Trends in Microbiology

CellPress

Review

Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning

Jonathan P. Allen [1,*] Evan Snitkin,[2] Nathan B. Pincus,[3] and Alan R. Hauser[3,4]

The advent of inexpensive and rapid sequencing technologies has allowed bacterial whole-genome sequences to be generated at an unprecedented pace. This wealth of information has revealed an unanticipated degree of strain-to-strain genetic diversity within many bacterial species. Awareness of this genetic heterogeneity has corresponded with a greater appreciation of intraspecies variation in virulence. A number of comparative genomic strategies have been developed to link these genotypic and pathogenic differences with the aim of discovering novel virulence factors. Here, we review recent advances in comparative genomic approaches to identify bacterial virulence determinants, with a focus on genome-wide association studies and machine learning.

Highlights

The plethora of bacterial whole-genome sequences generated in recent years has underscored the genetic diversity of strains within bacterial species, which has in turn suggested explanations for variable infectious manifestations caused by these strains.

A number of sophisticated comparative genomic strategies, such as genome-wide association studies and machine learning algorithms, have been developed to take advantage of bacterial genetic diversity to uncover novel bacterial virulence determinants.

Comparative genomic approaches have led to the identification of bacterial genes and polymorphisms linked to several disease endpoints, including cancer, invasive infection, mortality, cytotoxicity, and biofilm formation.

The Goal of Identifying Bacterial Virulence Factors

Approximately 1 billion bacterial species exist on Earth [1] but only a relative handful of these regularly cause infections in people. This paucity of pathogens is a testament to the human immune and defense systems which are quite effective in preventing the vast majority of bacteria from taking advantage of the rich milieu of nutrients present in human cells. Yet some bacterial species have successfully developed tools that subvert or evade these defenses to allow them to cause disease. These tools are referred to as virulence factors [2], and identifying them has been a central goal of the bacterial pathogenesis research community for decades. This goal has currently acquired new urgency as bacterial pathogens have gained resistance to conventional antibiotics. Recent efforts have focused on developing antivirulence therapies that disarm multidrug-resistant bacterial strains by targeting their virulence factors [3]. Identification of virulence factors, a prerequisite of such strategies, has been greatly facilitated by whole-genome sequencing. Since the publication of the *Haemophilus influenzae* genome in 1995 [4], dramatic advances have occurred that make whole-genome sequencing fast and affordable [5]. While sequencing of even a single genome provides a wealth of information about a bacterial isolate, recent decreases in cost and labor have made it feasible to sequence and compare large numbers of genomes. Here, we summarize advances in comparative genomic strategies aimed at identifying bacterial virulence factors. Our focus is on approaches that utilize genome-wide association studies and machine learning.

Whole-Genome Sequencing and Bacterial Populations

Many of the virulence factors we are familiar with today were originally discovered by applying biochemical and forward genetic approaches and, later, genome-wide mutagenesis techniques to a relatively small number of laboratory strains [6]. These approaches were remarkably successful but suffered from the shortcoming of interrogating a single strain and failing to capture the breadth of virulence diversity across the species. Global sequencing efforts have revealed the magnitude of genetic diversity across bacterial populations [7], and modern comparative

[1]Department of Microbiology and Immunology, Loyola University Chicago Stritch School of Medicine, Maywood, IL 60153, USA
[2]Department of Microbiology and Immunology, Department of Internal Medicine/Division of Infectious Diseases, University of Michigan, Ann Arbor, MI 48109, USA
[3]Department of Microbiology-Immunology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA
[4]Department of Medicine/Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA

*Correspondence:
jallen19@luc.edu (J.P. Allen).

genomic approaches have attempted to exploit this diversity to uncover genetic traits associated with virulence phenotypes [8]. Bacterial genomes contain a large number of genes that are conserved among nearly all members of the species, collectively referred to as the core genome. These core genes are interrupted by accessory genes whose presence varies among individual strains within that species. Pathogenic properties that are characteristic of the species as a whole tend to be encoded by virulence genes in the core genome. In contrast, virulence genes that are part of the accessory genome confer on some members of the species strain-specific pathogenic attributes [9]. Ultimately, the pathogenic potential of a bacterial strain is a multifaceted trait dependent upon the complement of core genome mutations and accessory genes present in that strain [10–12]. With the advent of modern DNA sequencing technologies, we now have an unprecedented ability to peer into bacterial genomes to gain insights into the genetic origins of virulence variability [13].

An Overview of Comparative Genomic Approaches to Virulence Gene Discovery

Association-based approaches to gene discovery attempt to identify common sequence variants across a sample population that associate with a particular trait [14]. These approaches are based on the principle that genetic variants responsible for a trait will occur at a higher frequency in members of the population with the trait relative to members that lack the trait. Such genetic variants are distinguished from the thousands of other genetic variations across a genome by their strong statistical associations with the trait. In the special case in which the trait of interest is bacterial virulence, comparative genomic approaches may differ substantially in the strategies they use, but most include several distinct steps (Figure 1, Key Figure). In this section, we discuss each of these steps in more detail. Although a relatively new field, the application of comparative genomic approaches to better understand bacterial virulence has been highly successful, and examples of recent studies are shown in Table 1.

Pathogen Collection

The first step in comparative genomic studies is to obtain a large collection of isolates. These isolates may be of the same clonal lineage, species, genus, or even phylum. However, further genetic distances between the isolates will result in a greater number of genetic differences, which in turn may complicate downstream analyses. Isolates may be obtained from patient cohorts, from the environment, or from archived collections [15].

Measurement of Virulence

Once a test population of bacterial isolates has been assembled, the next step is to measure the virulence for each isolate. The ability to gather experimentally consistent virulence data is a major challenge for association studies as data quality greatly impacts the ability to detect statistical relationships [14]. Measures of virulence depend on the particular pathogen but can be obtained from natural infections (e.g., human clinical outcomes) or laboratory studies (e.g., cell culture or animal/plant models). In some studies, the ability to infect humans has been used as a measure of virulence, such as sampling from clinical specimens vs. food sources for *Listeria monocytogenes* [16] or gastrointestinal vs. extraintestinal isolation of *Campylobacter jejuni* [17]. Association with a particular disease state is also a measure of virulence, for example, isolation from patients with gastric cancer or gastritis in *Helicobacter pylori* [18]. Virulence can also be quantified by patient disease severity scores such as the sequential organ failure assessment (SOFA) score, simplified acute physiologic score (SAPS), or the acute physiology and chronic health evaluation (APACHE) score in hospitalized patients. Finally, patient mortality is frequently a measure of infection severity in humans [19]. In the laboratory, virulence is typically assessed with cultured eukaryotic cells or animal infection models. The invasiveness of a bacterial strain can be modeled *in vitro* as the ability to internalize into eukaryotic cells [20] or translocate through

