Bacterial GWAS: A Comprehensive Assessment of Challenges, Methods, and Alternatives

Student: Sander Vermeulen

Examiner: dr. A.L. Zomer

Second reviewer: dr. A.C. Schürch

Plain language summary

Het ontdekken van welke genetische afwijkingen invloed hebben op fysieke eigenschappen (bijvoorbeeld oogkleur, haarkleur en lengte) en ziektes (bijvoorbeeld kanker, diabetes en kleurenblindheid) is een van de belangrijkste doelen van de genetica. Een onderzoeksmethode die genetische afwijkingen aan fysieke eigenschappen kan linken is een *genome-wide association study*, of GWAS in het kort. Tijdens een GWAS vergelijken wetenschappers de genetische eigenschappen van mensen die een bepaalde fysieke eigenschap of ziekte hebben met de genetische eigenschappen van mensen die dit niet hebben. Door de genetische eigenschappen van duizenden mensen met elkaar te vergelijken, kunnen genetische eigenschappen ontdekt worden die vaker voorkomen bij mensen die een bepaalde fysieke eigenschap of ziekte hebben. Dit hoeft overigens nog niet te betekenen dat mensen die een bepaalde genetische eigenschap hebben ook echt die fysieke eigenschap of ziekte hebben of in de toekomst krijgen, vaak geeft dit alleen aan dat er een verhoogde kans bestaat. Andere factoren zoals leefstijl, de omgeving waarin je je bevindt of andere genetische eigenschappen kunnen de kans verhogen of verlagen voor het ontwikkelen van fysieke eigenschappen of ziektes.

Net zoals bij mensen kunnen we GWAS ook voor bacteriën gebruiken. Het principe blijft hetzelfde; wetenschappers vergelijken de genetische eigenschappen van twee groepen bacteriën, één groep die een bepaalde eigenschap heeft en één groep die deze eigenschap niet heeft. Het doel is dan, net zoals bij menselijke GWAS, om genetische eigenschappen te vinden die wel voorkomen bij de groep met het onderzochte eigenschap maar niet bij de groep die dit eigenschap niet heeft. De eigenschappen die voor bacteriële GWAS interessant zijn bestaan bijvoorbeeld uit genen die resistentie tegen antibiotica veroorzaken, genen die zorgen dat bacteriën stoffen aanmaken waardoor mensen ziek worden en genen die zorgen dat bacteriën kunnen overleven in organismen. Als we weten hoe deze eigenschappen ontstaan, kunnen wetenschappers medicijnen ontwikkelen die specifiek gericht zijn op hoe deze genetische eigenschappen werken.

De technieken en computerprogramma's die gebruikt worden bij menselijke GWAS kunnen niet gebruikt worden voor bacteriële GWAS. Dit komt door de genetische verschillen tussen mensen en bacteriën. Daarom zijn er nieuwe technieken en computerprogramma's ontwikkeld waarmee wetenschappers wel GWAS met bacteriën kunnen uitvoeren. Deze computerprogramma's werken bijvoorbeeld met wiskundige modellen, bacteriële stambomen en *machine learning*.

Naast bacteriële GWAS zijn er nog andere methoden die gebruikt kunnen worden om de link tussen bepaalde genetische eigenschappen en bijvoorbeeld antibiotica resistentie aan te tonen. Deze methoden die we hier behandelen heten *Tn-seq* en *genome-scale metabolic models* (GSMMs). Tn-seq werkt door veel kleine stukjes DNA in het DNA van bacteriën te knippen en plakken, waardoor deze kleine stukjes DNA tussen genen komen die nodig zijn voor het groeien op bepaalde voedingsbodems. Door de groei van de aangepaste bacteriën te vergelijken met niet-aangepaste bacteriën kunnen wetenschappers ontdekken welke genen ervoor zorgen dat bacteriën kunnen groeien op de voedingsbodem.

GSMMs werken door wiskundige modellen van bacteriële stofwisseling te maken, waardoor gaten tussen voedingsstoffen kunnen worden opgevuld door te voorspellen welke genen ervoor kunnen zorgen dat een bepaalde voedingsstof naar een andere kan worden omgezet. Ook kan je in deze modellen bepaalde genen uitzetten, waardoor je met de computer de effecten kan voorspellen op de bacterie wanneer deze dit gen niet zou hebben.

Abstract

Genome-wide association studies (GWAS) have proven to be a successful method for identifying associations between human genotypes and phenotypes. Due to advances in sequencing technologies and the subsequent growth of bacterial datasets, bacterial GWAS is increasingly becoming a viable research method for identifying bacterial genotype-phenotype associations. However, bacterial GWAS cannot be performed using established methods used in human GWAS due to genomic differences. Specialized software to perform bacterial GWAS has been developed, utilizing regression models, phylogenetic trees, and machine learning to overcome the unique genomic challenges. Here, we will discuss these challenges of bacterial GWAS, the software methods that have been developed and our recommendations on their usage, and discuss alternative methods for identifying genotype-phenotype associations in bacteria.

Introduction

One of the major goals of genetics is linking phenotypic traits, such as eye color, height, and genetic diseases to the genes that affect them. Genome-wide association studies (GWAS) enable researchers to determine the underlying genetic variations of the phenotype of interest by analyzing a significant number of genetic variants (usually 10^3 to 10^6 for human GWAS) (Bush & Moore, 2012). The genetic variants examined for human GWAS are usually single nucleotide polymorphisms (SNPs), captured with genotyping microarrays designed to assay SNPs over the entire human genome, or identified using whole genome sequencing (WGS) data (Hasin et al., 2017).

In 2005, Klein et al. conducted the first successful GWAS, which established a link between age-related macular degeneration (AMD) and the *CFH* gene (Klein et al., 2005). Their study discovered that individuals who were homozygous for the risk allele had a 7.4-fold increased risk of developing AMD. Since then, GWAS successfully identified hundreds of other genetic variants for human diseases and traits, such as 45 loci associated with lung cancer (Bossé & Amos, 2018) and a variety of loci associated with skin and hair pigmentation (Pavan & Sturm, 2019).

Due to recent advances in DNA sequencing technologies, the cost-effectiveness of WGS of bacterial genomes has significantly improved, which resulted in a steep growth of available data on bacterial genomes (Kumar et al., 2019). Powered by this growth of data, bacterial GWAS offers exciting new opportunities for discovering the genetic causes of various relevant bacterial phenotypes, such as antimicrobial resistance, virulence, and host specificity. The discoveries of these genetic causes might lead to the identification of targets for vaccine and drug development. Furthermore, bacterial GWAS could also assist with tracking the spread of pathogenic bacteria within populations, for example during nosocomial outbreaks (Power et al., 2017).

Bacterial GWAS has not yet been as successful as human GWAS, which can be contributed to a number of factors (San et al., 2020). First, bacterial genomes are in strong linkage disequilibrium due to clonal reproduction as opposed to sexual reproduction where recombination occurs in every generation. Due to this strong linkage disequilibrium, true causal genetic variants are harder to detect (Chen & Shapiro, 2015; Read & Massey, 2014). Second, mutations occurring in ancestral branches can appear as causal in GWAS along with the true causal variant (Lees, 2017). Finally, designing genotyping microarrays for bacteria is challenging because of the significant genomic variation, even among strains of the same species (McInerney et al., 2017). However, the advent of next generation sequencing (NGS) and subsequent decrease of costs of WGS have overcome this obstacle, but prior to this, bacterial GWAS had an incredibly limited scope. (Read & Massey, 2014).

