Mycobacterial genomics and its application to disease control

Stephen V. Gordon

Abstract

The publication of the *Mycobacterium tuberculosis* genome sequence in 1998 heralded a new phase in mycobacterial research. The subsequent decade has seen an enormous increase in our understanding of the basic biology, virulence mechanisms, drug targets and evolution of the *M. tuberculosis* complex, an acceleration in knowledge that would not have been possible without the genome. This review summarises some key findings from mycobacterial genomics and highlights how this knowledge can be applied to disease control in the developing world. [*Ethiop.J.Health Dev.* 2008;22(Special Issue):119-122]

Introduction

From anywhere in the world with an internet connection, researchers can now access the genome sequences of all the major mycobacterial pathogens, such as *Mycobacterium tuberculosis*[1], *Mycobacterium leprae*[2], *Mycobacterium bovis*[3] and *Mycobacterium ulcerans*[4], as well as the vaccine strains *M. bovis* BCG[5] and *Mycobacterium microti* OV254[6]. Indeed, through the Wellcome Trust Sanger Institute[7], The Institute for Genome Research[8], and the Broad Institute[9] we currently have genome sequences for six different strains of *M. tuberculosis* (H37Rv, CDC1551, 210, C, F11, Haarlem), with many more to follow through projects underway at the Broad. The "TubercuList" *M. tuberculosis* H37Rv genome database regularly receives more than 70,000 "hits" per month from all over the world (S. Cole, personal communication), showing how the genome has become a cornerstone of mycobacterial research; we now all work in a post-genomic landscape.

How has the availability of this mass of genome data changed our understanding of the tubercle bacilli? How are we to apply this information to improve disease control? This review will briefly summarise key findings from the genome sequences of *M. tuberculosis*, *M. bovis* and *M. bovis* BCG, and highlight how genome findings are being applied in the fight against disease.

Evolution

The emergence of *M. tuberculosis* has often been viewed as a classic zoonotic infection, with the generalist animal pathogen *M. bovis* crossing the species barrier into man at the time of cattle domestication. Initial work on defining genetic diversity across the *M. tuberculosis* complex using single nucleotide polymorphisms (SNPs) suggested that *M. tuberculosis* arose ~15,000 years ago (1), a similar date to that proposed for cattle domestication (2), seemingly providing further support for the zoonosis model. However, subsequent attempts to date the clonal expansion of *M. tuberculosis* using SNP data gave a date of ~35,000 years ago (3). Indeed, using SNP data to define a molecular clock may not be appropriate for *M. tuberculosis*, given its small effective population size and distinctive DNA repair systems.

An alternative method to explore the evolution of the *M. tuberculosis* complex employed comparative genomics. With the *M. tuberculosis* genome as a starting point, the genomes of all members of the *M. tuberculosis* complex were scanned for deletion events (4). This defined a set of 12 deletions that could then be mapped across a reference collection that encompassed the diversity of the *M. tuberculosis* complex (5). This comparative analysis allowed a novel evolutionary scenario to be proposed, with *M. tuberculosis* closer to the common ancestor of the complex than *M. bovis* (Figure 1). This scenario appears to overturn conventional evolutionary thinking, suggesting that tuberculosis is in fact a reverse zoonosis given by man to domesticated animals. However, the host association of the common ancestor of the complex is unknown, and valid cases can be made for a human or animal associated ancestor.

Recent work from strains causing tuberculosis in East Africa has added a further twist to the evolution of the *M. tuberculosis* complex (6). These strains, classed as "smooth" tubercle bacilli due to their characteristic colony appearance, were characterised using multilocus sequencing and revealed to have an unexpected degree of sequence variation. However, most surprising was evidence for horizontal gene transfer among this population of strains, which goes against current data suggesting a purely clonal population structure for the *M. tuberculosis* complex. Based on these data the authors proposed that these in fact were isolates of the progenitor of the *M. tuberculosis* complex, *Mycobacterium prototuberculosis* (6). However, this latter suggestion has been challenged since recombination within the studied population would inflate the estimation of diversity of this group (7). Clearly, a greater sampling of strains causing tuberculosis on a worldwide level, and in particular from Africa, is needed to allow a better understanding of how

[1] (http://genolist.pasteur.fr/TubercuList)
[2] (http://genolist.pasteur.fr/Leproma)
[3] (http://genolist.pasteur.fr/BoviList)
[4] (http://genolist.pasteur.fr/BuruList)
[5] (http://genolist.pasteur.fr/BCGList)
[6] (http://www.sanger.ac.uk/Projects/M_ microti)
[7] (www.sanger.ac.uk)
[8] (ww.tigr.org)
[9] (www.broad.mit.edu)

TB Research Group, VLA Weybridge, Woodham Lane, New Haw, Addlestone, Surrey KT15 3NB, UK. Tel: (44) 01932 357 850, E-mail: s.v.gordon@vla.defra.gsi.gov.uk

smooth tubercle bacilli fit into the evolution of the *M. tuberculosis* complex.

