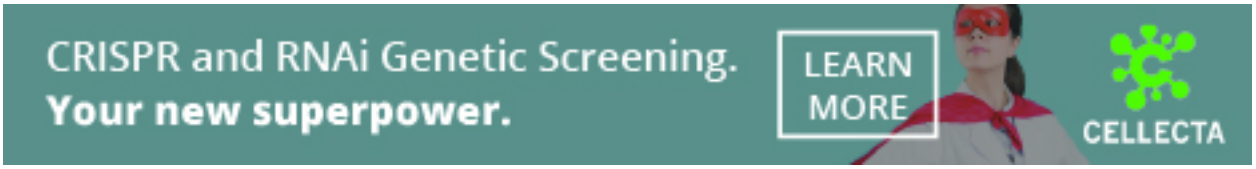
Tools for the Population Genomics of the Tubercle Bacilli

Alexander S. Pym and Roland Brosch

Genome Res. 2000 10: 1837-1839
Access the most recent version at doi:10.1101/gr.169200



Tools for the Population Genomics of the Tubercle Bacilli

Alexander S. Pym[1,2] and Roland Brosch[1,3]

[1]*Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 75724 Paris, CEDEX 15, France;* [2]*Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK*

Advances in sequencing technology have resulted in a rapidly increasing number of completed bacterial genome sequences (http://www.tigr.org/tdb/mdb/mdbcomplete.html, http://igweb.integratedgenomics.com/GOLD/). The relatively small size and limited gene content of these bacterial genomes make them readily amenable to functional genomic analysis. DNA microarrays in particular are proving practical and affordable tools for groups to study the global gene expression of particular organisms (Wilson et al. 1999). Most of the published studies using bacterial genome microarrays have used them to study alterations in gene expression caused by a targeted mutation of a specific regulatory gene or following an external stimulus. However, DNA microarrays also provide a means for complete genome comparisons, either between individual strains from the same species or between closely related species.

In essence, genomic DNA from the bacterial strain of interest is hybridized to the DNA microarray representing the entire genome of the sequenced reference strain and analyzed to determine if any genomic regions of the hybridizing strain are absent relative to the reference strain (Behr et al. 1999). These 'deleted' regions are then further analyzed by PCR and sequencing to define precisely the limits of any apparently deleted region. This currently underexploited use of microarrays will allow researchers to rapidly carry out whole-genome comparisons of large numbers of bacterial strains to determine intra- and interspecies genome variation. Such an analysis has the potential to provide important insights into bacterial evolution, horizontal gene transfer, speciation, and in the case of pathogenic bacteria, the genetic basis of interstrain variations of virulence.

This type of whole-genome deletion detection has already been successfully applied to members of the *Mycobacterium tuberculosis* complex, a single species as defined by DNA/DNA hybridization studies (Imaeda 1985). The *M. tuberculosis* complex includes *M. tuberculosis*, the causative agent in the vast majority of human tuberculosis cases, *M. microti*, an agent of tuberculosis in voles, *M. bovis*, which infects a wide variety of mammalian species including humans, and *M. bovis* BCG, an attenuated variant of *M. bovis*, used extensively since the 1920s as a vaccine against human tuberculosis. Hybridization of *M. bovis* BCG genomic DNA with the genome of *M. tuberculosis* H37Rv, a fully sequenced virulent reference strain (Cole et al. 1998), represented on either a spotted microarray (Behr et al. 1999) or on bacterial artifical chromosome (BAC)-arrays (Gordon et al. 1999), was able to identify up to 16 deletions in the *M. bovis* BCG genome relative to *M. tuberculosis*, ranging in size from 2 to 12.7 kb, extending previous subtractive hybridization studies (Mahairas et al. 1996). These genomic regions were predicted to code for a variety of potential virulence factors and antigens, which can now be systematically studied to determine the genetic basis of BCG's attenuation.