In this review, we will explore the differences between human GWAS and bacterial GWAS and discuss current methods and development of accompanying software for bacterial GWAS. Additionally, we will discuss the statistical principles of GWAS and how these were altered to consider the unique properties of bacterial GWAS.

Unique challenges of bacterial GWAS

The basic concept of both human GWAS and bacterial GWAS is the same: linking phenotypic traits, discrete (e.g. antibiotic resistance/sensitivity, eye color) or continuous (e.g. height), to genotypic variations. Human GWAS mainly use identification of (a group of) SNPs related to a phenotype of interest (Horwitz et al., 2019), but in bacterial GWAS, using just SNPs will not always account for all genetic variation. The genes in a set of bacterial genomes can be divided into two groups: the core genome, consisting of genes that are shared between all genomes, and the accessory genome, consisting of genes that are only present in a subset of the genomes (Lees et al., 2020). Genes that belong to the core genome are responsible for basic functions, such as cell growth or replication. Genes belonging to the accessory genome are responsible for specialization of niche environments, virulence and antibiotic resistance, among others (Kim et al., 2020). Thus, the presence or absence of certain genes in the accessory genome can have a major impact on the phenotype of bacteria which cannot be detected by SNPs alone.

In addition to the bacterial core- and accessory genome, another factor that has to be taken into consideration in bacterial GWAS methods is linkage disequilibrium. In eukaryotes, recombination of DNA occurs due to crossing-over events of a pair of homologous chromosomes during meiosis (Hillers et al., 2017). Since bacterial reproduction occurs clonally, no crossing-over events take place, resulting in complete linkage disequilibrium across the entire bacterial chromosome, if considering only bacterial reproduction.

Although crossing-over recombination events do not occur in bacteria, other types of recombination events unrelated to reproduction can take place in bacteria. These recombination events are transduction (transportation of DNA fragments from one bacterium to another by a bacteriophage), transformation (the process of incorporating foreign DNA from the environment in the bacterial genome), and conjugation (direct DNA transfer between bacterial cells) (Louha et al., 2021). Like eukaryotic recombination, bacterial recombination breaks the linkage but leaves behind a vastly different linkage pattern (Chen & Shapiro, 2015). In eukaryotes, only the region around a point of reference is in strong linkage disequilibrium, while in bacteria the entire chromosome is in linkage disequilibrium with 'patches' scattered throughout the chromosome from aforementioned bacterial recombination events (Fig. 1) (Chen & Shapiro, 2015). The distinction between eukaryotic and prokaryotic linkage is important for GWAS since genomic variants in regions that are in linkage disequilibrium with the causal genomic variant of a trait will also be detected as statistically significant for the trait. For human GWAS, these regions are often limited to a small region around the true causal genomic variant, but for bacterial GWAS these regions could be on the opposite side of the bacterial chromosome. Consequently, these distant regions obscure the location of the true causal genomic variant.

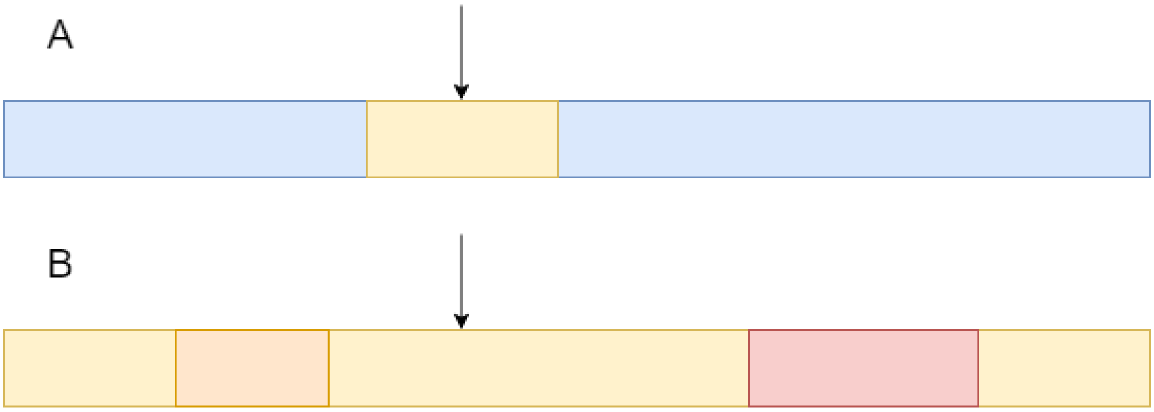

Figure 1. An example of an eukaryotic chromosome (A) and a bacterial chromosome (B). The arrow denotes the point of reference, for example a SNP. On the eukaryotic chromosome, we see that only a small DNA sequence around the point of reference, colored yellow, is in linkage disequilibrium relative to the point of reference. On the bacterial chromosome, we see that the entire chromosome is in linkage disequilibrium relative to the point of reference, with exceptions of the DNA sequences colored orange and red. These DNA sequencing could be genes acquired from for example transduction events.

The third factor, population structure, is closely related to the effects of linkage disequilibrium and clonal reproduction. A *de novo* mutation occurring on an ancestral branch of a subpopulation, which is the causal variant of a certain phenotype, will occur in the majority of subsequent strains due to clonal reproduction (Collins & Didelot, 2018). The true causal mutation will be positively associated with the phenotype, however, all other mutations that occurred on the ancestral branch will be positively associated with the phenotype as well (Lees, 2017) (Fig. 2). Due to strong linkage disequilibrium, these non-causal variants can be scattered over the entire genome.

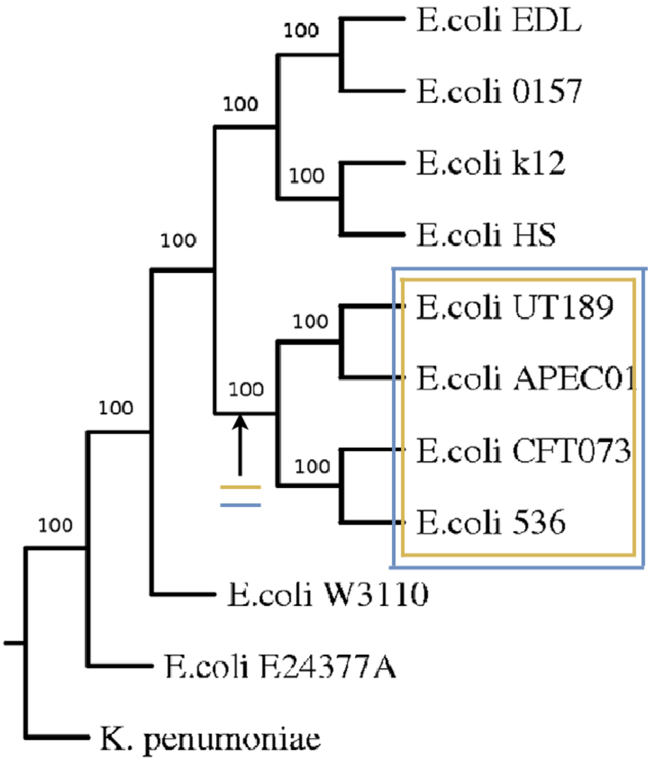

Figure 2. [Image source: (Vishnoi et al., 2010)] Phylogenetic tree of E. coli *showing two mutations on the ancestral branch of four strains. Consider the yellow mutation causative of a certain phenotype and the blue mutation a silent mutation. Since both mutations do not occur in the remaining* E. coli *strains, a naïve GWAS will consider both mutations to be causative of the given phenotype.*

Bacterial GWAS software and methodology

To overcome the challenges presented in the previous segment, various methods and software incorporating these methods were developed. This software can be loosely defined into three categories: phylogenetic, non-phylogenetic, and machine learning-based software (San et al., 2020).

