medical Microbiology, UMC Utrecht, Utrecht, The Netherlands
[5]Ineos Oxford Institute of Antimicrobial Research, Department of Biology, University of Oxford, Oxford, United Kingdom
[6]Melbourne Integrative Genomics, School of BioSciences and School of Mathematics & Statistics, University of Melbourne, Parkville, Victoria, Australia
[7]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom
[8]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, United Kingdom
[9]Parasites and Microbes, Wellcome Sanger Institute, Cambridge, UK
[10]Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[*]To whom correspondence should be addressed. Email: smallawaarahchi@gmail.com
Correspondence may also be addressed to Jukka Corander. Tel: +47 22 84 53 00; Fax: +47 22 84 53 01; Email: jukka.corander@medisin.uio.no
[†]The last two authors should be regarded as Joint Last Authors.

Abstract

Population genomics has revolutionized our ability to study bacterial evolution by enabling data-driven discovery of the genetic architecture of trait variation. Genome-wide association studies (GWAS) have more recently become accompanied by genome-wide epistasis and co-selection (GWES) analysis, which offers a phenotype-free approach to generating hypotheses about selective processes that simultaneously impact multiple loci across the genome. However, existing GWES methods only consider associations between distant pairs of loci within the genome due to the strong impact of linkage-disequilibrium (LD) over short distances. Based on the general functional organisation of genomes it is nevertheless expected that majority of co-selection and epistasis will act within relatively short genomic proximity, on co-variation occurring within genes and their promoter regions, and within operons. Here, we introduce LDWeaver, which enables an exhaustive GWES across both short- and long-range LD, to disentangle likely neutral co-variation from selection. We demonstrate the ability of LDWeaver to efficiently generate hypotheses about co-selection using large genomic surveys of multiple major human bacterial pathogen species and validate several findings using functional annotation and phenotypic measurements. Our approach will facilitate the study of bacterial evolution in the light of rapidly expanding population genomic data.

Introduction

The rapid rate of evolution of bacterial genomes has made them a popular target of studying selection in both experimental and natural populations. The emergence of affordable high-resolution population genomics a decade ago ushered us into a new era with improved possibilities to investigate both microscopic and macroscopic evolution of bacterial genomes, including for example codon bias (1), intergenic selection (2) and variation in gene content (3). Despite steady progress in genome-wide association study (GWAS) methodologies specifically designed for bacterial populations (4,5), the difficulty of measuring quantitative phenotypic variation in large numbers of isolates has restricted the use of GWAS mostly to the study of antibiotic resistance, with a few exceptions covering for example duration of colonization (6), genome-wide transcriptomics (7) and virulence (8). Traits such as reproductive rate, survival and transmissibility are of key interest in bacteria but remain difficult to measure in sufficient numbers in natural populations. Motivated by this obstacle, phenotype-free approaches to uncovering signals of selection have been introduced (9–12), based on the rationale that positively selected variation in complex traits would likely be caused by synchronized changes in multiple genes and/or regulatory elements that are detectable from excess linkage disequilibrium (LD) between distant sites across a sample of genomes. This provides a methodological toolkit complementary to GWAS enabling analyses termed genome-wide epistasis and co-selection studies (GWES), which has been recently used to unravel signals of selection due to epistasis for a wide diversity of bacteria (13–16). Arnold *et al.*, using positive, negative and sign epistasis models, demonstrated that under selection, even relatively weak epistasis is sufficient for

Received: February 11, 2024. **Revised:** April 15, 2024. **Editorial Decision:** May 13, 2024. **Accepted:** May 14, 2024

driving adaptation in moderately to highly recombinogenic bacteria, which provides a more theoretical justification for GWES analysis (17), see also the recent work on the possibly saltational role of epistasis in highly recombining bacterial species (18).

Existing GWES approaches are limited to discovering links between distant loci within the region where LD asymptotes towards its lower bound. However, many co-evolving loci are organized into clusters of genes (e.g. co-transcribed in operons). Therefore, studying only long-range links ignores most of the allelic co-variation occurring in genomes. A fine-scale haplotype structure analysis of *Neisseria gonorrhoeae* indeed revealed numerous likely examples of positive co-selection in different regions of the chromosome of this highly recombinant species, and extensive population genetic simulations suggested that such LD patterns could be explained by either directional selection on horizontally acquired alleles, or balancing selection maintaining the diversity (19). Motivated by these insights, we aimed at developing a scalable statistical approach that disentangles neutral LD from co-selection and epistasis at any genomic distance. Apart from co-selection and epistasis, LD can also be influenced by various other factors, including population structure, population expansion, mutation and admixture. While disentangling neutral LD remains a valuable approach for identifying co-selection and epistasis, sole reliance on it cannot distinguish between underlying evolutionary factors. Since wet-lab-based validation would typically be necessary to resolve causal factors underlying observed deviations from a baseline neutral LD, this serves as a useful starting point towards identifying important molecular drivers of success in bacterial populations.

GWES methods generally exploit the decay of LD as a function of genomic distance to label SNP pairs as 'outliers' with respect to the background distribution of LD estimated from population data. The intuition here, for example in the context of synergistic epistasis, is that when the combined effect on selection of two or more polymorphic loci is greater than the sum of their individual effects, this allele combination will be maintained in association in the population, giving rise to discernable patterns of LD. Similar to Arnold *et al.* (19), it is possible to extend the notion of 'outlier' LD level to SNPs in close proximity to each other by simulating the distribution of LD strength as a function of base pair distance using a neutral Wright-Fisher model. These simulations approximate the population level co-variation of alleles expected under neutrality and can be used to screen pairs of loci for outliers that may be due to co-selection. We show that this approach works well and maintains a low false positive rate. However, as fitting of the neutral model parameters and forward simulation of the fitted model in a sufficiently large number of replicates is computationally costly, we developed an empirical model-free approximation that is scalable to large population genomic datasets. The model-free method is motivated by the common assumption that a majority of the observed LD within a bacterial population reflects near-neutrality (20–23), which implies that a majority of observations are not strongly influenced by selection. As a result, it is possible to analyse and interpret LD patterns without relying on additional assumptions about underlying genetic models or selection pressures. By extension, it then becomes feasible to use the empirical LD decay distribution to call outliers. Moreover, the approach accounts for heterogeneity in evolutionary rates, such as mutation and recombination hotspots.

The model-free algorithm is implemented as an open-source R package 'LDweaver' (https://github.com/Sudaraka88/LDWeaver), which can be used to perform a comprehensive GWES in large-scale bacterial datasets. LDWeaver incorporates the functionality of the popular GWES package Spydrpick for long-range LD outlier detection (12) and extends this by allowing analysis of LD at any genomic distance. LDWeaver provides automated functional annotations on all putative co-selected SNPs and generates an array of visualizations to allow users to efficiently explore the results. We use published population genomic data for the major human pathogens *Streptococcus pneumoniae* (24,25), *Campylobacter jejuni* (26), *Escherichia coli* (27) and *Enterococcus faecalis* (28) to identify both known and novel signals of co-selection linked to the molecular basis of pathogenicity, survival and other key bacterial phenotypes.

Materials and methods

Measuring LD

By default, LDWeaver removes sites with MAF <0.01 and gap frequency >0.15. Sites with 'gap' as the second most common allele are also discarded from the analysis by default, but LDWeaver has a filtering option called 'relaxed' that retains these sites in the analysis. This option could be particularly useful for alignments with a limited number of SNPs, as gaps can reflect insertions or deletions with functional effects (80).

Following SpydrPick (12), we use MI to measure LD. The pairwise MI between two sites (modelled as discrete random variables) is given by:

$$\mathrm{MI}(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \qquad (1)$$

Where X and Y denote two sites with alleles $x \in X$ and $y \in Y$, respectively. For SNP data, the alphabet comprises nucleotides A, C, G, T and the gap character N. Here $p(x, y)$ denotes the joint probability of $x = X$ and $y = Y$ and $p(x)$, $p(y)$ are the corresponding marginal probabilities which are estimated from count data (12).

Let n_s denote the number of sequences in the alignment, then:

$$\hat{p}(x,y) = \frac{n(x,y) + 0.5}{n_s + r_x r_Y \times 0.5} \qquad (2)$$

where $n(x, y) = |S_{xy}|$ is the number of sequences with $X = x$ and $Y = y$, $r_X = |X|$ and $r_Y = |Y|$.

Strong population structure in bacteria presents a problem for analysis (80). We adopt a widely-used sequence-reweighting approach (10–12,81,82). The weight $w_i \in [1/n, 1]$ for sequence i is computed as the reciprocal of the number of sequences with mean per-site Hamming distance $< t$, where t is a dataset-dependent threshold that typically satisfies $t \in (0.1, 0.25)$. This population structure correction is applied to the LD structure estimation by substituting effective counts in Eq. (2) given by $n_s = \sum_{i=1}^{n} w_i$ and $n(x,y) = \sum_{i\in S_{xy}} w_i$.

In practice, the MI computation process is optimised using a sparse matrix representation and performed in blocks of SNPs (83), which can be feasible even in systems with relatively low

memory. LDWeaver requires genome-wide short-range links to be retained in memory (see below), but most long-range links with low MI values are discarded after processing each SNP block. The user specifies an approximate number of long-range links to be retained for downstream analysis.

Modelling short-range links.

By default, SpydrPick (12) uses $S = 10$ kb as the threshold for defining a short-range link, but in LDWeaver we chose $S = 20$ kb as the default threshold to better capture the region of rapid LD decay. The user can adjust this parameter as required.

Additionally, LDWeaver accounts for genome-wide variation in local LD patterns, which can arise due to varying mutation and/or recombination rates, by introducing a clustering and segregation step. First, for each coding sequence (CDS) segment in the annotations file, the per-site number of mismatches between the CDS and its reference sequence (i.e. the per-site Hamming distance) is computed. Next, the CDS are clustered using the k-means algorithm (84). Due to the challenges in accounting for heterogeneity stemming from population structure (80), LDWeaver avoids estimating this parameter and the choice of the number of clusters is a user modifiable (default $k = 3$). A user can avoid clustering by setting the $k = 1$. Generally, the LDWeaver output with CDS diversity and clustering (similar to Figure 1B) can be useful to determine k. In Supplementary Figures S2, S3, S6, S8 and S9, for each dataset analysed, we show our choice of k and the output CDS diversity and clustering plot. Generally, increasing k beyond a sensible value will have a limited impact on final analysis, provided that enough CDS remain in all clusters to approximately estimate the decay of background LD.

Since CDS regions exclude intergenic regions, each intergenic SNP is manually merged into the cluster closest to its genomic location.

We used the *C. jejuni* dataset to compare the modelling of short-range rapid LD decay between (1) the LDWeaver model-free approach and (2) computationally demanding neutral simulations. For each cluster, we used mcorr (85) to estimate the parameter triplet: mutation rate, recombination rate and the recombination tract length. For each parameter triplet, 100 neutral replicates were generated using bacmeta (86). In each simulation 20 subpopulations of bacteria, each comprising 1000 individuals with a 200 kb gen-me were included to generate a sufficiently large and diverse alignment. Migration rate probability was set at the default 0.01. Each simulation was performed for 20 000 generations to ensure convergence and then 1000 individuals were sampled. The mean LD decay was directly estimated from this neutral data as a function of bp-sep (base-pair separation).

LDWeaver directly models the LD-decay using the genomic alignment, separately for each cluster. At each discrete bp-sep, the 95th percentile (q_{95}) MI value is extracted from the genomic data. Here, q_{95} was chosen intuitively as a reasonable choice to model the background LD (87). Next, the linear model: $\log(q_{95}) \sim \log(\text{bp\_sep})$ is fitted and the exponent of the fitted values of this model (i.e. $\hat{q}_{95}$) is chosen as the bp-sep dependent short-range background-LD threshold. The 95th percentile is chosen to be modelled because it is close to the upper tail of the distribution, which is the region of interest, but little affected by large outliers so that the fitted curve is reasonably smooth.

