Comparative genomics of *Mycobacterium tuberculosis* and *Escherichia coli* for recombination (*rec*) genes

Despite the importance of *Mycobacterium tuberculosis* as a pathogen, the genetic basis of pathogenesis remains poorly understood. It is believed that allele exchange would facilitate understanding of the role(s) of specific genes that encode virulence determinants and help the conception of new therapeutic and prophylactic interventions. Although allele exchange has been achieved in *Mycobacterium smegmatis*, it has proved to be inefficient in *M. tuberculosis* due to the existence of an efficient illegitimate recombination (IR) system. The molecular mechanism for the intrinsic resistance of *M. tuberculosis* to allele exchange and increased frequency of IR remain obscure.

Classical genetic studies with *M. tuberculosis* have been hampered by the paucity of efficient selection procedures. As a result, no mutants with recombination-defective phenotypes similar in magnitude to that observed in *Escherichia coli* or other organisms have been isolated to date. The determination of the complete *M. tuberculosis* genome sequence (3) has opened the way for 'comparative genomics' and 'reverse genetics' by which the total genomic complement could be compared with the more intensively studied organisms. Here, we present several surprising insights gained from the analysis of the *M. tuberculosis* complete genome sequence for genes that encode components of homologous genetic recombination (HR).

Analysis of recombination-deficient mutants of *E. coli* and their suppressors in various genetic backgrounds led to the isolation of several genes that are thought to be involved in HR and to the concept of multiple genetic pathways of recombination (9). Central to all current models of HR, the initiation of recombination entails generating 3′-ended single-stranded DNA that can be acted upon by RecA (9). In wild-type *E. coli*, the RecBCD pathway is the major route for recombination and repair of double-strand breaks. The key component of this pathway, the RecBCD enzyme, encoded by *recB*, *recC* and *recD*, recognizes the blunt ends of double-stranded DNA generated by DNA damage or conjugative transfer of chromosomal DNA. After binding to this end, the helicase and nuclease activities of RecBCD convert the double-stranded DNA into a 3′ invasive single-stranded DNA. In the absence of RecBCD, several other gene products alone or in combination generate 3′-ended single-stranded DNA. In the RecE pathway, the *recE* gene product, a double-strand-specific 5′→3′ exonuclease is responsible for the production of single-stranded DNA with a 3′ tail, whereas in the RecF pathway, 3′-ended single-stranded DNA is generated by unwinding of duplex DNA by RecQ helicase with concomitant 5′→3′ resection of single-stranded DNA by RecJ nuclease. Finally, an alternative mechanism of initiation is mediated by RecQ helicase. Although it was believed that RecQ helicase is required only in the absence of RecBCD, the ability of RecQ helicase to begin recombination events at nicks or gaps suggests a distinct non-overlapping role for this enzyme in the initiation of HR in wild-type *E. coli* (12). The duplex DNA is first processed by a specific exonuclease/helicase to generate single-stranded DNA with 3′ ends. As the DNA ends are being resected, RecA polymerizes on the 3′-ended single-stranded DNA guided by single-stranded binding protein (SSB) to form a helical nucleoprotein filament. It has been established that a complex comprising RecF, RecO and RecR assists RecA in synapsis (5). The nucleoprotein filament then rapidly searches and aligns with homologous sequences in the duplex DNA to produce a joint molecule. The third step involves the extension of heteroduplex DNA by branch migration. The progressive expansion of heteroduplex DNA results in the formation of a Holliday junction involving four DNA strands. Finally, the Holliday junction is resolved by symmetrical cleavage by RuvC endonuclease to generate two heteroduplex DNA products (9, 11).

Examination of the complete genome sequence of *E. coli* and *M. tuberculosis* provided an opportunity to explore the functional genomic content and evolutionary relationship between them at a qualitative level. The complete genome sequence of *M. tuberculosis* was analysed by searching for homologues of *E. coli recA*, *recB*, *recC*, *recD*, *ssb*, *recF*, *recR*, *ruvA*, *ruvB*, *ruvC* and *recG*. A database search revealed that only one ORF structure similar to each gene was present in *M. tuberculosis*. The degree of sequence homology of *M. tuberculosis* Rec proteins with those of *E. coli* is very high. The conservation of the *E. coli* RecBCD pathway in *M. tuberculosis* implies that the tubercle bacillus can carry out processes such as recombinational repair of double-strand breaks and conjugational recombination. It must be noted that the ability of *M. tuberculosis* to actually perform the latter remains to be demonstrated, however. The RecBCD-like activity in *M. smegmatis* does not process the incoming linear duplex DNA (14). Assuming that the putative RecBCD enzyme behaves identically in non-pathogenic and pathogenic mycobacteria, we speculate that this feature might contribute to inefficient allele exchange in *M. tuberculosis*.