Phylogenetic methods

As the name suggests, phylogenetic methods depend on a phylogenetic tree (either provided by the user or constructed by the software), genetic data (e.g. a file containing gene presence/absence for each bacterial sample) and phenotypic data on each bacterial sample.

One example of a software package that integrates the phylogenetic method is Scoary (Brynildsrud et al., 2016). Scoary uses the gene presence-absence file from Roary (Page et al., 2015) to find associations between the observed phenotype and the pan-genome of the dataset. The software corrects the effects of population structure by implementing a pairwise comparison algorithm. The pairwise algorithm searches the phylogenetic tree to find the maximum number of genotypic and phenotypic contrasting pairs (e.g. positive for a certain gene and trait vs. negative for a certain gene and trait) that do not intersect with a non-contrasting individual. Effectively, finding the maximum number of contrasting pairs determines the minimum number of independent emergences of a certain genotype-phenotype combination. This method solves the problem of finding a strong correlation between a certain gene and trait due to population structure, since the emphasis is no longer on the number of times a trait correlates with a gene but instead the number of times a trait and a gene have emerged independently from each other (Fig. 3). Scoary validates the results using a permutation test, by switching the phenotype labels and calculating the maximum number of contrasting pairs for each permutation. According to the author’s own benchmarks, Scoary was able to correctly identify an association between the *cfr* gene and resistance to the antibiotic linezolid, as well as an association to the two plasmid genes *pinE* and *cueR*. Compared to PLINK (Purcell et al., 2007), which was developed for human GWAS, Scoary performed better in 7/12 power test comparisons using simulated data, equal in three, and slightly worse in two. Scoary has both a command line implementation and a graphical user interface, making it a viable choice for users that are not too familiar with the command line. However, Scoary is not ideal when analyzing continuous traits since these require binning into distinct categories.

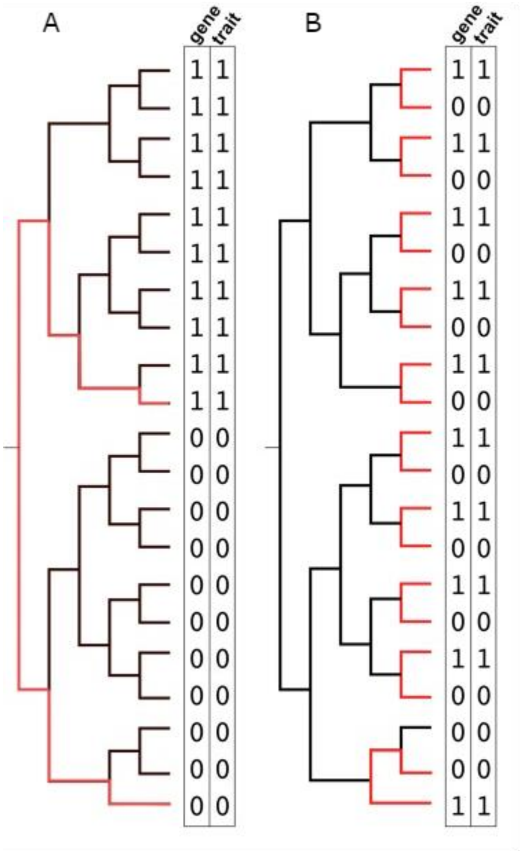

Figure 3. Example of pairwise comparisons. Both phylogenetic trees have the same number of 1-1 and 0-0 gene-trait associations, so a naïve method (e.g. a Fisher's exact test) would assume that both gene-trait associations would be highly significant (p-value 2.8E-6). However, phylogenetic tree A suggests that lineage-specific factors could play a role in the association between the gene and the trait, which violates the randomness requirement of the Fisher's exact test. For phylogenetic tree A the most parsimonious scenario is a single gene and trait introduction/loss in the root branch. This is illustrated by the pairwise comparison algorithm, which finds one contrasting gene-trait pair. Phylogenetic tree B has ten contrasting gene-trait pairs, which indicates a minimum of ten transitions between 1-1 and 0-0 gene-trait associations in the evolutionary history. This makes the association between this particular gene-trait association more convincing compared to phylogenetic tree A (Brynildsrud et al., 2016).

Another example of a software package that uses phylogeny is TreeWAS (Collins & Didelot, 2018). TreeWAS is able to construct the phylogenetic tree, however the authors recommend integrating ClonalFrameML (Didelot & Wilson, 2015) into TreeWAS and constructing the phylogenetic tree with ClonalFrameML. ClonalFrameML can produce more accurate phylogenetic trees, especially when analyzing species with a high recombination rate. Using the phylogenetic tree and the Fitch parsimony algorithm, the distribution of homoplasy in the tree is calculated. Using the input data, a simulated dataset is created to acquire a null distribution, after which the associations between genotype and phenotype are compared to those in the simulated dataset to identify significant hits.

The authors benchmarked their software on data from *Neisseria meningitidis* to identify penicillin resistance and invasive disease-associated variants. Penicillin resistance was measured both discretely (resistant vs. susceptible, categorized according to the minimum inhibitory concentration (MIC)) and continuously (by defining the ranks of the MIC values). For the discrete analysis, no significant genes were identified in the accessory genome, but 162 significant SNPs were identified in the core genome. These SNPs were all located in a known penicillin resistance gene (*penA)* (Matsuda et al., 2021). For the continuous analysis, 30 significant SNPs were identified. The majority of these SNPs were located in the *penA* gene, but the SNPs were also identified in three additional genes. These genes were not essential for antibiotic resistance, however, in the presence of antibiotics it has been shown that genes not associated with antibiotic resistance can still give a slight selective advantage (Read & Massey, 2014). For the analysis of invasive disease-associated variants, TreeWAS identified 12 genes and 7 SNPs, which the authors confirmed to be associated with invasive disease in *N. meningitidis*. The results conclude that TreeWAS is a powerful software package that is able to identify loci and genes associated with complex phenotypes, however a limit of the software is the implementation of TreeWAS in R. This requires users to have a basic understanding of R to successfully run the software.

Non-phylogenetic methods

Software packages that do not include phylogenetic trees often requires sequencing data (either raw reads or assemblies) to be provided to the software, use some sort of dimensionality reduction to control population structure and use a regression model to evaluate associations between phenotype and genotype.