Key Figure

General Approach for the Identification of Virulence Genes Using Comparative Genomic Strategies

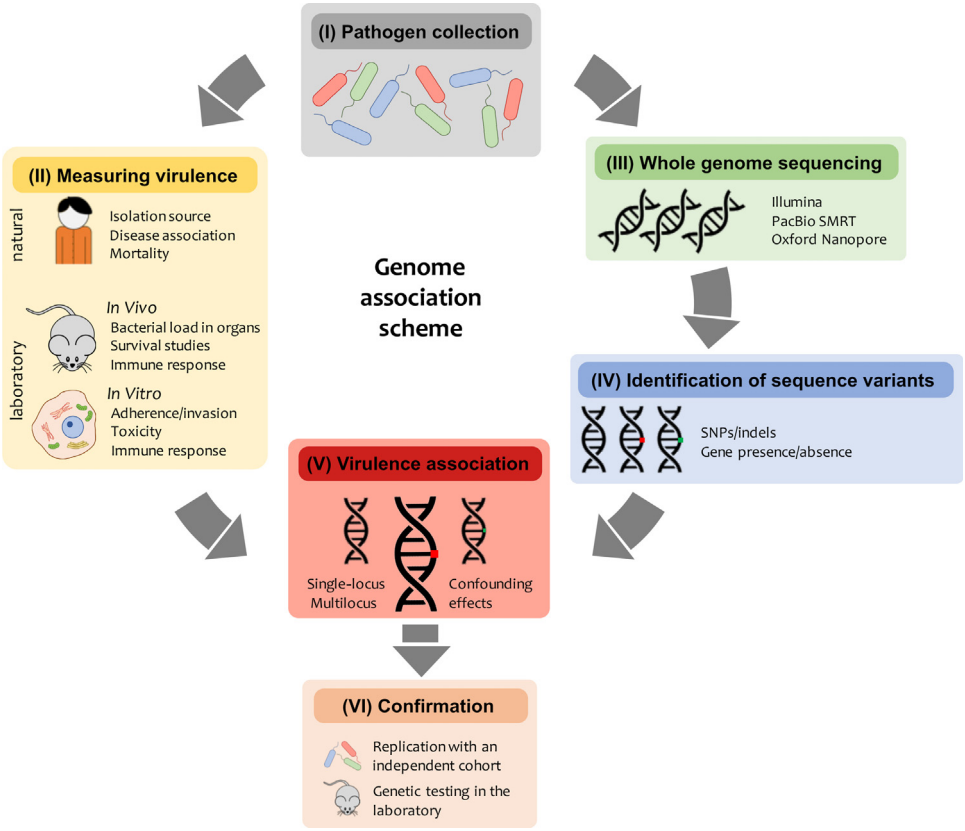

Figure 1. (I) Large numbers of isolates of a particular pathogen are collected. (II) The virulence of each isolate within the collection is determined from observations of natural infections or by using infection models in the laboratory. (III) In parallel, whole genome sequences are obtained for each isolate, and (IV) genetic differences (SNPs, indels or genes) are mapped for the entire collection (depicted by red and green points on the DNA helix). (V) The association of sequence variants with virulence (depicted by enlarged DNA helix with red point) can be bioinformatically determined while accounting for confounding effects to limit spurious associations. (VI) Virulence-associated genetic elements can be validated with an independent cohort of bacterial isolates or genetic testing to confirm loss or gain of virulence. Indels, insertions or deletions; SNPs, single-nucleotide polymorphisms.

cell monolayers [21]. The viability of infected cells can also be used as a measure of bacterial cytotoxicity [22]. Animal and plant infection models allow researchers to investigate complex interactions between a host and pathogen in the context of the host's immune system. Researchers often utilize rodent infection models, such as mice or rats, because of their anatomical, physiological, and immunological similarities to humans. However, lower cost and higher-throughput small-animal models such as the nematode (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), wax moth (*Galleria mellonella*), and zebrafish (*Danio rerio*) have also proven useful for quantifying bacterial virulence [23]. No matter the infection model, the bacterial burden at different anatomical sites [24] or lethality as determined by a 50% lethal dose (LD_{50}) [25] are both

Table 1. Examples of Comparative Genomics Studies for the Identification of Bacterial Virulence Factors

Organism	Virulence	*n*	Genetics	Method	Findings	Refs
Helicobacter pylori	Gastric cancer vs. gastritis	173	SNPs, *k*-mers	Linear mixed models (LMM) using BugWAS	Determined that alleles in several metabolic enzymes were associated with progression to gastric cancer	[18]
Staphylococcus aureus	Pyomyositis vs. Asymptomatic nasal carriage	518	*k*-mers	LMM using GEMMA	Revealed a strong relationship between the Panton–Valentine leukocidin and pyomyositis	[84]
Staphylococcus aureus	*In vitro* toxicity, biofilm formation, and patient outcomes	300	SNPs	ANOVA with Bonferroni correction and random forest machine learning to predict patient outcomes	Accounting for bacterial virulence with patient comorbidities assisted in predicting poor infection outcomes	[72]
Streptococcus pneumoniae	Invasive pneumococcal disease (IPD) vs. normal carriage	5892	SNPs, *k*-mers	LMM with FaST-LMM or burden testing with PYSEER	Identified nine alleles associated with IPD, including adhesins, an endonuclease, and a putative carboxypeptidase	[85]
Streptococcus pneumoniae	23 clinical manifestations of IPD	952	SNPs, *k*-mers, orthologous genes (OGs)	Fisher exact test in Plink or univariate χ^2 using SEER	Predictors of meningitis included a phospholipase and transmembrane protein, and of 30-day mortality included four prophage genes	[86]
Streptococcus agalactiae	Association with specific clonal complexes (CCs)	1988	Accessory genes	Fisher exact test using Scoary	CC-specific genes included metabolism, environmental information processing, and virulence functions	[87]
Pseudomonas aeruginosa	Association with *in vivo* virulence	100	Accessory genomic sequences	Spearman Rank test	Identified accessory genomic elements associated with virulence. Validated 11 of 15 by testing mutants for virulence attenuation *in vivo*	[88]
Campylobacter jejuni	Extraintestinal abortion vs. gastrointestinal carriage	193	SNPs	Machine learning using extreme gradient boosting (XGBoost)	Certain *porA* alleles, encoding an outer membrane protein, were strong predictors of abortion	[89]

n, number of strains.

common measures of virulence. Finally, immune activation following infection of primary immune cells or animal infection models has also been used as a disease trait [26]. While there is debate on how well *in vitro* or *in vivo* models reflect actual human disease [27,28], such approaches have the advantage of controlling for host genetic variabilities and inoculum size and therefore provide accurate, precise, and reproducible measures of virulence, which in turn increases the statistical power of association studies [29].