Using a combination of genome deletion events and SNPs it is possible to define a phylogeny for the members of the *M. tuberculosis* complex, with *M. tuberculosis* sitting closer to the common ancestor than *M. bovis*. The deletion events RD9 and RD4 serve as unequivocal markers of species; hence, strains deleted for RD9 and RD4 are defined as *M. bovis*. The mmpL6 SNP is at codon 561, where AAC is replaced by AAG; hence, *M. bovis* presents the AAG codon.

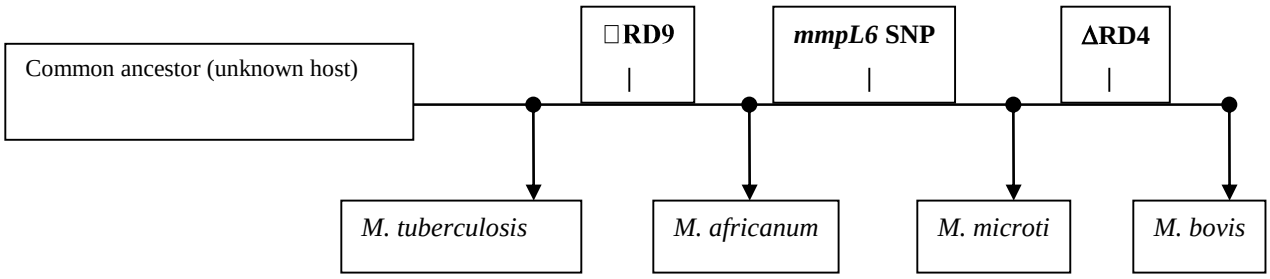

Figure 1: **An evolutionary scenario for the *M. tuberculosis* complex**.

Physiology

Analysis of the genome sequence of *M. tuberculosis* revealed a remarkable capacity for both catabolism and anabolism of lipids, reflecting complex synthetic requirements for the cell wall and the capacity to metabolise a wide variety of lipid substrates (8). The radiance of lipid metabolic genes is found across all members of the *M. tuberculosis* complex; however, comparative analysis of genes encoding key enzymes of central metabolism has uncovered the metabolic basis for notable phenotypic differences between the tubercle bacilli. For example, *M. bovis* requires pyruvate to be added to media where glycerol is the sole carbon source, while *M. tuberculosis* has no such requirement; the metabolic basis for this was unknown, but the answer presumably lay in the glycerol utilisation pathway. Once the *M. bovis* genome had been sequenced, it was relatively facile to compare all genes between *M. bovis* and *M. tuberculosis* that encoded proteins involved in glycerol utilisation. This revealed a nonsynonymous point mutation in the gene encoding pyruvate kinase which chaged a conserved glutamic acid residue to aspartic acid (9). Enzyme assays showed that *M. bovis* lacked pyruvate kinase activity, while complementation with the *M. tuberculosis* allele restored pyruvate kinase activity and the ability to grow on glycerol. Furthermore, the complemented strains of *M. bovis* switched from the typical "dysgonic" appearance of *M. bovis* colonies on glycerol-based media to abundant, "eugonic" colonies as seen with *M. tuberculosis*. As pyruvate kinase is one of the few irreversible steps in glycolysis, the lack of a functional pyruvate kinase in *M. bovis* blocks glycolytic intermediates from feeding into oxidative metabolism. Hence, *in vivo*, *M. bovis* must rely on amino acids or fatty-acids as a carbon source for energy metabolism; *in vitro*, media for the isolation of *M. bovis* must be supplemented with pyruvate.