Pyseer (Lees et al., 2018), an upgrade and reimplementation of SEER (Lees et al., 2016) in Python, calculates a pairwise distance matrix from the supplied genome assemblies using Mash (Ondov et al., 2016) and performs multi-dimensional scaling to control population structure. The program extracts k-mers of variable length from the genome assemblies and analyses their association with the phenotype of interest by fitting each k-mer to a generalized linear model. Alternatively, a linear mixed model can be used that is able to control population structure, but this method is computationally more expensive. Interactive visualizations of the results can be generated using Phandango (Hadfield et al., 2018). Saber & Shapiro benchmarked the linear mixed model of Pyseer using simulated genomes and phenotypes (Saber & Shapiro, 2020). In their benchmark, Pyseer was outclassed by other software on controlling false positives, but performed well when tasked with identifying genetic variants with low effect sizes. The linear mixed model of Pyseer performed about equal to other methods when tasked to detect genetic variants in simulated genomes with a low amount of linkage disequilibrium, but did not perform well when given samples with moderate or high linkage disequilibrium. However, the performance was still comparable to other methods since all software had issues with moderate and high linkage disequilibrium samples. Pyseer is easy to install via conda, but features no graphical user interface, thus the user has to have a basic understanding of the command line to execute the necessary commands.

HAWK (Rahman et al., 2018) uses sequencing reads to count k-mers of length 31 (longest k-mer that can be efficiently analyzed on 64-bit processors) using Jellyfish (Marçais & Kingsford, 2011). Population structure is controlled using a binary matrix and principal component analysis. The software counts the frequency each k-mer appears in the dataset, which are assumed to be Poisson distributed, to calculate the significance of each k-mer. Further correction of population structure and other confounding factors is done using a logistic regression method for discrete phenotypes and linear regression for continuous phenotypes. The k-mers that were found to be associated with the phenotype of interest are assembled using ABySS (Simpson et al., 2009) to acquire a larger sequence for the associated genomic region instead of multiple individual k-mers. These sequences can subsequently be analyzed by mapping them to a high-quality reference genome.

The authors benchmarked their software on data from *Escherichia coli*, with the goal of detecting genetics variants associated with ampicillin resistance. They found 5047 k-mers, which resulted in 16 sequences after assembly, to be significantly associated with ampicillin resistance. The obtained sequences mapped to several *E. coli* strains known for ampicillin resistance, and the strongest associations of the k-mers were found in the *blaTEM-1* gene, which is also known for conferring ampicillin resistance. Both installing and executing HAWK requires basic understanding of the command line, making it not ideal for novice users. HAWK implemented multi-threading support to speed up the analysis.

Machine learning methods

Software utilizing machine learning algorithms often requires, like non-phylogenetic methods, the user to supply sequencing reads or assembled genomes of the dataset to be analyzed. Most machine learning algorithms convert the supplied genomic and phenotype data to a vector, which is used for training the algorithm.

Kover (Drouin et al., 2016, 2019) is based on the set covering machine algorithm and adapted for genomic data. While most other non-machine learning methods separate feature selection and modeling, Kover integrates both steps in one, improving feature selection. The set covering machine algorithm also allows for multivariate testing, i.e. combinations of features that, together, can predict the phenotype of interest. The algorithm is trained using a set of genomes, which produces a set of rules that detect the presence or absence of k-mers in a genome. The rules are subsequently aggregated to form a prediction. The authors benchmarked Kover on datasets from *Clostridium difficile*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, and *Streptococcus pneumoniae*. Isoniazid and rifampicin resistance was identified in *M. tuberculosis*, erythromycin resistance in *S. pneumoniae* and azithromycin, clindamycin, and clarithromycin resistance in *C. difficile*. Analysis of the results revealed the genomic locations to be associated to the respective antibiotic. However, a third party benchmark by Verschuuren et al., also focused on identifying antibiotic phenotypes, noted that the performance of Kover was not accurate enough to pass the criteria of the U.S. Food and Drug Administration (FDA) (Verschuuren et al., 2022). Installation of Kover requires the user to install a number of dependencies themselves.

PhenotypeSeeker (Aun et al., 2018) consists of two modules, ‘PhenotypeSeeker modeling’, which takes the genome assemblies or raw sequencing data to build the prediction model, and ‘PhenotypeSeeker prediction’, which uses the generated prediction model to conduct phenotype predictions on input data. The modeling module counts all possible k-mers from the input data using GenomeTester4 (Kaplinski et al., 2015). Welch’s two-sample t-test (continuous phenotypes) or a chi-square test (discrete phenotypes) are used for selecting k-mers that are used for the model. Population structure correction can be performed by supplying the software with a pairwise distance matrix of the input data constructed by Mash (Ondov et al., 2016). The prediction module subsequently searches the input data for k-mers that are present in the model from ‘PhenotypeSeeker modeling.’ Predictions are made using the presence or absence of certain k-mers in the input data.

The software was benchmarked by the authors on a *Pseudomonas aeruginosa* dataset, with corresponding phenotype data on ciprofloxacin resistance measured by the MIC. A discrete model and a continuous model were constructed using the MIC data. Both models identified k-mers that were associated with mutations in quinolone resistance regions of the *parC* and *gyrA* genes. These genes encode target proteins for ciprofloxacin, and mutations in these regions are known to decrease sensitivity to quinolone antibiotics, such as ciprofloxacin (Jalal & Wretlind, 1998). The authors also compared their software with SEER (Lees et al., 2016) and Kover (Drouin et al., 2016). SEER was only able to detect the mutations in *gyrA* and *parC* in discrete phenotype mode, as the k-mers in continuous phenotype mode did not pass the p-value filtering step of SEER. Kover was able to detect the *gyrA* mutation, but not the *parC* mutation. PhenotypeSeeker was significantly faster compared to SEER and Kover, only taking 3.5 hours for the entire analysis vs 14 hours for Kover and 15 hours for SEER. Installation of PhenotypeSeeker requires basic understanding of the command line, but an online variant is available including 15 pre-trained models for clinically relevant bacteria.

Statistics and additional factors to consider

In addition to factors like population structure and linkage disequilibrium, multiple testing is an intrinsic source of false positives in both human GWAS and bacterial GWAS (San et al., 2020) (Dudbridge & Gusnanto, 2008). A p-value of $P \leq 0.05$ is usually considered to be statistically significant, however, with the thousands of genetic variants analyzed in a typical GWAS this p-value will lead to dozens of false positives purely by chance. This is why, for humans, a genome-wide significance threshold of P<5E-8 is typically used, based on the Bonferroni significance threshold for the number of SNPs that were normally analyzed during (early) human GWAS (Dudbridge & Gusnanto, 2008). The Bonferroni correction for multiple testing is utilized in some software packages previously mentioned (SEER and Scoary), but is often too stringent since the method assumes that genetic variants are independent, which is not the case due to linkage disequilibrium (Power et al., 2017). An alternative, less stringent method for multiple testing correction is the Benjamin Hochberg false discovery rate, implemented by Scoary and TreeWAS, but this method has also been found to be too stringent in some cases were tested genetic variants are not independent (San et al., 2020). The permutation test is a good alternative that does not suffer from stringency issues, but this test cannot be used in combination with linear mixed model-based methods and is computationally expensive (Joo et al., 2016).