It will now be of great interest to extend this analysis to individual *M. tuberculosis* strains. Tuberculosis is a complex disease with protean manifestations. Although the majority of individuals infected with *M. tuberculosis* remain asymptomatic, with only a small percentage subsequently developing a reactivation leading to overt disease, some individuals progress rapidly to severe disease. Tuberculosis is classically a pulmonary disease but can also present in a more disseminated form or with infections of other specific organs. Host factors are undoubtedly involved in these different disease courses and forms, but it is likely that interstrain differences in virulence are also important. This is further supported by reports of epidemic/hypertransmissible strains (Valway et al. 1998). Sequencing of a second *M. tuberculosis* genome and sequence analysis of structural genes have demonstrated that the genome of *M. tuberculosis* is highly conserved. The synonymous polymorphism rate has been estimated to be as low as one per 10,000 (Sreevatsan et al. 1997), suggesting that deletion or acquisition of genes might be a more important mechanism than point mutations for generating the genetic diversity to account for these phenotypic differences.

Deletions are likely to arise from different processes, but recombination between IS elements is one mechanism that has been well described (Fang et al. 1999, Brosch et al. 1999). Most *M. tuberculosis* clinical isolates contain multiple and variably spaced copies of the IS element IS*6110*, and if appropriately aligned and adjacent, their recombination leads to deletion of the intervening genomic segment. The number and distribution of these elements is sufficiently variable to use them as a basis for RFLP typing of clinical isolates (Small et al. 1994). This extensive diversity suggests that they may be an important

[3]Corresponding author.
E-MAIL rbrosch@pasteur.fr; FAX 33-1-45-68-89-53.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.169200.

mechanism for generating deletions. *M. tuberculosis* also contains >40 other insertion sequences and mobile genetic elements that could also mediate deletion.

Microarrays are powerful tools for determining the distribution of deletions within a population of strains. Although there is continuing progress in the technical aspects of their design and production, the analysis and interpretation of the enormous data set generated by a single hybridization experiment is still problematic. The type of analysis required is dependent not only on the design of the microarray but also on the experimental objectives. An experiment to analyze genome content will need a very different analysis, and probably microarray design, from one designed to determine differences in gene expression. In this issue, Salomon and colleagues (Salomon et al. 2000) have shown how an ingenious computational analysis can enhance the sensitivity of a *M. tuberculosis* Affymetrix GeneChip in the detection and accurate localization of small deletions in the hybridizing strain genome. Because the hybridizing sensitivities and specificities of each microarray probe are different, an analysis based only on individual-probe hybridizing intensities is associated with a high degree of noise. They therefore designed an algorithm to calculate the probability (*P* value) that a poorly or nonhybridizing probe corresponded to a deletion. These *P* values were derived by considering each probe's hybridization signal relative to its neighbors'. Probes with low hybridization scores were therefore only ascribed probabilities consistent with deleted DNA if their neighbors also provided supporting evidence of a deletion. They then elegantly demonstrated the efficacy of this algorithm by successfully detecting all the deletions identified in the fully sequenced strain *M. tubercu-losis* CDC1551 (http://www.tigr.org/tdb/CMR/gmt/htmls/SplashPage.html), one of which was as small as 454 bp, close to the algorithm's limit of detection (350 bp). In addition, they were able to identify and accurately localize three new deletions in *M. bovis* BCG that had not been detected in the previous studies, including one using a spotted microarray.

One limitation of deletion analysis is that it can only identify deletions relative to a fully sequenced reference strain. A single strain will not contain all the genetic material of a species, because the sequenced strain itself may be deleted relative to other members of the species and these additional genes may be responsible for specific phenotypes. This has been shown for *M. tuberculosis* H37Rv, which lacks at least five genomic regions identifiable in clinical isolates and other members of the *M. tuberculosis* complex (Brosch et al. 1999), though no phenotype has been demonstrated for these deletions. Although complete genome sequencing of multiple strains could describe the 'species genome' this is currently cost prohibitive. The techniques of subtractive hybridization could be applied to identify genes present in a test-isolate relative to the reference strain, but these are not yet adapted to analyzing large numbers of samples. Comparative genomics of the members of the *M. tuberculosis* complex and other closely related mycobacterial species provides an alternative strategy. The genome sequences of *M. bovis*, *M. microti*, *M. bovis* BCG and the closely related species *M. leprae*, *M. avium, M. paratuberculosis and M. ulcerans* are currently at different stages of completion (Table 1). These species are likely to have evolved from a common ancestor, and the combined genome sequences from these species may represent a complete mycobacterial gene set, at least for the slow-growing mycobacteria, which may encompass the individual species genomes. Evolution of individual species or subspecies can then be viewed in terms of the loss of portions of this gene pool, resulting in adaptation to specific hosts or niches. This assumes that horizontal transfer into this pool has not been an important process in recent mycobacterial evolution. Analysis of the GC content of the *M. tuberculosis* genome did not reveal any atypical base composition suggestive of a horizontally transferred pathogenicity islands, nor is there any other evidence of recent horizontal transfer (Cole et al. 1998).