Outlier calling and link ranking.

Outlier calling is performed at each discrete bp-sep value. Let $d \in [1, S]$ denote the bp-sep of interest, let $L(d)$ denote all the links that are d bp-sep apart and let $L^*(d)$ denote the subset with $\text{MI}(L(d)) \geq \hat{q}_{95}(d)$. First, the model: $\text{MI}(L(d)) - \hat{q}_{95}(d) \sim Beta(\alpha, \beta)$ is fitted (88) in order to compute an approximate, $\forall L^*(d)$, a short-range p-value (referred to as 'srp' in LDWeaver to denote short-range p-value). Links with $\text{MI} < \hat{q}_{95}(d)$ are always discarded and the user can choose a srp cut-off value (default $p = 1e-3$) to further reduce the set of links retained.

Although a model-free, permutation analysis is available to compute the srp, it would greatly add to the computational burden for large bacterial genomic datasets and is not required because the Beta distribution provides a good approximation. We confirmed this empirically by performing multiple trials comparing permutation-based and beta-approximation p-values using the descdist() function available in the R package, fitdistrplus (88).

The $\hat{q}_{95}$ values are modelled separately for each genome cluster. The srp for links between sites from the same cluster is computed using the LD-decay model fitted to data from that cluster. When a link spans two clusters, the maximum of the two srp values is used.

Filtering indirect links

Because of its success in inferring gene expression networks (89) and its utility in SpydrPick, we added ARACNE as a filtering step to the LDWeaver pipeline to overcome the inability of pairwise methods to distinguish 'direct' associations. For a dataset with 100 000 SNPs, MI values will be computed for approximately 5 billion unique links. When performing GWES, the interest is typically focussed on links with high MI values (i.e. with larger than normal linkage). However, many of these links will be driven by the same underlying association. Identifying the causal link is extremely challenging, especially for bacterial data due to factors such as clonality and genome plasticity (80).

Given multiple links that can be explained by the same underlying causal effect, we use ARACNE to retain only the link with the strongest signal (called a 'direct' link). This can reduce the number of links retained by several orders of magnitude, greatly reducing the manual curation task. ARACNE scans the entire LD landscape and a link (X, Y) is considered to be indirect if $\text{MI}(X, Y) \leq [\text{MI}(X, Z), \text{MI}(Z, Y)]$ for any SNP Z. A detailed explanation of the ARACNE algorithm and this specific filtering step is available in (12).

Detecting outliers in long-range links

Long-range (bp-sep $> S$) background LD varies little with bp-sep, so the genome clustering step that was used to model short-range links is not required. To determine a background-LD threshold (12), we adopt an approach similar to that of SpydrPick in which LDWeaver computes the Tukey (90) outlier threshold: $T_1 = Q_1 + 1.5 \times IQR$, where $IQR = Q_1 - Q_3$ and Q_1 and Q_3 are the first and third quartiles, respectively. Links with $MI > T_1$ are directly ranked based on the MI value. In cases where < 5000 links surpass T_1 , LDWeaver retains the top 5000 (default) links based on MI value.

The Tukey outlier detection approach is simple, but we prefer it to a permutation approach for two reasons. Firstly, this threshold has limited impact on the long-range GWES

analysis itself. While an outlier is defined based on whether its MI value passes this threshold, it has no bearing on link ranking itself. Downstream analysis and the eventual manual curation can be performed on the highest-ranked subset of ARACNE direct links without requiring a threshold. Secondly, due to the high degree of LD observed in bacteria, the background MI values computed from the observed MSA could be larger than the null MI values computed from a label-shuffled model. Therefore, the computed permutation threshold could potentially be too conservative.

While LDWeaver can perform GWES analysis on both short- and long-range links, it is fully compatible with the SpydrPick output for long-range link analysis. For a dataset that has already been analysed using SpydrPick, users have the option to first perform only the short-range analysis in LDWeaver, then present the SpydrPick output as input to LDWeaver to perform the downstream analyses of long-range links.

Downstream analyses and visualizing putative links and sites.

We have integrated several powerful R visualization tools into LDWeaver (91,92). To prioritize and perform wet-lab based modelling and validation, it is helpful to understand the functional annotations of the SNPs involved in high-MI links. Therefore, functional annotations are added to all such SNPs using SnpEff (29). Additionally, LDWeaver classifies each SNP as non-synonymous, synonymous, or intergenic. The non-synonymous versus synonymous distinction is based on the most common allele at multi-allelic sites. Based on this classification, LDWeaver generates an additional output comprising, by default, a set of 250 top ranked links excluding links between two synonymous sites.

GWES Manhattan plots were introduced to visualize the distribution of MI as a function of bp-sep (12). LDWeaver generates the long-range GWES plot introduced in SpydrPick, along with two plots for short-range links. The first short-range plot is similar to the long-range GWES plot but with points shaded according to the srp value. The second plot shows the segregation of links into genomic clusters. While these plots are less informative compared to Manhattan plots in genome-wide association studies (GWAS) due to the lack of genomic positions, they can be useful to assess the LD-decay fit between $\hat{q}_{95}$ and the genomic data.

LDWeaver also generates a linear tanglegram using the R package ChromoMap (93) to indicate the genomic positions of top ranked links in the short range. These tanglegrams span the whole genome and are broken down into segments for improved utility. Additionally, LDWeaver generates a figure depicting a genome-wide overview of the LD structure in the dataset. First, a sparse, SNP-level LD matrix is created using the MI values of the saved link data. Then, an averaging kernel is applied to reduce the dimensions of the matrix to approximately 1000×1000 (94). The resulting matrix is plotted in the form of a heatmap with SNP positions as x and y labels (95). This bird's-eye view can be useful in some analyses to quickly identify high-LD regions and blocks.

Additionally, LDWeaver generates a network plot for top ranked links using the R package ggraph (96). The gene-level regions of each site are extracted from the annotations file and the links between these regions are coerced into a network (97). To reduce clutter in the network plot, edges with only 1 link between nodes are removed. Afterwards, any nodes with no edges between them are also dropped from the plot. The edges are coloured to depict the number of links between genes. Furthermore, the edge width and transparency are also moderated to reflect the MI value of the highest ranked link between the two regions. Furthermore, LDWeaver provides the option to generate a similar gene network for any chosen gene. The user can decide whether to use short-range, long-range or both link lists to generate this plot.

Using a user-provided phylogeny, LDWeaver can generate a tree-plot of user-determined putative sites and phenotypes. This type of plot is inspired by some of the visualization options available in Microreact (98) and has also been widely used in the previous GWES literature (11,15). The phylogeny will be midpoint rooted by default (99,100) and the SNP data, user provided phenotypes are sorted in the same order as the phylogeny. Finally, the plot is generated using the R package ggtree (101).

Finally, LDWeaver generates the output required to dynamically visualize links using the R package GWESExplorer (https://github.com/jurikuronen/GWES-Explorer). This Node.js based shiny app can be used to generate the GWES Manhattan plot, circular tanglegram and the tree-plot for an arbitrarily chosen subset of putative links.

Runtime

A complete LDWeaver analysis with default parameters on a dataset comprising 2000 sequences with 80 000 SNPs on average requires 4756 s (~80 min) on a computer with 32GB of ram and 10 parallel CPU cores running R version 4.2.2 with openBLAS v0.3.21 support.

Results

Overview of LDWeaver

Performing GWES analysis using LDWeaver requires two inputs, a multiple sequence alignment (MSA) and the annotation file of the reference in Genbank or Gff3 format. Initially, LDWeaver filters out sites with low minor allele frequency (default: 0.01) and high gap frequency (default: 0.15), with an option for a 'relaxed' filter to retain sites with gap (N) as the second most common allele. Following SpydrPick (12), LD between each SNP pair is measured using mutual information (MI). To address population structure, a sequence-reweighting approach is applied, where the weight for each sequence is computed as the reciprocal of the number of sequences with a mean per-site Hamming distance below a user definable threshold (default: 0.1).

Generally, SNPs in genomic proximity tend to have very high LD, but LD levels rapidly decline with base-pair distance to a constant value for all long-range SNP pairs (see Figure 1A). First, LDWeaver uses a user-definable genomic distance threshold to classify short range links (default: links between sites < 20 kb apart are considered short range). To determine a threshold for outlier calling, it is necessary to model the decay in LD with genomic distance and the shape of this decay is determined by many factors. These may include the type of species, population structure, local mutation and recombination rates, variation in gene content and selection pressures. To account for this heterogeneity, LDWeaver measures the per-site mean Hamming distance within coding regions around the chromosome and clusters them using k-means based on