Considering the aforementioned factors, replication of a found association using an independent cohort is considered to be the gold standard (Chanock et al., 2007). Replication of results does not only aid in avoiding false positives, but it also allows accurate estimation of the effect size of the found variation due to increased statistical power (Power et al., 2017). Another more time consuming and expensive way to eliminate false positives unique to bacterial GWAS is testing identified genetic variants *in vitro*. By creating carriers of the identified genetic variant, researchers are able to observe the exact function of the variant to get a more detailed understanding than what can be done *in silico* (Power et al., 2017). However, possible unknown effects from the interaction of the identified genetic variant with other genetic variants (epistasis) limit this method somewhat (Power et al., 2017).

Calculating statistical power to ensure that the sample size is large enough to detect statistically significant genetic variants is an important aspect for any GWAS. For human GWAS, common statistical methods have been developed to calculate power, but this is not possible for bacterial GWAS due to major differences in population structure, recombination rates and homoplasy between bacterial species (Coll et al., 2022). Other variables more related to the genetic variants itself instead of the bacterial population are their minor allele frequencies and effect sizes (Coll et al., 2022). However, low effect size is usually not a significant problem when performing bacterial GWAS, since most genetic variants are under strong natural selection compared to human genetic variants, thus increasing their effect size and decreasing the need for large sample sizes (San et al., 2020). While developing a "one-size-fits-all" statistical method to calculate power for bacterial GWAS is not possible, Coll et al. developed two methods that utilize the samples used in a GWAS, which are both implemented in their software packages PowerBacGWAS (Coll et al., 2022). The first method uses a known genotype-phenotype relationship and subsampling of the original dataset. Apart from the reduced sample size, allele frequency and effect sizes are also reduced.

Because the genotype-phenotype relationship is known, a calculation can be made at which subsample size the known association can still be found. In the second method, phenotypes are simulated within a range of parameters (minor allele frequency, effect size and sample size) and a GWAS is performed on the simulated data. Based on the simulated phenotypes that are found, the minimum sample size can be calculated to find all simulated genotype-phenotype relationships. Because the bacterial genomes are not simulated or modified in both methods, PowerBacGWAS can be effectively used to calculate minimum sample size for unique datasets (Coll et al., 2022).

The results of GWAS are typically presented in Manhattan plots. These plots show each analyzed genetic variant, ordered according to the genomic position of the variants, on the x-axis and the associated -$\log_{10}$ transformed p-values of each variant on the y-axis (Uffelmann et al., 2021). In most cases, a horizontal line marks the genome-wide significance threshold above which identified genetic variants are statistically significant. Usually, the statistically significant variants are annotated with the gene or intergenic region they are located on, as shown in Fig. 4, were the CRyPTIC consortium analyzed the resistance of *Mycobacterium tuberculosis* to 13 antibiotics using 10,228 genomes (The CRyPTIC Consortium, 2022).

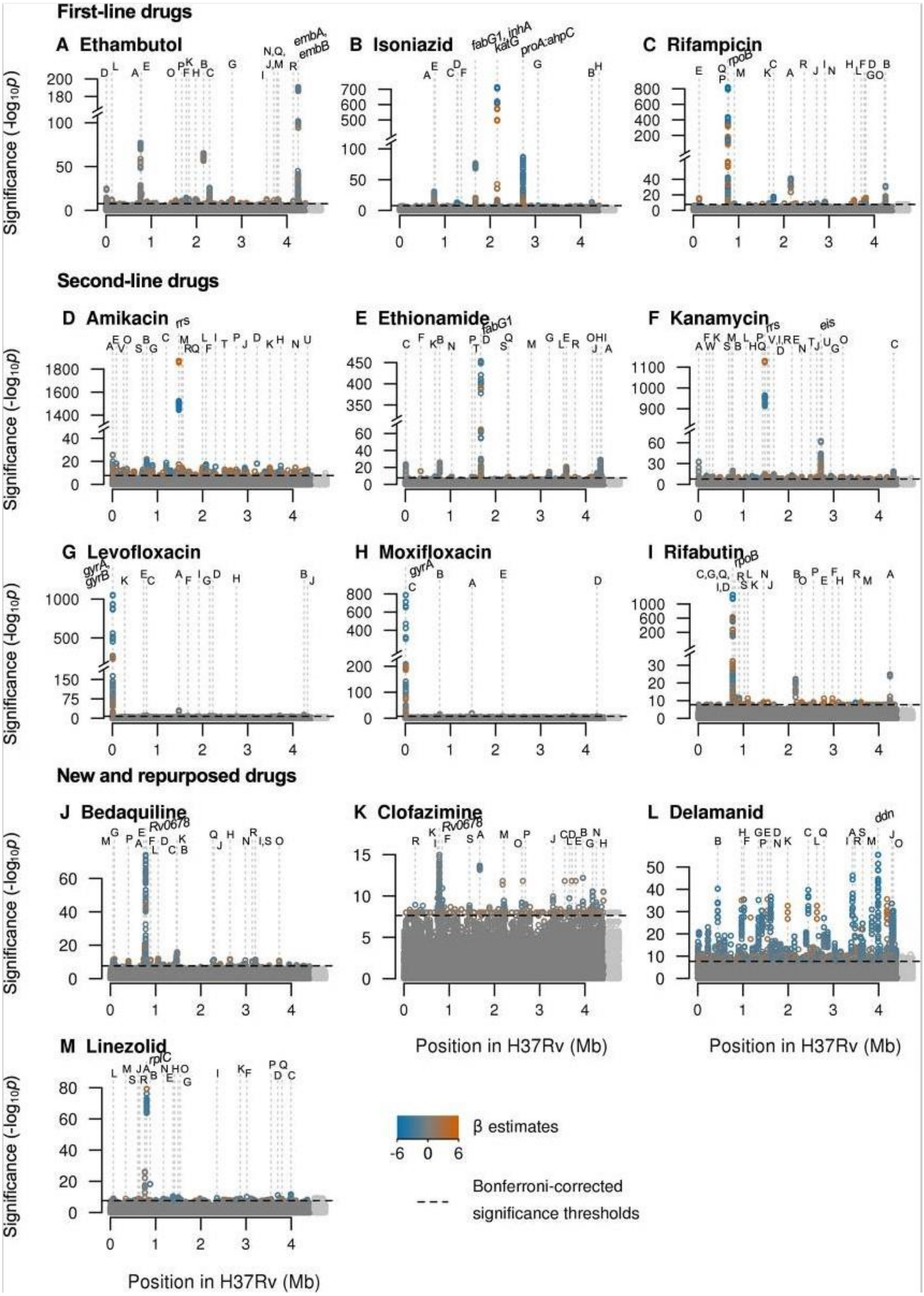

Figure 4. Manhattan plots of oligopeptide-containing regions that are associated with the minimum inhibitory concentration (MIC) of M. tuberculosis *across 13 antibiotics. Oligopeptides that increase the MIC are colored orange and oligopeptides that decrease the MIC are colored blue, with increased color intensity for higher effect sizes. The Bonferroni-corrected significance threshold is indicated with black dashed lines, and the most significant genes are annotated in the plots (full gene names can be found in Table 1 in the original article) (The CRyPTIC Consortium, 2022)*

Alternative methods for bacterial GWAS

While (bacterial) GWAS are a powerful tool to find associations between genotype and phenotype, they are not without flaws. Fortunately, other methods exist to find associations between genotype and phenotype in cases where a GWAS is insufficient or impractical. In this section, transposon insertion sequencing (Tn-seq) and genome-scale metabolic models (GSMMs) will be briefly discussed.

