Genome Medicine

REVIEW

Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology

Timothy D Read[1,2*] and Ruth C Massey[3]

Abstract

Genome-wide association studies (GWASs) have become an increasingly important approach for eukaryotic geneticists, facilitating the identification of hundreds of genetic polymorphisms that are responsible for inherited diseases. Despite the relative simplicity of bacterial genomes, the application of GWASs to identify polymorphisms responsible for important bacterial phenotypes has only recently been made possible through advances in genome sequencing technologies. Bacterial GWASs are now about to come of age thanks to the availability of massive datasets, and because of the potential to bridge genomics and traditional genetic approaches that is provided by improving validation strategies. A small number of pioneering GWASs in bacteria have been published in the past 2 years, examining from 75 to more than 3,000 strains. The experimental designs have been diverse, taking advantage of different processes in bacteria for generating variation. Analysis of data from bacterial GWASs can, to some extent, be performed using software developed for eukaryotic systems, but there are important differences in genome evolution that must be considered. The greatest experimental advantage of bacterial GWASs is the potential to perform downstream validation of causality and dissection of mechanism. We review the recent advances and remaining challenges in this field and propose strategies to improve the validation of bacterial GWASs.

Introduction

Genome-wide association studies (GWASs) involve testing large numbers of genetic variants, usually single nucleotide polymorphisms (SNPs) or insertions and deletions (indels), within a population of individual organisms for statistically significant associations with a given phenotype [1]. The first successful GWAS in humans, published in 2005, examined a set of 96 patients with age-related macular degeneration, a condition that leads to loss of vision in older adults, and 50 matched controls [2]. Out of 116,204 SNPs tested, two were statistically significantly associated with the condition. One of the SNPs was found in the complement factor H gene, encoding a protein integral to host immunity, and the condition has since then been linked to autoimmunity [3]. Although there is some controversy about specific aspects of the approach [4], many GWASs have now been published, making hundreds of associations between SNPs and important human diseases [5].

GWASs are clearly an important tool for genetic analysis but their use in microbiological research has been relatively slow to emerge [6]. Smaller-scale genetic association studies in bacteria have been performed for a number of years. Early research used PCR and limited sequence data (for example, data from multi-locus sequence typing [7]) or comparative genome hybridization [8] to link bacterial phenotypes with the presence or absence of specific genes or with the clonal background of an isolate [9-14]. In human genetics, high-throughput genotyping of panels of common SNPs using microarrays and bead-based assays have been a mainstay for GWASs for the past 10 years [15]. The creation of SNP-typing panels is, however, generally associated with high fixed costs and so few platforms were custom-designed for bacterial species. Those that were designed for bacteria were practically limited to species with low nucleotide diversity (such as *Bacillus anthracis* [16]). This reality began to change in 2010 with the advent of large-scale genome sequencing using affordable and accurate data produced by

* Correspondence: tread@emory.edu
[1]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA 30322, USA
[2]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA
Full list of author information is available at the end of the article

Illumina HiSeq and MiSeq instruments. These instruments made generation of the whole genome sequence of 50 or more bacterial strains a routine experiment and opened the door for bacterial GWASs (Figure 1).

The first successful application of a GWAS to bacteria using shotgun sequence data was published in 2013 [17] (see Table 1). Sheppard *et al.* [17] used a novel association approach to probe the genetic factors responsible for host adaptation in 192 shotgun-sequenced *Campylobacter jejuni* and *C. coli* strains. In another publication in the same year, mutations in *Mycobacterium tuberculosis* genes responsible for resistance to anti-tuberculosis drugs were detected on the basis of their recurrent appearance in resistant lineages of a whole-genome phylogenetic tree [18]. Three studies published in 2014 have extended the use of GWASs on bacterial shotgun data. Laabei *et al.* [19] studied a collection of 90 methicillin-resistant *Staphylococcus aureus* clinical isolates and identified more than 100 polymorphisms that associated with the ability of the bacteria to lyse human cells. Alam *et al.* [20], also studying *S. aureus*, used a GWAS to determine mutations in the RNA polymerase *rpoB* gene that are significantly associated with the clinically important vancomycin-intermediate-resistant phenotype. The first GWAS to use a number of cases and controls on the scale commonly seen in human genetic research was recently published by Chewapreecha *et al.* [21]; these researchers sequenced 3,701 *Streptococcus pneumoniae* isolates to identify polymorphisms associated with beta-lactam resistance.

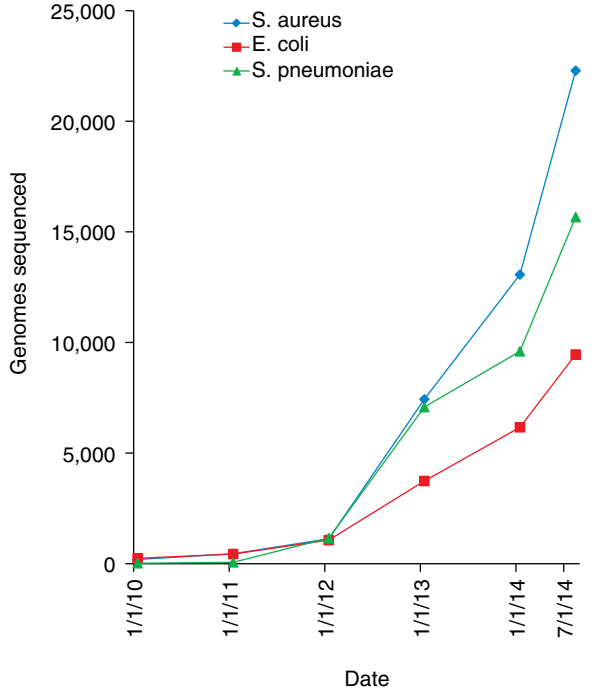

Figure 1 Growth in the cumulative number of next-generation sequencing runs in public databases for three important bacterial pathogens, *Staphylococcus aureus*, *Escherichia coli* and *Streptococcus pneumoniae*. The data were acquired by querying the National Center for Biotechnology Information Short Read Archive database and excluding datasets linked to RNA-seq experiments.

What is made clear by even these few early studies is that a GWAS is a powerful first step towards characterizing a phenotype at a population level. It is an unbiased screening approach to discover new loci that correlate with a specific phenotype. GWASs can form the basis of studies of the functionality of regulatory pathways and expression mechanisms and, when performed robustly, can be used to build predictive tools for the translation of genomic data into the clinical microbiology setting. Bridging the gap between genomics and traditional molecular genetics has the potential to uncover untapped levels of detail on how bacteria survive and cause disease. Discoveries could be used to personalize medicine so that treatments can be tailored for individual patients on the basis of the genome sequence of the infecting microbe. In this review, we discuss what should be taken into account when planning a bacterial GWAS, how to improve the validation of GWASs, how these studies are likely to impact on clinical microbiology in the future and what challenges remain.