Drug Targets

The current length of tuberculosis chemotherapy and paucity of frontline drugs are driving the emergence of MDR (multi-drug resistant) and XDR (extensively drug resistant) strains. There is clearly an immediate need for novel drug targets to be identified in *M. tuberculosis*. Genomics and allied functional genomic techniques have helped galvanize drug discovery initiatives. A key starting point was to identify which of the ~4000 genes are essential for the bacillus, as proteins essential for bacterial viability are considered the most promising targets for rational drug design. Using a combination of saturation transposon mutagenesis and microarray-based identification of insertion sites, Sassetti, Boyd and Rubin defined a minimal *in vitro* gene set for *M. tuberculosis* (10); 570 of these genes were also found to be essential by a similar experimental approach adopted by Bishai and colleagues (11). To prioritize these genes as targets for drug development further genomic criteria can be used, including the presence of druggable protein domains that bind potent compounds, situation at a unique position in a metabolic chain or "chokepoint", or phylogenetic restriction to avoid homologues in host or host flora.

In addition to identifying potential targets for drugs, genomics has greatly assisted in identifying the targets of drugs with known antibacterial activity. Previously this had to be done by laborious and time consuming genetic screens but the ease of full genome sequencing means whole genome comparisons can be used to locate

resistance-conferring mutations. For example the molecular target for the diaryquinoline R207910 was identified by comparing the genomes of resistant *M. tuberculosis* and *M. smegmatis* strains selected *in vitro* with the wild-type genome. Drug resistance conferring mutations localised in both species to the *atpE* genes, that encodes part of the ATP synthase, which was subsequently confirmed as the site of drug action (12). PA-824 is a nitroimidazole prodrug whose activation requires reduction via an F_{420}-dependent glucose-6-phosphate dehydrogenase; however, this activity is not sufficient to confer sensitivity. To identify further factors involved in sensitivity to PA-824, Barry and colleagues isolated PA-824 resistant strains with wild-type F_{420}-dependent glucose-6-phosphate dehydrogenase activity, and resequenced their genomes using high density oligonucleotide arrays (13). Resistant isolates were shown to harbour mutations in *Rv3547*, encoding a novel nitroimidazo-oxazine-specific nitroreductase, further elucidating the action of PA824. This approach of rapid full genome comparisons also represents a powerful new tool for characterizing the phenotypic consequences of single point mutations in other settings, such as the microevolution of *M. tuberculosis* within hosts or during transmission cycles.

BCG: a family of duplicating daughters

BCG is not a single vaccine, but rather a family of daughter strains derived from an original stock sent out by the Institute Pasteur to laboratories around the world (14). Accumulating evidence suggests that these daughter strains posses significant differences in terms of genome content, immunostimulatory activity, and protective efficacy. Indeed, despite being the most widely used vaccine in the world, the molecular basis for the attenuation of BCG has never been fully described, although the loss of the RD1 locus was a key attenuating event (15). The recent completion of the genome sequence of *M. bovis* BCG Pasteur, and comparative analyses across BCG daughter strains, exposed all the all genetic differences between BCG Pasteur and the virulent *M. bovis* and *M. tuberculosis*, and provided further evidence for strain specific genome content that may impact on vaccine efficacy (16).

A key finding from the BCG sequencing project was the presence of duplication events in the BCG genome. These duplications are the first to be described in a mycobacterial genome, and show variable configurations across the daughter strains. For example, DU1 is a duplication around the origin of chromosomal replication, *oriC*, that is only found in BCG Pasteur. On the other hand, DU2 defines a duplication that is present in all BCG strains, but in 4 different arrangements. Hence, after the original DU2 duplication event, internal regions were spliced out and duplicated a number of times, suggesting a selective pressure was driving the gene duplication events. Clues to what the in vitro selective pressure was are revealed in the core region of DU2 which is duplicated across the strains but that did not subsequently undergo deletion. This core region comprises three complete coding sequences (CDS), namely *glpD2* (encoding a putative glycerol 3-P dehydrogenase), *phoY1* (encoding a potential repressor of the high affinity phosphate uptake system), and *Rv3300c* (encoding a protein of unknown function). This duplication would serve to increase the production of glycerol 3-P dehyrogenase; glycerol was the carbon source used by Calmette and Guérin during the derivation of BCG from *M. bovis*, so it is possible that increased copy number of *glpD2* conferred a selective advantage for growth on glycerol. However, glycerol 3-P dehyrogenase is not a known metabolic bottleneck for glycerol growth; one might instead expect to have seen *glpK*, encoding glycerol kinase and a known rate limiting enzyme in glycerol utilisation, undergoing amplification. A second possibility is that increased production of the potential repressor of high affinity phosphate uptake, *phoY1*, was selected. BCG is known to have lesions in the high affinity phosphate uptake system compared to *M. bovis* and *M. tuberculosis*, suggesting that switching this system off may have been advantageous to the bacillus.