Tn-seq

The basic premise of Tn-seq is random genome-wide disruptions of loci by inserting transposons in a bacterial population and subsequently detecting mutants with transposons at precise insertion sites by deep sequencing (Mahmutovic et al., 2020). There are a number of different methods available, such as HITS (Gawronski et al., 2009), TraDIS (Langridge et al., 2009), IN-seq (Goodman et al., 2009), and Tn-seq (van Opijnen et al., 2009), but these are collectively known as Tn-seq (Burby et al., 2016). The main distinction between these methods is the protocol to amplify the transposon-genome junction for identification of transposon sites (Mahmutovic et al., 2020). As such, the basic steps taken in a Tn-seq experiment are comparable: random insertion of transposons, growth of mutant cells and wild-type cells in a selective environment, preparing the transposon-genome junction, sequencing, and data analysis (Mahmutovic et al., 2020).

There are two common data analysis methods for Tn-seq: calculating mutant fitness and calculating the ratio of mutant abundance for each locus. To calculate the mutant fitness, exponential growth data of the mutant strain and wild-type strain are collected and the relative population expansion of the mutant strain relative to the wild-type strain is calculated (Burby et al., 2016). Thus, a mutant strain and wild-type strain that grow at the same rate would result in the mutant strain having a fitness level of 1, and for a mutant strain that has an increased rate of growth compared to the wild-type strain the fitness level would be >1. The population expansion is estimated by determining the number of viable cells present before and after the experiment, after which the fitness value is calculated for every insertion mutant (van Opijnen et al., 2009). Since most loci will have multiple different insertion mutants spread over the population (i.e. transposons inserted in different regions of a locus), fitness values are averaged for each insertion mutant within a certain locus to estimate the fitness of a mutant that lacks gene(s) on this locus. (Burby et al., 2016).

Calculating mutant abundance for each gene is done by mapping the sequence reads of the Tn-seq experiment to the genome. By mapped the sequence reads to the genome, transposon insertion sites can be identified and tallied based on the locus they were located on (Mahmutovic et al., 2020). If a transposon insertion occurred on a site essential for survival in the tested environment, none or a limited number of sequencing reads will map to that site since the mutant population would have had a significant growth defect compared to the wild-type population. Consequently, when the transposon insertion resulted in a competitive advantage the number of mapped reads to the genome for a specific site would be higher compared to the wild-type strain (Mahmutovic et al., 2020).

Genome-scale metabolic models

Genome-scale metabolic models (GSMMs) are mathematic models that describe metabolic reactions in an organism using gene-protein reaction associations. These gene-protein reactions rely on accurate genome annotation data and experimentally obtained information to predict genes to fill in the gaps in metabolic pathways that are partially known or even predict entire metabolic pathways (Gu et al., 2019; Kotera & Goto, 2016). Construction of GSMMs includes four steps: 1) construction of the draft model, 2) refinement of the model, 3) model mathematization, and 4) verification of the model (Ye et al., 2022). Software to automatically construct GSMMs include RAVEN Toolbox (Wang et al., 2018) and Merlin (Dias et al., 2015), among others.

First-generation GSMMs constrained the simulation of the predicted metabolic pathway using the substrate uptake rate, but this had some drawbacks. For example, when the effects of increasing the rate of glucose uptake on cell growth was simulated, cell growth kept increasing, above what has been shown experimentally (Ye et al., 2022). Next-generation GSMMs integrated additional data in the models, such as transcriptomics, proteomics, metabolomics, and thermodynamics, to improve prediction accuracy (Ye et al., 2022). This integration of data allowed usage of additional parameters to simulate metabolic pathways, such as maximal growth rate, extracellular secretion rate and flux distribution (Ye et al., 2022). Using next-generation GSMM methods, Ye et al. improved the first-generation based *i*JO1366 GSMM for *Escherichia coli* and compared both models to experimental phenotype results from 24 cultures. They found that the Pearson correlation coefficient of the new model, *i*ML1515, was significantly higher, increasing from 0.20 (p-value 0.49) to 0.50 (p-value 0.07) (Ye et al., 2020). Consequently, the phenotype prediction accuracy increased from 89.8% for the *i*JO1366 model to 93.4% for the *i*ML1515 model (Ye et al., 2020).

Although phenotype prediction using GSMMs can be useful for industrial applications, such as the calculation of maximum growth rate and optimal substrate conditions for product production, it does not provide direct insight on genotype-phenotype relationships. For this task, researchers have developed single-gene deletion algorithms for GSMMs to determine the essentiality of specific genes or identify previously unknown enzyme functions under certain growth conditions (Ye et al., 2022). For example, a study by Guzmán et al. investigated a set of genes that were experimentally proven to be nonessential, but the *E. coli* GSMM *i*JO1366 predicted this set of genes to be essential (Guzmán et al., 2015). Guzmán et al. hypothesized that an unknown reaction may explain why nonessential genes became essential in the GSMM. Based on sequence homology analysis identifying high-confidence candidate isozymes, the *aspC, argD,* and *gltA* genes were chosen for further investigation from the set of ‘false negative’ genes. In an *E. coli* model, knocking out the previously mentioned genes revealed that the loss of aspartate aminotransferase, encoded by *aspC*, could be compensated by tyrosine aminotransferase, encoded by *tyrB* (Guzmán et al., 2015). Using the same knockout approach, potential isozymes that could function as alternative reaction enzymes to those encoded by *argD* and *gltA* were also identified.

Discussion

In the discussion, we will provide our suggested guidelines for the bacterial GWAS methods discussed earlier. Additionally, we will discuss the advantages and disadvantages of Tn-seq and genome-wide metabolic models (GSMMs) compared to bacterial GWAS.

First, it is important to consider the type of genetic variation that drives the phenotype of interest. If this is not known *a priori*, our recommendation is to use software utilizing k-mers (Pyseer, HAWK, Kover, and PhenotypeSeeker are examples discussed in this paper) since these methods are able to identify all genetic variants (San et al., 2020). However, using k-mers comes with certain drawbacks. These drawbacks include increased computational burden, especially for longer k-mers, and a higher likelihood of overfitting in machine learning-based methods due to the high number of genomic features that k-mers represent (Aun et al., 2018; Drouin et al., 2016). In addition, the majority of k-mers are uninformative, occur simultaneously, are highly correlated, and cannot be used for phenotype prediction (Drouin et al., 2016; San et al., 2020). To alleviate this problem, Pyseer and HAWK offer support for collapsing k-mers into unitigs, or uniquely assembleable contigs. In case genetic variation is known *a priori*, Scoary and TreeWAS can be used (in addition to the previously mentioned software), unless the analyzed bacterium has a high recombination rate. In such cases, we would not recommend using phylogenetic methods as the effects of recombination can decrease their effectiveness and power to detect associations (San et al., 2020).