Design considerations for bacterial GWASs

Bacterial GWAS is a brand new field. It is increasingly easy to generate genomic data, but there are challenges in identifying optimum GWASs strategies. Some of these challenges are also shared with eukaryotic GWASs, and, although there are many experiences and tools that can be drawn from eukaryotic studies (Table 2), caution should be used when translating approaches developed for different domains of life.

There are several prerequisites for a successful bacterial GWAS. There must be a testable phenotype and a set of bacterial strains with whole-genome sequences. Experimenters need to choose a statistical analysis strategy and perform power calculations to ensure that there are enough strains in their study to have a reasonable chance of successful association. None of these prerequisites are truly independent of one another.

Phenotypes

It is necessary to consider whether the phenotype to be tested by the GWAS is a continuously varying quantitative phenotype or a binary case versus control trait. A continuous phenotype can be subdivided into discrete categories, for instance using accepted breakpoints for antibiotic sensitivity to resistance [20]. Phenotypes for bacterial GWASs (such as host species, infection type, severity, or outcome) can be gleaned from metadata collected at the time of isolation of the strain or obtained by experimentation. It is important to make assessments about the consistency of

Table 1 Early bacterial genome-wide association studies based on whole-genome shotgun data

Organism	Sample size (isolates)	Phenotype	Finding	Genome-wide association study program used	Reference
Campylobacter jejuni and *C. coli*	192	Host adaptation	Vitamin B5 biosynthesis is important	30 bp 'word' searching [17]	[17]
Mycobacterium tuberculosis	123	Antibiotic resistance	39 novel resistance-associated loci	PhyC [18]	[18]
Staphylococcus aureus	75	Antibiotic resistance	Novel associated single nucleotide polymorphism in *rpoB* gene	ROADTRIPS [50]	[20]
S. aureus	90	Virulence	121 novel associated loci	PLINK [49]	[19]
Streptococcus pneumoniae	3,701	Antibiotic resistance	Multiple novel associated loci	PLINK [49]	[21]

the annotation, especially when the data come from multiple sources. In the case of experimental phenotypes, the need to perform the assays on very large numbers of strains will tend to limit experiments to those phenotypes that can be assayed in a simple and relatively inexpensive way. For these reasons, the early studies have concentrated on phenotypes such as antibiotic resistance [18,20,21] and *in vitro* toxicity [19].

In considering the genetic basis of the phenotype, it is important to have an idea of the effect sizes: a measure of the correlation of the variant with the phenotype. Effect sizes vary from 0 to 1, with 1 meaning that the phenotype is completely explained by the variant. Many bacterial variants (such as antibiotic-resistance mutations) are assumed to have very large effects, akin to a Mendelian trait in eukaryotes, because they are necessary for the survival of the cell. However, bacterial phenotypes that are influenced mainly by low-effect variants surely exist, and the use of GWASs is probably the only feasible approach to determining their genetic basis.

Genetic variation and population structure in bacterial strains

GWASs are dependent for their success on the way the genetic variants to be tested (for example, SNPs) are distributed among the genomes of the subject population. There are distinct differences in the dynamics of genetic variation between humans (and other higher diploids) and bacteria. In humans, genetic recombination and chromosome segregation, necessary for shuffling alleles, occurs each generation. A newly occurring mutation will be genetically linked to neighboring alleles as part of the same haplotype until a recombination event occurs to break the linkage. The extent that any two alleles within a population are on the same ancestral 'haplotype block' of DNA is termed their linkage disequilibrium (LD) and usually decreases with genetic distance on the chromosome. This mixing of alleles between different genetic backgrounds is important for distinguishing causal loci from passively linked mutations. Asexual bacterial reproduction does not offer the opportunity to exchange genetic information this frequently. There are instead three natural mechanisms

Table 2 Similarities and differences between bacterial and eukaryotic genome-wide association study approaches

Feature	Bacteria	Eukaryote
Ploidy	Haploid	Diploid
Genetic re-assortment	Infrequent short gene conversion and horizontal gene transfer events	Homologous recombination and chromosome segregation linked to reproduction
Accessory (non-core) genes	Variable numbers in different species	Rare
Linkage disequilibrium	Variable across the genome and between species	Variable across the genome
Population structure	Asexual, generally highly structured, except for relatively rare homologous recombination events	Sexual, variable allele frequencies in subpopulations owing to non-random mating, ancestral divergence, drift
Confounders in genome-wide association studies	Population structure	Population structure
How to move from association to causality	Genetic reconstruction of mutations in laboratory strains, transposon mutant screens	Forward genetics in animal models or cultured tissue systems; linkage to known genetic diseases; large monogenic association studies
Current burden of proof for causality	Molecular Koch's Postulates	Combined genetic and experimental evidence

that generate the variability needed for GWASs: gene acquisition through horizontal gene transfer (HGT) and non-homologous recombination, gene conversion through homologous recombination, and recurrent mutation (Figure 2). In each case, these processes can create homoplasy, which is the presence of a similar genetic locus (SNPs, indels, genes and so on) on different branches of the phylogeny.

Insertion of complete genes as a result of HGT can generate diversity for association testing in bacteria (Figure 2a) [22]. The three classical mechanisms of HGT are transduction by bacteriophages, transformation of DNA segments, and plasmid-mediated conjugation. Genome sequencing of multiple isolates within bacterial species has given rise to the concept of a 'pan-genome' [23], which consists of a core of genes present in every strain and all of the accessory genes (defined as those found in some but not all members of the sequenced population). Depending on the bacterial species, accessory genes may encode virulence factors, antibiotic resistance determinants, or other loci that contribute to the adaptation of the bacterium to its environment [24]. Ideally for GWASs, these genes should be acquired multiple times by different lineages. Deletion of accessory genes is a process that is effectively the reverse of HGT in creating the variable presence of accessory genes across strains and lineages of a species [25].

