MICROBIAL GENOMICS

TECHNICAL RESOURCE
Kobras *et al.*, *Microbial Genomics* 2026;12:001714
DOI 10.1099/mgen.0.001714

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

GeneScanner: profiling genetic variation across bacterial populations

Carolin M. Kobras[1,2,†], Seungwon Ko[3,†], Priyanshu S. Raikwar[3,†], Broncio Aguilar-Sanjuan[3], Keith A. Jolley[3] and Samuel K. Sheppard[3,*]

Abstract

Rapid, low-cost genome sequencing has transformed microbiology, advancing efforts to link genetic and phenotypic variation across diverse bacterial systems. Laboratory functional screens now uncover causal mechanisms underlying key traits in simplified systems, such as drug resistance, pathogenicity and metabolic adaptation, while population-scale comparative genomics reveal the immense natural diversity associated with these traits in real-world settings. Despite their complementary strengths, these approaches remain challenging to integrate, especially for researchers without advanced bioinformatics skills. This skills gap can constrain the capacity to reveal the mechanisms underlying microbial traits and evolutionary adaptations. We developed GeneScanner to aid user-friendly analyses of gene- and protein-level variation across large bacterial genome collections. GeneScanner detects genetic variants and amino acid substitutions in homologous sequences to improve functional interpretation of microbial variation. Using synthetic data and three case studies across different species and phenotypes, we show that GeneScanner reliably identifies nucleotide and protein-level variants associated with specific traits. The presented examples highlight the broad applicability of GeneScanner in microbial genomics, enabling research across diverse fields, such as antimicrobial resistance, host-pathogen interactions, microbial evolution, epidemiology and public health.

Impact Statement

Connecting laboratory functional genomics with the vast diversity of microbial populations is essential for understanding the genetic basis of key bacterial phenotypes. GeneScanner provides a scalable, accessible platform for characterizing gene- and protein-level variation across thousands of bacterial genomes. By integrating detection of diverse sequence variants with automated comparison of homologous regions, GeneScanner enables rapid evaluation of how laboratory-identified mutations correspond to population diversity. This capability strengthens genotype-phenotype association studies and broadens functional interpretation of microbial variation in real-world contexts, supporting advances in antimicrobial resistance research, virulence factor discovery and pathogen surveillance.

DATA SUMMARY

GeneScanner code and detailed requirements for each release version are publicly available (https://github.com/Sheppard-Lab/GeneScanner, DOI: 10.5281/zenodo.17495646). GeneScanner requires Python v3.6+. GeneScanner is also available as a plugin in the PubMLST online database (https://bigsdb.readthedocs.io/en/latest/data_analysis/genescanner.html). All genome assemblies

Received 15 January 2026; Accepted 20 April 2026; Published 02 June 2026
Author affiliations: [1]Sir William Dunn School of Pathology, University of Oxford, Oxford, UK; [2]Department of Microbes, Infection and Microbiomes, Institute of Microbiology and Infection, College of Medicine and Health, University of Birmingham, Birmingham, UK; [3]Department of Biology, Ineos Oxford Institute for Antimicrobial Research, University of Oxford, Oxford, UK.
***Correspondence:** Samuel K. Sheppard, samuel.sheppard@biology.ox.ac.uk
Keywords: antimicrobial resistance; biofilm; bioinformatics; comparative genomics; genetic variation; host association; mutation analysis; variant calling.
Abbreviations: GWAS, genome-wide association studies; SNP, Single nucleotide polymorphism; VCF, variant call format; WGS, whole-genome sequencing.
†These authors contributed equally to this work
All supporting data, code and protocols have been provided within the article or through supplementary data files. Nine supplementary files are available with the online version of this article.

used are available through the PubMLST database. Isolate IDs, query sequences and alignment and output files for designed and synthetic datasets can be found in the online supplementary files. All data are available in Figshare: https://doi.org/10.6084/m9.figshare.31750834[1].

INTRODUCTION

For decades, molecular microbiology has sought to understand the genetic basis of important bacterial traits, including pathogenicity, antimicrobial resistance and host interactions. Advances in sequencing technologies have transformed understanding of gene function, enabling the genome-wide analyses that define the field of functional genomics. For example, the principles of gene inactivation studies that compare the phenotypes of mutant and wild-type strains are now often augmented with genome-wide gene deletion [2–5] and transposon-insertion libraries [6–11], allowing rapid genotype–phenotypic screening under diverse selective conditions. Furthermore, in laboratory evolution experiments, where bacteria are subjected to selective pressures such as antibiotics or environmental stress [12–15], whole-genome sequencing (WGS) is often used to compare progenitor and adapted isolates. While scientifically rigorous, these laboratory studies are often performed using a single laboratory-adapted strain and may be limited in their ecological relevance. These approaches emphasize the importance of controlled experimental conditions and a precise understanding of the ancestral and derived genetic background of laboratory strains, enabling targeted investigation of genotype–phenotype association. However, a laboratory setting may not fully reflect the complexity of natural environments where bacteria evolve [16].

The impact of WGS in a laboratory setting is mirrored in natural populations. Here, the emphasis is often upon multi-strain comparative genomics for pathogen surveillance, phylogenetics and epidemiology. Millions of bacterial sequences are now freely available in public repositories [17, 18] and curated databases [19–21], with metadata such as isolation source, place and time. This has transformed understanding of microbial genomics in the 'wild' with approaches including genome-wide association studies (GWAS) and covariation analyses revealing genetic variation underlying phenotypes linked to host adaptation [22–25], biofilm and virulence [26–31] and antibiotic resistance [32–35]. However, while some of the limitations of laboratory studies are overcome by analysing multiple strains in bacterial populations, highly relevant population genomics studies lack the control and precision of laboratory functional genomics. This typically limits genotype–phenotype inference to association rather than cause-and-effect relationships.

Next-generation microbiology brings the rigour of the laboratory together with the relevance of population-scale studies and has major potential for understanding gene function in natural systems [36]. For example, studies of bacterial pathogens, including *Campylobacter* [24], *Streptococcus pneumoniae* [37] and *Staphylococcus aureus* [38], confirm laboratory observations in bacterial isolate populations and *vice versa*, suggesting that these genetic factors are relevant beyond laboratory conditions. While database resources and isolated genome collections are freely available for comparable analyses of other pathogen species, analysis pipelines are often aimed at bioinformaticians rather than laboratory microbiologists. This can require the installation of specific operating systems and environments and command-line experience for SNP calling [39–43].

To address these challenges, we developed GeneScanner, a user-friendly tool that enables researchers to explore genetic variation in specific genes or proteins across large datasets from bacterial populations. Implemented as both a stand-alone script on GitHub and a web plugin available through PubMLST [21], GeneScanner simplifies the comparison of gene loci across strains and the identification of variants that can be linked to phenotypic traits such as antibiotic resistance or pathogenicity. By providing an intuitive interface, it allows users without extensive bioinformatics expertise to analyse large datasets, contextualizing laboratory-observed genotypes within real-world bacterial diversity. While GeneScanner does not explicitly account for population structure, it provides a useful starting point for exploring genetic variation at specific loci in bacterial populations. This variation may include the presence or absence of genes across bacterial lineages or species, differences in the overall number of mutated residues or alleles, mutational hotspots or conserved regions, as well as overlap with functional features such as regulator binding sites or enzymatically important residues. The extent to which such patterns can be detected will depend on the locus under investigation and the diversity of the dataset analysed, but these insights can help identify candidates for further investigation.

In this study, we analyse synthetic data and three case studies across different bacterial species and phenotypes. We demonstrate how GeneScanner detects nucleotide and protein-level variation linked to specific traits, underscoring its broad applicability in microbial genomics, from antibiotic resistance studies to pathogen surveillance and laboratory evolution experiments.

METHODS

Technical specifications

All analyses were executed on PubMLST using the GeneScanner wrapper plugin [21]. Core software versions were Python 3.6+, Biopython 1.84 [44], pandas 2.2.2 [45], XlsxWriter 3.2.0 [46], MAFFT 7.490 [47] and SNP-sites 2.5.1 [39]. Detailed requirements for each release version and all code are available in the public GitHub repository (https://github.com/Sheppard-Lab/GeneScanner).