On the basis of sequence similarities noted among the components of the RecBCD pathway between the genomes of *E. coli* and *M. tuberculosis*, we reasoned that genes involved in other pathways of HR might be homologous. Analysis of the complete genome

Table 1. *rec* genes and homologous recombination proteins

ND, *E. coli* homologue not detectable. The nucleotide sequence of *M. tuberculosis* H37Rv was obtained from the Wellcome Trust Pathogen Genome Unit at the Sanger Centre, Cambridge, UK (http://www.sanger.ac.uk/Projects/M.tuberculosis/blast-server.shtml). Contiguous sequence database (TB.seq) was retrieved from ftp://ftp.sanger.ac.uk/pub/tb/sequences generated by Cole *et al.* (3). Sequence similarity matching to *E. coli* recombination proteins was performed by searching for homologues in the NCBI complete genome database (2) using TBLASTN of the GAPPED-BLAST search program (1). The nucleotide sequence was translated in all six reading frames. This program compares a given query sequence against all other proteins and nucleic acid sequences in the database to identify related proteins and present them in the order from highest to lowest similarity scores. The results of BLAST searches of the *E. coli* genome (2) were examined for homologues in *B. subtilis* (10) that scored highly. The query used for the search was either the sequence corresponding to the protein from *E. coli* and/or *B. subtilis* (obtained from GenBank). All the retrieved sequences having scores >70 were considered for further analysis. The sequences of homologous proteins of *M. tuberculosis* were retrieved and multiple sequence alignment was created using the CLUSTALW program (PCGENE software). Sequence alignments were visually inspected for signature sequences.

Rec proteins and functions	*rec* genes		
	E. coli	***M. tuberculosis***	***B. subtilis***
RecA protein (strand transfer, ATPase)	*recA*	*recA*	*recA*
Exonuclease V (ATPase, helicase)	*recB recC recD*	*recB recC recD*	*addA addB*
SbcCD exonuclease	*sbcC sbcD*	ND	*yirY sbcD*
ssDNA exonuclease (RecJ, ExoI)	*recJ sbcB*	ND	*yrvE yorK*
RecQ helicase	*recQ*	ND	*recQ*
RecE exonuclease	*recE*	ND	ND
RecT protein	*recT*	ND	*yqaK*
ssDNA-binding protein	*ssb*	*ssb*	*ssb*
RecF-RecO-RecR	*recF recO recR*	*recF recR* (*recO* absent)	*recF recR* (*recO* absent)
Holliday junction resolvases	*ruvA ruvB ruvC rusA recG*	*ruvA ruvB ruvC recG* (*rusA* absent)	*ruvA ruvB ylpB*

sequence of *M. tuberculosis* by the same approach for homologues of *E. coli sbcB*, *sbcC*, *sbcD*, *recJ*, *recO recQ*, *recE* and *recT* yielded unexpected results. The striking difference between the genomes of *M. tuberculosis* and *E. coli* is that homologues of the RecE and RecF pathways are not detectable in *M. tuberculosis* (Table 1). This is significant considering the association of RecE, SbcB, SbcCD and RecJ exonuclease and RecQ helicase activities in the generation of 3′-ended single-stranded DNA. We note that *recE* and *recT* are found on a cryptic prophage in *E. coli* and likely to be absent in the tubercle bacillus. However, the origin of the *recT* homologue in *Bacillus subtilis* is obscure. In wild-type *E. coli*, recombination between DNA molecules containing extensive stretches of homology requires the components of the RecF pathway and RecA (9). In addition, RecQ helicase plays a key role in disrupting aberrant recombination events and abolition of IR (8, 9). Thus, it is possible that the increased frequency of IR in *M. tuberculosis* is caused by the absence of genes that encode exonuclease and helicase activities required for the generation of substrates and processing of intermediates.

