A molecular biology approach to tuberculosis

Michel Tibayrenc*

Unité Mixte de Recherche, Institut de Recherche pour le Développement/Centre National de la Recherche Scientifique 2724 "Genetics and Evolution of Infectious Diseases," BP 64501, 34394 Montpellier Cedex 5, France

There is little doubt that infectious diseases are the major challenge of medicine at the opening of this new century (1). They are still by far the major population control factor of our species. At the very least, this peril of emerging and reemerging infectious diseases (2) has the beneficial side effect of boosting basic research. It is indeed a paradox that, for example, the basic biology of *Mycobacterium tuberculosis*, the causative agent of a major human disease, is far less well known than that of *Escherichia coli*, potentially because tuberculosis was more or less effectively controlled in the north. Facing the threat of tuberculosis epidemics throughout the world, scientists are currently erasing the somewhat artificial border between basic and applied research and are generating data that have both a strong added value in terms of basic science and an immediate usefulness for epidemiology and medicine. The two companion papers by Tsolaki *et al.* (3) and Hirsh *et al.* (4) in this issue of PNAS are perfect examples of this new scientific school, as well as of the opportunities offered by the progress of modern biotechnology when they are wisely used.

Like many organisms, the whole genome of *M. tuberculosis* has now been entirely sequenced (5). In itself, this is a major technological achievement that carried with it a great deal of raw material and crude data. The analysis of this bacterium's genome sequence has itself led to informative hypotheses on its biology (6). However, because only one reference strain (H37Rv) was originally sequenced (5), information on the species' genetic variability was lacking. Using this sequence and the powerful technology of microarrays, Tsolaki, Hirsh, and colleagues (3, 4) built a highly resolvent tool of comparative genomics to explore the genetic diversity of 100 *M. tuberculosis* isolates from tuberculosis patients in San Francisco. This subsample had been taken from a broader sample of 1,802 isolates collected in the same area between 1991 and 1999, which had been characterized by using two molecular methods: IS6110 and polymorphic G-C-rich sequence (PGRS). These well standardized typing tools are widely used by molecular epidemiologists working with *M. tuberculosis*. They are the basis for international electronic networks of researchers (7). These two markers are not perfect tools in terms of phylogenetic analysis and population genetics, because their target molecules are very specific DNA sequences that are not representative of the entire genome, and because they do not permit the analysis of a large number of independent genetic loci (8). However, they give useful information on the epidemiological relationships among strains. In the two papers analyzed here (3, 4), isolates exhibiting the same profile for both IS6110 and PGRS were considered as the same "strain," related to clustered (epidemiologically linked) cases. The final sample of 100 isolates was composed of 50 clustered and 50 nonclustered isolates.

Because only one strain was originally sequenced, information on the species' genetic variability was lacking.

Large deletions are assumed to play a major role in the molecular evolution of *M. tuberculosis* (9). The two companion papers used these deletions as markers to answer distinct, although complementary, questions. Tsolaki *et al.* (3) took the clear working hypothesis that these large sequence polymorphisms play a more important role than simple DNA base substitutions (single nucleotide polymorphisms) in the phenotypic expression of *M. tuberculosis*, including antibiotic resistance and pathogenicity. Through PCR amplification and sequencing of the regions that flank deletions, it was possible to map them to the base pair. Through statistical randomization procedures, the authors (3, 4) explored how far the localization of deletions departs from random expectations. They found two types of deletions: those that are limited to phylogenetically related isolates and those that are distributed throughout the species. Three deletions combined both types of distribution. Those deletions that are cluster-specific likely stem from a genetic event specific to this lineage, whereas widely distributed deletions reflect properties of the involved sequences that are general in the species.

The H37Rv reference strain was found to comprise slightly more than 4,000 genes (5). In Tsolaki *et al.*'s (3) study, a total of 224 genes were partially or totally deleted by comparison with the reference genome H37Rv. This percentage of 5.5% is significantly lower than the percentage of 22% found in other bacteria such as *Helicobacter pylori* (10) and *Staphylococcus aureus* (11). However, not all genes are equally expressed. Few mutated genes can have drastic effects. Here the authors (3) postulate that this is the case for the agent of tuberculosis and explore the effects of the deletions on the epidemiology of the disease. They find that, generally, deletions have slight deleterious effects. By comparing different functional gene categories, they discovered that some categories were overrepresented in deleted genes. This was the case for those genes involved in intermediary metabolism, respiration, and cell wall production. It is surprising that genes with such a crucial activity are frequently deleted. The hypothesis proposed is that they produce antigens and therefore undergo the selective pressure of the host's immunological system. Isolates that harbor deletions for these genes therefore have a short-term selective advantage, because they are able to better escape the host's immune defenses. However, because these genes are indispensable in the long run, their functionality is then restored. Other deletions had obvious advantages. One enhanced the resistance to isoniazid, a major antituberculosis antibiotic. Others disrupted the genes that allow the bacillus to survive in environments with little oxygen. This last property enables the bacteria to remain in a latency phase, without symptoms. When the genes are disrupted, the disease produces more severe symptoms, including cough, which obviously favors the dissemination of the bacterium (only pulmonary clinical forms are highly contaminating).