Application to disease control

In a collaborative project between the Armauer Hansen Research Institute (Ethiopia), Imperial College (UK), Trinity College (Ireland), the Swiss Tropical Institute (Switzerland), the Institute for Livestock Research (Kenya), and VLA (UK), we aim to measure the cost of bovine TB to Ethiopian society by assessing its impact on livestock productivity and human health. A necessary step in this project is to ensure that we can correctly culture and discriminate the members of the *M. tuberculosis* complex from both human and animal clinical samples. Findings from the mycobacterial genome projects feed directly into this endeavour. For example, from the *M. bovis* genome project we know that *M. bovis* can not use carbohydrates as a carbon source for energy production; pyruvate must be added to the media to allow *M. bovis* to grow (9). Hence, knowledge from the genome informs our choice of media for the improved isolation of *M. bovis* from clinical samples. Similarly, from comparative genomics we now know that deletion events provide unequivocal markers of species (5, 17). From the schematic in Figure 1, we can see that isolates that are deleted for RD9 and RD4 are *M. bovis*, while strains that have the RD9 and RD4 loci intact are *M. tuberculosis*. The presence or absence of these deletions can be determined using a simple PCR reaction; this is a significant improvement over biochemical tests (such as nitrate reduction or niacin production) which can prove misleading. Hence, genetic tests based on deletions provide confidence that the isolates are correctly identified as *M. bovis* or *M. tuberculosis*. Access to a specific, portable PCR for strain discrimination is an obvious benefit when one is trying to determine the burden of *M. bovis* infection in the human population.

Conclusions

Genomics has fundamentally altered our understanding of the members of the *M. tuberculosis* complex. The advances outlined above are also providing tools that can now be applied to disease control, with simple PCR-based strain identification offering a case in point. A key requirement is to ensure that mycobacterial genomics continues to drive efforts in disease control, and that novel findings are translated to the clinical and field setting where their application can have the greatest impact.

Acknowledgements

This work was funded by the Wellcome Trust and the UK Department for Environment, Food and Rural Affairs. The author wishes to acknowledge the guidance and support of Glyn Hewinson, Noel Smith, Roland Brosch, Thierry Garnier, and Stewart Cole.

References

1. Kapur V, Whittam TS, Musser JM. Is *Mycobacterium tuberculosis* 15,000 years old? J Infect Dis 1994;170(5):1348-9.
2. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature 2002;418(6898):700-7.
3. Hughes AL, Friedman R, Murray M. Genomewide pattern of synonymous nucleotide substitution in two complete genomes of *Mycobacterium tuberculosis*. Emerging Infectious Diseases 2002;8(11):1342-6.
4. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol 1999;32(3):643-55.
5. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A 2002;99(6):3684-9.
6. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog 2005;1(1):e5.
7. Smith NH. A re-evaluation of *M. prototuberculosis*. PLoS Pathog 2006;2(9):e98.
8. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998;393(6685):537-44.
9. Keating LA, Wheeler PR, Mansoor H, Inwald JK, Dale J, Hewinson RG, et al. The pyruvate requirement of some members of the *Mycobacterium tuberculosis* complex is due to an inactive pyruvate kinase: implications for in vivo growth. Mol Microbiol 2005;56(1):163-74.
10. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 2003;48(1):77-84.
11. Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, Grosset J, et al. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 2003;100(12):7213-8.
12. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. Science 2005;307(5707):223-7.
13. Manjunatha UH, Boshoff H, Dowd CS, Zhang L, Albert TJ, Norton JE, et al. Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 2006;103(2):431-6.
14. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 1999;284(5419):1520-3.
15. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. J Bacteriol 1996;178(5):1274-82.
16. Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, Valenti P, et al. Genome plasticity of BCG and impact on vaccine efficacy. Proc Natl Acad Sci U S A 2007;104(13):5596-601.
17. Parsons LM, Brosch R, Cole ST, Somoskovi A, Loder A, Bretzel G, et al. Rapid and simple approach for identification of *Mycobacterium tuberculosis* complex isolates by PCR-based genomic deletion analysis. J Clin Microbiol 2002;40(7):2339-45.