Identification of Sequence Variants

Sequencing data for comparative genomics projects can be easily generated using modern technologies (e.g., Illumina, PacBio SMRT, Oxford Nanopore Minion) [30], and many of the quality-control aspects to correct for sequencing and assembly errors have been thoroughly described elsewhere [31,32]. In contrast to humans, in which the primary genetic differences studied are single-nucleotide polymorphisms (SNPs) [33], genetic differences in bacteria include SNPs or small insertions and deletions (indels), sequence inversions, copy number variations, the presence or absence of accessory genes, and intragenic variants [34]. Examples of the latter are polymorphic toxins, in which a defined portion of an otherwise highly conserved gene differs markedly from strain to strain [35]. These additional genetic differences introduce unique challenges when performing association studies to identify likely causal variants. Sequence variants in the core genome can be detected by mapping SNPs/indels to a reference sequence. A major limitation of comparing to a reference sequence occurs when strains under investigation

contain genomic regions not found in the reference genome (Figure 2A). Alternative alignment-free methods using *k*-mers do not require a reference sequence and are becoming a preferred choice in association studies [8,36]. *k*-mers represent the complete and overlapping set of subsequences of length *k* nucleotides contained in a biological sequence (Figure 2B) [37]. *k*-mer-based approaches are able to capture the different types of variation present across many genomes discussed in the previous text and can be less computationally demanding than other methods (Figure 2C,D). However, single *k*-mers, especially when *k* is small, may be repeated at multiple loci within a single genome, which complicates analyses.

Determination of Genotypic and Phenotypic Associations

Once the virulence level and sequence variants of each bacterial isolate have been determined, associations between these two are sought. This analysis can be performed for each variant individually (single-locus) using allele counting, regression, or correlation measures. Genome-wide association studies (GWAS) are examples of such approaches that have been used for years with human genomes. Alternatively, a more holistic approach can be used in which all variants are analyzed together (multilocus) using machine learning-based methods [14]. These methodologies are discussed in more detail in the following sections, and a list of useful toolsets is provided in Table 2.

Validation of Associations

A final step used by some comparative genomic studies is confirmation that identified candidate virulence genes do indeed encode pathogenic factors. The validity of observed associations

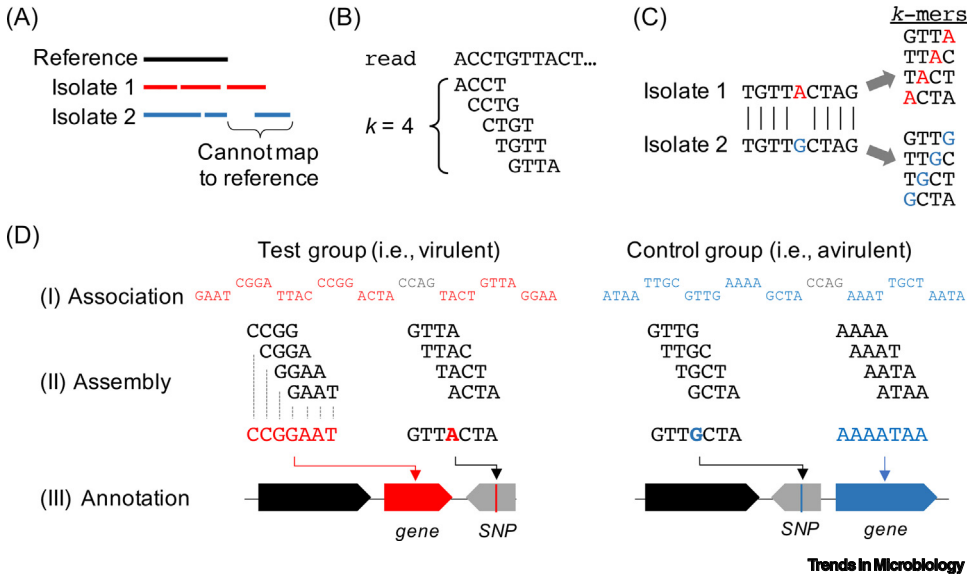

Figure 2. Use of *k*-mers to Define Genetic Differences. (A) The genomes of two separate isolates (represented as colored lines) are depicted aligned to a reference genome (black line). Mapping sequencing reads to a reference genome can only identify polymorphisms from aligned regions and excludes those regions unique to test strains. These limitations can be overcome with alignment-free methods using *k*-mers. (B) *k*-mers represent the complete and overlapping set of sequences of k nucleotides in length found in a specified genome. In this example, a sequencing read is broken down into overlapping *k*-mers of length $k = 4$. (C) Two identical genomes contain identical sets of *k*-mers, but a single SNP would generate slightly different sets of *k*-mers for each genome. Here, a genetic locus is shown in which Isolates 1 and 2 differ by a single SNP. Representative isolate-specific *k*-mers caused by this SNP are shown for the condition in which $k = 4$. (D) In comparative genomic analyses, complete sets of *k*-mers from each isolate are generated, and (I) *k*-mers associated with a test group (red) or control group (blue) are identified while k-mers found in both test and control groups (gray) can be excluded. (II) Overlapping *k*-mers can be assembled into larger contigs to facilitate annotation. (III) Polymorphisms associated with virulence, including the presence of entire genes and specific SNPs within a gene, can then be identified and further characterized. SNP, single-nucleotide polymorphism.