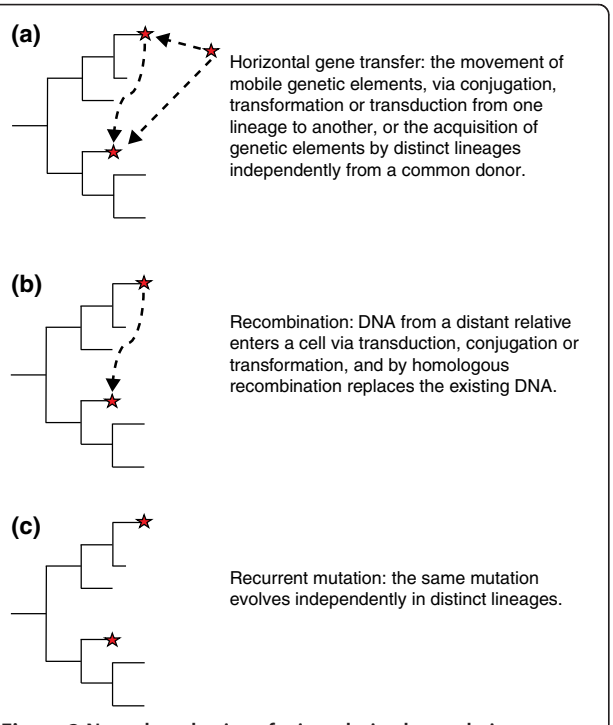

Figure 2 Natural mechanisms for introducing homoplasious mutations into the genomes of bacterial populations. Homoplasious mutations are necessary for association studies. The figure represents three mechanisms for forming an identical genetic variant (red star) on different lineages of a hypothetical phylogenetic tree of bacterial strains: **(a)** horizontal gene transfer, **(b)** recombination and **(c)** recurrent mutation.

In bacteria, homologous recombination happens after unidirectional transfer of DNA sequence into the recipient via HGT, leading effectively to gene conversion (Figure 2b) [26]. These events are rare, and generally do not occur at every generation, even in highly promiscuous bacterial species [27]. Exchanged DNA segments tend to be small (hundreds to a few thousand bp, although rarely larger events of more than 10 kb have been reported [28]), and typically create a patchwork of islands of introduced variation across the genome. Recombination results in a decay of LD across bacterial genomes that varies in rate in different species [29]. Several studies have shown recombination to be a mechanism used for adaptation. An example of this involves the mosaic *penA* allele XXXIV, derived from recombination between *Neisseria gonorrhoeae* and a commensal strain that confers resistance to cephalosporin antibiotics [30]. The *penA* XXXIV allele has been introduced by recombination into multiple *N. gonorrhoeae* lineages [31]. In another study that examined natural patterns of gene conversion, unidirectional transfer of DNA segments into diverse lineages was also found to be responsible for rapid adaptation to aquatic sub-niches by *Vibrio cyclitrophicus* [32].

Recurrent mutation of genetic variants within different lineages of a species as a response to selection offers a third way to create homoplasious genetic loci (Figure 2c). This can happen often in bacteria because of large local population sizes (sometimes billions of cells within a single infection). One example of a recurrent mutation is that which causes the H481Y codon change in the *rpoB* gene; this mutation has occurred in multiple *S. aureus* lineages and confers intermediate levels of resistance to vancomycin [20].

Bacterial species differ considerably in genetic diversity and show characteristic historical rates of recombination, HGT and recurrent mutation [26,27,29]. Many bacterial species are highly clonal, and exchange DNA through homologous recombination infrequently. In these species, recurrent mutation will be very important for genetic association [18]. *M. tuberculosis*, the causative agent of tuberculosis, is a classic example of a near-clonal species, with only 1.1% homoplasic SNPs within its core genome [33]. Rates of recombination (as measured by fixed events) also vary between species [27,34]. In one example, the Gram-negative pathogen *Chlamydia trachomatis*, gene conversion frequencies have been found to be higher in hotspots such as the *OmpA* major outer member protein gene [35], which is under diversifying selection for immune evasion. In *S. aureus*, horizontally transferred genes and regions surrounding them recombine at higher frequency than the core genome [36,37].

Another important aspect to consider when designing a bacterial GWAS is population structure. Populations

of a species are considered to be structured if they contain a non-random distribution of alleles within subpopulations. Population structure in humans can occur through mechanisms such as genetic drift, ancestral divergence [38] and non-random mating within subpopulations [39]. The stratification of human populations is reflected in complex patterns of LD in different parts of the chromosome and in different subgroups [40]. Importantly, population structure may confound GWASs, especially if it is not recognized, by causing the appearance of higher than expected allele frequencies within certain members of the study set [41]. Problems relating to structured genetic variation would be expected to be worse in bacterial strains than in human populations as bacteria are haploid and asexual. In the absence of recombination, all fixed genetic variants will be passed on to descendants and be in LD with other mutations that occur in that lineage. The separation of causative variants from passive linked loci is potentially a difficult problem.

The problem of population structure has been addressed in bacterial GWASs by using phylogenetic approaches [18,21], by using clustering followed by permutation [19], and by using databases of known variation to identify common mutations [20]. For future experimental design, it should also be possible not only to study variation in naturally occurring populations but also to utilize laboratory-induced mutation and recombination techniques to generate banks of strains that have artificial homoplasies [42].

Markers for bacterial GWASs

Whole genes, SNPs, indels or other loci such as mobile genetic elements [10] can be used as markers in GWASs. The quality of the DNA sequence data is an important consideration for experimental design. Because of the small genome size of bacteria it is now rare for Illumina shotgun projects to have average coverage (the number of sequence reads per base) of less than 20. At this level of redundancy, the confidence of the consensus base-calling accuracy is high [43,44]. Furthermore, the portion of the genome represented by multiple sequencing reads is also high, making the problem of imputation of missing genotypes small relative to human studies [45]. The increasing use of single molecule long-read sequencing technologies, which can produce complete or near-complete genome sequences following *de novo* assembly [46], will help to reduce the frequency of missing larger loci (such as genes or intergenic regions) in bacterial genomes.

SNPs are the most common units used as markers in GWASs. SNPs are commonly detected by comparison to a reference sequence, which can lead to ascertainment bias: the strains that are more genetically similar to the reference tend to have more accurate SNP calls. An alternative approach is to use 'reference-free' multiple alignment methods [47,48]. The penalty for these approaches, which use short sequence words (k-mers) for matching, is that multiple SNPs that occur in close proximity (less than the chosen word length) might not get reported. For convenience, early studies have focused on SNPs found in core regions of the genome (or in accessory genes that are found in all strains in the comparison set). Developing a strategy for the treatment of SNPs in accessory genes that are present in some strains but not in others will be important for bacterial GWASs. These are not missing data, as encountered in human projects with low sequence coverage [45]. One possible approach could be to run an association test for each accessory gene SNP using just the strains in which it occurs separate from the core genome GWAS.

An alternative to focusing on SNPs is to use k-mers. The *Campylobacter* GWAS by Sheppard *et al*. [17] used 30 bp 'words' extracted from the assembled genome sequences as the unit for association, each of which was tested against the species origin of isolation. The advantage of this approach was that it allowed discovery of multiple types of variants (SNP, indels and gene insertions) without requiring a genome alignment.

