Published in final edited form as:

Nat Genet. 2010 June ; 42(6): 476–478. doi:10.1038/ng0610-476.

Relics of selection in the mycobacterial genome

Christopher M Sassetti and

Howard Hughes Medical Institute and the University of Massachusetts Medical Center,
Worcester, Massachusetts, USA

Eric J Rubin

Harvard School of Public Health, Boston, Massachusetts, USA

Abstract

A new study reports the whole-genome sequences of 21 *Mycobacterium tuberculosis* complex strains, selected to represent geographically diverse isolates. Comparative genomic analyses identify surprising conservation of epitopes recognized by human T cells.

Tuberculosis, caused by infection with *M. tuberculosis*, is a prevalent disease that kills thousands each day. Although ~1.8 billion people are thought to be infected with this pathogen, only ~10% will ever develop disease¹. In the remaining individuals, host immunity is able to control bacterial growth and contain the infection in a clinically latent state. In animal models, the balance between latency and progressive disease has been shown to depend on the ability of host T cells to continually recognize the pathogen and suppress its growth². Many pathogens avoid this type of immunity through selection that increases the variation in the epitopes recognized by lymphocytes. This survival strategy leaves a mark on the genomes of these organisms in the form of hypervariable regions that become apparent when multiple genomes are compared³. On page 498 of this issue, Sebastien Gagneux and colleagues look for these telltale marks in the whole-genome sequences of 22 strains from the *M. tuberculosis* complex and discover a surprising conservation of T-cell epitopes, suggesting that more complex selective pressures may be at work on the genome of *M. tuberculosis*⁴.

Mycobacterial genomics

Tuberculosis is caused by a group of closely related bacteria known as the *M. tuberculosis* complex (MTBC). By analyzing the polymorphisms present in a relatively small number of genomic loci, previous studies have divided these organisms into a handful of distinct phylogenetic lineages, each associated with a particular geographic region⁵. Comas *et al.*⁴ now report the whole-genome sequences of a geographically and phylogenetically diverse group of 20 MTBC strains representative of the three known major lineages and one strain of *Mycobacterium canettii* used as an outgroup; they then compared these strains to an *M.*

tuberculosis reference genome. This type of study was made feasible by the application of next-generation sequencing. Using the Illumina Genome Analyzer, Comas *et al.*⁴ sequenced these strains at 40- to 90-fold coverage. Because this methodology provides only short read lengths, it is difficult to analyze the highly repetitive regions where much of the genetic variability is likely to reside. In the rest of the genome, rather limited variation was found, with ~1 base change per 3 kb. Nevertheless, because nearly complete genomes were analyzed, this amounted to more than 9,000 informative SNPs, which was enough to begin to probe the selective pressures acting on this organism's genome.

Because of the relatively unusual population structure of *M. tuberculosis*, selection acting on the pathogen's genome can be difficult to distinguish from stochastic variation. *M. tuberculosis* is an obligate pathogen; its only important natural niche is in humans, specifically those with active disease. Humans are also the only *M. tuberculosis* hosts capable of transmitting disease. During growth in a host, bacteria can accumulate mutations and diversify. Transmission likely occurs through infection by one organism or a very small number of organisms. Thus, the transmission event represents an extremely tight genetic bottleneck, and mutations can accumulate stochastically without a selective advantage (Fig. 1).

Despite this challenge, Comas *et al.* were able to detect evidence of selective pressure in the MTBC genomes⁴. They found that essential genes, encoding functions required for bacterial growth, contain a smaller number of variant nucleotides and a lower dN/dS ratio (ratio of nonsynonymous to synonymous changes) compared to other genes. This suggests that selection acts to maintain these processes. This is not entirely surprising, as these genes perform indispensable cellular functions, and a previous analysis of a single *M. tuberculosis* genome reached a similar conclusion⁶. However, these findings proved to be an excellent benchmark against which other, less anticipated results could be measured.

