Abstract

A bacterial species' genetic content can be remarkably fluid. The collection of genes found within a given species is called the pan-genome and is generally much larger than the gene repertoire of a single cell. A consequence of this pan-genome is that bacterial genomes are highly adaptable and thus variable.

The dominant paradigm for analysing genetic variation relies on a central idea: all genomes in a species can be described as minor differences from a single reference genome, which serves as a coordinate system. As an introduction to this thesis, we outline why this approach is inadequate for bacteria and describe a new approach using genome graphs.

In the first chapter, we present algorithms for *de novo* variant discovery within such genome graphs and evaluate their performance with empirical data. The remaining chapters address a question relating to a critical bacterial pathogen: can Nanopore sequencing of *Mycobacterium tuberculosis* provide high-quality public health information? We collect data from Madagascar, South Africa, and England to help answer this question. First, we assess outbreaks identified using single-reference and genome graph methods. Second, we evaluate AMR predictions and introduce a framework for using genome graphs to improve current methods. Lastly, we train an *M. tuberculosis*-specific Nanopore basecalling model with considerable accuracy improvement.

Together, this thesis provides general methods for uncovering bacterial variation and applies them to an important global public health question.