Bacterial GWAS statistical analysis approaches and software

There are many tools developed for human GWASs available for porting to bacterial datasets. Some consideration of the differences between bacterial and eukaryotic genetics will be needed when assigning parameters (Table 2). The popular PLINK [49] software for regression-based association of both quantitative and case versus control studies has been used (Table 1). In the study by Chewapreecha *et al*. [21], the Cochran-Mantel-Haenzel test was used to correct for genetic background in discovering SNPs that are associated with beta-lactam resistance in two genetically different *S. pneumoniae* population clusters. Alam *et al*. [20] used ROADTRIPS [50], a regression-based approach that incorporates corrections for both known and inferred population structure.

Two phylogeny-based approaches for association have been developed specifically for bacteria. In the Predict Phenotypes From SNPs package outlined by Hall [51], SNPs were associated with phenotypic changes inferred in internal branches of the whole-genome phylogeny. This method utilized template-free genome assembly and tree construction based on the kSNP software [47]. The phylogenetic convergence or 'PhyC' approach [18] looked at recurrent mutations on the tips and internal nodes of the phylogenetic tree, assuming that mutations occurred recently under strong selection. Significance was tested using a permutation approach to ask whether the number of times a SNP occurred on branch leading to an

antibiotic-resistant strain versus an antibiotic-sensitive strain was unusual in the population.

Calculation of statistical power

Software that estimates statistical power allows researchers to calculate the number of cases and controls needed to have a realistic chance of rejecting the null hypothesis (that there is no association between the variant and the phenotype) when the alternative hypothesis is indeed true. For example, a calculation may yield the number of strains necessary to have an 80% chance of detecting an association with an effect size of 0.5 or greater with a *P*-value threshold of 0.05. Power calculations have been important in human GWASs for improving the experimental design to increase the probability of obtaining a statistically meaningful result [52], and there are now a myriad of software packages available to researchers [40,53,54]. Commonly included variables that tend to increase power include larger effect-size cutoff, reduced population structure, and increased sequence quality [55].

The number of genetic loci to be tested is an important variable in statistical power calculations. Multiple tests of significance increase the chances of false-positive calls. For example, if 20 randomly selected loci are tested independently at the standard 0.05 significance threshold, one locus would be expected by chance to be a false positive. A conservative Bonferroni correction for the number of hypothesis tests in the study is usually imposed in order to reduce false-positive calls. Experimental designs that reduce the number of genetic variants tested serve to increase power. One way to reduce the number of tests is to select a subpopulation of the original set strains with a smaller number of total SNPs. Other strategies include disregarding low-frequency mutations and/or mutations that cause synonymous mutations or SNPs in intergenic regions, or treating all individual mutations within a genetic feature (a gene, intergenic region and so on) as having the same aggregate effect. The risk in removing rare mutations from the study is that they may be important for the phenotype, as has been found in several human diseases. This was also the case in the Laabei *et al.* study [19] where four novel toxicity-affecting intergenic loci were identified and their effect verified by mutagenesis. Permutation tests using scrambled cases and controls can also be used to increase statistical power [21,52]. Finally, false discovery rate could be used as an alternative to significance thresholds for identifying candidate loci [56].

Simple power models [52] may have value in offering a starting point when considering study size. The experience in human genetics is that the sophistication of power statistics has increased as knowledge of the population structure has improved [40]. Because of the immense variation in bacterial species genetics, empirical calculations using simulated genome datasets may be particularly important for experimental design. A software package for designing experiments based on recurrent mutations between matched pairs of cases and controls has recently been developed [57]. From the evidence of the early bacterial GWASs (Table 1), quite a small number of cases and controls (n =75) might be required to find variants associated with phenotype that have a large effect size. Future GWASs with experimental design informed by basic studies on bacterial species population structure and involving increasingly large collections of phenotypically characterized strains may be able to unearth larger numbers of small-effect variants.

Validating the results of GWASs: bridging the gap between genomics and traditional microbial molecular genetics

GWASs on bacteria has already yielded interesting new loci that are associated with clinically important phenotypes, but how can we be confident that these associations are causative or functionally linked? This question has been examined in depth in human studies (Table 2). Significance tests implemented in GWAS software necessarily rely on assumptions, such as a lack of cryptic population structure and consistent rates of mutation across evolutionary history, that may produce higher error rates than the *P*-values suggest [41]. Experimental errors in base-calling and phenotyping could also contribute to spurious results. We know from the experience of human GWASs that some loci found to be associated with a trait can turn out to have little or no functional significance [58]. Therefore, unless the associated locus has been previously shown to affect the phenotype, functional validation is desirable [19]. The questions that surround the strategy for functional validation are part of an ongoing dialog between two apparently diametrically opposed experimental philosophies in modern microbiology: the 'top down' unbiased, genomics-based approaches (which include GWASs and other experimental strategies [59-62]), and the 'bottom-up', gene-by-gene approach of classical molecular genetics (Figure 3) [6]. The disconnect is that, on the one hand, we will eventually have thousands of genome sequences of every bacterial pathogen, whereas on the other hand, the current *modus operandi* of molecular genetics is focused on fine-scale analysis of individual proteins in a very small number of isolates. The coming of GWASs will hopefully speed the genesis of a powerful synthesis between these two approaches.

Traditional molecular genetic approaches have been instrumental in carefully dissecting the functions of thousands of bacterial genes, sometimes down to the level of highly complex interactions between host cells and pathogens that lead to disease (such as Type III secretion or superantigens [63,64]). Typically, researchers seek to design systems to examine discrete phenotypes, where upon