Table 2. Comparative Genomics Toolsets

Software	Analysis	Description	Ref.
TreeWAS	SNPs, *k*-mers, genes	Phylogeny-based approach that performs three independent tests of association to all loci. Spurious associations are assessed using simulated datasets.	[41]
CCTSWEEP / VENN	SNPs	Phylogeny-based approach to identify significant correlations between SNPs and traits.	[90]
Scoary	genes	Correlates gene presence/absence from pangenome analysis with phenotypic traits.	[91]
BugWAS	SNPs, *k*-mers, genes	Performs association tests using linear mixed models to correct for population structure.	[92]
ROADTRIPS	SNPs	Performs a range of association tests and uses a covariance matrix to correct for population structure.	[93]
SEER	SNPs, *k*-mers, genes	Identifies *k*-mers using distributed string mining and performs robust regression analysis to associate phenotypic variations.	[44]
PySEER	SNPs, *k*-mers, genes	Python implementation of SEER that uses generalized linear models to test for associations.	[47]
GEMMA	SNPs	Computes exact association test statistics (Wald or likelihood ratio) using linear mixed models.	[46]
FAST-LMM	SNPs	Performs GWAS using an optimized linear mixed model approach.	[94]
HAWK	*k*-mers	Performs likelihood testing of *k*-mer counts to detect association with a given trait and logistic regression to account for confounding effects.	[95]
DBGWAS	*k*-mers	*k*-mer-based GWAS method using compacted De Bruijn graphs to produce interpretable genetic variants associated with distinct phenotypes	[96]
PLINK	SNPs	Open-source whole genome association toolset that efficiently performs a range of large-scale analyses	[43]
Phenotype seeker	*k*-mers	Machine learning approach that determines *k*-mer associations and builds a regression model to conduct phenotype predictions.	[97]
Kover	*k*-mers	Machine learning method using a Set Covering Machine (SCM) algorithm to identify *k*-mers with significant trait association.	[98]
Hogwash	SNPs, k-mers, genes	Implements three algorithms for convergence-based bGWAS	[40]

can be assessed in two ways: replication of the associations using an independent cohort of bacterial isolates, and genetic testing to confirm loss or gain of virulence [31]. In the latter approach, a gene of interest is deleted from or complemented into a bacterial isolate, and a change in virulence is sought using laboratory assays. This validation approach has the advantage of demonstrating whether the variants associated with virulence actually play a causal role.

Specific Considerations for GWAS Approaches

Most studies using GWAS methods have used single-locus approaches, whereby each variant of interest (e.g., SNPs, *k*-mers, etc.) is individually evaluated for association with a phenotype. When evaluating the strength of genotype–phenotype associations it is critical to take into account biases that can lead to spurious associations. Two issues that have been found to be critical for bacterial GWAS are nonrandom association of genetic alleles (i.e., linkage disequilibrium) and the existence of subpopulations that have broad differences in the prevalence of both the phenotype and allele frequencies (i.e., population structure) (Figure 3).

Accounting for Linkage Disequilibrium

In an extreme setting of no recombination and high linkage disequilibrium (LD), it can be extremely challenging to identify causal genetic variation in the background of passenger alleles that arise in the same lineage. Recent simulation studies suggest that popular GWAS methods perform poorly in the context of high LD [38,39]. In settings of high LD, it is therefore important for

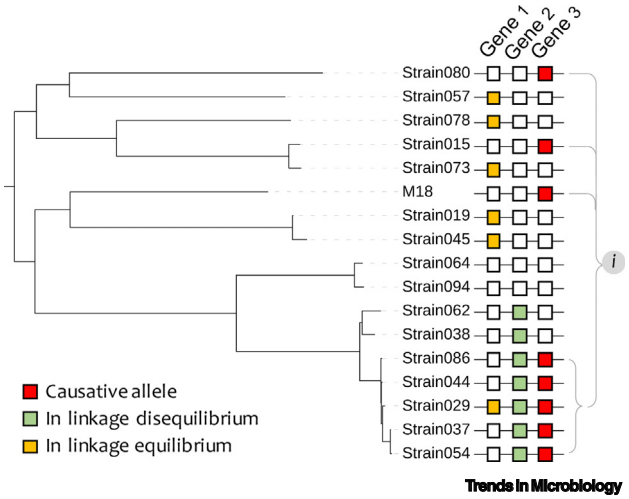

Figure 3. Confounding Due to Population Stratification Can Result in Spurious Genotype–Phenotype Associations. In this example, alleles in three individual genes (squares in a row) are differentially colored. Without recombination, fixed, noncausal genetic variants (green) will be passed onto bacterial descendants and be in linkage disequilibrium with other causal mutations that occur in that lineage (red). (*i*) Homoplasy counting methods, such as phyC, search for causal associations (red) occurring in separate lineages and exclude lineage-specific noncausal events (green).

investigators to characterize patterns of phenotypic evolution before selecting a GWAS approach and/or assessing whether GWAS is likely to be effective. If, for instance, there is evidence of frequent phenotypic convergence, then convergence-based GWAS methods that rely on homoplasy may still be effective, even in the presence of high LD [40–42].

Accounting for Population Structure

In contrast to LD, popular GWAS methods appear to be relatively robust to bacterial population structure, although to varying degrees [39]. All standard GWAS methods circumvent the issue of bacterial population structure by controlling for the genetic relatedness of sequenced isolates. However, methods differ in how they control for genetic relatedness, which likely influences their precision. The three most common strategies for incorporating genetic relatedness into models are delineation of strain clusters (e.g., PLINK) [43], using dimension reduction methods to genotype matrices (e.g., SEER, BugWAS) [44,45], and incorporation of the complete genotype matrix (e.g., GEMMA, Fast-LMM) [46,47]. Even within a given class of methods, there are additional decisions that can influence performance. For instance, with strain clusters one must decide how loosely to group strains; loose grouping (e.g., small number of strains) likely decreases precision, and overgrouping (e.g., all isolates belonging to their own strain) likely decreases recall. With dimension-reduction techniques, one must decide the number of dimensions to include in the model, which can influence model performance. Even when including the entire genotype matrix, model performance may differ depending on whether the matrix is based on SNP distances, *k*-mer distances, or patristic distances extracted from phylogenetic reconstructions [39].

Accounting for Multiple Testing

A final consideration with standard GWAS approaches is multiple-testing burden. In particular, standard GWAS methods use single-locus tests, with a separate association test being performed for each genotype of interest. Thus, in order to avoid unacceptably high false-positive rates a test correction must be performed to take into account the large number of tests being performed and the expectation that some number of significant results will be observed by chance. For phenotypes modulated by variants of large effect (e.g., antibiotic resistance), multiple test correction has proved not to be a major hindrance in detection of associations. However, for more complex phenotypes like virulence, that may be influenced by large numbers of variants of small effect sizes, multiple testing burden can hinder detection of

significant associations [48]. The two levers available to counteract multiple testing burden are to increase sample size or decrease the number of tests. One approach for decreasing numbers of tests is by grouping correlated variants (e.g., building *k*-mers into unitigs), although by definition one would have no insight into which of the correlated variants was likely causal. A second strategy for decreasing the number of tests is to perform burden testing, whereby variants are grouped together by functional units (e.g., genes, pathways), and tests are then performed at the level of these functional units [49]. Burden testing also has the potential advantage of increasing power in situations where a phenotype has emerged multiple times via distinct genetic events that impacted a common set of genes or pathways. It is important to note that pre-test grouping is only possible for variant detection methods where functional annotation is known (e.g., reference-based variant mapping) [40].