GeneScanner pipeline and analytical framework

Multiple-sequence alignment FASTA files are the primary input for GeneScanner. Two quality filters are applied relative to a designated reference sequence: a default minimum of 80% ungapped coverage across the alignment length and at least 80% pairwise identity. If requested, a variant call format variant call format (VCF) file is generated via SNP-sites [39]. In nucleotide analysis mode, GeneScanner iterates through the alignment position-by-position while preserving codon phase relative to the reference, which can be user-defined or automatically set as the first sequence in the alignment. For each non-gap codon triplet, the translated codon is compared with the reference translation to classify synonymous and non-synonymous mutations. Insertions, deletions and premature stop codons are recorded independently. When protein analysis mode is activated, GeneScanner first removes all the gaps within the original alignment file, translates in the user-specified frame and realigns peptides with MAFFT [47]. To prevent alignment disruption caused by early stop codons, the programme temporarily fills post-stop-codon regions with reference residues, which are removed after realignment to maintain alignment quality.

Optional grouping analyses are triggered when the user provides a two-column CSV file specifying isolate names and their categories. GeneScanner partitions the dataset by category and re-runs the full analytical pipeline for each subset, producing parallel mutation analysis reports and mutation matrix worksheets that share column definitions but are restricted to their respective groups. When distinct reference sequences are supplied for individual groups, GeneScanner automatically reassigns the appropriate reference prior to variant calling. After all analyses are complete, GeneScanner compiles the results into a spreadsheet (.xlsx) by using pandas [45] and XlsxWriter [46]. All parameters are user-configurable via command-line flags, and run metadata, as well as quality-control filtering logs, can be written to separate files. The command line version of GeneScanner does not impose intrinsic limits on the number of samples analysed in a run, provided sufficient RAM and storage are available. Test runs of the software using a dataset of 1,982 sequences for single-gene analysis (test file available in the GitHub repository: *icaA* gene analysis) completed in 12.2 s on a standard laptop with 16 GB RAM and an Intel i7 (13th generation) processor or 8.3 s on an Apple M2 laptop with 16 GB RAM.

Using GeneScanner through the PubMLST database

For improved user accessibility, GeneScanner has been implemented as a plugin within the publicly available PubMLST database [21]. Integration with PubMLST allows users to easily select bacterial isolate genomes from its genome databases and either choose specific loci via a dropdown menu or alternatively paste a query sequence. Additional settings, including nucleotide or protein analysis, selection of the alignment tool, reference sequence options and grouping parameters, can be configured directly through the web interface. PubMLST either uses the sequences of alleles designated in isolate records for pre-defined loci or performs an initial BLAST [48] search to identify the relevant sequences in the selected genomes when a pasted sequence is supplied. A multi-FASTA alignment file is subsequently generated. This is then used as input for GeneScanner, allowing analyses to be performed without requiring command-line interaction while producing the same output files as the command-line version in a single step. Due to server constraints, analyses conducted through the PubMLST web interface are currently limited to 5,000 isolates per run, as the platform also performs the sequence alignment.

More information on how to run GeneScanner through PubMLST can be found here: https://bigsdb.readthedocs.io/en/latest/data_analysis/genescanner.html.

Synthetic sequence design for GeneScanner validation

To create a dataset with explicitly designed mutations, we generated a 300-bp reference coding sequence and designed point mutations and short indels using custom Python code from the GeneScanner repository. To evaluate coding-effect categorization, we placed synonymous substitutions at alignment positions 12 and 147 with allele frequencies of 0.20 and 0.30, respectively, nonsynonymous substitutions at positions 46 and 181 with frequencies of 0.20 and 0.30, and a nonsense stop codon substitution at position 210 with frequency 0.10 to confirm that GeneScanner treats stop-gains distinctly from other variants. To assess frameshift handling, we designed two indel scenarios with single-nucleotide changes: (i) frameshift leading to premature termination via a deletion at nucleotide position 4 (10%) and an insertion at position 69 (10%) and (ii) a frame-restoring pair consisting of an insertion at position 153 (5%), followed by a downstream deletion at position 219 (5%) such that the reading frame is restored after a segment of shifted translation. All coordinates are based on the aligned nucleotide sequence and amino-acid coordinates refer to the translation of the reference.

In silico sequence evolution simulation for GeneScanner validation and benchmarking

The sequence evolution of two subpopulations was simulated under opposing selection regimes using SLiM v5.0 [49] with nucleotide-based models and phylogeny data. The model script was generated by applying the previously published code [50]. The simulated genome comprised a 300-bp random ancestral coding sequence. Mutations followed a symmetric Jukes-Cantor model with a per-site mutation rate of $\mu=1\times10^{-6}$ per generation. Recombination occurred at a rate of $\rho=1/3\times10^{-4}$ per site per generation along the locus with 30 bp as the average recombination block length. At generation 1, we split the population into two panmictic subpopulations of 5,000 haploid individuals each. Selection was restricted to a 31-bp window spanning nucleotide positions

100–130. Fitness was calculated by multiplying across sites and across the two populations. We assigned opposite selection coefficients in the two subpopulations: In subpopulation p1, mutations within the window were beneficial (s=+0.005), whereas in subpopulation p2, the same mutations were deleterious (s=−0.005). Mutations outside the window mimicked neutral mutation with slightly deleterious fitness cost (s=−0.0001). Total fitness is calculated by $1 + \sum s$ comparing to 1 at the starting reference gene. To limit the extent of linkage hitchhiking, we chose an elevated recombination rate, as theory predicts that the influence of a selective sweep on linked diversity decreases with recombination distance relative to selection strength [51]. Density-dependent fitness adjustment was added to keep the population size near the original 5,000 haploid genomes per population, while it changes the sequence frequency within populations. Simulations were run for 1,000 generations under SLiM's default non-Wright–Fisher framework. At the final generation, 500 sequences were sampled from each subpopulation. Sampled sequences were aligned using MAFFT, and mutational patterns were subsequently characterized using GeneScanner in dual nucleotide–protein mode. In addition, we used SNP-sites 2.5.1 [39] with default settings to generate a VCF output file of the simulation dataset. All data used in this study are archived on FigShare [1].

Case study 1: genetic variation underlying ciprofloxacin resistance in *Escherichia coli*

The *E. coli* genome assemblies (*n*=1,509, Material S1) from a previous study [52] were accessed on the PubMLST database [21]. Information on the ciprofloxacin susceptibility of these isolates was taken from supplementary material. Nucleotide sequences of query genes (*gyrA*, *gyrB*, *parC* and *parE*) were downloaded from GenBank and used to query against isolate assemblies using blastn (BLAST 2.12.0+ with settings: word size: 20; reward: 2; penalty: −3; gapopen: 5; gapextend: 2). The identified matching sequences were extracted and aligned using MAFFT v7.505 with default parameters to create the nucleotide alignment input file. GeneScanner was run in 'nucleotide+protein' mode, selecting strain *E. coli* K12 MG1655 (PubMLST *Escherichia* ID: 1167) as reference for both groups. A CSV file for group 2 (resistant isolates) was created. Selected data, such as mutation frequency, were normalized by the isolate number and were visualized using GraphPad Prism (version 10.5.0)

Case study 2: conservation and diversity of vitamin B5 biosynthesis genes in *Campylobacter*

The genome assemblies of *Campylobacter jejuni* ST-45 complex isolates were accessed on the PubMLST database [21], using all available isolates resulting from the search term 'cattle' (*n*=277), and randomly selecting an equal number of isolates containing the term 'chicken' in the search field (Material S1). Nucleotide sequences of the *panBCD* locus intergenic were downloaded from GenBank and used to query against isolate assemblies using blastn (BLAST 2.12.0+ with settings: word size: 20; reward: 2; penalty: −3; gapopen: 5; gapextend: 2). The identified matching sequences were extracted and aligned using MAFFT v7.505 with default parameters to create the nucleotide alignment input file. GeneScanner was run in 'nucleotide+protein' mode, selecting strain *C. jejuni* NCTC11168 (PubMLST *C. jejuni/coli* ID: 48) as reference. A CSV file for group 2 (isolates from cattle) was created. The number of unique alleles was determined using the BIGSdb plugin GenomeComparator [53] using the same isolates as input and default settings. Selected data, such as mutation frequency, were normalized by the isolate number and visualized using GraphPad Prism (version 10.5.0).