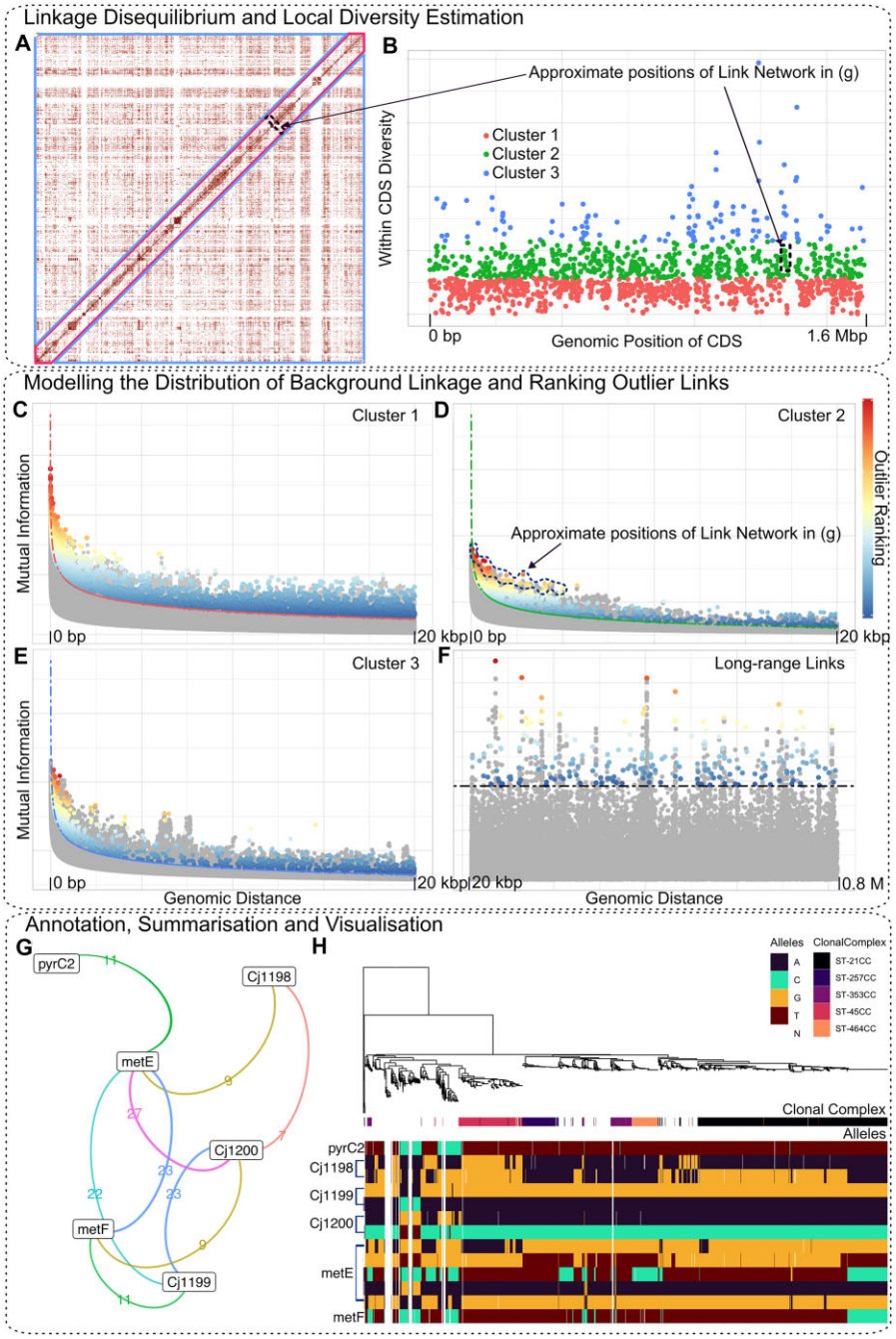

Figure 1. Overview of the LDWeaver pipeline. (**A**) Genome-wide linkage disequilibrium (LD) for 1480 *Campylobacter jejuni* isolates (26). LD is measured using mutual information (MI) and a weighting strategy is employed to account for population structure. The axes correspond to genomic positions in the NCTC 11168 reference genome, and the colour intensity reflects the strength of LD. Blue triangles outline regions of long-range LD, while the short-range high-LD region is outlined in red. (**B**) Genomic diversity (measured using Hamming distance compared to the reference) within each coding region (CDS) is used to account for local variation in short-range background LD. Each point corresponds to a CDS, and the vertical axis shows the average number of sites that differ from the reference. K-means clustering is used to divide the CDS into three clusters (see legend). Sites from intergenic regions are allocated to the nearest cluster. (**C–F**) GWES Manhattan plots show the distribution of LD measured using weighted MI (*y*-axis) at varying genomic distances (*x*-axis). For better visualisation in this overview figure, numeric values are removed from the *y*-axis. All links (between synonymous, non-synonymous and intergenic sites) are included in these plots. Plots are shown for short-range links in the three clusters (C–E) and long-range links (F). Modelled background LD is shown respectively using red, green, blue, and black dashed lines, respectively. The colour shading of each point indicates the ranking given to outlier links (see the colour bar in the rightmost panel - numeric values removed to reduce clutter). The topmost (bright red) colour signifies rank 1, highlighting the most extreme outlier. Decreasing ranks follow the colour bar from top-bottom. For short- and long-range analyses, link ranking is based on the estimated short-range *P*-value and the measured MI, respectively (see Materials and methods). Links that are either inferred as indirect or with MI below the background LD are shown in grey and not ranked. In (F) the background LD is invariant to genomic distance and computed using the Tukey criteria (dashed black line). (**G**) The LDWeaver network plot generated for *metE* summarises all the outlier links involving a site in the gene. Here, the edges are coloured based on the number of links between linked genomic region nodes (see *Campylobacter* results section). (**H**) Investigating the allele distribution within the linked region (alleles panel) shows that several deletions and mutations observed within several clonal complexes are driving the high LD signal picked up by LDWeaver. This and the subsequent phylogenetic trees shown in this manuscript were generated using FastTree 2 (102).

the estimated local diversity (see Figure 1B). Based on the output in Figure 1B (generated by LDWeaver), the user has the option to select the most appropriate number of clusters for the dataset (default: 3).

Background decay in LD is modelled separately for each cluster. Here, a linear model is fitted to log transformed 95th percentile MI values and the corresponding base pair separation. Afterwards, the exponent of fitted values is used as the base pair separation dependent background LD (see Figure 1C–E). Since long-range background LD is uniform (see Figure 1F), the Tukey outlier approach in SpydrPick can be used to estimate the background LD.

After calling outliers based on the estimated background LD, LDWeaver provides a list of locus pairs ranked in order of strength of evidence for co-selection. This ranking is based on the outlier short-range *P*-value (see Materials and Methods). LDWeaver includes several options to ease the task of generating a list of potential epistatic SNP-pairs for expert manual curation and wet lab validation. First, all outlier links are annotated using SnpEff (29) and links that include a non-synonymous substitution are given a higher priority. Afterwards, this SNP classification is leveraged to generate an additional output comprising a set of (default: 250) top-ranked links after discarding links between two synonymous sites. All short-range results presented in this manuscript using real data are based on analysing these top 250 most significant links. Furthermore, LDWeaver summarizes links into networks (see Figure 1G), which helps to prioritize the most promising genome regions. Finally, LDWeaver can be used to visualize the allele distributions within these networks (see Figure 1H).

LDWeaver detects co-selection signals in simulated genomes.

With the detection of long-range links validated previously (12), the primary objective here is to detect links short-range under epistasis by looking for signs of co-selection between SNPs. We validated LDWeaver using Wright-Fisher models representing several evolutionary scenarios that were simulated using SLiM version 4 (30). Although the original design of SLiM does not support simulating bacterial evolution, recent advances (31) now enable considering circular genomes, horizontal gene transfer, and bacterial recombination. Notably, SLiM is one of the only available options with the ability to simulate epistasis in bacterial populations and generate full genome alignments as output, which is necessary for LDWeaver analysis.

To make the simulations as realistic as possible, we chose the first 200 kb from the ATCC 700669 (*S. pneumoniae*) reference genome as the ancestral sequence. Each simulation comprised 10 000 isolates, equally distributed across 10 subpopulations. Recombination tract lengths were drawn from a geometric distribution with a mean of 500 bp, while mutation and recombination rates were fixed at 2e-7 (per bp, per generation) and 8e-7 (per bp, per generation), respectively. The simulations allowed migration between all 10 subpopulations at a rate of 5e-2 (per bp, per generation).

Introducing positive or negative *synergistic* epistasis alone led to a loss of genetic diversity through fixations after several generations, which stands in contrast with observations in real bacterial populations (32). While such epistatic interactions are undoubtedly present in bacterial populations, these simulations lack the complexity needed to model the processes that continuously shape the LD within populations. To address this and to maintain more realistic levels of genetic diversity, we opted to simulate a balancing selection scenario employing a version of negative multiplicative (synergistic) epistasis.

We identified the 22 longest coding regions (CDS) from within the first 200kb of the ATCC700669 reference genome, each with length >1500 bp, as *potential* 'target regions' for epistatic interactions. The distribution of these 22 regions on the genome is illustrated in Figure 2A. Initially, each CDS was randomly allocated to one of five groups represented here using different colours: magenta, green, blue, purple and orange (see Figure 2). Utilizing these groups, we constructed five simulation scenarios, denoted as **s1**-**s5**. In **s1**, all 22 CDS from the five groups were selected as *targets*, covering a combined target region length of 23.8% of the total genome. For **s2**, four groups were chosen, resulting in 18 target CDS covering 20.8% of the genome. Similarly, **s3** involved three groups with 14 CDS with 17.5% coverage, **s4** involved two groups with 10 CDS with 13.7% coverage, and **s5** involved only one group with six CDS and 9.5% coverage. This design ensures that **s5** is considerably more challenging than **s1** because only two orange regions near 70 Kb (see Figure 2A) can contribute to 'between target-region' short-range links. The remaining target links in **s5** are from 'within target-regions'.

For all simulation scenarios, three types of mutations were introduced to the population. All regions *outside* the target CDS were assigned mutation type **m1**, which had a *slightly* deleterious selection coefficient of −0.00001, reflecting the assumed near-neutrality in bacterial genomes. The target CDS were randomly allocated either mutation type **m2** or **m3**, both were beneficial with an additive selection coefficient of 0.001. Since both mutation types had the same effect, in the absence of epistasis, they would fix after several generations. An epistatic interaction was introduced between **m2** and **m3**; carrying both will result in a 5% reduction in fitness.

The simulation was continued for 20000 generations and 1000 genotypes were sampled at the end. Each scenario (**s1–s5**) was replicated 100 times, totalling 500 simulations. At the end of each replicate, the extracted genome alignment was used for LDWeaver analysis. Importantly, since this simulation does not account for the complexities of amino acid modifications, SnpEff annotations were avoided, and no distinctions were made between synonymous and non-synonymous mutations.

In **s1**, only 2.4% of examined links were true target links (i.e. a link from within or between two target CDS regions), which reduced to 1.3% for **s5** (see Supplementary Table S1). Since LDWeaver is primarily a link ranking algorithm, its performance was evaluated based on its ability to include target links in the top subset of ranked links. This best aligns with the analysis method used for real biological data presented in the manuscript.

For scenarios **s1– s5**, 70.4%, 62.6%, 61.8%, 52.4% and 49.8% of *top 5* ranked links were target links, respectively. Next, examining the *top 20* ranked links revealed that at least one target link appeared in all 400 replicates for **s1– s4**, and in 97 of 100 **s5** replicates (see Figure 2B). Furthermore, comparing the *top 25* link ranking to a random allocation revealed that LDWeaver performs approx. 30 times better across all scenarios (see Supplementary Figure S1).

Finally, to assess the possibility of a target link being tagged by an alternative site in genomic proximity due to LD, we cal-