Recent Advances in GWAS Approaches

While most applications of GWAS to date have used the single-locus testing framework described in the previous text, there have been recent innovations that seek to expand upon this paradigm to elucidate more complex genotype–phenotype associations. An active area of research is in the development and application of methods to detect epistatic interactions that influence phenotypes of interest. While searching for pairwise interactions is even more fraught by multiple testing than single-locus tests, there have been recent innovations that seek to circumvent these issues to identify epistatic interactions of large effect [50,51]. A second area of active research is the use of approaches to identify multiple loci whose influence on a phenotype may differ in their magnitude, thereby obfuscating the contribution of loci of small effect. Two recently applied approaches to identify independently contributing loci are elastic net regularization and conditional regression, with both methods found to be capable of finding secondary loci associated with a phenotype of interest [38,52]. Lastly, there is a clear need to generate more explicitly testable hypotheses for the molecular bases underlying genotype–phenotype associations. Attempts to meet this need have largely focused on genetic variation linked to microbial metabolism due to the existence of turnkey systems biology toolkits for metabolic networks [53]. These metabolic models allow for incorporation of both genetic variation of different types, as well as other omics data such as RNA-seq and metabolomics, thereby facilitating insight into the direct and indirect impact of genetic variation on metabolic function [54–56].

Special Considerations for Machine Learning Approaches

Types of Machine Learning Approaches

Machine learning, defined as the development and use of computer algorithms that improve with experience [57], is ideally suited for applications involving large datasets such as genomic sequences. It has been widely applied to the identification of bacterial species [58], the prediction of antibiotic resistance [59], and the interpretation of transcriptomic data [60], and is now increasingly used to predict virulence. In a 'supervised' machine learning approach, computational models are 'fit' to a training dataset that contains a number of 'samples' (e.g., bacterial genomes), each with 'features' (e.g., SNPs, accessory genes, *k*-mers) and associated 'labels' (e.g., levels of virulence). The generated models can then be applied to a new and uncharacterized dataset to predict labels using the associated features (Figure 4). Features that are nondiscriminatory, such as two core genome SNPs that occur together in all genomes, can be combined in a process called 'feature reduction' to improve the model or at least allow for more rapid calculations. Supervised machine learning algorithms are divided further into classification methods, in which the labels are a categorical variable (e.g., infection vs. colonization), and regression methods, in which the labels are a continuous variable (e.g., percent cytotoxicity) [57,61,62]. Examples of supervised machine learning algorithms include logistic regression, random forest, support vector machine, and neural network [61,62]. In contrast, many

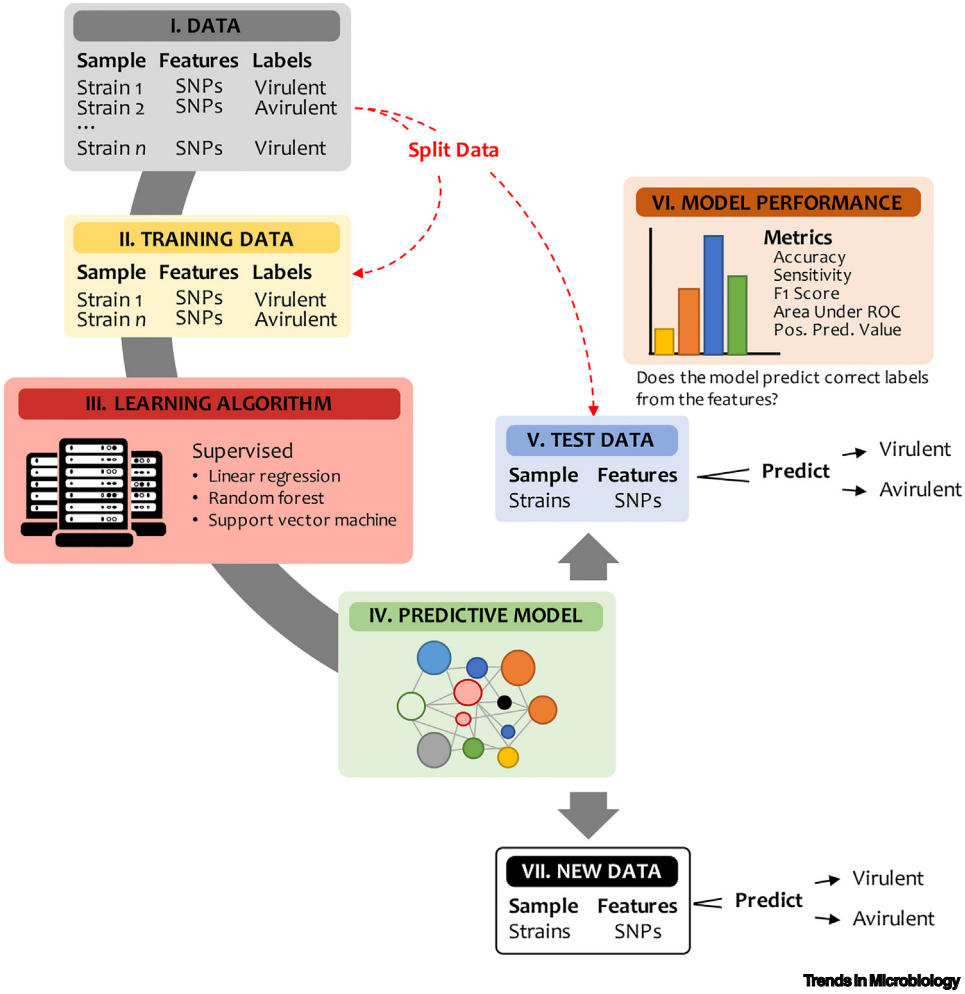

Figure 4. General Supervised Machine Learning Approach for Identifying Virulence Genes. Previously acquired data (I) can be split into separate datasets used for (II) 'training' a predictive model and (V) 'testing' its performance. Both features (e.g., SNPs) and labels (e.g., virulence) from the training set considered by a learning algorithm (III) to generate a predictive model (IV) that best fits the virulence labels. Performance of the model (VI) is then assessed by using SNPs from the 'test' dataset (V) to predict strain virulence and comparing to the true labels for each strain. The model can then be used to predict the virulence of other sequenced strains based on their genomes (VII). As more data become available, this process can be repeated and models refined with the goal of improving performance. SNPs, single-nucleotide polymorphisms.