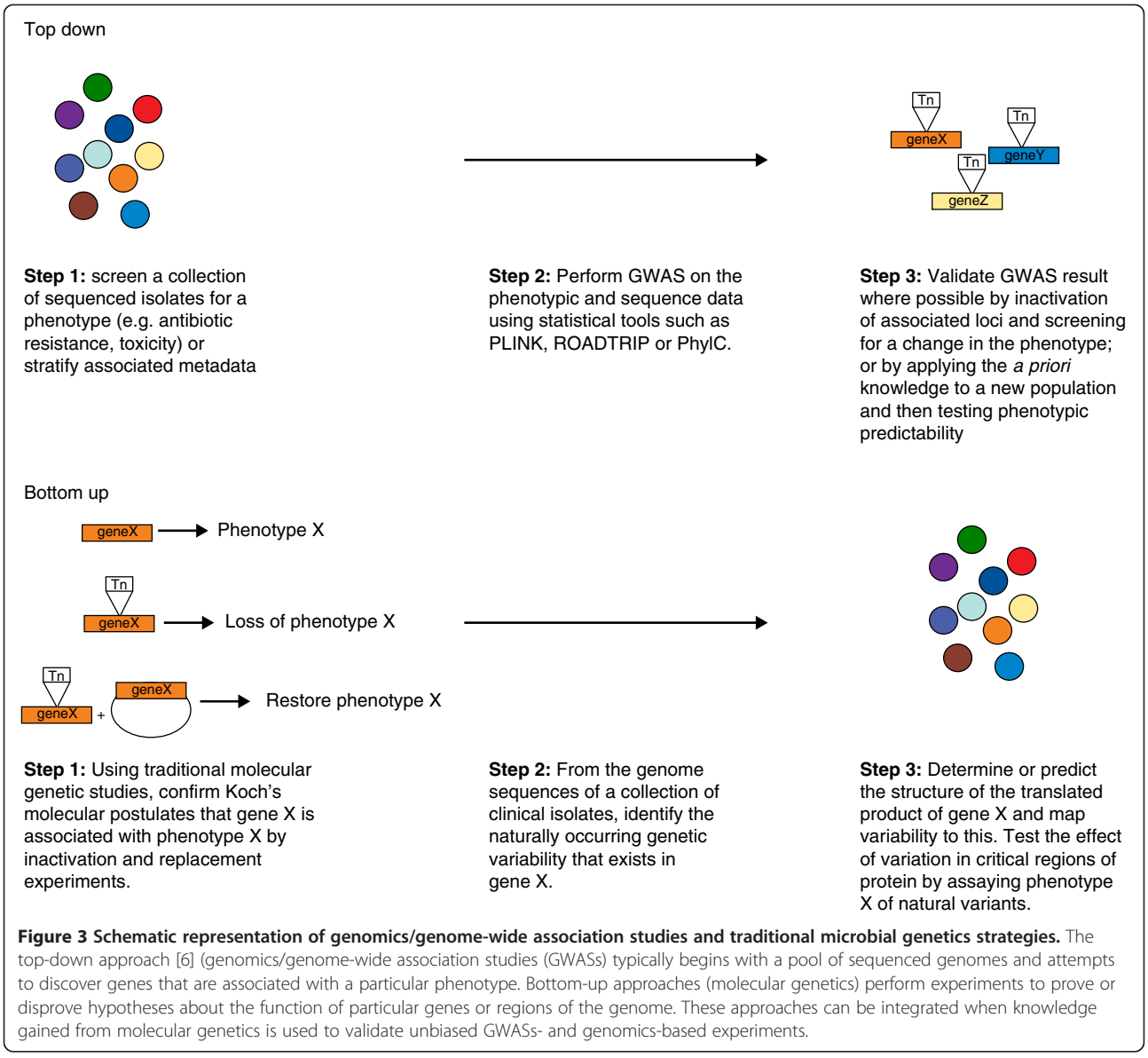

Figure 3 Schematic representation of genomics/genome-wide association studies and traditional microbial genetics strategies. The top-down approach [6] (genomics/genome-wide association studies (GWASs) typically begins with a pool of sequenced genomes and attempts to discover genes that are associated with a particular phenotype. Bottom-up approaches (molecular genetics) perform experiments to prove or disprove hypotheses about the function of particular genes or regions of the genome. These approaches can be integrated when knowledge gained from molecular genetics is used to validate unbiased GWASs- and genomics-based experiments.

mutation (directed or random), the loss or gain of a specific phenotype can be efficiently screened or selected. Depending on the activity of the gene in question, further specific molecular or cellular experiments follow to characterize the mechanisms in detail. This approach is tremendously powerful in manipulating the microorganism and the environment to test precise hypotheses within the artificial confines of the laboratory. Since the 1980s, the dominant paradigm for linking genes to phenotype in microbiology has been based on the Molecular Koch's Postulates, outlined by Falkow [65]. These state that disruption and reconstruction of the gene under investigation coupled with loss and regain of the phenotype is needed for firm proof of a functional role. Molecular Koch's Postulates are often used as a stringent standard for validation, although the original article offered a nuanced discussion of some of the difficulties in their application to all situations [65].

Validation by genetic disruption and reconstruction can be applied to GWASs results, especially for microorganisms for which genome-wide transposon mutant libraries are available, such as *S. aureus*, *Escherichia coli*, *Streptococcus pneumoniae*, *Pseudomonas aeruginosa*, *Yersinia pseudotuberculosis* and *Salmonella enterica* [60,66,67]. Nevertheless, there can be situations in which laboratory genetics are more challenging or even impossible, for example when the identified polymorphism is in an essential gene, or when the species being studied is not amenable to genetic manipulation. We are also increasingly sampling beyond where the traditional microbiology laboratory can venture, sequencing single cells [68], and reconstructing genomes directly from environmental DNA [69,70]. In these circumstances, it may

be possible to use a model genetic organism such as *E. coli* to test for the phenotypic effect of a mutation, but any result may not be considered a direct validation under the Molecular Koch's Postulates rules.

There is also the problem of potential epistatic interactions between genes and the contribution of non-core, accessory genes to the phenotype. If a reconstructed mutant strain does not have the expected phenotype, this could result from the lack of a specific interacting allele in the host strain, or possibly a missing non-core gene. No single strain can ever represent a species, but the strains commonly used for genetic reconstruction may be especially poor choices because of their long history of laboratory adaptation [71]. Laboratory strains are chosen because they are locally available and have familiar, useful properties: generally fast growth and easy genetic manipulation. As a consequence, laboratory strain phenotypes often do not represent the majority of the species. The quixotic properties of certain laboratory strains have misled generations of scientists about the true nature of their subject organisms. For example, the ubiquitous genetic workhorse, *Bacillus subtilis* 168 is a very rare naturally transformable strain within its species (it is also a non-swarming tryptophan auxotroph, amongst other unusual features [72]), and the *S. aureus* genetic strain 8325-4 has a mutation in the *sigB* locus that causes an enhanced toxic profile [73].

If the one-at-a-time genetic reconstruction method is unlikely to work for all variants discovered through GWASs, and in some cases may produce misleading results because of complex gene interactions, statistical modeling may also be able to provide an alternative type of validation. Commonly, machine-learning techniques such as support vector machines and random forests [74] can be trained on a reserved portion of the dataset and then tested on the remainder. Random forests were used to make reliable predictions of an individual isolates' level of toxicity and vancomycin-intermediate phenotype [19,20]. Although a successful model would not be able to explain the mechanistic contribution of the loci, it would inform that sufficient information on the genetic basis of the phenotype for sensitive prediction had been learned.

Ultimately, it is likely that combining molecular genetic and statistical modeling approaches will be fruitful. In a hypothetical situation in which GWASs results in more than 200 loci that are significantly associated with a complex phenotype, validating the effect of the top 20 most important mutations might allow the statistical model to predict the phenotype accurately in more than 95% of unknown strains. There has been interest in developing methods to prioritize variants discovered in human GWASs [75], and potentially some of these approaches can be applied to the bacterial realm. Further on in the future, systems biology and systems genetics approaches to high dimensional data integration may offer an alternative to 'one gene at a time' genetic validation [76,77].