Case study 3: genetic variation in biofilm-regulating intergenic region in staphylococci

The genome assemblies of *S. aureus* (*n*=1984), *Staphylococcus epidermidis* (*n*=1000), *Staphylococcus hominis* (*n*=13), *Staphylococcus haemolyticus* (*n*=44), *Staphylococcus pseudintermedius* (*n*=178) and *Staphylococcus chromogenes* (*n*=49) were accessed on the PubMLST database [21] (Material S1). Nucleotide sequences of the *icaR_icaA* intergenic region of *S. aureus* and *S. epidermidis* were downloaded from GenBank and used to query against isolate assemblies using BLASTN (BLAST 2.12.0+ with settings: word size: 20; reward: 2; penalty: −3; gapopen: 5; gapextend: 2). The identified matching sequences were extracted and aligned using MAFFT v7.505 with default parameters to create the nucleotide alignment input file. GeneScanner was run in 'nucleotide' mode, selecting strain USA300_FPR3757 (PubMLST *S. aureus* ID: 37463) and N13018T (PubMLST *S. epidermidis* ID: 41156) as reference for *S. aureus* and *S. epidermidis*, respectively. As the NucleotideAnalysis sheet will still assume a coding sequence, the descriptions *synonymous*, *non-synonymous* and *stop codons* should be ignored and mutation frequencies added together. Mutation frequencies were normalised to the isolate number and visualised using GraphPad Prism (version 10.5.0). Information on the regulator binding sites of *icaR_icaA* intergenic region was taken from previous studies [54–58].

RESULTS

GeneScanner is a user-friendly tool for exploring genetic variation in bacterial populations

GeneScanner is a Python-based workflow, available on GitHub (https://github.com/Sheppard-Lab/GeneScanner), that accepts pre-aligned nucleotide or protein FASTA files. It screens every alignment column for sequence variation and collates the results into a multi-sheet workbook in .xlsx format. Even without a bioinformatics background, GeneScanner can be used as a plugin on PubMLST (https://bigsdb.readthedocs.io/en/latest/data_analysis/genescanner.html) [21], supporting isolate selection and input file generation directly from organism-specific databases and eliminating the need for local installation or command-line operation.

GeneScanner operates in four sequential stages (Fig. 1). In stage 1, the input alignment undergoes rigorous coverage and identity filtering, ensuring that only sequences meeting the predefined thresholds for ungapped alignment coverage and sequence identity to the reference are retained for subsequent analysis. In stage 2, the filtered sequences are analysed according to the user-selected mode (nucleotide or protein). In nucleotide mode, GeneScanner traverses the alignment both position-by-position and in codon triplets. Assuming no insertion or deletion events, each translated codon is compared with the corresponding reference codon to identify synonymous and non-synonymous substitutions, while insertions, deletions and premature stop codons are recorded separately. In protein mode, the algorithm examines each aligned amino-acid column to quantify substitutions, gaps and stop codons and to calculate both mutation frequency and the remaining number of sequences following the stop codons. When the user requests protein analysis using nucleotide input, it automatically translates and realigns the sequences before proceeding as if native protein data had been provided.

In stage 3, GeneScanner can optionally perform grouping analyses when the user defines categories for sequences. This categorization helps subdivide the population, for example, by phenotype or other defining traits. The programme partitions the sequence dataset according to these categories and re-runs the full analytical pipeline for each subset. As the population structure is not explicitly controlled for, quantitative analyses should be interpreted carefully. Nevertheless, they provide a useful starting point for exploring patterns of genetic variation.

In stage 4, GeneScanner generates comprehensive nucleotide and amino acid mutation analysis spreadsheets that include detailed mutation categories. It also produces sparse mutation matrices in both nucleotide and protein modes, in which rows represent isolates, columns correspond to alignment positions and each filled cell contains the variant symbol when it differs from the reference. Each matrix is written to a dedicated worksheet to facilitate downstream visualization and inspection.

GeneScanner also calculates and exports summary statistics to a dedicated worksheet. For nucleotide analyses, the summary includes the total alignment length, the number of mutated sites, positions with mutation frequencies exceeding 20% and cumulative counts of synonymous and non-synonymous substitutions and insertions, deletions and stop-codons. For protein analyses, the corresponding statistics include substitutions, insertions, deletions, stop codons and position-wise mutation frequencies. When multiple isolated groups are defined, GeneScanner generates separate worksheets and mutation matrices for each group, ensuring consistent column definitions and analytical parameters across all outputs. If distinct reference sequences are specified for individual groups, the programme automatically reassigns the appropriate reference before performing variant calling.

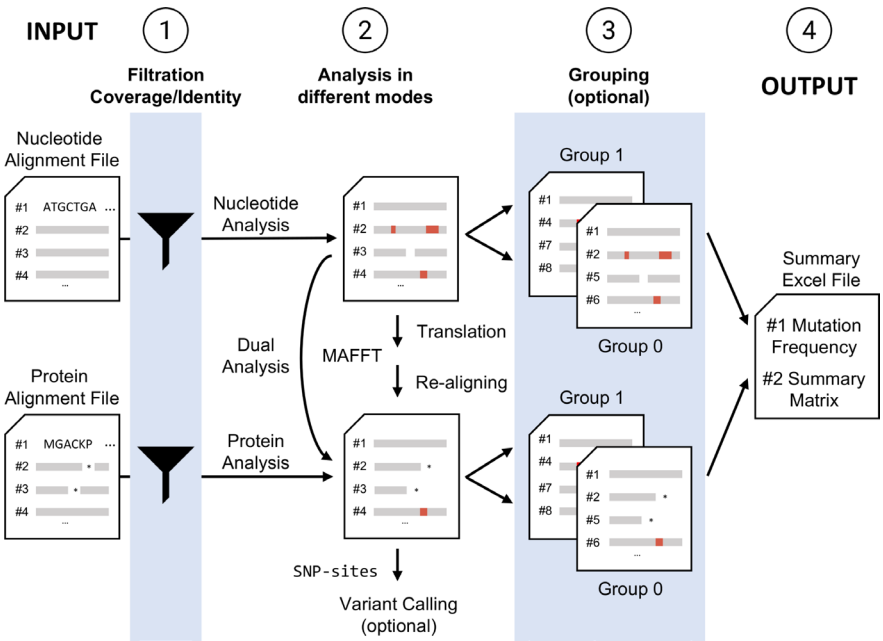

Fig. 1. The GeneScanner workflow consists of four distinct steps. Schematic illustration of the GeneScanner workflow: (1) input parsing and quality filtering of alignment files (input), (2) codon-aware nucleotide or residue-aware protein variant calling, (3) optional per-group re-analysis, (4) output: xlsx format report generation with analysis, matrix and summary worksheets. External tools MAFFT and snp-sites are invoked when translation, and realignment or VCF output is requested.

Testing GeneScanner using a controlled synthetic dataset

To demonstrate the functionality of GeneScanner, we first tested it on a controlled synthetic dataset of 100 to 300 bp sequences in which the locations, types and frequencies of mutations were explicitly designed (Material S2). The dataset reproduced the intended patterns (Fig. 2a, b), and GeneScanner successfully captured variation at both nucleotide and protein levels (Material S3). At the nucleotide level, 20 and 30% synonymous variants occurred at positions 12 and 147, respectively, while 20 and 30% non-synonymous variants appeared at positions 46 and 181, alongside a 10% nonsense mutation at position 210. When translated, synonymous sites produced no residue changes, whereas non-synonymous mutations manifested as 20% Q16E and 30% S61P substitutions. The stop-gain mutation was correctly assigned to amino acid position 70, with GeneScanner distinguishing it from other non-synonymous substitutions.