'unsupervised' machine learning methods seek patterns in unlabeled samples that allow them to be clustered into groups. While clustering approaches may reveal important patterns within a dataset, they may identify groups for which it is difficult to associate phenotypes. Examples of unsupervised machine learning algorithms include *k*-means, divisive hierarchical clustering, and latent Dirichlet allocation [63–65].

Assessing Model Performance

For a machine learning model to be useful one must have an estimation of how well it performs. To assess supervised model performance, the generated model can be applied to a separate test dataset in which the labels are empirically known. Using the model to assess features within the test dataset, one can compare the calculated labels with the known labels to determine how well the model performs. The most robust assessment occurs when using a test dataset

that was compiled independently of the dataset used to train the model. In the absence of an external dataset, one can perform a process termed 'cross-validation' on the training dataset. In this approach, the training dataset is randomly split into separate nonoverlapping subsets (termed 'cross-validation folds'). A single cross-validation fold is excluded from the dataset and the remaining folds are used to train a model. The generated model is then used to predict the labels of the excluded cross-validation fold and assess model performance. The process is then repeated in a similar manner with each of the cross-validation folds. By considering performance across all cross-validation folds, one can estimate how well a model built using the training dataset will generalize to new data [61,62].

When training and evaluating a model, it is also important to have a performance metric in mind. Accuracy may seem a natural choice, but in some cases another metric may be more important (e.g., high sensitivity, so as to not miss patients with a rare disease in a medical screening test). Additionally, accuracy can be a poor metric of model performance when there is substantial class imbalance in the dataset. Alternative metrics include F1 score, the harmonic mean of sensitivity and positive predictive value, and area under the receiver operating characteristic curve [61,62].

Applying Machine Learning to Virulence Gene Discovery

Machine learning has been applied to genomic sequences in several ways to better understand bacterial virulence. These include distinguishing bacterial strains that are capable of causing clinical infections in humans from those that infect animals, are found in the natural environment, or are nonpathogenic [66–69]. Machine learning algorithms have been used to categorize pathogenic bacterial strains based on the type or severity of infections they cause [69–71]. It is often possible to rank the features that are most important to the performance of the model (e.g., permutation importance, Gini importance), so highly ranked features in genomic-based models of virulence identify candidates for pathogenic genes and alleles [72]. However, correlated nonpathogenic features will also be identified by this approach [70] so care must be taken to genetically confirm candidate virulence genes.

A different strategy has been to apply machine learning methods to individual genes rather than to whole genome sequences. In these approaches, machine learning algorithms are trained to identify signatures of virulence factors. In several studies, machine learning algorithms were trained on known type III and type IV secretion system effector proteins. To accomplish this, an ensemble of classification algorithms, including naive Bayes, Bayesian networks, support vector machine, and random forest, were trained on specific features that encompassed five general traits: homology, physical properties, genome organization, taxonomic grouping, and regulatory characteristics [73–77]. The classifiers were put through three successive rounds of training. After each round, the top predicted effectors were experimentally validated *in vitro* and added to the training set for the next round. The final classification score of an open reading frame (ORF) was a weighted mean of its scores from all classifiers. This approach led to the discovery of 40 new type IV secretion effectors in *Legionella pneumophila* [73] and 13 in *Coxiella burnetii* [74], as well as two new type III secretion system effectors in *Pseudomonas aeruginosa* [75], seven in *Xanthomonas euvesicatoria* [76], and 17 in *Pantoea agglomerans* [77]. Similar machine learning approaches have been used to identify pathogenic proteins across bacterial species [78,79]. As the number of sequenced bacterial genomes increases, it is anticipated that machine learning will play an expanding role in the identification of virulence genes.

Concluding Remarks and Future Perspectives

Comparative genomic studies are emerging as a powerful tool for identifying novel virulence factors and important disease processes in pathogenic bacteria. Nevertheless, several challenges remain

Outstanding Questions

Can comparative genomic approaches incorporate bacterial genetic biomarkers with complex human clinical and host factor data to successfully predict patient outcomes in the context of the broad diversity of strains that cause infections?

How important are the effects of strain-to-strain differences in bacterial virulence relative to host factors in predicting the outcomes of human infections?

Can metagenomic sequence information be combined with pathogen whole-genome sequences to better predict and define the occurrence and course of infection?

Can statistical tests for comparative genomic approaches be optimized to better balance detection with reproducibility?

When applied to the same dataset, how do the performances of the many different GWAS and machine learning methodologies compare?

(see Outstanding Questions). Researchers entering the field face disparate available tools, each with their unique strengths and weaknesses. To make the field more accessible, experts in biology and software engineering must collaborate on more universal yet flexible approaches [8]. Open-source programming and data sharing (both genotypic and phenotypic) continue to be championed by researchers in the field. In this regard, the numbers of publicly available bacterial genomes have increased at an impressive rate, but the phenotypic characteristics (type and severity of infection, virulence phenotypes in infection models) associated with these genomes remain limited. Large and multidimensional datasets associated with whole-genome sequences are needed to facilitate both initial comparative genomic studies and their validation.

Despite these challenges, the future of comparative genomics in bacterial pathogenesis is extremely promising. New approaches are being developed to leverage additional information against association signals to compensate for limited sample sizes and small effect sizes of disease-associated variants. These include incorporating transcriptomic, proteomic, and metabolomic data [80], computational reconstruction of host–pathogen protein–protein interaction networks [81], and network-based gene prioritization to integrate predicted interactions between microbial and host proteins together with host molecular networks [82,83]. These types of studies will undoubtedly accelerate our understanding of the mechanisms used by pathogenic bacteria to cause disease.