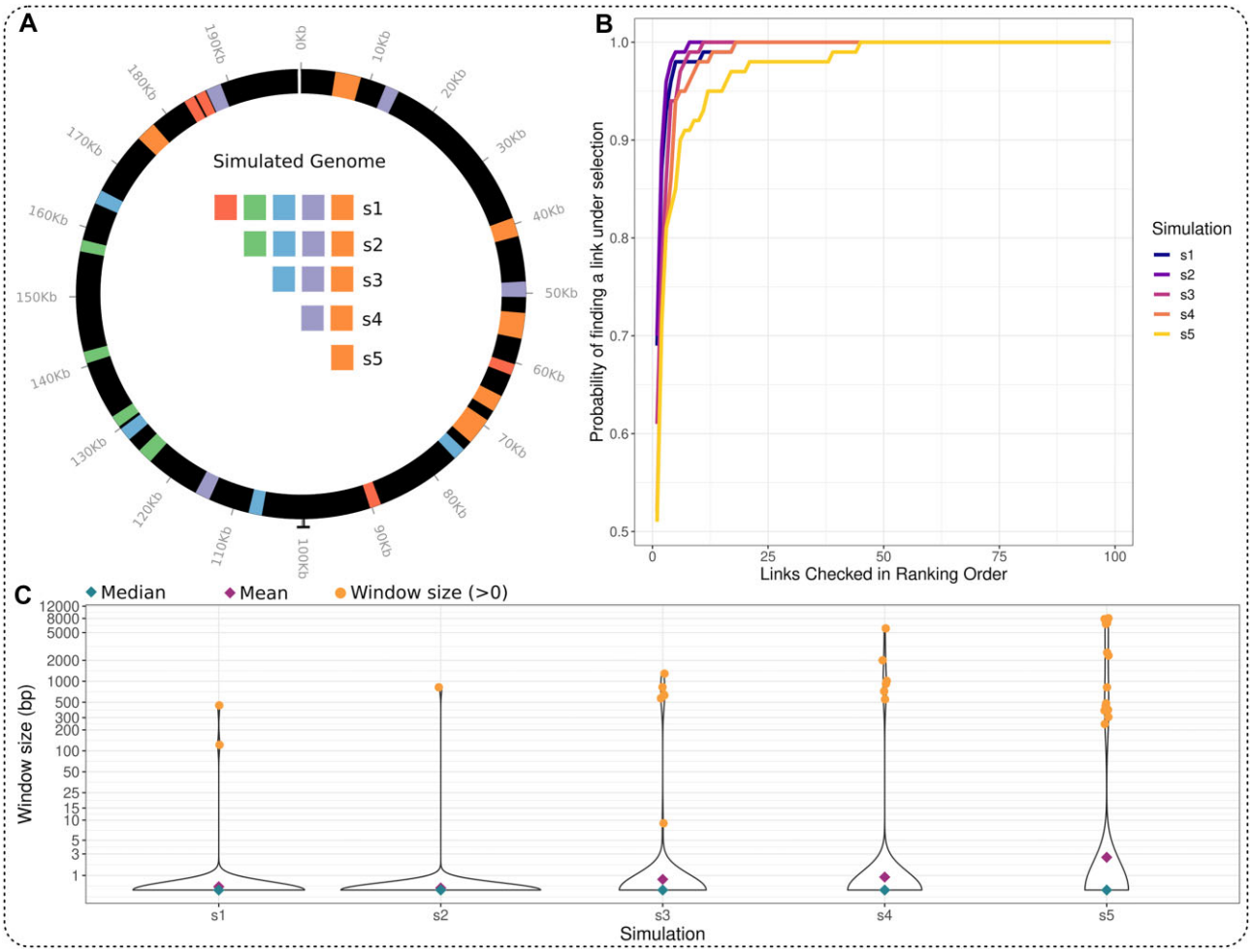

Figure 2. Validation of LDWeaver using bacterial population simulations. (**A**) Simulated genomic region (first 200kb of ATCC700669) showing the target Coding Sequences (CDS). In each simulation scenario, grouped target CDS (depicted by colours) were chosen for epistatic interactions. In **s1**, all 22 CDS selected and in **s5**, only the 6 CDS in the orange group were selected (see main text for a detailed breakdown). (**B**) For each simulation (see legend), each curve shows the variation between the number of links checked in ranking order (*x*-axis) and the probability of detecting a 'target link', a link from within or between two target regions (*y*-axis). This reaches 1 when all replicates contain a target link. All replicates from **s1–s4** and 97% of replicates from the most challenging **s5** contain a target link within the *top 20 ranked*. The challenge in **s5** is 2-fold: only < 10% of the genome is under epistasis, and only two orange regions near position 70 Kb (panel a) can contribute to 'between target-region' short-range links. (**C**) For each simulation scenario (*x*-axis), *y*-axis shows the genomic distance between sites in the top 5 ranked links and the closest target region. Most replicates in **s1–s5** contain at least one target link in the *top 5 ranked* (blue diamond shows median = 0, purple diamond shows mean < 3, which increases from **s1** to **s5**). Each orange dot corresponds to a replicate that does not contain a target link, and the *y*-axis shows the minimum genomic distance from a target region to a site. To elaborate, two replicates in s1 do not contain a target region link in the *top 5 ranked*, and the closest site in each case is approx. 100 and 500 bp to a target region.

culated the genomic distance between all 10 detected sites and their closest target regions for the top 5 ranked links in each replicate (see Figure 2C). In replicates containing a target link within the top 5 ranked, this distance is 0. For others, it represents the distance to the nearest target region for the site that is in the closest genomic proximity to a target region. The analysis revealed that 94.2% of replicates contain a target link in the top 5 ranked, as indicated by a median distance of 0.

LDWeaver detects co-evolutionary links in multiple pathways in *S. pneumoniae.*

S. pneumoniae is a naturally transformable nasopharyngeal commensal and respiratory pathogen. It causes a substantial global disease burden in humans, representing a major cause of pneumonia, meningitis, and otitis media. There are >100 immunologically distinct capsule types, termed serotypes, of *S. pneumoniae*. Serotype replacement after the introduction of pneumococcal conjugate vaccines (PCVs) is a serious concern, particularly given the association of many pneumococcal genotypes with multidrug resistance (33).

The LDWeaver analysis of pneumococci focussed on two populations isolated from carriage in contrasting settings: Mae La, Thailand and Massachusetts, USA. The Mae La sample comprised 2663 high quality assemblies (accessions available in Supplementary information) (32) collected from mother and infant pairs (34). The Massachusetts sample (25) comprised 616 draft genomes (accessions available in Supplementary information) of similar quality (32). Both datasets were aligned to the ATCC 700669 (accession code FM211187) reference genome (35), and after filtering out sites with minor allele frequency (MAF) < 0.01 and gap frequency > 0.15, respectively 88603 and 89386 SNPs were

retained for analysis (see Supplementary Figures S2 and S3 for LDWeaver plot panels). Analysed and detected link counts are summarised in Supplementary Table S2.

Long-range interactions included strong signals of co-evolution between pbp2b and pbp2x in both populations. These genes are key in determining resistance to beta lactam antibiotics and these interactions were reported previously (12). Another interaction conserved across both populations was that between three loci encoding immunogenic surface-exposed degradative enzymes (36): the beta-galactosidase BgaA, the immunoglobulin A protease ZmpA, and PabB, encoded by a gene directly downstream of that for the ZmpA paralogue, ZmpB. These co-evolutionary signals may arise through direct interactions on the surface, or indirect effects emerging through immune selection for particular combinations of antigens (37).

The short-range interactions were more similar between the two populations. Adjacent genes functioning in the same metabolic pathway included links consistently identified between the neighbouring coding sequences cdsH (SPN23F08030) and thiI (SPN23F08040), which are both involved in thiamine biosynthesis (38). Similarly, the adjacent genes SPN23F02450 and SPN23F02460 both encode proteins predicted to function as N-acetyltransferases. Signals were also identified in both populations between the overlapping genes SPN23F08250 and SPN23F08260, encoding subunits of a transporter of unknown function (39). Hence, LDweaver can identify signals of proteins likely to be co-evolving as participants in the same metabolic pathways.

Another regulatory locus involved in short range interactions across both datasets is that encoding the competence regulator TfoX. Multiple sites were in strong LD in the subset of isolates encoding this locus, which was absent in a minority of isolates (see Figure 3A). Yet these sites are not in the gene itself, but in the flanking intergenic regions, or the adjacent alaDH pseudogene that encodes an apparently functionally unrelated alanine dehydrogenase. This suggests that these paired sites do not interact functionally. A search for the functional sequence of the alaDH gene identified intact versions in other streptococci. Further alignments indicated that the tfoX-alaDH pairing was intact in *Streptococcus anginosu*s. Hence, the variable distribution of this locus in *S. pneumoniae* is likely the consequence of one, or more, interspecies transfers through homologous recombination introducing the gene cassette into pneumococci, followed by degradation of the alaDH gene into a non-functional form (40). Hence the elevated LD identified in this case may be the consequence of a relatively recent introgression (41) from another streptococcal species, demonstrating LDWeaver can identify loci under sufficiently strong selection to drive interspecies transfers (40).

Another transporter (encoded by SPN23F03500) was linked to an adjacent pseudogene (SPN23F03510) in both populations (see Supplementary Figure S4). The undisrupted sequence of SPN23F03510 is predicted to function as a lantibiotic synthesis protein, suggesting the associated transporter is likely to be a self-immunity protein. Hence, this pairing likely represents an example of a non-producing, immune-only 'cheater' bacteriocin phenotype (42), common in pneumococci (43). The most variable bacteriocin-encoding locus in the pneumococcal genome is the *blp* gene cluster (44), in which multiple interactions between genes were detected. These included interactions between the genes encoding the BlpRH quorum sensing two component system, and the gene encoding the cognate peptide pheromone, BlpC (45), identified in both populations. Further links were found in the Massachusetts sample only, which likely represent relationships between the synthesized bacteriocins and corresponding immunity proteins (44).

In addition to performing biological analyses, we used the Massachusetts dataset to compare the effect of varying the CDS clustering parameter (k). We repeated the LDWeaver analysis using $k = 1$, 3 and 5 and fitted the decay model to the 95th percentile of the empirical distribution. Fitted parameters for each case are shown in Supplementary Table S3. Each parameter pair corresponds to the slope and intercept of the fitted model (see Materials and methods). Corresponding decay curves are shown in Supplementary Figure S5a. Next, we examined the variation between the top 250 ranked links from each analysis. Links between the same coding regions were pooled together irrespective of the ranking and the counts are shown in Supplementary Figure S5(b). Here, the qualitative difference between the choice of $k = 1$ and $k = 3$ is clear, however, choosing between $k = 3$ and $k = 5$ only results in a marginal difference.

LDWeaver identifies patterns of co-evolution of cytolethal distending protein toxin subunits in *C. jejuni*

C. jejuni is a leading cause of food-borne bacterial gastroenteritis worldwide, associated with the consumption of contaminated poultry meat. It is well-adapted to colonize the gut of the majority of mammalian and avian host species and has been isolated from many environmental sources. The successful colonization of strains is often host-specific for the majority of lineages (specialist clonal complexes) and this is reflected in the population structure during phylogenetic comparison of genomes. Certain clonal complexes (CCs) are also known to be host generalists, where the same lineage is well-adapted to colonize and to survive in multiple different hosts.

The population of *C. jejuni* is structured by well-defined, genetically similar clusters of isolates (clonal complexes) which are documented to be maintained over time (26), despite the very frequent homologous recombination occurring across the known lineages (26,46). This high frequency of recombination, which is not limited to any particular region of the genome, would generally break down the LD observed in *C. jejuni* lineages and therefore, potentially disrupt co-selected/epistatic functional groups of genes throughout the genome. Despite the high recombination, the population structure has remained stable over long periods of time (26), making *C. jejuni* an excellent candidate species for the analysis of short- and long-range epistasis and co-selection links.

A collection of 1480 previously published *C. jejuni* genomes (26) consisting of 18 different human, animal, and environmental sources and 37 CCs were selected for LDWeaver analysis (accessions available in Supplementary information). These were aligned against the NCTC 11168 reference genome (47) using snippy. After removing sites with MAF < 0.01 and gap frequency > 0.15, 102591 SNPs were retained for LDWeaver analysis (see Supplementary Figure S6 for the LDWeaver plot panel).

Analysis of short-range interactions identified strong signals of co-evolution between genes with functions mostly related to virulence such as: amino acid ABC transporters, flagellar biosynthesis, periplasmic and outer membrane pro-