Whereas Tsolaki *et al.* (3) focused on the phenotypic effects of large deletions, Hirsh *et al.* (4) used them as phylogenetic markers corresponding to unique

See companion articles on pages 4865 and 4871.

*E-mail: michel.tibayrenc@mpl.ird.fr.

© 2004 by The National Academy of Sciences of the USA

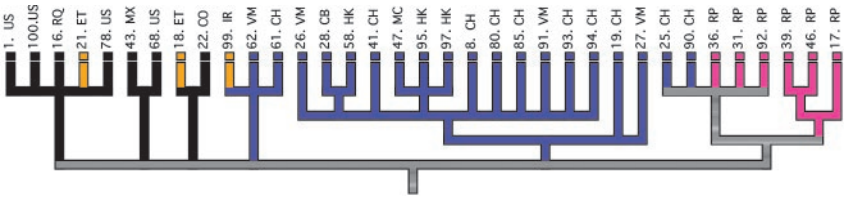

Fig. 1. Correspondence between molecular phylogeny of *M. tuberculosis* strains isolated in San Francisco and birth regions of the corresponding patients. Four large geographic regions are indicated by color: blue, East Asia; pink, The Philippines; black, the Americas; orange, Africa, Europe, and the Middle East. Here and in figure 3 b and c of ref. 4, country of origin is indicated by the abbreviations at the top of the tree: CB, Cambodia; CH, China; CO, Colombia; ET, Ethiopia; HK, Hong Kong; IR, Iran; MC, Macau; MX, Mexico; RP, Philippines; RQ, Puerto Rico; U.S., United States; VM, Vietnam; YO, Yugoslavia. [Reproduced with permission from ref. 4 (Copyright 2004, National Academy of Sciences)].

event polymorphisms, that is to say, evolutionary events that occur only once for all and cannot revert to their former state. Here we have again a very clean working hypothesis: the authors (4) postulated that many tuberculosis patients analyzed in the urban San Francisco area harbor *M. tuberculosis* strains that infected them in their mother countries. Based on this initial idea, it becomes possible to see whether different kinds of *M. tuberculosis* lineages circulate in different regions of the world. Data fully support the working hypothesis. By analyzing the phylogenetic trees and birth countries of the patients, the authors (4) evidenced four major clades (Fig. 1): one including mainly strains from Southeast Asia patients, one including mainly strains from Filipino patients, and the last two composed chiefly of strains having the dominant North American type. Interestingly, Hirsh *et al*. (4) showed that even patients recently contaminated in San Francisco (clustered cases) tend to be infected by those strains that are more specific to their regions of origin. The authors attributed this last result to sociological and epidemiological parameters. The transmission of the bacterium usually requires extensive contact. Moreover, like many large western cities, San Francisco is divided into ethnic districts. Immigrants from the same countries tend to settle together and have, of course, more contact with one another than with other immigrants.

This approach is an extreme example of what modern molecular biology can offer when it depends on a thorough knowledge of the epidemiological and sociological reality. The implications in terms of public health could be immense; for example, it has been hypothesized that the variable efficacy of bacillus Calmette–Guérin in different parts of the world could be due to the circulation of different *M. tuberculosis* strains in these areas. In terms of basic science and evolutionary biology, Hirsh *et al*.'s results show a very fine example of coevolution: the pathogen can be considered a character of the host and vice versa.

These two works are major contributions to our knowledge of the evolution of *M. tuberculosis* and illustrate the power of new technologies in modern biology. The results and hypotheses they contain have obvious implications in terms of medical research as well, because they tell us much about the evolutionary strategies utilized by *M. tuberculosis* to resist antibiotics, disseminate, and infect different human populations. Apart from satisfying our legitimate desire for pure knowledge, this kind of good science builds the foundation for the antibiotics and vaccines of tomorrow.

1. Tibayrenc, M. (2001) *Infect. Genet. Evol.* **1,** 1–2.
2. Anonymous (1994) *Report* (Center for Disease Control and Prevention, Atlanta).
3. Tsolaki, A. G., Hirsh, A. E., DeRiemer, K., Enciso, J. A., Wong, M. Z., Hannan, M., Goguet de la Salmoniere, Y.-O. L., Aman, K., Kato-Maeda, M. & Small, P. M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 4865–4870.
4. Hirsh, A. E., Tsolaki, A. G., DeRiemer K., Feldman, M. W. & Small, P. M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 4871–4876.
5. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Chuercher, C., Harris, D., Gordon, V., Eiglmeier, K., Gas, S., Barry, C. E., III, *et al.* (1998) *Nature* **393,** 537–544.
6. Bishai, W. (1998) *Trends Microbiol.* **6,** 464–465.
7. Behr, M. & Small, P. M. (1997) *Clin. Infect. Dis.* **25,** 806–810.
8. Mazars, E., Lesjean, S., Bañuls, A. L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C. & Supply, P. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 1901–1906.
9. Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3684–3689.
10. Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L. & Falkow, S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14668–14673.
11. Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R. & Musser, J. M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 8821–8826.