Acknowledgments

This project was funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health awards F32 AI108247 (to J.P.A.), and RO1 AI118257, U19 AI135964, K24 AI04831, R21 AI129167, and R21 AI153953 (all to A.R.H.), and RO1 AI148259-01 (to E.S.). Financial support was also provided by the American Heart Association under Contract No. 15POST25830019 (to J.P.A.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Dykhuizen, D. (2005) Species numbers in bacteria. *Proc. Calif. Acad. Sci.* 56, 62–71
2. Casadevall, A. and Pirofski, L.A. (2003) The damage-response framework of microbial pathogenesis. *Nat. Rev. Microbiol.* 1, 17–24
3. Dickey, S.W. *et al.* (2017) Different drugs for bad bugs: antivirulence strategies in the age of antibiotic resistance. *Nat. Rev. Drug Discov.* 16, 457–471
4. Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512
5. Didelot, X. *et al.* (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* 13, 601–612
6. Cain, A.K. *et al.* (2020) A decade of advances in transposon-insertion sequencing. *Nat. Rev. Genet.* 21, 526–540
7. Medini, D. *et al.* (2020) The pangenome: a data-driven discovery in biology. In *The Pangenome* (Tettelin, H. and Medini, D., eds), pp. 51–68, Springer
8. San, J.E. *et al.* (2019) Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. *Front. Microbiol.* 10, 3119
9. Ho Sui, S.J. *et al.* (2009) The association of virulence factors with genomic islands. *PLoS One* 4, e8094
10. Lee, D.G. *et al.* (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol.* 7, R90
11. Paauw, A. *et al.* (2010) Evolution in quantum leaps: multiple combinatorial transfers of HPI and other genetic modules in Enterobacteriaceae. *PLoS One* 5, e8662
12. Siena, E. *et al.* (2018) Interplay between virulence and variability factors as a potential driver of invasive meningococcal disease. *Comput. Struct. Biotechnol. J.* 16, 61–69
13. Olsen, R.J. *et al.* (2012) Bacterial genomics in infectious disease and the clinical pathology laboratory. *Arch. Pathol. Lab. Med.* 136, 1414–1422
14. Dutilh, B.E. *et al.* (2013) Explaining microbial phenotypes on a genomic scale: GWAS for microbes. *Brief Funct. Genomics* 12, 366–380
15. Ochman, H. and Selander, R.K. (1984) Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* 157, 690–693
16. Maury, M.M. *et al.* (2016) Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat. Genet.* 48, 308–313
17. Wheeler, N.E. *et al.* (2019) Genomic correlates of extraintestinal infection are linked with changes in cell morphology in *Campylobacter jejuni*. *Microb. Genom.* 5, e000251
18. Berthenet, E. *et al.* (2018) A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biol.* 16, 84
19. Hifumi, T. *et al.* (2020) Clinical characteristics of patients with severe sepsis and septic shock in relation to bacterial virulence of beta-hemolytic *Streptococcus* and *Streptococcus pneumoniae*. *Acute Med. Surg.* 7, e513
20. Raju, D. *et al.* (2012) Cell culture-based assays to test for bacterial adherence and internalization. *Methods Mol. Biol.* 921, 69–76
21. Cruz, N. *et al.* (1994) The Caco-2 cell monolayer system as an *in vitro* model for studying bacterial–enterocyte interactions and bacterial translocation. *J. Burn. Care Rehabil.* 15, 207–212
22. Riss, T. *et al.* (2004) Cytotoxicity assays: *in vitro* methods to measure dead cells. In *Assay Guidance Manual* (Sittampalam, G.S. *et al.*, eds), Eli Lilly & Company and the National Center for Advancing Translational Sciences
23. Lopez Hernandez, Y. *et al.* (2015) Animals devoid of pulmonary system as infection models in the study of lung bacterial pathogens. *Front. Microbiol.* 6, 38
24. Becavin, C. *et al.* (2014) Comparison of widely used *Listeria monocytogenes* strains EGD, 10403S, and EGD-e highlights

