Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

The advent of genome-wide association studies for bacteria

Peter E Chen and B Jesse Shapiro

Significant advances in sequencing technologies and genome-wide association studies (GWAS) have revealed substantial insight into the genetic architecture of human phenotypes. In recent years, the application of this approach in bacteria has begun to reveal the genetic basis of bacterial host preference, antibiotic resistance, and virulence. Here, we consider relevant differences between bacterial and human genome dynamics, apply GWAS to a global sample of *Mycobacterium tuberculosis* genomes to highlight the impacts of linkage disequilibrium, population stratification, and natural selection, and finally compare the traditional GWAS against phyC, a contrasting method of mapping genotype to phenotype based upon evolutionary convergence. We discuss strengths and weaknesses of both methods, and make suggestions for factors to be considered in future bacterial GWAS.

Address
Département de sciences biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada

Corresponding author: Shapiro, B Jesse (jesse.shapiro@umontreal.ca, jesse.shapiro@gmail.com)

Current Opinion in Microbiology 2015, **25**:17–24

This review comes from a themed issue on **Environmental microbiology**

Edited by **Nicole King** and **Susann Müller**

http://dx.doi.org/10.1016/j.mib.2015.03.002

Introduction

A central goal of biology is to understand how DNA, the primary sequence, gives rise to observable traits. Historically, much effort has gone into deciphering the primary sequence of eukaryotes, primarily *Homo sapiens*. As of August 8, 2014, the National Human Genome Research Institute (NHGRI) reported 1961 publications of genome-wide association studies (GWAS). Within these studies, a total of 14,014 single nucleotide polymorphisms (SNPs) are associated with over 600 phenotypes. The advent of GWAS in bacteria has mainly occurred in the last two years [1••,2••,3••,4••,5••,6••], and provides an unbiased 'top–down' framework [7] to dissect the genetic basis of bacterial phenotypes. In principle, any measurable bacterial phenotype (or archaeal phenotype, although here our focus is on bacteria) can be dissected with a GWAS approach. To date, bacterial GWAS have focused on clinically-relevant phenotypes such as virulence and antibiotic resistance, but there is also great potential to investigate environmentally or industrially relevant phenotypes as well.

Bacterial genomes experience strong linkage, strong stratification and strong selection

Are bacterial genetic mapping studies any different from eukaryotic studies? Although there are many fundamental differences, this review highlights three features that are most germane to GWAS. The impact of the first two differences, in linkage and population stratification, have been recognized before [6••,7], but we identify the strength of natural selection relative to drift as a third and underappreciated factor to consider in bacterial GWAS.

First, unlike eukaryotic recombination which occurs predominantly via the crossing-over of two homologous chromosomes during meiosis, bacterial recombination occurs via gene conversion of relatively short stretches of DNA. In bacteria, recombination is not coupled with reproduction, and can occur multiple times within a cell's lifespan, or not at all. Without any recombination, purely clonal transmission of DNA leaves the entire bacterial chromosome in complete linkage (in strong linkage disequilibrium; LD). As with eukaryotic genomes, bacterial recombination events break this linkage, but the landscape of LD is markedly different from that seen in eukaryotes; gene conversion events leave a 'patchwork' of recombined tracts on top of a genomic background of linked regions called a clonal frame [8]. In contrast to eukaryotic LD patterns, all regions of the clonal frame are in complete linkage, and these regions may be quite distant from one another. The clonal frame phenomenon limits the utility of classic genetic mapping methods mainly by obscuring the true causal variant from the rest of the linked sites in the clonal frame. Here, we define a variant as causal if it plays a functional role in the phenotype of interest, as opposed to only being correlated with the phenotype.

Second, as with eukaryotes, bacterial genomic diversity may be shaped by population stratification. Stratification refers to a 'situation in which the population of interest includes subgroups of individuals that are on average more related to each other than to other members of the wider population' [9]. These subpopulations give rise to spurious associations when 'cases' (with phenotype A) are on average more closely related with each other than

with 'controls' (without phenotype A); in other words, associations due to genetic relatedness rather than causality for the phenotype of interest. The problem of population stratification is particularly acute in highly clonal (rarely recombining) bacteria, and in those with separate geographic or host-associated subpopulations [6••].