Because host T-cell responses control *M. tuberculosis* growth in the majority of infected individuals⁷, it might have been logical to assume that genomic loci encoding the epitopes recognized by these cells would be under selection for greater diversity. This is the case for many viruses that evade adaptive immunity^{8,9}. In contrast, Comas *et al.*⁴, surprisingly, found that, as a group, the peptides recognized by human T cells were highly conserved across these strains. They surveyed 78 experimentally confirmed antigens (from the Immune Epitope Database and Analysis Resource (IEDB)¹⁰) and found that the dN/dS ratio in the genes encoding these peptides was strikingly low, much like that for essential genes. This unexpected finding demonstrated that these antigenic proteins are highly conserved among MTBC strains and, therefore, that instead of selection favoring increased diversity, these peptides might be under selection against variability (Fig. 1).

Explaining selection

The observed conservation of these T-cell epitopes might be explained if the stability of antigens recognized by the host were beneficial to the pathogen. If so, how could immune recognition benefit *M. tuberculosis*? There are a few possibilities. The most interesting, and that discussed by Comas *et al.*⁴, is that immune recognition might benefit the bacterium by

promoting tissue destruction, allowing the infected lesions to erode into the airway and increasing transmission via the aerosol route. The transmission between individual hosts is a driving selective pressure on the pathogen, and immune-mediated tissue destruction is critical for this process.

However, the immune pressures on this pathogen are likely more complex, as individuals with either active or latent disease produce superficially similar T-cell responses but show markedly different disease outcomes. This could be explained if recognition of certain antigens is protective and recognition of others promotes transmission. The epitopes in the IEDB database were identified in numerous studies that used differing methodologies, and the database includes peptides recognized by epitopes in individuals with latent infection, active disease or both. Thus, these epitopes might consist of a mix of those that are beneficial for the host and those beneficial for the pathogen. Although, overall, the epitopes showed conservation, a small number of them showed evidence of increased diversity, consistent with the idea that not all antigens are under the same selective pressures in all individuals. In addition, many of the epitopes selected from the IEDB database were initially discovered by testing peptides derived from a single bacterial strain against T cells from individuals infected with strains of different lineages. This process would bias toward more evolutionarily stable peptides and could also explain at least part of the conservation of the epitopes observed by Comas *et al.*⁴

The attenuated live vaccine *Mycobacterium bovis* BCG is given to infants in many parts of the world as a preventative against tuberculosis. Unfortunately, this strain offers adults very poor protection against developing the disease. The work of Comas *et al.*⁴ suggests that some epitopes, including many present in *M. bovis* BCG, might augment transmission rather than protect against disease. If this work can be extended to define epitopes that are protective, it could provide a valuable tool for developing new, more efficacious vaccines. Given the limited variation among natural MTBC isolates, this is likely to require better characterization of existing epitopes, the discovery of new immune recognition sites and, perhaps, harnessing the known variation among *M. bovis* BCG strains.

References

1. Corbett EL, et al. Arch. Intern. Med. 2003; 163:1009–1021. [PubMed: 12742798]
2. Maartens G, Wilkinson RJ. Lancet. 2007; 370:2030–2043. [PubMed: 17719083]
3. Palmer GH, Bankhead T, Lukehart SA. Cell. Microbiol. 2009; 11:1697–1705. [PubMed: 19709057]
4. Comas I, et al. Nat. Genet. 2010; 42:498–503. [PubMed: 20495566]
5. Gagneux S, et al. Proc. Natl. Acad. Sci. USA. 2006; 103:2869–2873. [PubMed: 16477032]
6. Plotkin JB, Dushoff J, Fraser HB. Nature. 2004; 428:942–945. [PubMed: 15118727]
7. Kaufmann SH. Nat. Rev. Immunol. 2001; 1:20–30. [PubMed: 11905811]
8. Streeck H, et al. PLoS Med. 2008; 5:e100. [PubMed: 18462013]
9. Neumann-Haefelin C, et al. J. Virol. 2008; 82:3438–3451. [PubMed: 18216107]
10. Vita R, et al. Nucleic Acids Res. 2010; 38:D854–D862. [PubMed: 19906713]

Figure 1.

The population dynamics of *Mycobacterium tuberculosis* are influenced by immune-related selection during growth within a host. An individual human is infected by a single bacterium (or small group of bacteria), and growth within the human host may bring some diversity in the mycobacterial population. **(a)** In the absence of selection, each clone has an equal likelihood of transmission to a new host. **(b)** If the immune system recognizes early clones and destroys them, variant strains will have an advantage and outgrow the original strains. **(c)** Alternatively, strains with specific epitopes might generate an inflammatory response that increases their probability of transmission.