The choice of k-mer length is an important parameter in k-mer-based software packages, since it can significantly affect the accuracy and computational burden; longer k-mer lengths increase accuracy at the expense of significantly increased processor and memory usage (Aun et al., 2018). However, excessively large k-mers will generate highly specific k-mers that are only present in a few genomes (Drouin et al., 2016). Drouin et al., the authors of Kover, recommend using shorter k-mers for bacteria with high recombination rates and longer k-mers for bacteria with low recombination rates, with k = 31 as a default value (Drouin et al., 2016). Aun et al. benchmarked PhenotypeSeeker using k-mers ranging in length up to 32. The authors recommend a k-mer length of 13 based on their findings that such k-mers were able to detect mutations in the *parC* and *gyrA* genes in the previously discussed *Pseudomonas aeruginosa* ciprofloxacin dataset. However, the authors do not entirely rule out longer k-mer lengths in certain situations (Aun et al., 2018). Jaillard et al., authors of the software package DBGWAS (not discussed in this paper), recommend a minimum k-mer length of 11 and a maximum k-mer length of 100, with the optimal k-mer length for their investigated dataset being 31 (Jaillard et al., 2018). Based on the literature, we recommend users to experiment with k-mer lengths ranging from 11 to 100, starting at k = 31 and decreasing or increasing based on recombination rate of the investigated bacterium.

Bacterial GWAS and Tn-seq are approaches that can both identify genes associated with bacterial phenotypes, but with vastly different methodologies. Compared to bacterial GWAS, Tn-seq does not suffer from the effects of linkage disequilibrium and population structure. Furthermore, Tn-seq can be used in cases where high-quality genotype-phenotype association data is not available or in cases with an insufficient number of genomes to satisfy statistical power requirements. However, Tn-seq has some disadvantages that must be taken into consideration. Compared to bacterial GWAS, Tn-seq may not be able to capture the full range of genetic variants that contribute to bacterial phenotypes (e.g. SNPs with low effect sizes) and is limited by genes that can be disrupted by transposon insertion (Kobras et al., 2021). Additionally, sources of false-positive and false-negative results include random birth-death processes and the effects of sampling events/population bottlenecks (Mahmutovic et al., 2020). Tn-seq cannot be used on all bacteria; some species have mechanisms that prevent uptake of foreign DNA, which makes it difficult to insert transposons (Wetmore et al., 2015). Therefore, the selection between bacterial GWAS and Tn-seq should be considered on a case-by-case basis, also taking into account the laboratory's capabilities and preferences. However, it should be noted that combining both methods would provide more robust evidence for testing genotype-phenotype associations.

GSMMs can, in some cases, also act as an alternative or complement bacterial GWAS. Like Tn-seq, GSMMs are not negatively affected by linkage disequilibrium and population structure. GSMMs are useful for predicting metabolic pathways and genes essential for growth in specific conditions by performing *in silico* gene knockout experiments (Gu et al., 2019; Ye et al., 2022). Furthermore, GSMMs can be used to optimize production of chemicals by disabling redundant or competing metabolic pathways (Gu et al., 2019). The potential applications of GSMMs are diverse, but their utility is limited by the availability of high-quality models. Although GSMMs are available for thousands of bacterial species and new models can be generated with software packages, manually curated models are only available for model organisms that have scientific, industrial, or medical value (Gu et al., 2019). Additionally, GSMMs focus solely on metabolic pathways which may not capture all relevant genetic variants affecting phenotypes (Ye et al., 2022). Therefore, it is worth considering to validate the results of a bacterial GWAS on an organism with high-quality GSMM(s) available by performing additional *in silico* experiments, such as a gene knockout experiment, to confirm the impact of identified genes on the phenotype.

References

Aun, E., Brauer, A., Kisand, V., Tenson, T., & Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLOS Computational Biology*, *14*(10), e1006434. https://doi.org/10.1371/journal.pcbi.1006434

Bossé, Y., & Amos, C. I. (2018). A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *27*(4), 363–379. https://doi.org/10.1158/1055-9965.EPI-16-0794

Brynildsrud, O., Bohlin, J., Scheffer, L., & Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, *17*, 238. https://doi.org/10.1186/s13059-016-1108-8

Burby, P. E., Nye, T. M., Schroeder, J. W., & Simmons, L. A. (2016). Implementation and Data Analysis of Tn-seq, Whole-Genome Resequencing, and Single-Molecule Real-Time Sequencing for Bacterial Genetics. *Journal of Bacteriology*, *199*(1), e00560-16. https://doi.org/10.1128/JB.00560-16

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni, J. F., Freimer, N. B., Gerhard, D. S., Gunter, C., Guttmacher, A. E., … NCI-NHGRI Working Group on Replication in Association Studies. (2007). Replicating genotype–phenotype associations. *Nature*, *447*(7145), Article 7145. https://doi.org/10.1038/447655a

Chen, P. E., & Shapiro, B. J. (2015). The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, *25*, 17–24. https://doi.org/10.1016/j.mib.2015.03.002

Coll, F., Gouliouris, T., Bruchmann, S., Phelan, J., Raven, K. E., Clark, T. G., Parkhill, J., & Peacock, S. J. (2022). PowerBacGWAS: A computational pipeline to perform power calculations for bacterial genome-wide association studies. *Communications Biology*, *5*, 266. https://doi.org/10.1038/s42003-022-03194-2

Collins, C., & Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Computational Biology*, *14*(2), e1005958. https://doi.org/10.1371/journal.pcbi.1005958

Dias, O., Rocha, M., Ferreira, E. C., & Rocha, I. (2015). Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Research*, *43*(8), 3899–3910. https://doi.org/10.1093/nar/gkv294

Didelot, X., & Wilson, D. J. (2015). ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Computational Biology*, *11*(2), e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., Bourgault, A.-M., Laviolette, F., & Corbeil, J. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, *17*(1), 754. https://doi.org/10.1186/s12864-016-2889-6

Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., & Laviolette, F. (2019). Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-40561-2

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V., & Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences*, *106*(38), 16422–16427. https://doi.org/10.1073/pnas.0906627106

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., & Gordon, J. I. (2009). Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. *Cell Host & Microbe*, *6*(3), 279–289. https://doi.org/10.1016/j.chom.2009.08.003

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*(1), 121. https://doi.org/10.1186/s13059-019-1730-3

Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O., & Feist, A. M. (2015). Model-driven discovery of underground metabolic functions in Escherichia coli. *Proceedings of the National Academy of Sciences*, *112*(3), 929–934. https://doi.org/10.1073/pnas.1414218112

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M., & Harris, S. R. (2018). Phandango: An interactive viewer for bacterial population genomics. *Bioinformatics (Oxford, England)*, *34*(2), 292–293. https://doi.org/10.1093/bioinformatics/btx610

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*, 83. https://doi.org/10.1186/s13059-017-1215-1

Hillers, K. J., Jantsch, V., Martinez-Perez, E., & Yanowitz, J. L. (2017). Meiosis. *Wormbook*, 1–43. https://doi.org/10.1895/wormbook.1.178.1