genomic variations underlying differences in pathogenicity. *mBio* 5, e00969-00914
25. Reed, L.J. and Muench, H. (1938) A simple method of estimating fifty per cent endpoints. *Am. J. Epidemol.* 27, 493–497
26. Sela, U. *et al.* (2018) Strains of bacterial species induce a greatly varied acute adaptive immune response: The contribution of the accessory genome. *PLoS Pathog.* 14, e1006726
27. van der Worp, H.B. *et al.* (2010) Can animal models of disease reliably inform human studies? *PLoS Med.* 7, e1000245
28. Colby, L.A. *et al.* (2017) Considerations for infectious disease research studies using animals. *Comp. Med.* 67, 222–231
29. Flint, J. and Eskin, E. (2012) Genome-wide association studies in mice. *Nat. Rev. Genet.* 13, 807–817
30. Bansal, V. and Boucher, C. (2019) Sequencing technologies and analyses: where have we been and where are we going? *iScience* 18, 37–41
31. Power, R.A. *et al.* (2017) Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* 18, 41–50
32. Carrico, J.A. *et al.* (2018) A primer on microbial bioinformatics for nonbioinformaticians. *Clin. Microbiol. Infect.* 24, 342–349
33. National Institutes of Health (US) (2007) Understanding human genetic variation. In *Biological Sciences Curriculum Study. NIH Curriculum Supplement Series* (National Institutes of Health, ed.), National Institutes of Health
34. Clark, D.P. *et al.* (2019) *Molecular Biology*, Academic Cell
35. Ruhe, Z.C. *et al.* (2020) Polymorphic toxins and their immunity proteins: diversity, evolution, and mechanisms of delivery. *Annu. Rev.Microbiol.* 74, 497–520
36. Bernard, G. *et al.* (2018) *k*-mer similarity, networks of microbial Genomes, and taxonomic rank. *mSystems* 3 e00257-18
37. Ren, J. *et al.* (2018) Alignment-free sequence analysis and applications. *Annu. Rev. Biomed. Data Sci.* 1, 93–114
38. Lees, J.A. *et al.* (2020) Improved prediction of bacterial genotype–phenotype associations using interpretable pangenome-spanning regressions. *mBio* 11 e01344-20
39. Saber, M.M. and Shapiro, B.J. (2020) Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb. Genom.* 6, 1–15
40. Saund, K. and Snitkin, E.S. (2020) Hogwash: Three Methods for Genome-Wide Association Studies in Bacteria. *Microb. Genom.* 6, 1–10
41. Collins, C. and Didelot, X. (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.* 14, e1005958
42. Farhat, M.R. *et al.* (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 45, 1183–1189
43. Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575
44. Lees, J.A. *et al.* (2016) Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* 7, 12797
45. Una, R. *et al.* (1995) Ebstein's anomaly. Anesthetic alternatives in non-cardiac surgery. *Rev. Esp. Anestesiol. Reanim.* 42, 35
46. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824
47. Lees, J.A. *et al.* (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* 34, 4310–4312
48. Laabei, M. *et al.* (2014) Predicting the virulence of MRSA from its genome sequence. *Genome Res.* 24, 839–849
49. Li, B. and Leal, S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321
50. Schubert, B. *et al.* (2019) Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nat. Microbiol.* 4, 328–338
51. Skwark, M.J. *et al.* (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet.* 13, e1006508
52. Ma, K.C. *et al.* (2020) Increased power from conditional bacterial genome-wide association identifies macrolide resistance mutations in *Neisseria gonorrhoeae*. *Nat. Commun.* 11, 5374
53. Fang, X. *et al.* (2020) Reconstructing organisms *in silico*: genome-scale models and their emerging applications. *Nat. Rev.Microbiol* 18, 731–743
54. Oyas, O. *et al.* (2020) Model-based integration of genomics and metabolomics reveals SNP functionality in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U. S. A.* 117, 8494–8502
55. Bosi, E. *et al.* (2016) Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* 113, E3801–E3809
56. Kavvas, E.S. *et al.* (2020) A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.* 11, 2580
57. Libbrecht, M.W. and Noble, W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332
58. Qu, K. *et al.* (2019) Application of machine learning in microbiology. *Front. Microbiol.* 10, 827
59. Su, M. *et al.* (2019) Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.* 57, e01405-18
60. Razaghi-Moghadam, Z. and Nikoloski, Z. (2020) Supervised learning of gene-regulatory networks based on graph distance profiles of transcriptomics data. *NPJ Syst. Biol. Appl.* 6, 21
61. Müller, A.C. and Guido, S. (2016) *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly Media
62. Baştanlar, Y. and Özuysal, M. (2014) Introduction to machine learning. In *miRNomics: MicroRNA Biology and Computational Analysis* (Yousef, M. and Allmer, J., eds), pp. 105–128, Humana Press
63. Blei, D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022
64. MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Statistics*, pp. 281–297, University of California Press
65. Maimon, O. and Rokach, L. (2005) Clustering methods. In *Data Mining and Knowledge Discovery Handbook* (Maimon, O. and Rokach, L., eds), pp. 151–183, Springer
66. van der Ploeg, T. and Steyerberg, E.W. (2016) Feature selection and validated predictive performance in the domain of *Legionella pneumophila*: a comparative study. *BMC Res. Notes* 9, 147
67. Lupolova, N. *et al.* (2017) Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb. Genom.* 3, e000135
68. Andreatta, M. *et al.* (2010) *In silico* prediction of human pathogenicity in the gamma-proteobacteria. *PLoS One* 5, e13680
69. Barash, E. *et al.* (2019) BacPaCS-bacterial pathogenicity classification via sparse-SVM. *Bioinformatics* 35, 2001–2008
70. Pincus, N.B. *et al.* (2020) A genome-based model to predict the virulence of *Pseudomonas aeruginosa* isolates. *mBio* 11, e01527-20
71. Obolski, U. *et al.* (2019) Identifying genes associated with invasive disease in *S. pneumoniae* by applying a machine learning approach to whole genome sequence typing data. *Sci. Rep.* 9, 4049
72. Recker, M. *et al.* (2017) Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nat. Microbiol.* 2, 1381–1388
73. Burstein, D. *et al.* (2009) Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* 5, e1000508
74. Lifshitz, Z. *et al.* (2014) Identification of novel *Coxiella burnetii* Icm/Dot effectors and genetic analysis of their involvement in modulating a mitogen-activated protein kinase pathway. *Infect. Immun.* 82, 3740–3752
75. Burstein, D. *et al.* (2015) Novel type III effectors in *Pseudomonas aeruginosa*. *mBio* 6, e00161
76. Teper, D. *et al.* (2016) Identification of novel *Xanthomonas euvesicatoria* type III effector proteins by a machine-learning approach. *Mol. Plant Pathol.* 17, 398–411
77. Nissan, G. *et al.* (2018) Revealing the inventory of type III effectors in *Pantoea agglomerans* gall-forming pathovars using draft genome sequences and a machine-learning approach. *Mol. Plant Pathol.* 19, 381–392
78. Garg, A. and Raghava, G.P. (2008) A machine learning based method for the prediction of secretory proteins using amino

acid composition, their order and similarity-search. *In Silico Biol.* 8, 129–140
79. Gupta, A. *et al.* (2014) MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS One* 9, e93907
80. Kachroo, P. *et al.* (2019) Integrated analysis of population genomics, transcriptomics and virulence provides novel insights into *Streptococcus pyogenes* pathogenesis. *Nat. Genet.* 51, 548–559
81. Mei, S. and Zhang, K. (2020) *In silico* unravelling pathogen–host signaling cross-talks via pathogen mimicry and human protein–protein interaction networks. *Comput. Struct. Biotechnol. J.* 18, 100–113
82. Kim, C.Y. *et al.* (2018) Network-based genetic investigation of virulence-associated phenotypes in methicillin-resistant *Staphylococcus aureus*. *Sci. Rep.* 8, 10796
83. Andrighetti, T. *et al.* (2020) MicrobioLink: an integrated computational pipeline to infer functional effects of microbiome–host interactions. *Cells* 9, 1278
84. Young, B.C. *et al.* (2019) Panton–Valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial GWAS. *eLife* 8, e42486
85. Lees, J.A. *et al.* (2019) Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat. Commun.* 10, 2176
86. Cremers, A.J.H. *et al.* (2019) The contribution of genetic variation of *Streptococcus pneumoniae* to the clinical manifestation of invasive pneumococcal disease. *Clin. Infect. Dis.* 68, 61–69
87. Gori, A. *et al.* (2020) Pan-GWAS of *Streptococcus agalactiae* highlights lineage-specific genes associated with virulence and niche adaptation. *mBio* 11, e00728-20
88. Allen, J.P. *et al.* (2020) A comparative genomics approach identifies contact-dependent growth inhibition as a virulence determinant. *Proc. Natl. Acad. Sci. U. S. A.* 117, 6811–6821
89. Bandoy, D.D.R. and Weimer, B.C. (2020) Biological machine learning combined with *Campylobacter* population genomics reveals virulence gene allelic variants cause disease. *Microorganisms* 8, 549
90. Habib, F. *et al.* (2007) Large scale genotype-phenotype correlation analysis based on phylogenetic trees. *Bioinformatics* 23, 785–788
91. Brynildsrud, O. *et al.* (2016) Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17, 238
92. Earle, S.G. *et al.* (2016) Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* 1, 16041
93. Thornton, T. and McPeek, M.S. (2010) ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* 86, 172–184
94. Lippert, C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835
95. Rahman, A. *et al.* (2018) Association mapping from sequencing reads using k-mers. *eLife* 7, e32920
96. Jaillard, M. *et al.* (2018) A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* 14, e1007758
97. Aun, E. *et al.* (2018) A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.* 14, e1006434
98. Drouin, A. *et al.* (2016) Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genom.* 17, 754