Evolutionary Genomics of Pathogenic Mycobacteria

Josephine M. Bryant

Darwin College
University of Cambridge
Wellcome Trust Sanger Institute
November 2014
This dissertation is submitted for the degree of Doctor of Philosophy

This dissertation is the result of my own work. Any work that is the outcome of work done in collaboration is specifically indicated in the text.

No part of this dissertation has been submitted for any other qualification and it does not exceed the word limit stipulated by the Biological Sciences Degree committee.

Abstract

The genus *Mycobacterium* includes many species pathogenic to human health. This thesis concentrates on two of these species: *Mycobacterium tuberculosis*, the causative agent of tuberculosis and an obligate intracellular human pathogen; and *Mycobacterium abscessus,* an environmental bacterium that can opportunistically cause respiratory and soft tissue infections in humans. Whole genome sequencing was carried out on large sample collections of these two species in order to understand how they transmit in addition to their evolutionary dynamics over small to large evolutionary scales.

For *M. tuberculosis,* a very low substitution rate of ~0.3 single nucleotide polymorphisms (SNPs) per genome per year was observed in the context of patient-patient transmission. This low genetic turnover presents challenges to our ability to use whole genome sequencing to infer direct transmission of tuberculosis, and highlights the continuing importance epidemiology will play in strengthening these inferences. Whole genome sequencing was also applied to recurrent tuberculosis disease, where patients had had a second disease episode within two years of being cured of the first. This enabled the clear differentiation of those caused by relapse and those by re-infection. In addition mixed infections were detected and deconvoluted, which would not have been possible using traditional genotyping methods. Finally the highly variable PE and PPE genome families were studied in detail using both mapping and *de novo* assembly approaches. The functions of these gene families are unknown, but they are often cell-surface associated and antigenic, so have been speculated to play a role in within-patient antigenic diversification. This analysis found that although these genes were more variable than the rest of the genome, this variability was not generated within patients, suggesting another role for these gene families.

Compared to *M. tuberculosis, M. abscessus* is poorly understood, with little genomic data or an understanding of population structure available prior to this study. This thesis concentrates on the infection of cystic fibrosis patients with *M. abscessus*,

which is causing concern due to its high level of antibiotic resistance and rising incidence. Whole genome sequencing was carried out on a collection from a single cystic fibrosis clinic collected over four years. For most patients, their isolates were distantly related, a pattern consistent with independent acquisition from the environment. This was expected as transmission between patients was previously assumed to be impossible or rare. Surprisingly there were some patients however who shared identical or near identical isolates, which fell into two phylogenetic clusters. This suggested transmission between patients had occurred, a conclusion supported by both epidemiological evidence and Bayesian dating methods. In addition to transmission, this dataset also provided the opportunity to capture within-patient diversity through the detection of minority variants. These minority variants were correlated with clinical outcome and treatment, revealing fluctuations in genetic diversity over time with associated changes in phenotype.

Whole genome sequencing has allowed the analysis of the evolution of two important mycobacterial pathogens over different timescales: within patient, within outbreaks and across the species. These analyses have not only provided us with greater insights into how they evolve, and at what rate, but also have had a significant clinical impact. This work has highlighted the power of the whole genome approach, especially when applied to organisms with a low mutation rate, which will be essential for furthering our understanding of mycobacteria.

Acknowledgements

This work wouldn't have been possible without Julian Parkhill, who has always made time for me, and is the most informed supervisor one could hope for in the field of bacterial genomics. Simon Harris, has not only been a good friend, but has developed many of the bioinformatic pipelines used in this thesis, and has provided advice on countless occasions. Stephen Bentley has provided guidance throughout my PhD, particularly in the early years. I'd like to thank him for encouraging me to join this group and getting me working on mycobacteria. I will miss E212 and the rest of the pathogen genomics team, who have provided a friendly and fun environment to work in. My thesis committee: James Wood, Paul Kellam and Jeff Barratt, have provided valuable advice and critique.

My work has depended on countless collaborators and support teams both within and outside Sanger. Within the Sanger, I am indebted to the DNA pipeline and Pathogen Informatics teams who have provided sequencing and bioinformatic support, allowing my research to be less about the practicalities and more about the biology. Outside Sanger, I would like to thank Ed Feil who got me interested in pathogen evolution in the first place and encouraged me to apply for this PhD; Stephen Gillespie who has always made time for me; Andres Floto and Dot Grogono who have been fantastic collaborators; and all those that provided and prepared samples, I'm aware mycobacteria can be awkward buggers to grow.

I'd like to thank my family, who has provided unwavering support and limitless pride. In particular: my dad for reminding me that science is a creative process; my mum for showing me how a toilet cistern worked at an early age, and getting me interested in the world; Liam for feeding me; and finally my grandfather Maurice for setting the bar high.

Contents

Abbreviations

CF – cystic fibrosis

CRP – C reactive protein

FEV – forced expiratory volume

HPD – higher posterior density

MDR – multi drug resistant

MGIT – mycobacterial growth indicator tube

MIC – minimum inhibitory concentration

MIRU – mycobacterial interspersed repeat unit

MLSA – multi-locus sequence analysis

MLST – multi-locus sequence typing

MTBC – *Mycobacterium tuberculosis* complex

NTM – non-tuberculous mycobacteria

PCR – polymerase chain reaction

PFGE – pulsed field gel electrophoresis

RFLP – restriction fragment length polymorphism

VNTR – Variable number tandem repeat

XDR – extensively drug resistant