The correct detection of insertion and deletion (indel) events further highlighted the flexibility of the workflow (Fig. 2c, d, Materials S4 and S5). A single-nucleotide deletion at position 4 produced a frameshift with an early stop at amino acid position 14 (10%), and a single-nucleotide insertion at position 69 generated a frameshifted protein truncated at position 29 (10%). In the frame-restoration scenario, a paired insertion at position 153 (5%) and deletion at position 219 (5%) created a transient frameshift with amino acid substitutions between residues 52–73, after which the original reading frame was recovered. These results illustrate how GeneScanner's dual nucleotide-protein mode can trace both the disruption and recovery of coding frames, preserving alignment quality while recording complex mutational events.

Validation and benchmarking of *in silico* sequence evolution across two populations

To validate the functionality of GeneScanner on datasets with unknown outcomes and to test the grouping feature for analysing distinct subpopulations, we simulated evolution with differential selection. We created two populations of 5,000 bacterial genomes, starting with a random 300-bp coding sequence. Mutations occurring in a specific region (positions 100–130) were set to have opposite fitness effects in the two populations, while mutations elsewhere were neutral. After 1,000 generations, we sampled 500 isolates from each population (p1 and p2) and ran GeneScanner (Fig. 3a, Materials S6 and S7). The imposed selection produced clear divergence within the targeted 100–130 bp window (Fig. 3b, c). Using the original sequence before simulation as a reference, subpopulation p2 exhibited minor variation across the locus, consistent with neutral dynamics outside the selected window. In contrast, subpopulation p1 displayed a concentration of both synonymous and non-synonymous differences restricted to positions 100–130, reflecting the impact of opposing selection pressures. Outside of the region under simulated selection, both p1 and p2 showed similar mutational patterns. GeneScanner successfully recapitulated these differences at the nucleotide level, confirming the tracking of mutational events across divergent evolutionary trajectories.

To further benchmark GeneScanner, we analysed the *in silico* evolution dataset using the variant-calling tool SNP-sites [38]. SNP-sites identified the same variants as GeneScanner (Materials S8 and S9) confirming the accuracy of our pipeline. Although SNP-sites efficiently produce a VCF file summarizing variable positions and indicating the presence or absence of variants, it does not further classify variant types. GeneScanner adds this layer of interpretation by annotating mutations (e.g. synonymous,

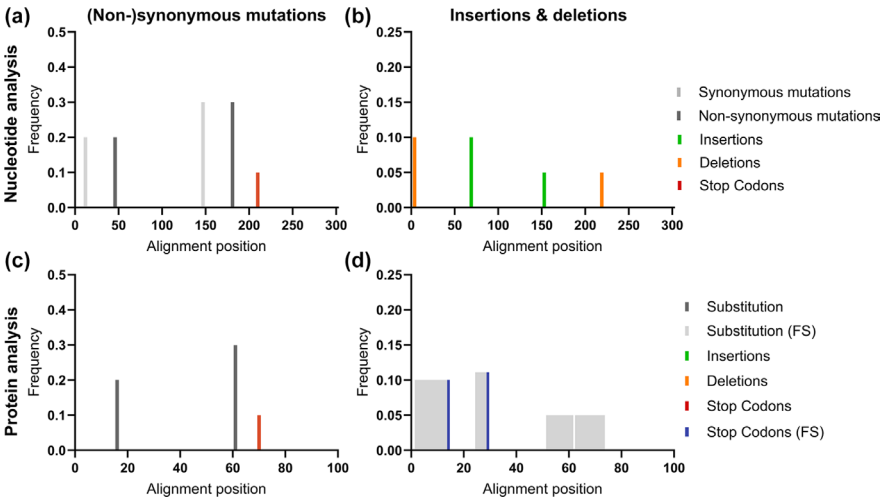

Fig. 2. GeneScanner can reliably detect sequence variation on a nucleotide and protein level. Analysis of a synthetic dataset demonstrates and validates the functionality of GeneScanner. The tool successfully captures synonymous and non-synonymous mutations, including those leading to stop codons at both nucleotide (a) and protein (b) levels. Insertions and deletions are recognized as events by nucleotide analysis (c) and as frameshifts (FS), premature stop codons or frame-restoration, where applicable, at the protein level (d).

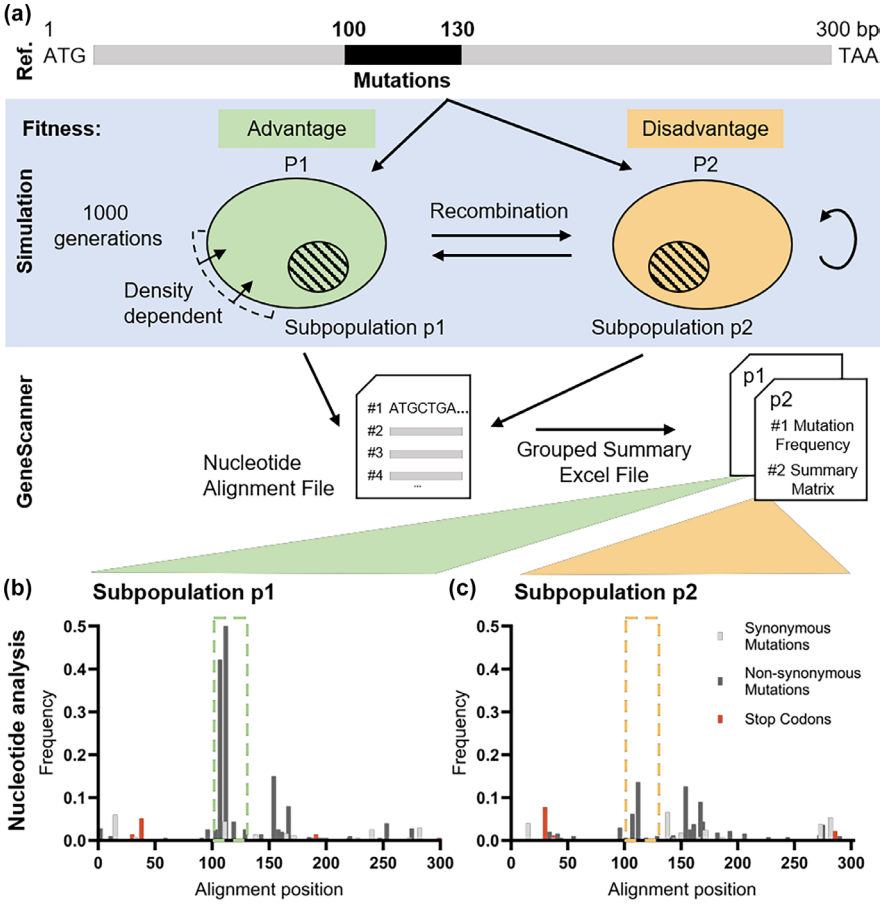

Fig. 3. GeneScanner reliably detects sequence variation from *in vitro* sequence evolution under differential selection. (a) Simulated sequence evolution produced divergent outcomes between two subpopulations exposed to opposing selection pressures. (b, c) Using the original reference sequence, GeneScanner detected only scattered minor variation across the locus in p2, whereas p1 exhibited concentrated synonymous and non-synonymous substitutions within the 100–130 bp window. These differences were captured for both subpopulations, demonstrating the tool's ability to resolve selection-driven divergence.

non-synonymous and indels) and by extending analysis to protein sequences. It also supports dataset grouping to facilitate the identification of population-specific polymorphisms and allows the use of custom reference sequences rather than defaulting to the first sequence in the alignment. Moreover, GeneScanner provides a user-friendly interface in addition to a command-line option, making it accessible to users with varying levels of computational expertise. These features collectively offer greater flexibility for comparative analyses.

Together, these analyses demonstrate the utility of GeneScanner for detecting, classifying and contextualizing genetic variation in bacterial populations. Using a controlled synthetic dataset, the workflow reproduced expected mutation patterns, and in simulations of *in silico* sequence evolution across different subpopulations, it accurately traced selection-driven divergence. This proof-of-concept evaluation established the four-stage GeneScanner pipeline, from alignment parsing to mutation classification and summary statistic generation. Building on this foundation, GeneScanner was used in three case studies spanning distinct bacterial species, phenotypes and research contexts.