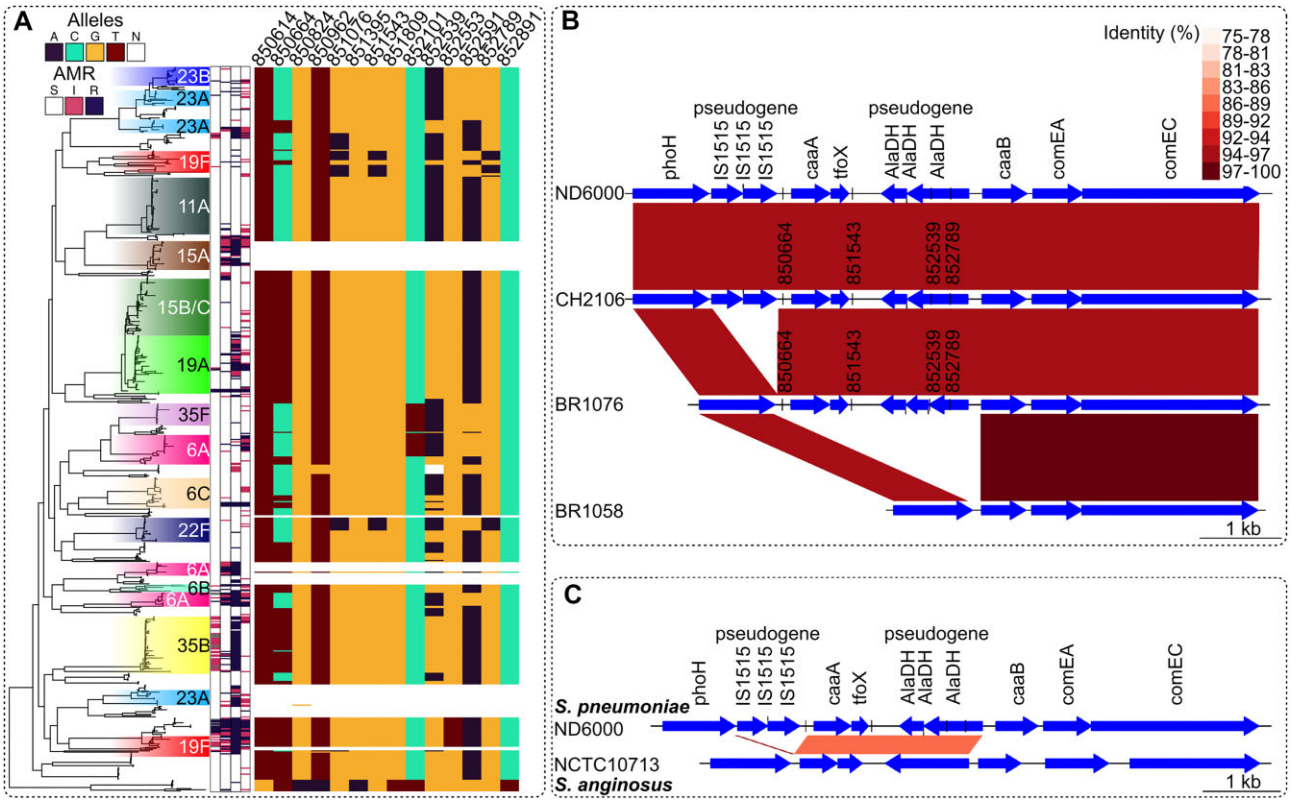

Figure 3. Overview of genomic variations of the flanking region of the TfoX competence regulator demonstrated using the Massachusetts S. pneumoniae dataset. (**A**) The phylogenetic tree ($n = 616$) is coloured according to the serotype shown to the right of the tree. The 4 bars immediately to the right denote the antimicrobial resistance data for ceftriaxone, erythromycin, benzyl penicillin and trimethoprim, respectively. The key above indicates the colour shadings for S – sensitive, I – intermediate and R – resistant. Allelic variation in the tfoX locus is shown by the rightmost heatmap and the key above shows the colour for each nucleotide with N indicating an ambiguous base. SNP positions above the heatmap are based on the ATCC 700669 reference. This suggests that the tfoX locus is present in most pneumococci and can be divided into three genotypes: that containing the major alleles at each polymorphic site; that containing the minor alleles at sites 850664 and 852539, and that containing the minor alleles at these two sites, as well as sites 851543 and 852789. (**B**) The alignment of representatives of the four observed genotypes in the dataset, demonstrated here using sample data from selected isolates. Colour shading indicates the identity between regions. In ND6000 (ERR129187) and CH2106 (ERR129095), the flanking region is intact with multiple insertion sequences, two functional genes, a non-coding region and an alaDH pseudogene. The locus is intact, but the insertion sequences are not observed in BR1076 (ERR129048). In contrast, the entire locus is missing in BR1058 (ERR129043). With reference to the phylogenetic tree in (A): ND6000 is a serotype 7C isolate between regions 19F and 11A, CH2106 is a 19F isolate, BR1076 is 6C isolate and BR1058 is a 7C isolate. (**C**) Shows the alignment between FM211187 and *S. anginosus* NCTC10713 genomes. For the region of interest, this is the most similar locus among streptococcal species. Despite the general divergence between *S. pneumoniae* and *S. anginosus*, the tfoX-alaDH gene pair is intact in both species with high similarity (same key as in (B) to show region identity). However, the dissimilarity between the flanking regions demonstrates the typical level of divergence between the genomes. Hence, the localized similarity indicates a possible recent introgression into the pneumococcus.

teins, antibiotic efflux genes, among others. One of the most promising findings was multiple highly significant links located within the cytolethal distending toxin (CDT) genomic region. The CDT is a protein toxin composed of three subunits: CdtA, CdtB and CdtC encoded by a cdtABC operon, and is considered one of the most important virulence factors for Campylobacter pathogenesis. CDT acts to halt host cell division by cell cycle arrest at the G_2 stage occurring before mitosis (48). CdtA and CdtC are anchored into the membrane and act to deliver CdtB to the host cell which arrests the cell cycle. CdtB has the toxin activity but is reliant on CdtA and CdtC for its binding and delivery to the host cell (49).

Our results showed that the three subunits were well conserved throughout the dataset, except for three distinct clusters of isolates identified in wild birds (see Figure 4). There was high variability of the presence/absence and allelic variation of the three CDT subunits across these three clusters of isolates. For example, all three subunits were absent from the wild bird-associated ST-1287 CC. Another wild bird-associated clade, ST-1034 CC (mixed) consisted of synonymous/non-synonymous nucleotide changes in all subunits compared with other sources. Finally, isolates belonging to the third wild bird-associated clade, ST-1325 CC consisted of a mix of both the same synonymous/non-synonymous nucleotide changes as seen in ST-1034 CC, some of the isolates in this cluster also exhibited the same variation as observed in the other sources and CCs within the dataset, while some isolates had an absence of various CDT subunits (see Figure 4). A recent study comparing the gene sequences of the *cdtABC* operon of wild bird, broiler chicken and human sources (50) confirms the LDWeaver-generated hypothesis of significant co-selection occurring within this operon. The study identified high variability of *cdtABC* alleles in wild birds with several alleles producing no functional CDT. Sequence conservation outside of wild bird sources, such as broiler chickens and humans, was also observed, suggesting that the variation of the *cdtABC* operon may play a role in the host range of *Campylobacter* (50).

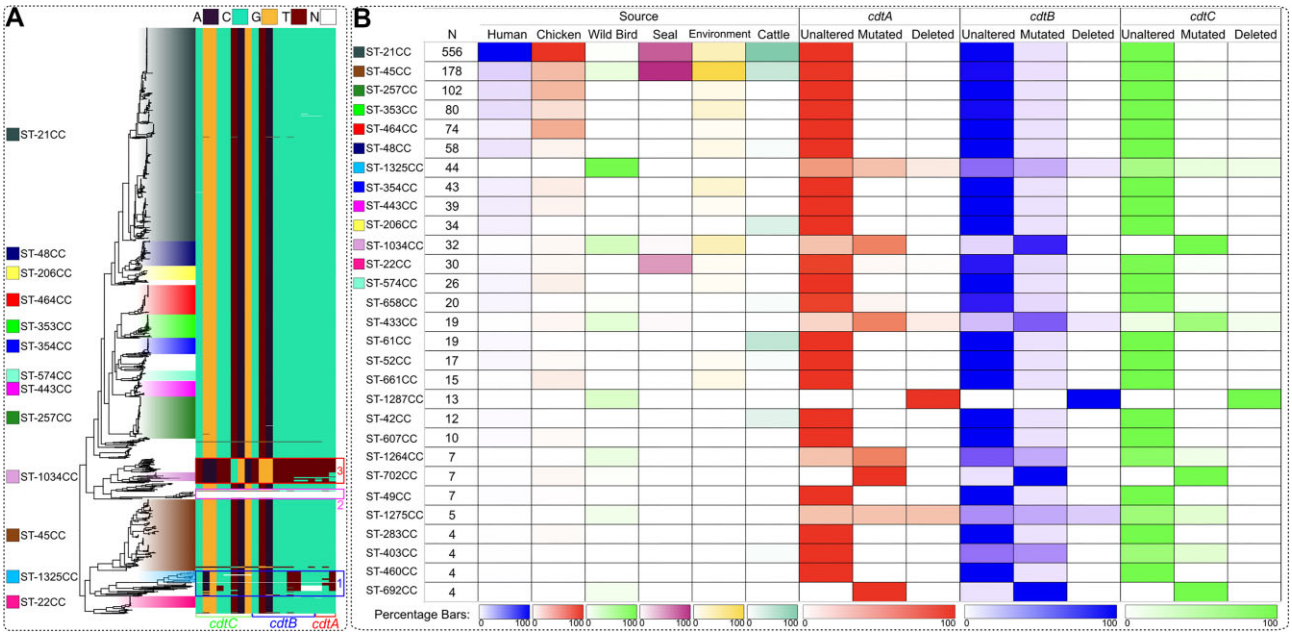

Figure 4. Overview of allelic variation of CDT subunits in the *C. jejuni* dataset. (**A**) The phylogenetic tree ($n = 1480$) is coloured according to the clonal complex and the key to the left of the tree is labelled according to the corresponding topology of each clonal complex. Allelic variation of the three CDT subunits (CdtA, CdtB and CdtC (*x axis*)) encoded by the *cdtABC* operon across the tree is represented by the heatmap to the right of the tree. (**B**) The association of a particular clonal complex with a source is highlighted by the first panel. The darker the shading (explained by percentage keys at the bottom of the panel), the higher the percentage of isolates from each source belonging to the particular clonal complex (rows). The rows are ordered by abundance (column *N*) in the *C. jejuni* dataset. Three panels to the right represent the allelic variation within CdtA (red), CdtB (blue) and CdtC (green) subunits. The shading represents the percentage of isolates associated with a particular clonal complex/source with either: 1) the unaltered allele; 2) a nucleotide change (mutated); or, 3) a deletion of the CDT subunit. Shaded keys at the bottom of the three panels represent the percentage of isolates.

Potential co-selection was also identified in the gene cluster containing genes *metE* and *metF* (see Figure 1). These genes are located on an operon involved in methionine synthesis which has been proven to have a vital role in the colonisation of *C. jejuni* to the gastrointestinal tract of different hosts (Kelley *et al.*, 2021). Similar patterns of the presence/absence and mutated versions of this locus were observed within the same wild bird clusters as the *cdtABC* operon (Figures 1H and 3) while also remaining relatively conserved in the remaining sources and CCs.

In addition to performing biological analyses, in *C. jejuni*, we also explored the impact of the choice of background LD modelling (LDWeaver approximate vs. using a neutral simulation) on short-range outlier ranking (Supplementary Figure S7). Link ranking was robust to the modelling choice for generally high LD links (MI > 0.5) (Supplementary Figure S7a), however, the LDWeaver model allocates systematically higher ranks to links that are further apart (Supplementary Figure S7b). Given the primary goal of short-range epistasis analysis is to accurately rank high LD outliers in genomic proximity, these findings indicate that the choice of background LD modelling has a minimal impact. Furthermore, both approaches on average allocate the same score to links between same site pairs (Supplementary Figure S7c).

LDWeaver recapitulates co-evolving links involved in clade evolution and success of the *E. coli* pandemic clone ST131

To investigate epistasis and co-selection in *E. coli*, we considered a dataset consisting of 2156 ST131 genome assemblies (accessions available in Supplementary information) aligned against the EC958 reference genome (51) using Snippy. Loci with MAF ≥0.01 and gap frequency <0.15 were included, leading to 44 092 SNPs (see Supplementary Figure S8 for the LDWeaver plot panel). It is noted that the *E. coli* dataset had the highest amount of LD among all analysed datasets, and the asymptote of Supplementary Figure S8b cluster 1 has the largest intercept among all LDWeaver panel plots.

The *E. coli* ST131 lineage belongs to the phylogroup B2 that emerged globally around 20 years ago and is associated with urinary tract (UTI) and bloodstream infections (BSI) (52–54). Large epidemiological studies have identified major differences in the virulence and antibiotic resistance between the three main clades of ST131 (A, B and C) (55). Clade C, which is associated with fluoroquinolone resistance arising from mutations in *gyrA* and *parC* has further been split into two sub-lineages (C1 and C2) with a distinct pattern of mobile genetic elements (MGEs) and associated antimicrobial resistance (AMR) genes (56,57).

The long-range loci pairs in *E. coli* ST131 corresponded to clade C specific SNPs differentiating sub-lineages C1 and C2 (58). These included links between sites in *sbmA* (EC958_0513), a transporter involved in the internalisation of peptide antibiotics into the cytoplasm (59) and identified as a virulence factor in avian extraintestinal *E. coli* (APEC) (60), and sites in (i) *nikA* (EC958_3870), a periplasmic protein from the ATP-binding cassette type nickel transport system acting as the initial receptor of nickel (61), (ii) *acrF* (EC958_4822) encoding an efflux pump with homology to the major pump AcrB (62), (iii) *lepA* (EC958_2875) encoding a conserved GTPase with a role in the initiation phase of translation (63) and (iv) *iscS* (EC958_2841), encoding a cysteine desulfurase implicated in the activity of a number of

Fe-S proteins (64). These clade-specific SNP pairs are spread across the *E. coli* chromosome, separated by at least 1 Mbp, and may have contributed to the recent expansion of the *E. coli* sub-lineage C2.