How will GWASs affect clinical microbial diagnostics?

Bacterial GWASs have the potential to deepen our understanding of phenotypic variation across pathogenic species. This information will be particularly useful in the future as we attempt to interpret genome sequences that are routinely produced by clinical microbiology laboratories. There is great interest in the development of whole-genome sequencing for clinical diagnostics of pathogens [78-81] because it is possible to envisage genomics technology maturing to the extent that *de novo* sequencing becomes a relatively cheap and rapid assay. Whole-genome sequence data have numerous advantages over the directed PCR-based tests that currently dominate this arena. Unlike shotgun genomics, PCR relies on the presence of highly conserved DNA sequences for primer binding and yields false-negative results when these are mutated, as happened, for example, with a plasmid-borne marker for *C. trachomatis* [82]. Importantly, the whole-genome sequence also allows unbiased discovery of other information about the strains that the clinician may not have considered, such as the unexpected presence of antibiotic-resistance genes.

To take advantage of our ability to acquire the genome sequence of a pathogen rapidly ahead of the results of a laboratory-based phenotypic test, such as an antibiotic minimal inhibitory concentration (MIC) test, we must be able to not only call drug sensitivity on the basis of the genome sequence alone but also know the reliability of the assignment. Several schemes for predicting drug resistance have already been developed, based on knowledge obtained from early comparative genomics and genetic knockout studies [83,84]. Further development of these diagnostic tests will necessitate understanding how the activities of well known genes are influenced by epistatic interactions within the pathogen species. For the reasons we have outlined earlier, GWASs provide the natural training set data to build statistical models that predict phenotypes by integrating genetic variation across the entire genome. Another advantage of a test that is based on trained genomic data is that variability in how the phenotype is measured is no longer a problem. Many clinically relevant phenotypes are ascertained using a plethora of different technologies and are variable across different conditions. MIC, for example, can be determined by disk diffusion, test strips, spiral plating, or several other methods. GWASs performed on a genetically diverse set of strains measured using gold-standard phenotypic assays could be used to train models that effectively replace much routine clinical antimicrobial-resistance testing.

Large-scale clinical sequencing could provide a pool of thousands of new genomes for GWASs that could discover

variants that have ever-smaller effect. Existing statistical models could also be tested and refined with the new clinical data. For this feedback cycle to occur, we will need to improve and make more efficient our collection of metadata (time and place of isolation, clinical manifestations, phenotype tests and so on). Several schema for organizing bacterial strain metadata have been proposed [85,86]. Even today, when it is possible to sequence 96 or more strains each day on a bench-top instrument, it is a feat of organization to manually gather metadata retrospectively for submission with the genomes to public databases. For us to keep up with future throughput, we need systems that facilitate information storage at the time of isolation and phenotypic testing. This will be a challenge, particularly in the high-throughput, time-pressured environment of the clinical microbiology laboratory. There is also an issue with access to collections of sequenced isolates. Many organizations make sequence data available in public databases, but either do not maintain the bacterial collections from which the sequenced DNA was extracted or are unable to bear the costs of making large sets of strains available to the research community. The solution is to have regular accession of large numbers of sequenced isolates with high-quality metadata from clinical and academic laboratories into public strain collections, but this will need new organization and funding.

Conclusions and perspectives

GWAS in bacteria is a new research opportunity that is being driven forward by advances in genome-sequencing technology. Although in its infancy, the early studies have shown it to be not only a reliable method to identify loci that affect a phenotype but also a powerful tool to uncover new levels of complexity in the expression of clinically important bacterial traits. The approaches and tools used to do this are likely to adapt and develop as we sample ever-greater numbers of bacterial genomes that are associated with high-quality metadata. What is clear is that GWASs represent a versatile and highly productive approach to maximizing the usefulness of the genomic data available to us from both laboratory and clinical settings.

Abbreviations

GWASs: Genome-wide association studies; HGT: Horizontal gene transfer; indel: Insertion and deletion; LD: Linkage disequilibrium; MIC: Minimal inhibitory concentration; PCR: Polymerase chain reaction; SNP: Single nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The manuscript resulted from a visit by TDR to RCM's laboratory that was funded by the Raymond Schinazi International Scientific Exchange Program (SIEP). The SIEP had no role in writing or editing this manuscript. Thanks to Jesse Shapiro and Sandeep Joseph for reading and commenting on draft versions.

Author details

[1]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA 30322, USA. [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA. [3]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.

Published online: 22 November 2014

References

1. Corvin A, Craddock N, Sullivan PF: **Genome-wide association studies: a primer.** *Psychol Med* 2010, **40:**1063–1077.
2. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration.** *Science* 2005, **308:**385–389.
3. Anderson DH, Radeke MJ, Gallo NB, Chapin EA, Johnson PT, Curletti CR, Hancox LS, Hu J, Ebright JN, Malek G, Hauser MA, Rickman CB, Bok D, Hageman GS, Johnson LV: **The pivotal role of the complement system in aging and age-related macular degeneration: hypothesis re-visited.** *Prog Retin Eye Res* 2010, **29:**95–112.
4. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery.** *Am J Hum Genet* 2012, **90:**7–24.
5. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** *Nucleic Acids Res* 2014, **42:** D1001–D1006.
6. Falush D, Bowden R: **Genome-wide association mapping in bacteria?** *Trends Microbiol* 2006, **14:**353–355.
7. Enright MC, Spratt BG: **Multilocus sequence typing.** *Trends Microbiol* 1999, **7:**482–487.
8. Zhou J: **Microarrays for bacterial detection and microbial community analysis.** *Curr Opin Microbiol* 2003, **6:**288–294.
9. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, O'Neill G, Day NPJ: **Virulent combinations of adhesin and toxin genes in natural populations of Staphylococcus aureus.** *Infect Immun* 2002, **70:**4987–4996.
10. Bille E, Zahar J-R, Perrin A, Morelle S, Kriz P, Jolley KA, Maiden MCJ, Dervin C, Nassif X, Tinsley CR: **A chromosomally integrated bacteriophage in invasive meningococci.** *J Exp Med* 2005, **201:**1905–1913.
11. Howard SL, Gaunt MW, Hinds J, Witney AA, Stabler R, Wren BW: **Application of comparative phylogenomics to study the evolution of *Yersinia enterocolitica* and to identify genetic differences relating to pathogenicity.** *J Bacteriol* 2006, **188:**3645–3653.
12. Herron-Olson L, Fitzgerald JR, Musser JM, Kapur V: **Molecular correlates of host specialization in *Staphylococcus aureus*.** *PLoS One* 2007, **2:**e1120.
13. Harrison OB, Evans NJ, Blair JM, Grimes HS, Tinsley CR, Nassif X, Kriz P, Ure R, Gray SJ, Derrick JP, Maiden MCJ, Feavers IM: **Epidemiological evidence for the role of the hemoglobin receptor, hmbR, in meningococcal virulence.** *J Infect Dis* 2009, **200:**94–98.
14. Bessen DE, Kumar N, Hall GS, Riley DR, Luo F, Lizano S, Ford CN, McShan WM, Nguyen SV, Dunning Hotopp JC, Tettelin H: **Whole-genome association study on tissue tropism phenotypes in group A *Streptococcus*.** *J Bacteriol* 2011, **193:**6651–6663.
15. Ragoussis J: **Genotyping technologies for genetic research.** *Annu Rev Genomics Hum Genet* 2009, **10:**117–133.
16. Zwick ME, Thomason MK, Chen PE, Johnson HR, Sozhamannan S, Mateczun A, Read TD: **Genetic variation and linkage disequilibrium in *Bacillus anthracis*.** *Sci Rep* 2011, **1:**169.
17. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J, Falush D: **Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*.** *Proc Natl Acad Sci U S A* 2013, **110:**11923–11927.
18. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M: **Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*.** *Nat Genet* 2013, **45:**1183–1189.

19. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L, Peacock SJ, Holden M, Wilson D, Read TD, van den Elsen J, Priest NK, Feil EJ, Hurst LD, Josefsson E, Massey RC: **Predicting the virulence of MRSA from its genome sequence.** *Genome Res* 2014, **24:**839–849.
20. Alam MT, Petit RA 3rd, Crispelll EK 3rd, Thornton TA, Conneely KN, Jiang Y, Satola SW, Read TD: **Dissecting vancomycin intermediate resistance in *Staphylococcus aureus* using genome-wide association.** *Genome Biol Evol* 2014, **6:**1174–1185.
21. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, Turner P, Bentley SD, Parkhill J: **Comprehensive Identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes.** *PLoS Genet* 2014, **10:**e1004547.
22. Thomas CM, Nielsen KM: **Mechanisms of, and barriers to, horizontal gene transfer between bacteria.** *Nat Rev Microbiol* 2005, **3:**711–721.
23. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, *et al*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102:**13950–13955.
24. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome.** *Trends Genet* 2009, **25:**107–110.
25. Mira A, Ochman H, Moran NA: **Deletional bias and the evolution of bacterial genomes.** *Trends Genet* 2001, **17:**589–596.
26. Didelot X, Lawson D, Darling A, Falush D: **Inference of homologous recombination in bacteria using whole-genome sequences.** *Genetics* 2010, **186:**1435–1449.
27. Didelot X, Maiden MC: **Impact of recombination on bacterial evolution.** *Trends Microbiol* 2010, **18:**315–322.
28. Somboonna N, Wan R, Ojcius DM, Pettengill MA, Joseph SJ, Chang A, Hsu R, Read TD, Dean D: **Hypervirulent *Chlamydia trachomatis* clinical strain is a recombinant between lymphogranuloma venereum (L(2)) and D lineages.** *MBio* 2011, **2:**e00045–11.
29. Shapiro BJ, David LA, Friedman J, Alm EJ: **Looking for Darwin's footprints in the microbial world.** *Trends Microbiol* 2009, **17:**196–204.
30. Ito M, Deguchi T, Mizutani K-S, Yasuda M, Yokoi S, Ito S-I, Takahashi Y, Ishihara S, Kawamura Y, Ezaki T: **Emergence and spread of *Neisseria gonorrhoeae* clinical isolates harboring mosaic-like structure of penicillin-binding protein 2 in central Japan.** *Antimicrob Agents Chemother* 2005, **49:**137–143.
31. Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, Weinstock H, Parkhill J, Hanage WP, Bentley S, Lipsitch M: **Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study.** *Lancet Infect Dis* 2014, **14:**220–226.
32. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ: **Population genomics of early events in the ecological differentiation of bacteria.** *Science* 2012, **336:**48–51.
33. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S: **Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans.** *Nat Genet* 2013, **45:**1176–1182.
34. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D: **Efficient inference of recombination hot regions in bacterial genomes.** *Mol Biol Evol* 2014, **31:**1593–1605.
35. Joseph SJ, Didelot X, Rothschild J, de Vries HJC, Morré SA, Read TD, Dean D: **Population genomics of *Chlamydia trachomatis*: insights on drift, selection, recombination, and population structure.** *Mol Biol Evol* 2012, **29:**3933–3946.
36. Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, Feil EJ: **The impact of recombination on dN/dS within recently emerged bacterial clones.** *PLoS Pathog* 2011, **7:**e1002129.
37. Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, Charlesworth J, Golubchik T, Ip CLC, Godwin H, Fung R, Peto TEA, Walker AS, Crook DW, Wilson DJ: **Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*.** *Nat Commun* 2014, **5:**3956.
38. Price AL, Helgason A, Palsson S, Stefansson H, St. Clair D, Andreassen OA, Reich D, Kong A, Stefansson K: **The impact of divergence time on the nature of population structure: an example from Iceland.** *PLoS Genet* 2009, **5:**e1000505.
39. Sebro R, Hoffman TJ, Lange C, Rogus JJ, Risch NJ: **Testing for non-random mating: evidence for ancestry-related assortative mating in the Framingham heart study.** *Genet Epidemiol* 2010, **34:**674–679.
40. Spencer CCA, Su Z, Donnelly P, Marchini J: **Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.** *PLoS Genet* 2009, **5:**e1000477.
41. Price AL, Zaitlen NA, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies.** *Nat Rev Genet* 2010, **11:**459–463.
42. Nguyen BD, Valdivia RH: **Forward genetic approaches in *Chlamydia trachomatis*.** *J Vis Exp* 2013, **23:**e50636.
43. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms.** *Nat Biotechnol* 2012, **30:**434–439.
44. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marçais G, Pop M, Yorke JA: **GAGE: a critical evaluation of genome assemblies and assembly algorithms.** *Genome Res* 2012, **22:**557–567.
45. Porcu E, Sanna S, Fuchsberger C, Fritsche LG: **Genotype imputation in genome-wide association studies.** *Curr Protoc Hum Genet* 2013, **Chapter 1:** Unit 1.25.
46. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM: **Reducing assembly complexity of microbial genomes with single-molecule sequencing.** *Genome Biol* 2013, **14:**R101.
47. Gardner SN, Hall BG: **When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes.** *PLoS One* 2013, **8:**e81760.
48. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E: **Automated reconstruction of whole-genome phylogenies from short-sequence reads.** *Mol Biol Evol* 2014, **31:**1077–1088.
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
50. Thornton T, McPeek MS: **ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure.** *Am J Hum Genet* 2010, **86:**172–184.
51. Hall BG: **SNP-associations and phenotype predictions from hundreds of microbial genomes without genome alignments.** *PLoS One* 2014, **9:**e90490.
52. Sham PC, Purcell SM: **Statistical power and significance testing in large-scale genetic studies.** *Nat Rev Genet* 2014, **15:**335–346.
53. Purcell S, Cherny SS, Sham PC: **Genetic power calculator: design of linkage and association genetic mapping studies of complex traits.** *Bioinformatics* 2003, **19:**149–150.
54. Feng S, Wang S, Chen C-C, Lan L: **GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits.** *BMC Genet* 2011, **12:**12.
55. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing: implications for design of complex trait association studies.** *Genome Res* 2011, **21:**940–951.
56. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100:**9440–9445.
57. Farhat MR, Shapiro BS, Sheppard SK, Colijn C, Murray M: **A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens.** *Genome Med* 2014, **6:**101.
58. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C: **Guidelines for investigating causality of sequence variants in human disease.** *Nature* 2014, **508:**469–476.
59. Dyer M, Murali T, Sobral B: **The landscape of human proteins interacting with viruses and other pathogens.** *PLoS Pathog* 2008, **4:**e32.
60. Van Opijnen T, Bodi KL, Camilli A: **Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.** *Nat Methods* 2009, **6:**767–772.

61. Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using RNA-seq.** *Curr Opin Microbiol* 2010, **13:**619–624.
62. Kharchenko PV, Tolstorukov MY, Park PJ: **Design and analysis of ChIP-seq experiments for DNA-binding proteins.** *Nat Biotechnol* 2008, **26:**1351–1359.
63. Xu SX, McCormick JK: **Staphylococcal superantigens in colonization and disease.** *Front Cell Infect Microbiol* 2012, **2:**52.
64. Moest TP, Méresse S: **Salmonella T3SSs: successful mission of the secret (ion) agents.** *Curr Opin Microbiol* 2013, **16:**38–44.
65. Falkow S: **Molecular Koch's Postulates applied to microbial pathogenicity.** *Rev Infect Dis* 1988, **10:** S274–S276.
66. Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, Bayles KW: **A genetic resource for rapid and comprehensive phenotype screening of nonessential *Staphylococcus aureus* genes.** *MBio* 2013, **4:**e00537–12.
67. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM: **An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants.** *Proc Natl Acad Sci U S A* 2006, **103:**2833–2838.
68. McLean JS, Lombardo M-J, Ziegler MG, Novotny M, Yee-Greenbaum J, Badger JH, Tesler G, Nurk S, Lesin V, Brami D, Hall AP, Edlund A, Allen LZ, Durkin S, Reed S, Torriani F, Nealson KH, Pevzner PA, Friedman R, Venter JC, Lasken RS: **Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform.** *Genome Res* 2013, **23:**867–877.
69. Sharon I, Banfield JF: **Microbiology. Genomes from metagenomics.** *Science* 2013, **342:**1057–1058.
70. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF: **Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla.** *Science* 2012, **337:**1661–1665.
71. Hobman JL, Penn CW, Pallen MJ: **Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully?** *Mol Microbiol* 2007, **64:**881–885.
72. Zeigler DR, Prágai Z, Rodriguez S, Chevreux B, Muffler A, Albert T, Bai R, Wyss M, Perkins JB: **The origins of 168, W23, and other *Bacillus subtilis* legacy strains.** *J Bacteriol* 2008, **190:**6983–6995.
73. Gertz S, Engelmann S, Schmid R, Ohlsen K, Hacker J, Hecker M: **Regulation of sigmaB-dependent transcription of sigB and asp23 in two different *Staphylococcus aureus* strains.** *Mol Gen Genet* 1999, **261:**558–466.
74. Hastie T, Tibishrani R, Friedman J: *The Elements of Statistical Learning.* New York: Springer; 2009.
75. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS results: a review of statistical methods and recommendations for their application.** *Am J Hum Genet* 2010, **86:**6–22.
76. Civelek M, Lusis AJ: **Systems genetics approaches to understand complex traits.** *Nat Rev Genet* 2014, **15:**34–48.
77. Priest NK, Rudkin JK, Feil EJ, van den Elsen JMH, Cheung A, Peacock SJ, Laabei M, Lucks DA, Recker M, Massey RC: **From genotype to phenotype: can systems biology be used to predict *Staphylococcus aureus* virulence?** *Nat Rev Microbiol* 2012, **10:**791–797.
78. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ: **Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak.** *N Engl J Med* 2012, **366:**2267–2275.
79. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW: **Transforming clinical microbiology with bacterial genome sequencing.** *Nat Rev Genet* 2012, **13:**601–612.
80. Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ: **Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*.** *N Engl J Med* 2013, **369:**290–292.
81. Pallen MJ, Loman NJ, Penn CW: **High-throughput sequencing and clinical microbiology: progress, opportunities and challenges.** *Curr Opin Microbiol* 2010, **13:**625–631.
82. Seth-Smith H, Harris S, Persson K, Marsh P, Barron A, Bignell A, Bjartling C, Clark L, Cutcliffe L, Lambden P, Lennard N, Lockey S, Quail M, Salim O, Skilton R, Wang Y, Holland M, Parkhill J, Thomson N, Clarke I: **Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain.** *BMC Genomics* 2009, **10:**239.
83. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain J-M: **ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.** *Antimicrob Agents Chemother* 2014, **58:**212–220.
84. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TEA, Crook DW, Walker AS, Golubchik T: **Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing.** *J Clin Microbiol* 2014, **52:**1182–1191.
85. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, *et al*: **The minimum information about a genome sequence (MIGS) specification.** *Nat Biotechnol* 2008, **26:**541–547.
86. Dugan VG, Emrich SJ, Giraldo-Calderón GI, Harb OS, Newman RM, Pickett BE, Schriml LM, Stockwell TB, Stoeckert CJ Jr, Sullivan DE, Singh I, Ward DV, Yao A, Zheng J, Barrett T, Birren B, Brinkac L, Bruno VM, Caler E, Chapman S, Collins FH, Cuomo CA, Di Francesco V, Durkin S, Eppinger M, Feldgarden M, Fraser C, Fricke WF, Giovanni M, Henn MR, *et al*: **Standardized metadata for human pathogen/vector genomic sequences.** *PLoS One* 2014, **9:**e99979.

doi:10.1186/s13073-014-0109-z
Cite this article as: Read and Massey: **Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology.** *Genome Medicine* 2014 **6**:109.