Case study 1: genetic variation of DNA replication genes is closely linked to ciprofloxacin resistance

Antibiotic resistance is a major global health concern, with *E. coli* being among the most problematic pathogens [59]. The fluoroquinolone antibiotic ciprofloxacin is a commonly used treatment that targets essential enzymes involved in bacterial DNA replication, specifically DNA gyrase and topoisomerase IV [60, 61]. However, *E. coli* can rapidly develop resistance to ciprofloxacin through mutations in genes linked to DNA replication [62]. In this case study, we used GeneScanner to highlight the most common amino acid substitutions associated with ciprofloxacin resistance in a population of *E. coli* isolates (n=1,509). Using GeneScanner, we analysed the DNA gyrase subunits A and B (GyrA and GyrB) and the topoisomerase IV subunits A and B (ParC and ParE) [62], across ciprofloxacin susceptible (n=1,229) and resistant (n=280) isolates (Fig. 4) . To ensure accurate detection of group-specific

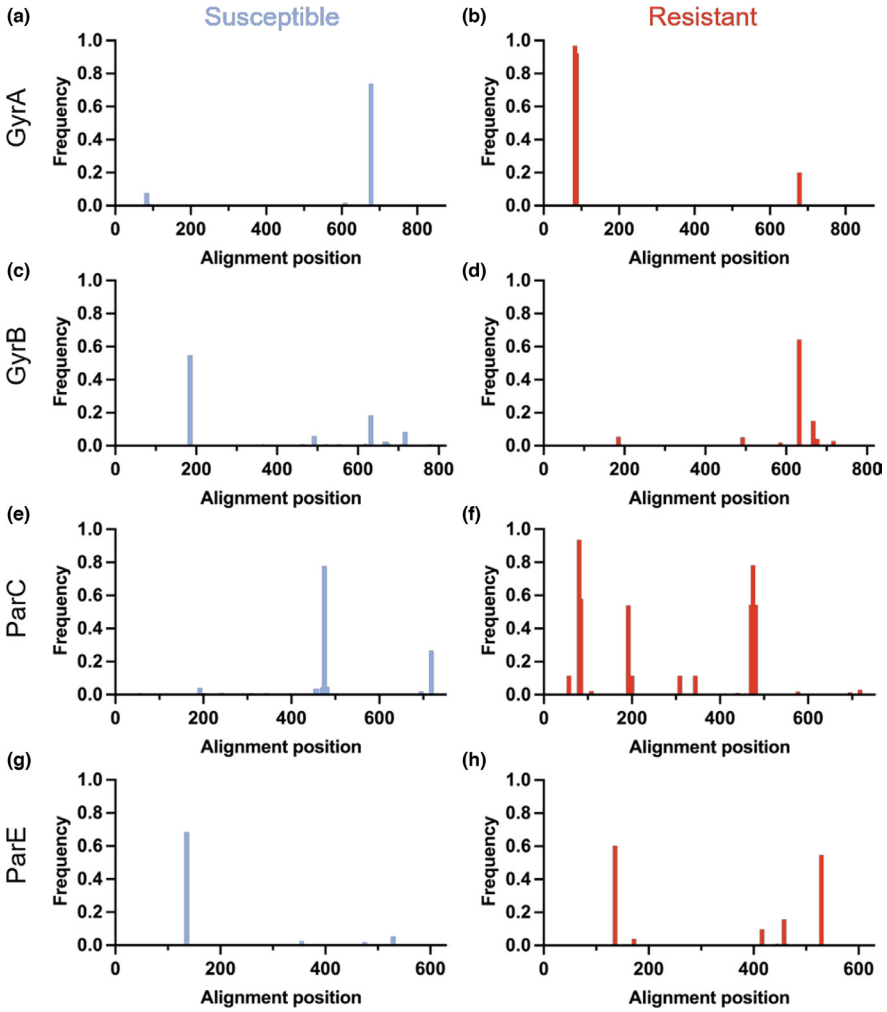

Fig. 4. Ciprofloxacin resistance is linked to genetic variation in an *E. coli* isolate population. Genetic variation of ciprofloxacin resistance-related proteins was analysed in a population of 1,509 *E. coli* isolates. Amino acid changes at each protein alignment position for GyrA (a+b), GyrB (c+d), ParC (e+f) and ParE (g+h) are given in blue or red to indicate the ciprofloxacin susceptible (n=1,229) and ciprofloxacin resistant (n=280) subpopulations, respectively.

variations, GeneScanner performs nucleotide and amino acid alignments before the populations are divided into susceptible and resistant groups, allowing the tool to accommodate insertions that may be unique to one group, and the resulting difference in alignment. Comparing the ProteinAnalysis sheets between groups, we identified several well-known amino acid changes at high frequencies in the resistant population. For example, over 97% of ciprofloxacin-resistant isolates had GyrA substitutions at S83 and ~92% at D87 [62]. However, ~8% of the susceptible isolates also carried the S83 substitution, indicating that this change on its own may only confer small increases in resistance. Similarly, known resistance-associated ParC substitutions, such as S57, S80 and E84, were clearly overrepresented in our resistant population, suggesting that most isolates carried several resistance mutations. Although substitutions at GyrA position D678 (to A or E) have previously been described in fluoroquinolone-resistant *E. coli*, experimental studies have demonstrated that it is neutral with respect to resistance [63, 64]. Consistent with this, we observed the substitutions in both susceptible and resistant isolates in our dataset, with a high frequency among susceptible isolates (Fig. 4a,b). We further identified ParE^{I529L} in 153 resistant isolates, a substitution previously associated with *E. coli* ST131 [65]. Interestingly, we also identified ParC substitution A192V in 151 of these isolates by comparing the ProteinMatrix, indicating co-variation between these otherwise well-conserved genes, and further highlighting the utility of GeneScanner.

Case study 2: vitamin B5 biosynthesis genes are conserved in *C. jejuni* isolated from cattle

C. jejuni is a leading cause of bacterial gastroenteritis in humans, most often transmitted through the consumption of contaminated food, especially poultry [66–68]. *C. jejuni* colonizes multiple hosts, including chickens, cattle and wild birds, and distinct lineages frequently display host preference, consistent with niche specialisation through adaptation. One such adaptation involves the pantothenate (vitamin B5) biosynthesis pathway encoded by the *panBCD* locus. The first formal bacterial GWAS [22] revealed