The genome wide distribution of LD estimated from LDWeaver revealed multiple interesting patterns (see Figure 5). For example, the short-range co-evolving SNP pairs highly ranked by LDWeaver correspond to regions involved in the synthesis of the *E. coli* capsule polysaccharide (*kpsM* EC958_3343, *kpsC* EC958_3337, *kpsS* EC958_3338) and type II secretion system located downstream in the chromosome (*gspL* EC958_3345, *gspM* EC958_3344). The observed tight linkage in the *E. coli* capsular region might be critical for having a functional system since the capsule plays an important role as a major virulence factor contributing to

the colonization of different eukaryotic host niches, reducing the efficacy of the immune system by complement inactivation and shaping the horizontal gene transfer mediated by MGEs (65–67). These results indicate that the variation within the capsule region is also

linked to SNPs present in the conserved type II secretion system. Variation in these regions could thus alter the capsule expression in *E. coli*, contributing to a non-capsulated state that allows the introduction of a new pool of MGEs (67).

LDWeaver detects novel interactions between sites associated with virulence in *E. faecalis*.

E. faecalis represents a classical generalist microorganism, with little phylogenetic divergence and limited host specialization over the population (68). Few hospital-associated *E. faecalis* clusters have been identified, with some overrepresentation of virulence factors and AMR genes (28,69). However, these lineages as well as the traits potentially underlying their success predate the modern hospital settings (28,70). The *E. faecalis* dataset analysed using LDWeaver comprised 2027 isolates (accessions available in Supplementary information) aligned against the V583 reference genome (71) using Snippy. After filtering out sites with MAF < 0.01 and gap frequency > 0.15, we analysed 85982 SNPs (see Supplementary Figure S9 for the LDWeaver plot panel).

Inspecting the short-range links within the *E. faecalis* population revealed multiple top-ranking links in known enterococcal virulence genes. Particularly, the elr operon was represented in several links (see Supplementary Figure S10). The genes in the *elr* operon code for putative surface proteins and their overexpression have been associated with increased virulence and ability to evade host immune defence by resistance to phagocytosis (72). Three *elr* genes, *elrB, elrC* and *elrR*, were linked to each other, and the positive regulator *elrE* (73) was also linked to an ATP-binding cassette transporter.

Another cluster of links involved *ace*, a widespread virulence gene in *E. faecalis*, coding for adherence factor (74,75). In addition to a gene coding for a hypothetical protein, it was linked to a bacteriocin-encoding *entV*. This bacteriocin has been shown to act against fungal *Candida albicans* co-infection by the inhibition of biofilm formation (76,77). Intriguingly, the absence of ace has also been shown to result in reduced biofilm formation *in vivo*, but despite both *ace* and *entV* being partly involved in biofilm-associated enterococcal infections (78), we are not aware of any report of a direct link between the functions of the two. These findings demonstrate the ability of LDWeaver to highlight both known and putative functional links between enterococcal genes. Supplementary Figure S10 further illustrates that the minor allele haplotypes at the candidate sites under co-selection are not enriched in hospital-associated multi-drug resistant lineages. Given that the ages of these lineages have been estimated as 50–150 years (28), and that they all share the major alleles at these variable sites, the co-selective pressure may have acted more recently in some other ecological setting outside hospitals.

Discussion

Bacterial population genomics research is rapidly moving towards an era where hundreds of thousands of whole-genome sequences will be available for many species. These data represent an unprecedented opportunity to seek signals of selection in natural populations and to unravel genomic clues for adaptation under changing ecology, which can contribute towards improved understanding about evolution, dissemination and maintenance of antibiotic resistance and virulence traits. Existing GWES methodology has already enabled discoveries of co-selection affecting polymorphisms across distant genomic sites in a variety of human pathogens (13–15). By extending GWES to joint screening of polymorphisms in close proximity, we increase the potential of data-driven molecular discovery for bacterial populations, which are particularly challenging for GWAS due to the difficulty of large-scale measurement of traits.

Applying our methodology across a diverse spectrum of bacterial species: *S. pneumoniae*, *C. jejuni*, *E. coli* and *E. faecalis*, has revealed fresh insights into genomic interactions between proximate variants. Notably, our results were obtained without use of phenotype data and would not be detected via conventional genome-wide association (GWAS) methods as they would typically be discarded due to the influence of short-range linkage disequilibrium. Our results included findings associated with host range, antibiotic resistance, virulence, and immune evasion. Furthermore, our top results often represented links from genomic islands: *tfoX* in *S. pneumoniae*, *cdtABC* in *C. jejuni*, capsular locus in *E. coli* and *elr* genes in *E. faecalis*, all of which are self-contained units that may evolve differently to the rest of the chromosome due to reduced functional integration. While these novel potential interactions require experimental validation, our approach drastically reduces the number of pairwise relationships that need to be considered. Recent advances in computational protein structure prediction could further reduce the need for time-consuming wet-lab experiments by rapidly considering the impacts of the identified intragenic interactions on the resulting protein structure.

A potential target for further development of genome-wide short-range LD analysis is to consider genomic variation beyond reference-based core genome alignments. With the increasing availability of long-read based assemblies of chromosomes and plasmids, it would be attractive to consider detection of co-selection both within plasmids and between plasmid and host chromosome polymorphisms, to potentially uncover either compensatory evolution or pre-adaptation to stable carriage of particular plasmids (79).

Furthermore, while wet-lab-based validation will remain the gold standard to verify combined effects of mutations, future developments should incorporate information beyond the DNA sequence, such as gene expression levels, protein

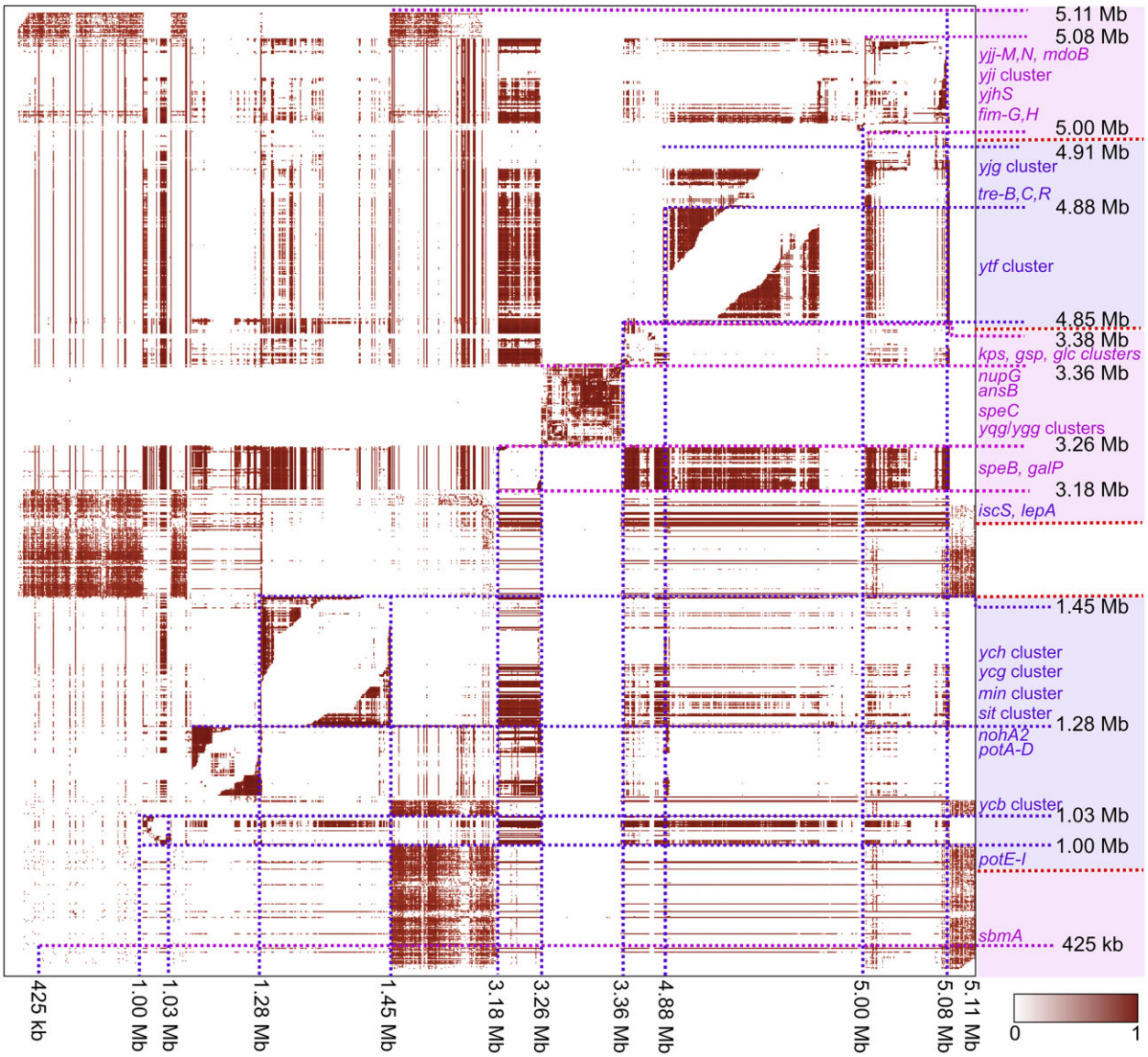

Figure 5. Genome-wide distribution of excess LD in the *E. coli* ST131 dataset ($n = 2156$) measured using LDWeaver. Approximate genomic positions are marked to the right and bottom of the LD map as per the EC958 reference genome. Brown shading indicates the amount of LD between sites (see key at bottom right for MI values scaled between 0 and 1). Entire genomic regions without excess LD are dropped from the figure to enhance the visibility of variation within high LD regions and each new region is coloured in an alternating shade of blue and purple in the right panel. Additionally, the right panel shows genomic regions comprising short-range excess-LD links.

structure alterations and epigenetic variations, which has the potential to greatly contribute towards improved detection. Using only DNA sequence information about LD is limiting because, in addition to selection acting on combined sets of mutations, LD signals can reflect evolutionary factors such as population expansion, elevated mutation rates, and hitchhiking.

Given its existing and future potential for enabling molecular discoveries, we anticipate that GWES will continue to attract a wide interest from both methodological and applied perspectives.

Data availability

LDWeaver v.1.5 is available as an R package under a GNU General Public License (Version 3) on GitHub (https://github.com/Sudaraka88/LDWeaver) and the source code is available on Zenodo (https://zenodo.org/records/10016711). Genomic data accessions used in this analysis are available via figshare (https://doi.org/10.6084/m9.figshare.24079491). An interactive phylogenetic tree (Figure 3A) clearly marking the detected isolates can be accessed on microreact (https://microreact.org/project/s94ZeKRZUSkz7JZrkwsuFY-spnmschtree).

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

J.C. and G.T.H. were funded by NFR [299941]; S.M., N.M., S.A.-A., R.A.G. and A.K.P. were funded by the AMR grant from Trond Mohn Foundation; S.M., N.M. and S.A.-A. were additionally funded by Marie Skłodowska-Curie Actions

[801133]; N.J.C. was supported by a Sir Henry Dale Fellowship, jointly funded by Wellcome and the Royal Society [104169/Z/14/A; https://wellcome.org/; https://royalsociety.org/]. The funders had no role in study design, method development and analysis nor interpretation of the results, writing of the report, and the decision to submit the paper for publication.

Conflict of interest statement

None declared.

