Available online at www.sciencedirect.com
SCIENCE DIRECT
Research in Microbiology 157 (2006) 282–290
Research in Microbiology
www.elsevier.com/locate/resmic

A genome-wide sequence-independent comparative analysis of insertion–deletion polymorphisms in multiple *Mycobacterium tuberculosis* strains

Tatyana Azhikina [a,*], Nikolay Gvozdevsky [a], Anna Botvinnik [a,1], Alexey Fushan [a], Igor Shemyakin [b], Valentina Stepanshina [b], Mikhail Lipin [b], Clifton Barry 3rd [c], Eugene Sverdlov [a]

[a] *Laboratory of Structure and Functions of Human Genes, Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, 16/10 Miklukho-Maklaya st., Moscow 117997, Russia*

[b] *State Research Center for Applied Microbiology, 1-149 Biologov st., Obolensk 142279, Russia*

[c] *Tuberculosis Research station, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD 20852, USA*

Received 1 April 2005; accepted 12 August 2005

Available online 15 September 2005

Abstract

We applied an enhanced version of subtractive hybridization for comparative analyses of indel differences between genomes of several *Mycobacterium tuberculosis* strains widespread in Russian regions, and the H37Rv reference strain. A number of differences were detected and partially analyzed, thus demonstrating the practicality of the approach. A majority of the insertions found were shared by all Russian strains, except for strain 1540 that revealed the highest virulence in animal tests. This strain possesses a number of genes absent from other clinical strains. Two of the differential genes were found to encode putative membrane proteins and are presumed to affect mycobacterial interaction with the host cell, thus enhancing virulent properties of the isolate. The method used is of general application, and enables the elaboration of a catalogue of indel polymorphic genomic differences between closely related strains.

Keywords: *Mycobacterium tuberculosis* strains; Subtractive hybridization; Comparative genomics

1. Introduction

Tuberculosis remains one of the 10 most important causes of premature mortality worldwide, claiming over 2 million lives every year. Recent increases in the incidence of tuberculosis (particularly drug-resistant forms of the disease) demonstrate that control of tuberculosis requires a concerted effort at a global level. Such control will require significant advances in the basic science underlying diagnosis and treatment, together with prevention by effective vaccination [26]. In Russia in particular, tuberculosis has been exacerbated by recent social and economic changes. An increase in the tuberculosis incidence in Russia began in 1991, and by 1997 it had increased by 117% to reach 73.9 per 100 000 of the total population. The mortality rate reached 17.0 per 100 000 in 1997 [27].

The tenacity of the causative pathogen, *Mycobacterium tuberculosis*, poses a formidable obstacle to control strategies. Great efforts have been made to decipher the molecular causes of virulence; however, very few specific virulence factors have been uniquely associated with the ability to cause disease [4,8, 11,30,42–44]. One factor that represents a considerable challenge to improvement in the diagnosis and treatment of the disease is the variability of *M. tuberculosis,* which in turn causes variability in pathogenic properties of various strains of this bacterial species. The identification of genomic differences among strains of *Mycobacteria* can be used for the development of rapid, reliable methods for identification of subpopulations

* Corresponding author.
E-mail address: tanya@humgen.siobc.ras.ru (T. Azhikina).
[1] Current address: Max-Planck Institute for biophysical Chemistry, Department Membrane Biophysics, Am Fassberg 11, D-37077 Goettingen, Germany.

0923-2508/$ – see front matter
doi:10.1016/j.resmic.2005.08.002

of circulating strains and may help to pinpoint key genomic regions and functional differences related to pathogenesis and disease outcome or presentation.

Recent progress in the sequencing of various mycobacterial species (*M. leprae* [14]); *M. bovis* [17] and 3 strains of *M. tuberculosis* (H37Rv [7], CDC1551 [15]) and 210 (uncompleted)) have enabled detailed comparison of mycobacterial genomes. Amongst *M. tuberculosis* strains, this comparison revealed a high level of genomic polymorphism, including differences resulting from large-sequence polymorphism (LSPs) (greater than 10 bp) and single-nucleotide polymorphism (SNPs) [15]. In contrast to previous findings examining the variation predicted from a small number of loci [34], these data indicate a high level of intraspecies variability, with a substantial contribution from insertion–deletion (indel) polymorphism [15].

However, it is unrealistic to expect massive routine sequencing of numerous *M. tuberculosis* strains in the near future. Therefore, the current problem is how to use the completely sequenced microorganisms as a reference to reveal structural features of other strains or isolates of the same or related species. One possible way to solve this problem involves the parallel comparison of genomes by hybridization to DNA or oligonucleotide microarrays representing the whole genome of a sequenced microorganism. Such an approach was recently successfully used for a comprehensive comparison of the genomes of 100 *M. tuberculosis* clinical strains with the sequenced genome of H37Rv, resulting in the identification of 68 different LSPs that are present in H37Rv but absent from one or more clinical strains [40]. Another important result achieved with the use of microchips shows rather stable associations between host and pathogen populations [18].