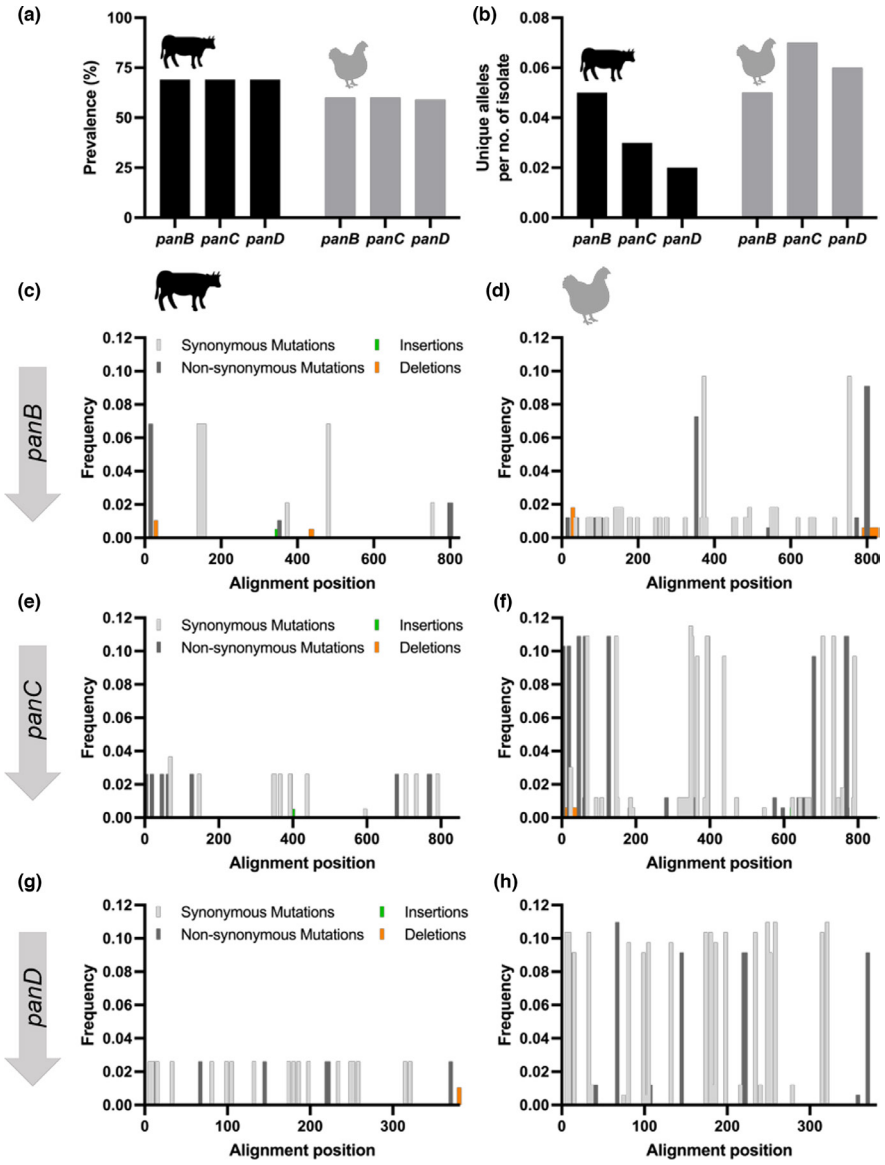

Fig. 5. Conservation and diversity of vitamin B5 biosynthesis genes in *C. jejuni* from cattle and chickens. (a) The *panBCD* locus is present in a higher proportion of cattle isolates compared to chicken isolates. (b) Chicken-associated isolates harbour a greater number of unique *panBCD* alleles relative to cattle isolates. (c–g) Patterns of genetic variation across the *panBCD* locus show reduced diversity in cattle isolates compared with chicken isolates.

that host association signals in *C. jejuni* arose from two major forms of genetic variation. First, genes within the primary host-associated region were more commonly absent from chicken isolates. Second, when present, their sequences differed between cattle and chicken isolates, with cattle alleles exhibiting reduced homologous sequence variation, consistent with gene conservation. Accordingly, isolates from cattle grew better in low vitamin B5 broth than isolates from chickens [22].

We used GeneScanner in a second case study to validate these findings in a larger collection of *C. jejuni* ST-45 complex isolates, comprising 277 from cattle and an equal number from chickens (Fig. 5). Consistent with previous results [22], the *panBCD* locus was more common in cattle isolates (69%) than in those from chickens (60%), as indicated on the output sheet. Additionally, chicken-associated isolates generally exhibited a greater diversity of unique alleles for the *panBCD* locus (Fig. 5b) . This pattern was confirmed by GeneScanner, which confirmed the elevated overall genetic variation among chicken isolates (Fig. 5c-h). For instance, although the number of unique *panB* alleles was similar between cattle and chicken isolates, the 826 nt *panB* alignment contained 18 variable sites in isolates from cattle, whereas chicken isolates had 80. A similar trend was observed for *panC,* with 21 vs. 88 variable sites, with only *panD* showing more comparable values with 26 vs. 35

variable sites in cattle and chicken isolates, respectively. These statistics are readily accessible in the summary statistics tab of the output files.

Overall, GeneScanner here refined observations from a GWAS [22] by analysing host-associated variation and conserved coding regions at the previously identified loci. This higher-resolution view confirmed the enrichment of *panBCD* in cattle-associated lineages while revealing greater allelic diversity in chicken-associated lineages. Although lineage effects were not explicitly controlled for, GeneScanner provided additional insights into host-associated patterns of variation in *C. jejuni* populations.

Case study 3: intergenic variation suggests strain-level modulation of biofilm expression

The third GeneScanner case study focused on cross-species comparison of a non-coding but significant regulatory region in staphylococci. The intercellular adhesion (*ica*) locus has an important role in biofilm formation in *S. aureus* [38, 54, 69], contributing to persistent infections and treatment failure. The locus comprises *icaADBC* genes, which encode enzymes involved in polysaccharide intercellular adhesin production, and the divergently transcribed repressor *icaR*. *icaADBC* expression is regulated by multiple elements beyond canonical transcription and translation initiation sites. Hence, the *icaR_icaA* intergenic region contains binding sites for several regulatory proteins, including the repressors IcaR [55], TcaR [56, 57] and Rob [58]. The presence of the 163-nt-long intergenic region is highly conserved across a diverse *S. aureus* isolate population [70], but despite nucleotide-resolution analysis of regulator binding sites in model strains, the genetic variation of the *icaR_icaA* intergenic region remains understudied.

GeneScanner nucleotide analysis revealed multiple loci with high levels of genetic variability within the *ica* promoter region, particularly near the start codons of *icaR* and *icaA*, as well as within the IcaR binding site (Fig. 6). Notably, we also observed variation within the Rob binding site, including modifications and complete loss of the TATTT motif (Fig. 6, orange box), which is essential for Rob recognition and binding [58]. Such changes are likely to impair Rob-mediated repression, leading to increased *ica* expression and enhanced biofilm formation in the laboratory [58], and are associated with mucoid *S. aureus* isolates from patients with cystic fibrosis [71, 72]. These patterns of variation suggest the potential for strain-specific modulation of *ica* operon activity and biofilm production.

Consistent with previous reports, the *ica* operon was absent in more distantly related staphylococcal species [54]. In contrast, 482 out of 1,000 *S. epidermidis* isolates harboured the *ica* operon. However, GeneScanner summary statistics revealed only five variable sites within the *icaR_icaA* intergenic region, compared to 58 in *S. aureus*. This indicates a markedly higher degree of regulatory intergenic variation in *S. aureus*. Together, these findings demonstrate the utility of GeneScanner for interspecies comparison of non-coding regions, providing insights into the evolution of regulatory elements and their potential impact on phenotypic diversity.

DISCUSSION

Extensive bacterial genome datasets provide major opportunities to study microbial diversity, evolution and adaptation [73]. However, connecting mutations discovered in controlled laboratory experiments with the naturally occurring variation present in bacterial populations remains challenging [16, 36]. This methodological disconnect is particularly pronounced for researchers without extensive bioinformatics expertise, as most available tools require complex computational skills and

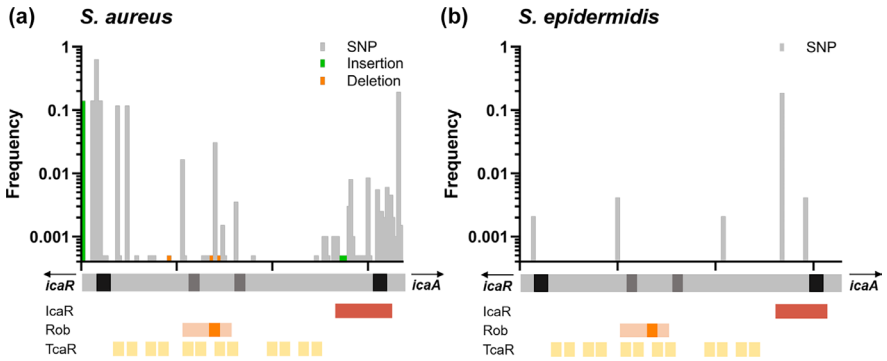

Fig. 6. Genetic variation at the non-coding regulatory regions of the biofilm-associated *ica* operon in *S. aureus* and *S. epidermidis*. SNPs (grey), insertions (green) and deletions (orange) are indicated along the nucleotide alignment (grey bar below graph) of the non-coding region between *icaR* and *icaA* in *S. aureus* (a) and *S. epidermidis* (b). Graphical representation of regulatory elements to scale: black, Shine–Dalgarno sequence for *icaR* and *icaA*, respectively; dark grey, likely −35 and −10 regions for transcription initiation of *icaADBC* operon; red, IcaR-binding site; orange, Rob-binding site with TATTT motif; yellow, TcaR-binding sites.

resources. GeneScanner is designed to bridge this gap by offering an accessible and powerful platform for analysing gene- and protein-level variation of specific loci across large bacterial genome collections via the PubMLST database or when used as a locally installed programme. The tool efficiently identifies both nucleotide variation and protein-altering amino acid changes in genes and proteins associated with phenotypic traits, including antibiotic resistance and pathogen surveillance. Demonstrating the functionality and broad applicability of GeneScanner, we applied it to designed and simulated datasets as well as three distinct bacterial datasets encompassing different species, phenotypes and gene loci. In each case, the tool successfully captured nucleotide-level and protein-altering changes, illustrating GeneScanner's utility as a comprehensive and versatile platform for investigating genetic variation across a wide range of bacterial systems, traits and experimental contexts.