Third, the phenotypes of most interest in bacterial GWAS are largely different from many human disease phenotypes. In particular, bacterial phenotypes tend to be shaped by strong natural selection (e.g. positive directional selection driving drug resistance), while many human disease phenotypes evolve largely by genetic drift owing to historically small effective population sizes (e.g. due to population bottlenecks); in this scenario, drift overpowers purifying selection and leaves slightly deleterious alleles in the population that underlie disease traits [10,11]. This is not to say that bacteria do not experience genetic drift (particularly in frequently bottlenecked populations), but simply that many traits of interest (e.g. resistance, virulence, host-association) have evolved recently and under strong positive selection. These bacterial traits might also be controlled by mutations with large effect sizes on the phenotypes of interest. If this is the case, relatively small samples of bacterial genomes should be sufficient to identify causal mutations [11,12].

Units of genetic and phenotypic variation

The two basic requirements for GWAS are genotypic and phenotypic measurements from a sample of organisms. Phenotypes are usually broken into either discrete units (e.g. resistance/sensitive or high/low virulence) or continuous traits (e.g. human height). Phenotypes must be reproducible, and easy to measure, ideally in high-throughput if hundreds or thousands of samples are being studied. At the genotypic level, a set of bacterial genomes can be broken down into a 'core' genome shared among nearly all members and an 'accessory' genome composed of elements present in some strains but not others (typically including genes involved in environmental adaptation) [13,14]. The genetic units of a GWAS may be variants in the core (e.g. single nucleotide polymorphisms (SNPs) or small indels) [2••,3••,4••,5••] or in the flexible genome (e.g. presence/absence of larger pieces of DNA including genes or operons [1••,15,16,17] (Table 1). While most bacterial GWAS to date have studied either SNPs or gene presence/absence, Sheppard *et al.* [1••] described a method that uses n-mers ('words' of DNA) as the basic unit of association, allowing them to study both the core and flexible genome simultaneously.

Allele counting and homoplasy counting approaches to GWAS

GWAS approaches for bacteria can be broadly broken down into allele counting [1••,3••,4••,5••] and homoplasy counting [2••,12] methods (Table 1 and graphical abstract). The primary association signal for allele counting methods is derived from an over-representation of an allele at the same site in cases relative to controls, which can later be corrected for population stratification. In contrast, homoplasy counting methods (in this case, phyC [2••]) derives its evidence of association by counting repeated and independently emerged mutations occurring more often on branches of cases relative to controls. Homoplasy, as an indicator of convergent evolution, is a well-known signal of positive selection [28]. Combining this signal of selection with phenotypic associations (e.g. convergent mutations that occur only in cases and not in controls) provides the basis for homoplasy-based association tests.

Architecture of a strong association signal

GWAS signals from allele counting and homoplasy counting methods are not expected to perfectly overlap because each method represents different strengths and weaknesses. However, with a sufficiently large sample size, allele counting methods theoretically can detect all convergent sites (identified by homoplasy counting methods) as well as non-convergent sites. Still, ever-increasing sample size does not directly address the confounding effects of both population stratification and LD on allele counting methods. Homoplasy counting intrinsically accounts for these effects by virtue of its phylogenetic convergence criterion. In contrast, allele counting methods have no such phylogenetic requirement. Thus, a monophyletic group containing many cases with the same over-represented allele at the same site may provide a strong signal for allele counting while providing no signal for homoplasy counting. Conversely, homoplasy counting requires a smaller count of homoplasy events (versus allele counts) in order to reach statistical significance; thus, a relatively small sample size with a strong paraphyletic structure may provide homoplasy counting with a much stronger signal than allele counting.

A genome-wide association study of antibiotic drug resistance in *Mycobacterium tuberculosis*

To examine the impacts of clonal frames (strong LD) and population stratification, we performed a 'traditional' GWAS using PLINK on a population of 123 *M. tuberculosis* (MTB) genomes that had been previously analyzed by phylogenetic convergence (phyC) [2••]. Of the 123 strains, 47 (cases) are resistant to at least one antibiotic and 76 strains are sensitive to all antibiotics (controls). This dataset contains 11 'gold standard' experimentally-verified antibiotic resistance alleles, all of which were identified by phyC, along with 39 new phyC hits in nonsynonymous coding sites and intergenic regions, and seven hits in synonymous sites. We chose this particular MTB dataset as it allows a comparison of the results from traditional GWAS and phyC, and also because MTB genomes possess extensive LD and strong population structure, making them challenging subjects for traditional GWAS.