With all its power, one of the main drawbacks to the microarray approach is that it does not allow the evaluation of genomic regions absent from the sequenced reference genome in comparison to the genome in question. In addition, microarray technology is inefficient at genomic regions consisting largely of repetitive sequences due to cross-hybridization.

An efficient approach for comparative genomics is subtractive hybridization (SH). This technique allows one to isolate genomic differences between related but phenotypically distinct bacterial strains and/or species [6,25,35,37,38]. Various versions of SH have been used for comparisons of different mycobacterial species (*M. bovis BCG* [25]; *M. ulcerans–M. marinum* [22]); however, to our knowledge it has not been applied to intraspecific *M. tuberculosis* comparisons. One of the problems with successful application of SH for comparison of closely related genomes may arise from difficulties in distinguishing between fragments where the difference is a small deletion or insertion. Such fragments can cross-hybridize and thus avoid detection due to SH. To solve this problem, one can use so-called RFLP SH, that allows for the comparison of genomic DNA restriction fragments from two genomes independently separated by agarose gel electrophoresis [31] and a closely related "In-Gel Competitive Reassociation" method [20]. In this report we describe an enhanced version of SH for comparison of entire bacterial genomic sequences that is capable of providing a comprehensive catalogue of genomic differences. The technique enables reliable isolation of indel genomic polymorphism that is one of the major factors in bacterial variability.

2. Materials and methods

Growth and transformation of *Escherichia coli* cells, preparation of plasmid DNA, agarose gel electrophoresis and other nucleic acid manipulations were performed according to standard molecular biology techniques [33] or recommendations by the manufacturers. The cells for plasmid extraction were grown overnight at 37 °C in 5 ml of Luria–Bertani (LB) medium supplemented with ampicillin (0.1 mg/ml). Plasmid DNA was isolated using a Wizard Plus Minipreps DNA Purification System (Promega) according to the manufacturer's recommendations. Clone inserts were sequenced with an Amersham Biosciences/Molecular Dynamics MegaBACE4000 Capillary Sequencer. Genomic DNA of *M. tuberculosis* strains was obtained as described in [39]. Oligonucleotides were synthesized using an ASM-102U DNA synthesizer (Biosset Ltd., Russia).

2.1. Subtractive hybridization

Purified genomic DNA samples of the two comparable strains (0.5 μg each) were digested with *Alu*I (MBI Fermentas, Lithuania). An adapter I consisting of the non-phosphorylated oligonucleotides Not1Alu (AGCGTGGTCGCGGCCGAGAG) and Not2Alu (CTCTCGGC) (tenfold excess of adapter) was ligated to the obtained fragments. After ligation the purified reaction mix was incubated with 0.5 unit of Taq polymerase (SibEnzyme, Russia) and dNTPs at a final concentration of 0.2 mM at 75 °C for 5 min to extend the adapters ("fill-in"). PCR amplification using the primer Not1Alu was performed immediately after that (95 °C, 15 s; 68 °C, 30 s; 72 °C, 1 min; 10–13 cycles). The amplified products were separated by gel electrophoresis in 0.8% agarose (LMPA, USB). The "zone" representing DNA fragments below 600 bp in size was cut and eluted from the gel. The fraction of fragmented genomic DNA of the reference strain was produced in preparative quantities (5 μg) by PCR with Not1Alu and used as the driver.

Purified genomic DNA from the comparative strain was treated with *Alu*I to remove the Not1Alu adapter. After subsequent purifying, an adapter II comprised of oligonucleotides St19 (AGGGCGTGGTGCGGAGGGCGGT) and St20 (ACCGCCCTCCG) was ligated to 5′-termini of double-stranded DNA for use as the tracer.

Purified DNA samples in a molar ratio of tracer/driver (1:100) were combined, lyophilized and dissolved in 4 μl of 10 mM EEPS, 1 mM EDTA (pH 8.0) buffer. The mixture was overlaid with approximately 30 μl of mineral oil and denatured at 99 °C for 3 min. 1 μl of 5 M NaCl solution was added to the reaction mix. Hybridization was carried out at 68 °C overnight. Following co-hybridization, tracer-specific targets were selectively amplified with primer St19 by PCR under the following conditions: "fill-in" (72 °C, 3 min), followed by 20 cycles of (95 °C, 15 s; 68 °C, 30 s; 72 °C, 1 min).