References

1. Rocha,E.P.C. and Feil,E.J. (2010) Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLoS Genet.*, **6**, e1001104.
2. Thorpe,H.A., Bayliss,S.C., Hurst,L.D. and Feil,E.J. (2017) Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, **206**, 363–376.
3. Tonkin-Hill,G., MacAlasdair,N., Ruis,C., Weimann,A., Horesh,G., Lees,J.A., Gladstone,R.A., Lo,S., Beaudoin,C., Floto,R.A., *et al.* (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.*, **21**, 180.
4. Lees,J.A., Galardini,M., Bentley,S.D., Weiser,J.N., Corander,J. and Stegle,O. (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, **34**, 4310–4312.
5. Lees,J.A., Mai,T.T., Galardini,M., Wheeler,N.E., Horsfield,S.T., Parkhill,J. and Corander,J. (2020) Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *mBio*, **11**, e01344-20.
6. Lees,J.A., Croucher,N.J., Goldblatt,D., Nosten,F., Parkhill,J., Turner,C., Turner,P. and Bentley,S.D. (2017) Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife*, **6**, e26255.
7. Kachroo,P., Eraso,J.M., Beres,S.B., Olsen,R.J., Zhu,L., Nasser,W., Bernard,P.E., Cantu,C.C., Saavedra,M.O., Arredondo,M.J., *et al.* (2019) Integrated analysis of population genomics, transcriptomics and virulence provides novel insights into Streptococcus pyogenes pathogenesis. *Nat. Genet.*, **51**, 548–559.
8. Lees,J.A., Ferwerda,B., Kremer,P.H.C., Wheeler,N.E., Serón,M.V., Croucher,N.J., Gladstone,R.A., Bootsma,H.J., Rots,N.Y., Wijmega-Monsuur,A.J., *et al.* (2019) Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat. Commun.*, **10**, 2176.
9. Cui,Y., Yang,X., Didelot,X., Guo,C., Li,D., Yan,Y., Zhang,Y., Yuan,Y., Yang,H., Wang,J., *et al.* (2015) Epidemic clones, oceanic gene pools, and eco-LD in the Free living marine pathogen Vibrio parahaemolyticus. *Mol. Biol. Evol.*, **32**, 1396–1410.
10. Skwark,M.J., Croucher,N.J., Puranen,S., Chewapreecha,C., Pesonen,M., Xu,Y.Y., Turner,P., Harris,S.R., Beres,S.B., Musser,J.M., *et al.* (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet.*, **13**, e1006508.
11. Puranen,S., Pesonen,M., Pensar,J., Xu,Y.Y., Lees,J.A., Bentley,S.D., Croucher,N.J. and Corander,J. (2018) SuperDCA for genome-wide epistasis analysis. *Microb. Genom.*, **4**, e000184.
12. Pensar,J., Puranen,S., Arnold,B., MacAlasdair,N., Kuronen,J., Tonkin-Hill,G., Pesonen,M., Xu,Y., Sipola,A., Sánchez-Busó,L., *et al.* (2019) Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res.*, **47**, e112.
13. Schubert,B., Maddamsetti,R., Nyman,J., Farhat,M.R. and Marks,D.S. (2019) Genome-wide discovery of epistatic loci affecting antibiotic resistance in Neisseria gonorrhoeae using evolutionary couplings. *Nat. Microbiol.*, **4**, 328–338.
14. Top,J., Arredondo-Alonso,S., Schürch,A.C., Puranen,S., Pesonen,M., Pensar,J., Willems,R.J.L. and Corander,J. (2020) Genomic rearrangements uncovered by genome-wide co-evolution analysis of a major nosocomial pathogen, Enterococcus faecium. *Microb. Genom.*, **6**, mgen000488.
15. Chewapreecha,C., Pensar,J., Chattagul,S., Pesonen,M., Sangphukieo,A., Boonklang,P., Potisap,C., Koosakulnirand,S., Feil,E.J., Dunachie,S., *et al.* (2022) Co-evolutionary signals identify Burkholderia pseudomallei survival strategies in a hostile environment. *Mol. Biol. Evol.*, **39**, msab306.
16. Posada-Reyes,A.-B., Balderas-Martínez,Y.I., Ávila-Ríos,S., Vinuesa,P. and Fonseca-Coronado,S. (2022) An epistatic network describes and as relevant genes for. *Front Mol. Biosci.*, **9**, 856212.
17. Arnold,B.J., Gutmann,M.U., Grad,Y.H., Sheppard,S.K., Corander,J., Lipsitch,M. and Hanage,W.P. (2018) Weak epistasis may drive adaptation in recombining bacteria. *Genetics*, **208**, 1247–1260.
18. Taylor,A.J., Yahara,K., Pascoe,B., Mageiros,L., Mourkas,E., Calland,J.K., Puranen,S., Hitchings,M.D., Jolley,K.A., Kobras,C.M., *et al.* (2023) A two-hit epistasis model prevents core genome disharmony in recombining bacteria. bioRxiv doi: https://doi.org/10.1101/2021.03.15.435406, 21 April 2023, preprint: not peer reviewed.
19. Arnold,B., Sohail,M., Wadsworth,C., Corander,J., Hanage,W.P., Sunyaev,S. and Grad,Y.H. (2020) Fine-scale haplotype structure reveals strong signatures of positive selection in a recombining bacterial pathogen. *Mol. Biol. Evol.*, **37**, 417–428.
20. Rocha,E.P.C. (2018) Neutral theory, microbial practice: challenges in bacterial population genetics. *Mol. Biol. Evol.*, **35**, 1338–1347.
21. Arnold,B.J., Huang,I.-T. and Hanage,W.P. (2022) Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.*, **20**, 206–218.
22. Baumdicker,F., Hess,W.R. and Pfaffelhuber,P. (2012) The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.*, **4**, 443–456.
23. Kimura,M. (1985) In: *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
24. Turner,P., Turner,C., Jankhot,A., Helen,N., Lee,S.J., Day,N.P., White,N.J., Nosten,F. and Goldblatt,D. (2012) A longitudinal study of Streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. *PLoS One*, **7**, e38271.
25. Croucher,N.J., Finkelstein,J.A., Pelton,S.I., Mitchell,P.K., Lee,G.M., Parkhill,J., Bentley,S.D., Hanage,W.P. and Lipsitch,M. (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.*, **45**, 656–663.
26. Calland,J.K., Pascoe,B., Bayliss,S.C., Mourkas,E., Berthenet,E., Thorpe,H.A., Hitchings,M.D., Feil,E.J., Corander,J., Blaser,M.J., *et al.* (2021) Quantifying bacterial evolution in the wild: a birthday problem for Campylobacter lineages. *PLoS Genet.*, **17**, e1009829.
27. Blackwell,G.A., Hunt,M., Malone,K.M., Lima,L., Horesh,G., Alako,B.T.F., Thomson,N.R. and Iqbal,Z. (2021) Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLoS Biol.*, **19**, e3001421.
28. Pöntinen,A.K., Top,J., Arredondo-Alonso,S., Tonkin-Hill,G., Freitas,A.R., Novais,C., Gladstone,R.A., Pesonen,M., Meneses,R., Pesonen,H., *et al.* (2021) Apparent nosocomial adaptation of Enterococcus faecalis predates the modern hospital era. *Nat. Commun.*, **12**, 1523.
29. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *fly.*, **6**, 80–92.
30. Haller,B.C. and Messer,P.W. (2023) SLiM 4: multispecies eco-evolutionary modeling. *Am. Nat.*, **201**, E127–E139.
31. Cury,J., Haller,B.C., Achaz,G. and Jay,F. (2022) Simulation of bacterial populations with SLiM. *Peer Community J.*, **2**, e7.

32. Harrow,G.L., Lees,J.A., Hanage,W.P., Lipsitch,M., Corander,J., Colijn,C. and Croucher,N.J. (2021) Negative frequency-dependent selection and asymmetrical transformation stabilise multi-strain bacterial population structures. *ISME J.*, **15**, 1523–1538.
33. Løchen,A., Croucher,N.J. and Anderson,R.M. (2020) Divergent serotype replacement trends and increasing diversity in pneumococcal disease in high income settings reduce the benefit of expanding vaccine valency. *Sci. Rep.*, **10**, 18977.
34. Chewapreecha,C., Harris,S.R., Croucher,N.J., Turner,C., Marttinen,P., Cheng,L., Pessia,A., Aanensen,D.M., Mather,A.E., Page,A.J., *et al.* (2014) Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.*, **46**, 305–309.
35. Croucher,N.J., Walker,D., Romero,P., Lennard,N., Paterson,G.K., Bason,N.C., Mitchell,A.M., Quail,M.A., Andrew,P.W., Parkhill,J., *et al.* (2009) Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniaeSpain23F ST81. *J. Bacteriol.*, **191**, 1480–1489.
36. Croucher,N.J., Campo,J.J., Le,T.Q., Liang,X., Bentley,S.D., Hanage,W.P. and Lipsitch,M. (2017) Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E357–E366.
37. Callaghan,M.J., Buckee,C.O., Jolley,K.A., Kriz,P., Maiden,M.C.J. and Gupta,S. (2008) The effect of immune selection on the structure of the meningococcal opa protein repertoire. *PLoS Pathog.*, **4**, e1000020.
38. Palmer,L.D., Leung,M.H. and Downs,D.M. (2014) The cysteine desulfhydrase CdsH is conditionally required for sulfur mobilization to the thiamine thiazole in Salmonella enterica. *J. Bacteriol.*, **196**, 3964–3970.
39. Kumar,R., Shah,P., Swiatlo,E., Burgess,S.C., Lawrence,M.L. and Nanduri,B. (2010) Identification of novel non-coding small RNAs from Streptococcus pneumoniae TIGR4 using high-resolution genome tiling arrays. *Bmc Genomics [Electronic Resource]*, **11**, 350.
40. D'Aeth,J.C., van der Linden,M.P., McGee,L., de Lencastre,H., Turner,P., Song,J.-H., Lo,S.W., Gladstone,R.A., Sá-Leão,R., Ko,K.S., *et al.* (2021) The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. *eLife*, **10**, e67113.
41. Racimo,F., Sankararaman,S., Nielsen,R. and Huerta-Sánchez,E. (2015) Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.*, **16**, 359–371.
42. Lehtinen,S., Croucher,N.J., Blanquart,F. and Fraser,C. (2022) Epidemiological dynamics of bacteriocin competition and antibiotic resistance. *Proc. Biol. Sci.*, **289**, 20221197.
43. Corander,J., Fraser,C., Gutmann,M.U., Arnold,B., Hanage,W.P., Bentley,S.D., Lipsitch,M. and Croucher,N.J. (2017) Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.*, **1**, 1950–1960.
44. Miller,E.L., Abrudan,M.I., Roberts,I.S. and Rozen,D.E. (2016) Diverse ecological strategies are encoded by Streptococcus pneumoniae bacteriocin-like peptides. *Genome Biol. Evol.*, **8**, 1072–1090.
45. de Saizieu,A., Gardès,C., Flint,N., Wagner,C., Kamber,M., Mitchell,T.J., Keck,W., Amrein,K.E. and Lange,R. (2000) Microarray-based identification of a novel Streptococcus pneumoniae regulon controlled by an autoinduced peptide. *J. Bacteriol.*, **182**, 4696–4703.
46. Wilson,D.J., Gabriel,E., Leatherbarrow,A.J.H., Cheesbrough,J., Gee,S., Bolton,E., Fox,A., Hart,C.A., Diggle,P.J. and Fearnhead,P. (2009) Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. *Mol. Biol. Evol.*, **26**, 385–397.
47. Gundogdu,O., Bentley,S.D., Holden,M.T., Parkhill,J., Dorrell,N. and Wren,B.W. (2007) Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *Bmc Genomics [Electronic Resource]*, **8**, 162.
48. Whitehouse,C.A., Balbo,P.B., Pesci,E.C., Cottle,D.L., Mirabito,P.M. and Pickett,C.L. (1998) Campylobacter jejuni cytolethal distending toxin causes a G2-phase cell cycle block. *Infect. Immun.*, **66**, 1934–1940.
49. Lara-Tejero,M. and Galán,J.E. (2001) CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect. Immun.*, **69**, 4358–4365.
50. Guirado,P., Iglesias-Torrens,Y., Miró,E., Navarro,F., Attolini,C.S.-O., Balsalobre,C. and Madrid,C. (2022) Host-associated variability of the cdtABC operon, coding for the cytolethal distending toxin, in Campylobacter jejuni. *Zoonoses Public Health*, **69**, 966–977.
51. Forde,B.M., Ben Zakour,N.L., Stanton-Cook,M., Phan,M.-D., Totsika,M., Peters,K.M., Chan,K.G., Schembri,M.A., Upton,M. and Beatson,S.A. (2014) The complete genome sequence of Escherichia coli EC958: a high quality reference sequence for the globally disseminated multidrug resistant E. coli O25b:H4-ST131 clone. *PLoS One*, **9**, e104400.
52. Forde,B.M., Roberts,L.W., Phan,M.-D., Peters,K.M., Fleming,B.A., Russell,C.W., Lenherr,S.M., Myers,J.B., Barker,A.P., Fisher,M.A., *et al.* (2019) Population dynamics of an Escherichia coli ST131 lineage during recurrent urinary tract infection. *Nat. Commun.*, **10**, 3643.
53. Gladstone,R.A., McNally,A., Pöntinen,A.K., Tonkin-Hill,G., Lees,J.A., Skytén,K., Cléon,F., Christensen,M.O.K., Haldorsen,B.C., Bye,K.K., *et al.* (2021) Emergence and dissemination of antimicrobial resistance in Escherichia coli causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. *Lancet Microbe.*, **2**, e331–e341.
54. Kallonen,T., Brodrick,H.J., Harris,S.R., Corander,J., Brown,N.M., Martin,V., Peacock,S.J. and Parkhill,J. (2017) Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res.*, **27**, 1437–1449.
55. Petty,N.K., Ben Zakour,N.L., Stanton-Cook,M., Skippington,E., Totsika,M., Forde,B.M., Phan,M.-D., Gomes Moriel,D., Peters,K.M., Davies,M., *et al.* (2014) Global dissemination of a multidrug resistant Escherichia coli clone. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 5694–5699.
56. Price,L.B., Johnson,J.R., Aziz,M., Clabots,C., Johnston,B., Tchesnokova,V., Nordstrom,L., Billig,M., Chattopadhyay,S., Stegger,M., *et al.* (2013) The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-rx. *mBio*, **4**, e00377-13.
57. Johnson,T.J., Danzeisen,J.L., Youmans,B., Case,K., Llop,K., Munoz-Aguayo,J., Flores-Figueroa,C., Aziz,M., Stoesser,N., Sokurenko,E., *et al.* (2016) Separate F-type plasmids have shaped the evolution of the H30 subclone of Escherichia coli sequence type 131. *mSphere*, https://doi.org/10.1128/mSphere.00121-16.
58. Ben Zakour,N.L., Alsheikh-Hussain,A.S., Ashcroft,M.M., Khanh Nhu,N.T., Roberts,L.W., Stanton-Cook,M., Schembri,M.A. and Beatson,S.A. (2016) Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of Escherichia coli ST131. *mBio*, **7**, e00347-16.
59. Ghilarov,D., Inaba-Inoue,S., Stepien,P., Qu,F., Michalczyk,E., Pakosz,Z., Nomura,N., Ogasawara,S., Walker,G.C., Rebuffat,S., *et al.* (2021) Molecular mechanism of SbmA, a promiscuous transporter exploited by antimicrobial peptides. *Sci. Adv.*, **7**, eabj5363.
60. Li,G., Laturnus,C., Ewers,C. and Wieler,L.H. (2005) Identification of genes required for avian Escherichia coli septicemia by signature-tagged mutagenesis. *Infect. Immun.*, **73**, 2818–2827.