As with most genome analysis tools, the reliability of results depends heavily on data quality and sampling design. While GeneScanner includes quality control, it does not explicitly correct for population structure. Poor assemblies, incomplete genomes or incorrect annotations may introduce artefacts. Furthermore, biassed sampling can distort allele frequencies; for example, overrepresentation of certain lineages can generate genotype–phenotype associations driven by population structure rather than true causality. To address this, analyses can incorporate population subsampling [29, 74], linear mixed models [75, 76] and phylogenetic approaches [77] to account for the clonal structure of the population. For example, identified genetic variants can be mapped onto a phylogenetic tree to assess whether they cluster within a particular lineage or have emerged independently on multiple branches [37]. Additionally, using high-quality, well-annotated genomes and including metadata on isolation source, geographic origin and collection date can help control for potential confounding factors.

Uncovering the structural and functional impacts of naturally occurring genetic variation will be essential for advancing our understanding of microbial adaptation and phenotype variation. GeneScanner provides a powerful starting point by efficiently identifying candidate SNPs and amino acid changes across large bacterial genome collections. Of course, understanding how these alterations influence protein conformation, stability or activity will require additional analysis. Integrating structural bioinformatics methods, such as domain conservation analysis [78, 79], structure-function prediction [80–82] and molecular dynamics [83, 84], alongside experimental validation, could achieve this. Advancing microbiology in this direction will move future research beyond merely detecting variation, towards mechanistically explaining how it drives microbial phenotypes, maximizing the impact of bacterial genomic data on surveillance, outbreak response and therapeutic development.

Funding information
This work is supported by an Ineos Oxford Institute grant; Wellcome Trust grant 088786/C/09/Z and UK Research and Innovation grants MR/L015080/1, MR/V001213/1, MR/S009264/1 and MR/T030062/1 to S.K.S. C.M.K. was funded by an Edward Penley Abraham Junior Research Fellowship from Linacre College, University of Oxford; a University of Oxford Medical Sciences Internal Fund Pump-Priming Award (0015060); and a Biotechnology and Biological Sciences Research Council (BBSRC) Fellowship (UKRI905). K.A.J. was funded by a Wellcome Trust Biomedical Resource Grant (218205/Z/19/Z). P.S.R. is funded through an Ineos Oxford Institute (IOI) DPhil Studentship. The funders had no role in study design, data collection and interpretation or the decision to submit the work for publication.

Acknowledgements
Icons in Fig. 5 were designed by PLANBstudio and sourced from flaticon.com.

Author contributions
S.K.S. and C.M.K. conceptualized the study. S.K., P.S.R., C.M.K., S.K.S. and B.A.S. designed the methodology. S.K., P.S.R., B.A.S. and K.J. implemented the work. C.M.K., S.K., P.S.R. and S.K.S. analysed the data. C.M.K., S.K., P.S.R., B.A.S. and K.J. curated the data. C.M.K., S.K., P.S.R. and S.K.S. wrote the original draft, and all authors reviewed and edited the manuscript. C.M.K. and S.K. created the visualizations. S.K.S. and C.M.K. supervised and administered the project and acquired the funding.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Kobras CM, Ko S, Raikwar PS, Aguilar-Sanjuan B, Jolley KA, *et al*. GeneScanner: profiling genetic variation across bacterial populations. 2026.
2. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, *et al*. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2006;2:2006.0008.
3. de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, *et al*. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Mol Syst Biol* 2008;4:174.
4. Porwollik S, Santiviago CA, Cheng P, Long F, Desai P, *et al*. Defined single-gene and multi-gene deletion mutant collections in *Salmonella enterica* sv Typhimurium. *PLoS One* 2014;9:e99820.
5. Koo B-M, Kritikos G, Farelli JD, Todor H, Tong K, *et al*. Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Syst* 2017;4:291–305..
6. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* 2009;6:767–772.
7. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, *et al*. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res* 2009;19:2308–2316.
8. Gawronski JD, Wong SMS, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci USA* 2009;106:16422–16427.
9. Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, *et al*. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* 2009;6:279–289.
10. Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* 2013;11:435–442.

11. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, *et al*. A decade of advances in transposon-insertion sequencing. *Nat Rev Genet* 2020;21:526–540.
12. Conrad TM, Lewis NE, Palsson BØ. Microbial laboratory evolution in the era of genome-scale science. *Mol Syst Biol* 2011;7:509.
13. Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, *et al*. Spatiotemporal microbial evolution on antibiotic landscapes. *Science* 2016;353:1147–1151.
14. Lenski RE. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J* 2017;11:2181–2194.
15. Maeda T, Iwasawa J, Kotani H, Sakata N, Kawada M, *et al*. High-throughput laboratory evolution reveals evolutionary constraints in *Escherichia coli*. *Nat Commun* 2020;11:5970.
16. Ascensao JA, Desai MM. Experimental evolution in an era of molecular manipulation. *Nat Rev Genet* 2026;27:81–95.
17. O'Cathail C, Ahamed A, Burgin J, Cummins C, Devaraj R, *et al*. The European nucleotide archive in 2024. *Nucleic Acids Res* 2025;53:D49–D55.
18. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, *et al*. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* 2025;53:D20–D29.
19. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. The enterobase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 2020;PubMed PMID:138.
20. Dyer NP, Päuker B, Baxter L, Gupta A, Bunk B, *et al*. EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens. *Nucleic Acids Res* 2025;53:D757–D762.
21. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
22. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, *et al*. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad Sci USA* 2013:11923–11927.
23. Yahara K, Méric G, Taylor AJ, de Vries SP, Murray S, *et al*. Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environ Mirobiol* 2016.
24. Taylor AJ, Yahara K, Pascoe B, Ko S, Mageiros L, *et al*. Epistasis, core-genome disharmony, and adaptation in recombining bacteria. *mBio* 2024;15.
25. Hwang W, Yong JH, Min KB, Lee KM, Pascoe B, *et al*. Genome-wide association study of signature genetic alterations among *Pseudomonas aeruginosa* cystic fibrosis isolates. *PLoS Pathog* 2021;17:e1009681.
26. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, *et al*. Predicting the virulence of MRSA from its genome sequence. *Genome Res* 2014;PubMed PMID:839–849.
27. Laabei M, Uhlemann A-C, Lowy FD, Austin ED, Yokoyama M, *et al*. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus*. *PLoS Biol* 2015;13:e1002229.
28. Berthenet E, Yahara K, Thorell K, Pascoe B, Meric G, *et al*. A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biol* 2018;16:84.
29. Méric G, Mageiros L, Pensar J, Laabei M, Yahara K, *et al*. Disease-associated genotypes of the commensal skin bacterium *Staphylococcus epidermidis*. *Nat Commun* 2018;9:5034.
30. Monteith W, Pascoe B, Mourkas E, Clark J, Hakim M, *et al*. Contrasting genes conferring short- and long-term biofilm adaptation in *Listeria*. *Microb Genom* 2023;9:001114.
31. Pascoe B, Méric G, Murray S, Yahara K, Mageiros L, *et al*. Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in *Campylobacter jejuni*. *Environ Microbiol* 2015;17:4779–4789.
32. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, *et al*. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet* 2013;45:1183–1189.
33. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, *et al*. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet* 2014;10:e1004547.
34. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, *et al*. GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat Commun* 2019;10:2128.
35. Ma KC, Mortimer TD, Duckett MA, Hicks AL, Wheeler NE, *et al*. Increased power from conditional bacterial genome-wide association identifies macrolide resistance mutations in *Neisseria gonorrhoeae*. *Nat Commun* 2020;11:1.
36. Kobras CM, Fenton AK, Sheppard SK. Next-generation microbiology: from comparative genomics to gene function. *Genome Biol* 2021;22:123.
37. Kobras CM, Monteith W, Somerville S, Delaney JM, Khan I, *et al*. Loss of Pde1 function acts as an evolutionary gateway to penicillin resistance in *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A* 2023;120:e2308029120.
38. Rudolph E, Li S, Aguilar-Sanjuan B, Ko S, Raikwar PS, *et al*. Functional and comparative genomic characterization of biofilm formation in *Staphylococcus aureus*. *Biofilm* 2026;11:100341.
39. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, *et al*. *SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016;2:e000056.
40. Lindenbaum P. JVarkit: java-based utilities for bioinformatics. *Bytes* 2015;347403.
41. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
42. Lischer HEL, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 2012;28:298–299.
43. Swofford DL. Paup*. Phylogenetic Analysis Using Parsimony (* and Other Methods).; 2002. https://cir.nii.ac.jp/crid/1370285712570623744 [accessed 26 February 2025].
44. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25:1422–1423.
45. McKinney W. Data Structures for Statistical Computing in Python. In. Austin, Texas; 2010 [cited 2025 Oct 14]. p. 56–61; (n.d.). https://doi.curvenote.com/10.25080/Majora-92bf1922-00a doi:10.25080/Majora-92bf1922-00a
46. PyPI [Internet]. xlsxwriter; 2025. https://pypi.org/project/xlsxwriter/ [accessed 18 June 2025].
47. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* 2002;30:3059–3066.
48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
49. Messer PW. SLiM: simulating evolution with selection and linkage. *Genetics* 2013;194:1037–1039.
50. Cury J, Haller BC, Achaz G, Jay F. Simulation of bacterial populations with SLiM. *Peer Commun J* 2022;2.
51. Stephan W. Genetic hitchhiking versus background selection: the controversy and its implications. *Philos Trans R Soc Lond B Biol Sci* 2010;365:1245–1253.
52. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, *et al*. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res* 2017;27:1437–1449.

53. Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:1–11.
54. Cramton SE, Gerke C, Schnell NF, Nichols WW, Götz F. The intercellular adhesion (*ica*) locus is present in *Staphylococcus aureus* and is required for biofilm formation. *Infect Immun* 1999;67:5427–5433.
55. Conlon KM, Humphreys H, O'Gara JP. *icaR* encodes a transcriptional repressor involved in environmental regulation of *ica* operon expression and biofilm formation in *Staphylococcus epidermidis*. *J Bacteriol* 2002;184:4400–4408.
56. Jefferson KK, Pier DB, Goldmann DA, Pier GB. The teicoplanin-associated locus regulator (TcaR) and the intercellular adhesin locus regulator (IcaR) are transcriptional inhibitors of the *ica* locus in *Staphylococcus aureus*. *J Bacteriol* 2004;186:2449–2456.
57. Chang Y-M, Jeng W-Y, Ko T-P, Yeh Y-J, Chen CK-M, *et al*. Structural study of TcaR and its complexes with multiple antibiotics from *Staphylococcus epidermidis*. *Proc Natl Acad Sci U S A* 2010;107:8617–8622.
58. Yu L, Hisatsune J, Hayashi I, Tatsukawa N, Sato'o Y, *et al*. A novel repressor of the ica locus discovered in clinically isolated super-biofilm-elaborating *Staphylococcus aureus*. *mBio* 2017;8:e02282-16.
59. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, *et al*. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022;399:629–655.
60. Khodursky AB, Zechiedrich EL, Cozzarelli NR. Topoisomerase IV is a target of quinolones in *Escherichia coli*. *Proc Natl Acad Sci U S A* 1995;92:11801–11805.
61. Drlica K. Mechanism of fluoroquinolone action. *Curr Opin Microbiol* 1999;2:504–508.
62. Hopkins KL, Davies RH, Threlfall EJ. Mechanisms of quinolone resistance in escherichia coli and salmonella: recent developments. *Int J Antimicrob Agents* 2005;25:358–373.
63. Fisher LM, Lawrence JM, Josty IC, Hopewell R, Margerrison EE, *et al*. Ciprofloxacin and the fluoroquinolones. New concepts on the mechanism of action and resistance. *Am J Med* 1989;87:2S–8S.
64. Heisig P, Schedletzky H, Falkenstein-Paul H. Mutations in the gyra gene of a highly fluoroquinolone-resistant clinical isolate of *Escherichia coli*. *Antimicrob Agents Chemother* 1993;37:696–701.
65. Paltansing S, Kraakman MEM, Ras JMC, Wessels E, Bernards AT. Characterization of fluoroquinolone and cephalosporin resistance mechanisms in *Enterobacteriaceae* isolated in a Dutch teaching hospital reveals the presence of an *Escherichia coli* ST131 clone with a specific mutation in *parE*. *J Antimicrob Chemother* 2013;68:40–45.
66. Sheppard SK, Dallas JF, Strachan NJC, MacRae M, McCarthy ND, *et al*. *Campylobacter* genotyping to determine the source of human infection. *Clin Infect Dis* 2009;48:1072–1078.
67. Arning N, Sheppard SK, Bayliss S, Clifton DA, Wilson DJ. Machine learning to predict the source of campylobacteriosis using whole genome data. *PLoS Genet* 2021;17:e1009436.
68. Pascoe B, Futcher G, Pensar J, Bayliss SC, Mourkas E, *et al*. Machine learning to attribute the source of *Campylobacter* infections in the United States: a retrospective analysis of national surveillance data. *J Infect* 2024;89:106265.
69. O'Gara JP. Ica and beyond: biofilm mechanisms and regulation in *Staphylococcus epidermidis* and *Staphylococcus aureus*. *FEMS Microbiol Lett* 2007;270:179–188.
70. Young BC, Wu C-H, Charlesworth J, Earle S, Price JR, *et al*. Antimicrobial resistance determinants are associated with *Staphylococcus aureus* bacteraemia and adaptation to the healthcare environment: a bacterial genome-wide association study. *Microb Genom* 2021;7:000700.
71. Schwartbeck B, Birtel J, Treffon J, Langhanki L, Mellmann A, *et al*. Dynamic in vivo mutations within the ica operon during persistence of *Staphylococcus aureus* in the airways of cystic fibrosis patients. *PLoS Pathog* 2016;12:e1006024.
72. Lennartz FE, Schwartbeck B, Dübbers A, Große-Onnebrink J, Kessler C, *et al*. The prevalence of *Staphylococcus aureus* with mucoid phenotype in the airways of patients with cystic fibrosis-A prospective study. *Int J Med Microbiol* 2019;309:283–287.
73. Hall N. Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* 2007;210:1518–1525.
74. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, *et al*. Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. *Nat Commun* 2021;12:765.
75. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, *et al*. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun* 2016;7:12797.
76. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, *et al*. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* 2016;1:16041.
77. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput Biol* 2018;14:e1005958.
78. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, *et al*. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 2007;35:D237–40.
79. Yang M, Derbyshire MK, Yamashita RA, Marchler-Bauer A. NCBI's conserved domain database and tools for protein domain analysis. *Curr Protoc Bioinformatics* 2020;69.
80. Yang J, Yan R, Roy A, Xu D, Poisson J, *et al*. The I-TASSER suite: protein structure and function prediction. *Nat Methods* 2015;12:7–8.
81. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, *et al*. Improved protein structure prediction using potentials from deep learning. *Nature* 2020;577:706–710.
82. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–589.
83. Cheatham TE, Kollman PA. Molecular dynamics simulation of nucleic acids. *Annu Rev Phys Chem* 2000;51:435–471.
84. Khalid S, Brandner AF, Juraschko N, Newman KE, Pedebos C, *et al*. Computational microbiology of bacteria: advancements in molecular dynamics simulations. *Structure* 2023;31:1320–1327.