Session GWAS for prokaryotes

Daniel Wilson
Titre: Identifying lineage effects when controlling for population structure improves power in bacterial association studies
Summary:

Bacteria pose unique challenges for genome-wide association studies (GWAS) because of strong structuring into distinct strains and substantial linkage disequilibrium across the genome. While methods developed for human studies can correct for strain structure, this risks considerable loss- of-power because genetic differences between strains often contribute substantial phenotypic variability. We propose a new method that captures lineage-level associations even when locus-specific associations cannot be fine-mapped. We demonstrate its ability to detect genes and genetic variants underlying resistance to 17 antimicrobials in 3144 isolates from four taxonomically diverse clonal and recombining bacteria: Mycobacterium tuberculosis, Staphylococcus aureus, Escherichia coli and Klebsiella pneumoniae. Strong selection, recombination and penetrance confer high power to recover known antimicrobial resistance mechanisms, and reveal a candidate association between the outer membrane porin nmpC and cefazolin resistance in E. coli. Hence our method pinpoints locus-specific effects where possible, and boosts power by detecting lineage-level differences when fine-mapping is intractable

Zamin Iqbal
Titre: Graph representations of bacterial genetic variation provide substrate for prediction and association
Summary:

The levels of diversity and recombination in bacterial species lead to considerable challenges for genomic analysis of an entire species (as opposed to microevolution or transmission studies) that are based on a single reference genome. Further, incorporation of accessory genes into a robust statistical framework that is consistent with that used for chromosomal mutations is hard. We show how an annotated graph representation of both SNP and gene population variants can be used for antibiotic resistance prediction, applying to a species with non-trivial accessory genome (*S. aureus,* N=~1100*)* and one with no recombination but appreciable levels of mixture, either through in-host evolution or multiple transmission (*M. tuberculosis,* N=~6600*)*. Genotypic predictions of resistance are shown to be as sensitive/specific (>99%) as two standard clinical phenotyping methods for *S. aureus.* The genetic basis for drug resistance in *M. tuberculosis* is less well understood, so our method explains only ~82.5% of resistance. However we also demonstrate that minor alleles significantly increase power to detect resistance in 2nd line drugs without affecting specificity, which could have significant impact on monitoring and diagnosis of XDR-TB (Extensively Drug Resistant TB), a major concern for the World Health Organisation. Finally we discuss how a graph representation of genetic variation provides a better substrate for association analysis than a pure kmer-based approach.

Alexandre Drouin
Titre: Modelling Antibiotic Resistance with Sparse Machine Learning and Reference-Free Genome Comparisons
Summary:

Case-control studies compare groups of related individuals with the aim of identifying genomic variations that are biomarkers of a phenotype. Recent advances in next-generation sequencing have led to a tremendous increase in the scale of such studies. This trend will persist, motivating the need for efficient computational biomarker discovery and validation methodologies. In this context, we present a novel reference-free method for genomic biomarker discovery that learns uncharacteristically sparse models from whole genome sequences. It relies on a k-mer representation of the genomes and on the Set Covering Machine, a learning algorithm that yields sparse and interpretable classifiers. We provide theoretical and empirical results that demonstrate that the method is well suited to learn from feature spaces of extremely high dimension. The method successfully predicted the antibiotic resistance of four common human pathogens: Clostridium difficile, Mycobacterium tuberculosis, Pseudomonas aeruginosa and Streptococcus pneumoniae, and yielded computational models for 17 commonly used antibiotics.The obtained models are accurate, faithful to the biological pathways targeted by the antibiotics and they provide insight into the process of resistance acquisition. Kover, an out-of-core implementation of our method, can readily scale to large genomic datasets. It is open-source and can be obtained from http://github.com/aldro61/kover.