61. Navarro,C., Wu,L.F. and Mandrand-Berthelot,M.A. (1993) The nik operon of Escherichia coli encodes a periplasmic binding-protein-dependent transport system for nickel. *Mol. Microbiol.*, **9**, 1181–1191.
62. Pugh,H.L., Connor,C., Siasat,P., McNally,A. and Blair,J.M.A. (2023) E. coli ST11 (O157:H7) does not encode a functional AcrF efflux pump. *Microbiology*, **169**, 001324.
63. Balakrishnan,R., Oman,K., Shoji,S., Bundschuh,R. and Fredrick,K. (2014) The conserved GTPase LepA contributes mainly to translation initiation in Escherichia coli. *Nucleic Acids Res.*, **42**, 13370–13383.
64. Schwartz,C.J., Djaman,O., Imlay,J.A. and Kiley,P.J. (2000) The cysteine desulfurase, IscS, has a major role in *in vivo* Fe-S cluster formation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9009–9014.
65. Cross,A.S., Gemski,P., Sadoff,J.C., Orskov,F. and Orskov,I. (1984) The importance of the K1 capsule in invasive infections caused by Escherichia coli. *J. Infect. Dis.*, **149**, 184–193.
66. Opal,S., Cross,A. and Gemski,P. (1982) K antigen and serum sensitivity of rough Escherichia coli. *Infect. Immun.*, **37**, 956–960.
67. Haudiquet,M., Buffet,A., Rendueles,O. and Rocha,E.P.C. (2021) Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae. *PLoS Biol.*, **19**, e3001276.
68. Palmer,K.L., Godfrey,P., Griggs,A., Kos,V.N., Zucker,J., Desjardins,C., Cerqueira,G., Gevers,D., Walker,S., Wortman,J., *et al.* (2012) Comparative genomics of enterococci: variation in Enterococcus faecalis, clade structure in E. faecium, and defining characteristics of E. gallinarum and E. casseliflavus. *mBio*, **3**, e00318-11.
69. Raven,K.E., Reuter,S., Gouliouris,T., Reynolds,R., Russell,J.E., Brown,N.M., Török,M.E., Parkhill,J. and Peacock,S.J. (2016) Genome-based characterization of hospital-adapted Enterococcus faecalis lineages. *Nat. Microbiol.*, **1**, 15033.
70. Lebreton,F., Manson,A.L., Saavedra,J.T., Straub,T.J., Earl,A.M. and Gilmore,M.S. (2017) Tracing the enterococci from paleozoic origins to the hospital. *Cell*, **169**, 849–861.
71. Paulsen,I.T., Banerjei,L., Myers,G.S.A., Nelson,K.E., Seshadri,R., Read,T.D., Fouts,D.E., Eisen,J.A., Gill,S.R., Heidelberg,J.F., *et al.* (2003) Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. *Science*, **299**, 2071–2074.
72. Cortes-Perez,N.G., Dumoulin,R., Gaubert,S., Lacoux,C., Bugli,F., Martin,R., Chat,S., Piquand,K., Meylheuc,T., Langella,P., *et al.* (2015) Overexpression of Enterococcus faecalis elr operon protects from phagocytosis. *BMC Microbiol.*, **15**, 112.
73. Dumoulin,R., Cortes-Perez,N., Gaubert,S., Duhutrel,P., Brinster,S., Torelli,R., Sanguinetti,M., Posteraro,B., Repoila,F. and Serror,P. (2013) Enterococcal rgg-like regulator ElrR activates expression of the elrA operon. *J. Bacteriol.*, **195**, 3073–3083.
74. Rich,R.L., Kreikemeyer,B., Owens,R.T., LaBrenz,S., Narayana,S.V., Weinstock,G.M., Murray,B.E. and Höök,M. (1999) Ace is a collagen-binding MSCRAMM from Enterococcus faecalis. *J. Biol. Chem.*, **274**, 26939–26945.
75. Lebreton,F., Riboulet-Bisson,E., Serror,P., Sanguinetti,M., Posteraro,B., Torelli,R., Hartke,A., Auffray,Y. and Giard,J.-C. (2009) ace, which encodes an adhesin in Enterococcus faecalis, is regulated by Ers and is involved in virulence. *Infect. Immun.*, **77**, 2832–2839.
76. Graham,C.E., Cruz,M.R., Garsin,D.A. and Lorenz,M.C. (2017) Enterococcus faecalis bacteriocin EntV inhibits hyphal morphogenesis, biofilm formation, and virulence of Candida albicans. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 4507–4512.
77. Cruz,M.R., Cristy,S., Guha,S., De Cesare,G.B., Evdokimova,E., Sanchez,H., Borek,D., Miramón,P., Yano,J., Fidel,P.L. Jr, *et al.* (2022) Structural and functional analysis of EntV reveals a 12 amino acid fragment protective against fungal infections. *Nat. Commun.*, **13**, 6047.
78. Ch'ng,J.-H., Chong,K.K.L., Lam,L.N., Wong,J.J. and Kline,K.A. (2019) Biofilm-associated infection by enterococci. *Nat. Rev. Microbiol.*, **17**, 82–94.
79. Kloos,J., Gama,J.A., Hegstad,J., Samuelsen,Ø. and Johnsen,P.J. (2021) Piggybacking on niche adaptation improves the maintenance of multidrug-resistance plasmids. *Mol. Biol. Evol.*, **38**, 3188–3201.
80. Mallawaarachchi,S., Tonkin-Hill,G., Croucher,N.J., Turner,P., Speed,D., Corander,J. and Balding,D. (2022) Genome-wide association, prediction and heritability in bacteria with application to. *NAR Genom Bioinform*, **4**, lqac011.
81. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1293–E301.
82. Ekeberg,M., Lövkvist,C., Lan,Y., Weigt,M. and Aurell,E. (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **87**, 012707.
83. Bates,D., Maechler,M. and Maechler,M.M. (2022) Package 'Matrix'.
84. Hartigan,J.A. and Wong,M.A. (1979) Algorithm AS 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.*, **28**, 100.
85. Lin,M. and Kussell,E. (2019) Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods*, **16**, 199–204.
86. Sipola,A., Marttinen,P. and Corander,J. (2018) Bacmeta: simulator for genomic evolution in bacterial metapopulations. *Bioinformatics*, **34**, 2308–2310.
87. Vos,P.G., Paulo,M.J., Voorrips,R.E., Visser,R.G.F., van Eck,H.J. and van Eeuwijk,F.A. (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.*, **130**, 123–135.
88. Delignette-Muller,M.L. and Dutang,C. (2015) fitdistrplus: an R package for fitting distributions. *J. Stat. Softw.*, **64**, 1–34.
89. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Dalla Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf.*, **7**, S7.
90. Stuart,M., Hoaglin,D.C., Mosteller,F. and Tukey,J.W. (1984) Understanding robust and exploratory data analysis. *Statistician*, **33**, 320.
91. Bunn,A. and Korpela,M. (2018) Crossdating in dplR. http://cran.nexr.com/web/packages/dplR/vignettes/xdate-dplR.pdf.
92. Wickham,H., Chang,W. and Wickham,M.H. (2016) Package 'ggplot2'. *Create Elegant Data Visualisations Using the Grammar of Graphics. Version*, **2**, 1–189.
93. Anand,L. and Rodriguez Lopez,C.M. (2022) ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes. *BMC Bioinf.*, **23**, 33.
94. Takeda,H., Farsiu,S. and Milanfar,P. (2007) Kernel regression for image processing and reconstruction. *IEEE Trans. Image Process.*, **16**, 349–366.
95. Zhao,S., Guo,Y., Sheng,Q. and Shyr,Y. (2014) Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinf.*, **15**, P16.
96. Thomas,M. and Pedersen,L. (2022) Package 'ggraph'.
97. Csardi,M.G. (2013) Package 'igraph'.
98. Argimón,S., Abudahab,K., Goater,R.J.E., Fedosejev,A., Bhai,J., Glasner,C., Feil,E.J., Holden,M.T.G., Yeats,C.A., Grundmann,H., *et al.* (2016) Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genom.*, **2**, e000093.
99. Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217–223.

100. Paradis,E. and Schliep,K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
101. Yu,G., Smith,D.K., Zhu,H. and Guan,Y. (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.
102. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Received: February 11, 2024. **Revised:** April 15, 2024. **Editorial Decision:** May 13, 2024. **Accepted:** May 14, 2024