GENOME WATCH

Bacterial GWAS: not just gilding the lily

John A. Lees and Stephen D. Bentley

This month's Genome Watch discusses recent improvements to bacterial genome-wide association studies (GWAS) that were inspired by techniques developed for human genetics.

GWAS have become a staple of modern genetics. This approach is a powerful tool that is used to identify genetic variants that influence a given phenotype, which in general may be binary (usually cases of disease versus healthy controls) or continuous. Common single-nucleotide-polymorphisms (SNPs; more than 5% minor allele frequency) were the first form of variation to be studied using this technique. After appropriate quality control of the data, a simple statistical test of each variant yields the probabilities of whether they are linked to a specific phenotype. Recent applications of GWAS in bacteria have overcome some of the difficulties that are associated with small sample sizes and issues with clonal population structure, and have successfully associated genetic variation in the pan-genome with phenotypes such as antimicrobial resistance or host specificity.

In contrast to human populations, bacterial populations are highly related owing to their clonal method of reproduction. Consequently, when comparing lineages with

different phenotypes, all variants that separate the lineages seem to be associated with a phenotype when several samples with essentially the same genome are treated as independent, even though there is no causal link. One approach to tackling this issue is to select only those pairs of strains from the same lineage that have different phenotypes for sequencing. Maury *et al.* applied this rationale to a large collection of *Listeria monocytogenes* isolates, and the authors successfully identified and validated a gene cluster that causes increased virulence¹.

In human populations, samples that have divergent ancestry or samples from related individuals can also bias associations with common variants. Applying a linear mixed model (LMM) to all samples can obtain unbiased results, without losing as much statistical power as if these ancestrally divergent or related samples were removed entirely. Following this reasoning, Earle *et al.* have successfully applied a LMM to numerous bacterial genomes with a range of resistance phenotypes². In particular, for phenotypes in which highly penetrant common variants have a major role, the authors showed that this method improves the ability to find causal variants by leveraging the largest sample size possible.

The association of rare variants with a phenotype requires a different approach compared with common variation, because even in very large sample sizes there are not enough observations to achieve statistical significance. However, as rare variants occur in only a few bacterial strains (near the tips of the phylogeny in the context of bacteria), issues with lineage effects are less of a problem. To solve the issue of few observations, variants that affect the same gene or pathway can be pooled in a burden test. To reduce the rate of false positives,

only variants that are known to cause a loss of function in the protein are considered by these tests. Desjardins *et al.* used this approach to identify new resistance-conferring variants in *Mycobacterium tuberculosis*³. They found 10 different frameshift mutations and one nonsense mutation that caused a loss of function of the gene that encodes L-alanine dehydrogenase. Moreover, the authors showed that these mutations correlated with phenotypic resistance and that they evolved exclusively in multidrug-resistant strains.

In these studies, techniques that were developed for human genetics have been successfully adapted to study genomic variation in bacteria and have provided crucial insights into antimicrobial resistance that would probably have taken much longer to attain through hypothesis-driven study designs.

Bacterial GWAS are established techniques that are now entering a stage of refinement. So what might the future hold for these studies? Methodologically, the integration of more complex genomic variation and the inclusion of other large data sources (for example, the transcriptome) remains challenging. Biologically, GWAS will help to increase our knowledge of the genetic architecture of more complex phenotypes such as virulence.

John A. Lees and Stephen D. Bentley are at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro.2016.82
Published online 31 May 2016

1. Maury, M. M. *et al.* Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat. Genet.* **48**, 308–313 (2016).
2. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041 (2016).
3. Desjardins, C. A. *et al.* Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate *ald* in D-cycloserine resistance. *Nat. Genet.* **48**, 544–551 (2016).

Competing interests statement
The authors declare no competing interests.

NPG