Deletion analysis is also not capable of detecting genetic rearrangements and duplication. Gene duplication undoubtedly played an important role in the evolution of the mycobacteria, as proteome analysis of the H37Rv genome suggested that at least 50% of proteins resulted from gene duplication or do-

Table 1. Genome Sequencing Projects for Slow-Growing Mycobacteria

Finished or almost finished

Mycobacterium tuberculosis H37Rv (Cole et al. 1998); http://genolist.pasteur.fr/TubercuList/
M. tuberculosis CDC1551; http://www.tigr.org/tdb/CMR/gmt/htmls/SplashPage.html
M. bovis; http://www.sanger.ac.uk/Projects/M_bovis/
M. bovis BCG; http://www.pasteur.fr/recherche/unites/Lgmb
M. leprae; http://genolist.pasteur.fr./Leproma/
M. avium; http://www.tigr.org/tdb/mdb/mdb.html#progress
M. paratuberculosis; http://www.cbc.umn.edu/ResearchProjects/AGAC/Mptb/Mptbhome.html

Initiated

M. microti
M. ulcerans

main shuffling events (Tekaia et al. 1999). Evidence that this could be important for the ongoing evolution of mycobacterial species is suggested by the observation that two large tandem duplications have arisen in strains of *M. bovis* BCG (Brosch et al. 2000).

Despite these limitations, microarrays are an attractive technique for the study of population genomics. The Affymetrix Genechip in the study by Salomon et al. (2000) was designed for gene expression profiling and therefore was not optimized for deletion analysis. As pointed out by the authors, optimization of the algorithm and probe size and genomic distribution could further enhance the resolution of this technique. This would provide a remarkable tool for high-resolution genome scanning, which will keep population genomicists busy for some time to come.

REFERENCES

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. 1999. *Science* **284:** 1520–1523.

Brosch, R., Philipp, W., Stavropolous, E., Colston, M.J., Cole, S.T., and Gordon, S.V. 1999. *Infect. Immun.* **67:** 5768–5774.

Brosch, R., Gordon, S.V., Buchrieser, C., Pym, A.S., Garnier, T., and Cole S.T. 2000. *Yeast* **17:** 111–123.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. 1998. *Nature* **393:** 537–544.

Fang, Z., Doig, C., Kenna, D.T., Smittipat, N., Palittapongarnpim, P., Watt, B., and Forbes, K.J. 1999. *J. Bacteriol.* **181:** 1014–1020.

Gordon, S.V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., and Cole, S.T. 1999. *Mol. Microbiol.* **32:** 643–656.

Imaeda, T. 1985. *Int. J. Syst. Bacteriol.* **35:** 147–150.

Mahairas, G.G., Sabo, P.J., Hickey, M.J., Singh, D.C., and Stover, C.K. 1996. *J. Bacteriol.* **178:** 1274–1282.

Salomon, H., Kato-Maeda, M., Small, P.M., Drenkow, J., and Gingeras T.R. 2000. *Genome Res.* **10:** 2044–2054.

Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., and Schoolnik, G.K., et al. 1994. *N. Engl. J. Med.* **330:** 1703–1709.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., and Musser, J.M. 1997. *Proc. Natl. Acad. Sci.* 1997. **94:** 9869–9874.

Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G., and Cole, S.T. 1999. *Tubercle Lung Dis.* **79:** 329–342.

Valway, S.E., Sanchez, M.P., Shinnick, T.F., Orme, I., Agerton, T., Hoy, D., Jones, J.S., Westmoreland, H., Onorato, I.M. 1998 *N. Engl. J. Med.* **338:** 633–639.

Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown, P.O., and Schoolnik, G.K. 1999. *Proc. Natl. Acad. Sci.* **96:** 12833–12838.