Horwitz, T., Lam, K., Chen, Y., Xia, Y., & Liu, C. (2019). A Decade in Psychiatric GWAS Research. *Molecular Psychiatry*, *24*(3), 378–389. https://doi.org/10.1038/s41380-018-0055-z

Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., & Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genetics*, *14*(11), e1007758. https://doi.org/10.1371/journal.pgen.1007758

Jalal, S., & Wretlind, B. (1998). Mechanisms of quinolone resistance in clinical strains of Pseudomonas aeruginosa. *Microbial Drug Resistance (Larchmont, N.Y.)*, *4*(4), 257–261. https://doi.org/10.1089/mdr.1998.4.257

John Lees. (2017). *The background of bacterial GWAS*. https://doi.org/10.6084/m9.figshare.5550037.v1

Joo, J. W. J., Hormozdiari, F., Han, B., & Eskin, E. (2016). Multiple testing correction in linear mixed models. *Genome Biology*, *17*(1), 62. https://doi.org/10.1186/s13059-016-0903-6

Kaplinski, L., Lepamets, M., & Remm, M. (2015). GenomeTester4: A toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, *4*(1), s13742-015-0097-y. https://doi.org/10.1186/s13742-015-0097-y

Kim, Y., Gu, C., Kim, H. U., & Lee, S. Y. (2020). Current status of pan-genome analysis for pathogenic bacteria. *Current Opinion in Biotechnology*, *63*, 54–62. https://doi.org/10.1016/j.copbio.2019.12.001

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh., J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, N.Y.)*, *308*(5720), 385–389. https://doi.org/10.1126/science.1109557

Kobras, C. M., Fenton, A. K., & Sheppard, S. K. (2021). Next-generation microbiology: From comparative genomics to gene function. *Genome Biology*, *22*(1), 123. https://doi.org/10.1186/s13059-021-02344-9

Kotera, M., & Goto, S. (2016). Metabolic pathway reconstruction strategies for central metabolism and natural product biosynthesis. *Biophysics and Physicobiology*, *13*, 195–205. https://doi.org/10.2142/biophysico.13.0_195

Kumar, K. R., Cowley, M. J., & Davis, R. L. (2019). Next-Generation Sequencing and Emerging Technologies. *Seminars in Thrombosis and Hemostasis*, *45*(7), 661–673. https://doi.org/10.1055/s-0039-1688446

Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., & Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, *19*(12), 2308–2316. https://doi.org/10.1101/gr.097097.109

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310–4312. https://doi.org/10.1093/bioinformatics/bty539

Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J., & Corander, J. (2020). Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *MBio*, *11*(4), e01344-20. https://doi.org/10.1128/mBio.01344-20

Lees, J. A., Vehkala, M., Välimäki, N., Harris, S. R., Chewapreecha, C., Croucher, N. J., Marttinen, P., Davies, M. R., Steer, A. C., Tong, S. Y. C., Honkela, A., Parkhill, J., Bentley, S. D., & Corander, J. (2016). Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms12797

Louha, S., Meinersmann, R. J., & Glenn, T. C. (2021). Whole genome genetic variation and linkage disequilibrium in a diverse collection of Listeria monocytogenes isolates. *PLoS ONE*, *16*(2), e0242297. https://doi.org/10.1371/journal.pone.0242297

Mahmutovic, A., Abel zur Wiesch, P., & Abel, S. (2020). Selection or drift: The population biology underlying transposon insertion sequencing experiments. *Computational and Structural Biotechnology Journal*, *18*, 791–804. https://doi.org/10.1016/j.csbj.2020.03.021

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Matsuda, K., Fujita, K., & Wakimoto, T. (2021). PenA, a penicillin-binding protein-type thioesterase specialized for small peptide cyclization. *Journal of Industrial Microbiology & Biotechnology*, *48*(3–4), kuab023. https://doi.org/10.1093/jimb/kuab023

McInerney, J. O., McNally, A., & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, *2*(4), Article 4. https://doi.org/10.1038/nmicrobiol.2017.40

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Pavan, W. J., & Sturm, R. A. (2019). The Genetics of Human Skin and Hair Pigmentation. *Annual Review of Genomics and Human Genetics*, *20*(1), 41–72. https://doi.org/10.1146/annurev-genom-083118-015230

Power, R. A., Parkhill, J., & de Oliveira, T. (2017). Microbial genome-wide association studies: Lessons from human GWAS. *Nature Reviews Genetics*, *18*(1), Article 1. https://doi.org/10.1038/nrg.2016.132

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Rahman, A., Hallgrímsdóttir, I., Eisen, M., & Pachter, L. (2018). Association mapping from sequencing reads using k-mers. *ELife*, *7*, e32920. https://doi.org/10.7554/eLife.32920

Read, T. D., & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: A new direction for bacteriology. *Genome Medicine*, *6*(11), 109. https://doi.org/10.1186/s13073-014-0109-z

Saber, M. M., & Shapiro, B. J. (2020). Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microbial Genomics*, *6*(3), e000337. https://doi.org/10.1099/mgen.0.000337

San, J. E., Baichoo, S., Kanzi, A., Moosa, Y., Lessells, R., Fonseca, V., Mogaka, J., Power, R., & de Oliveira, T. (2020). Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Frontiers in Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fmicb.2019.03119

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

The CRyPTIC Consortium. (2022). Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biology*, *20*(8), e3001755. https://doi.org/10.1371/journal.pbio.3001755

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9 van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, *6*(10), Article 10. https://doi.org/10.1038/nmeth.1377

Verschuuren, T., Bosch, T., Mascaro, V., Willems, R., & Kluytmans, J. (2022). External validation of WGS-based antimicrobial susceptibility prediction tools, KOVER-AMR and ResFinder 4.1, for Escherichia coli clinical isolates. *Clinical Microbiology and Infection*, *28*(11), 1465–1470. https://doi.org/10.1016/j.cmi.2022.05.024

Vishnoi, A., Roy, R., Prasad, H. K., & Bhattacharya, A. (2010). Anchor-Based Whole Genome Phylogeny (ABWGP): A Tool for Inferring Evolutionary Relationship among Closely Related Microorganims. *PLOS ONE*, *5*(11), e14159. https://doi.org/10.1371/journal.pone.0014159

Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J., & Kerkhoven, E. J. (2018). RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Computational Biology*, *14*(10), e1006541. https://doi.org/10.1371/journal.pcbi.1006541

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. *MBio*, *6*(3), e00306-15. https://doi.org/10.1128/mBio.00306-15

Ye, C., Luo, Q., Guo, L., Gao, C., Xu, N., Zhang, L., Liu, L., & Chen, X. (2020). Improving lysine production through construction of an Escherichia coli enzyme-constrained model. *Biotechnology and Bioengineering*, *117*(11), 3533–3544. https://doi.org/10.1002/bit.27485

Ye, C., Wei, X., Shi, T., Sun, X., Xu, N., Gao, C., & Zou, W. (2022). Genome-scale metabolic network models: From first-generation to next-generation. *Applied Microbiology and Biotechnology*, *106*(13), 4907–4920. https://doi.org/10.1007/s00253-022-12066-y