What is the biological role of recombination in bacteria under normal growth conditions? In *E. coli*, recent observations underscore a key role for HR in repair of double-strand breaks and re-establishment of stalled replication forks (11). The RecF pathway, which plays a limited role in HR in wild-type *E. coli*, seemingly is required for reactivation of stalled replication forks at lesions in the template strand (4). The most important genetic components of the RecF pathway are *recFOR*, *recJ*, *recQ* and *recN*. Among these components, *recFOR* are crucial for the re-establishment of stalled replication forks (4, 11). Comparative genomic analysis discloses that *recJ*, *recQ* and *recO* of *E. coli* are missing in *M. tuberculosis*. Thus, we speculate that *M. tuberculosis* is likely to be defective in recombinational repair, especially in the reactivation of stalled replication forks at DNA lesions.

An important factor that might affect the validity of these comparisons relates to the differences in the G+C content between the genomes of these two organisms. Gram-positive bacteria can be divided into two major classes based on G+C content of the genome and signature sequences in different proteins. One class of species, represented by *B. subtilis*, is characterized by a low G+C content. Species in the second category, of which *M. tuberculosis* is one, are characterized by a high G+C content. Phylogenetic analyses have indicated that the latter group is more similar to Gram-negative bacteria (7). The available data and existing models suggest that the differences in protein coding sequences are confined to changes in the third codon position (7). To determine whether the absence of key components of HR is common to all Gram-positive species, we searched for homologues in the genome sequence of *B. subtilis*. Despite the fact that *B. subtilis* is distantly related to Gram-negative bacteria, the presence of almost all the components of HR in *B. subtilis* provides a strong reaffirmation to the differences observed between *E. coli* and *M. tuberculosis* (Table 1).

From the perspective of comparative genomics, we note that the genome of *M. tuberculosis* lacks key components of the RecE and RecF pathways. It is possible that the absence of these genes renders HR inefficient and thereby allows the integration of newly introduced DNA at random sites. It has also been noted that homologues of the proteins

involved in the *E. coli* mismatch-directed repair pathway are also missing in *M. tuberculosis* (6, 13). The question is whether the absence of certain genes of HR in *M. tuberculosis* is due to the overall genomic compaction? The genome size (4·41 Mb) of *M. tuberculosis* is slightly larger than that of *B. subtilis* (4·22 Mb) and smaller than that of *E. coli* (4·6 Mb). This observation argues against the notion of the genome evolving through a series of modifications, but rather for the idea of genes subjected to evolutionary constraints. Consequently, some genes are retained and others are expended. Currently, it is not known whether this has any significance in the biology of *M. tuberculosis*.

In summary, the functional genomics of *M. tuberculosis* is still in its infancy. The question is whether the identification of homologues of HR from the complete *M. tuberculosis* genome sequence would enable us to predict the basis for low levels of HR and high frequencies of IR in this organism. The answer, of course, is an emphatic no. It is important to examine the roles of each of these components systematically and then apply this information to enhance the efficiency of allele exchange and decrease random integration events. Thus, we anticipate that such an approach will be useful in providing a valuable tool in facilitating genetic manipulation of mycobacteria which has immense relevance to public health.

Acknowledgements

This work was supported by grants from the Council of Scientific and Industrial Research, New Delhi, and the Wellcome Trust, UK. We thank the reviewer for thoughtful suggestions.

K. Muniyappa, M. B. Vaze, N. Ganesh, M. Sreedhar Reddy, N. Guhan & R. Venkatesh

Department of Biochemistry, Indian Institute of Science, Bangalore 560012, India

Author for correspondence: K. Muniyappa. Tel: +91 80 309 2235. Fax: +91 80 360 0683/0814.
e-mail: kmbc@biochem.iisc.ernet.in

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

2. Blattner, F. R., Plunkett, G. I., Bloch, C. A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K12. *Science* **277**, 1453–1474.

3. Cole, S. T., Brosch, R., Parkhill, J. & 39 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

4. Courcelle, J., Carswell-Crumpton, C. & Hanawalt, P. C. (1977). *recF* and *recR* are required for the resumption of replication at DNA replication forks in *Escherichia coli*. *Proc Natl Acad Sci USA* **94**, 3714–3719.

5. Cox, M. M., Goodman, M. F., Kruezer, K. N., Sherrat, D. J., Sandler, S. J. & Marians, K. J. (2000). The importance of repairing stalled replication forks. *Nature* **404**, 37–41.

6. Eisen, J. A. & Hanawalt, P. C. (1999). A phylogenetic study of repair genes, proteins, and processes. *Mutat Res* **435**, 171–213.

7. Gupta, R. S. (1998). Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435–1491.

8. Hanada, K., Ukita, T., Kohno, Y., Saito, K., Kato, J. & Ikeda, H. (1997). RecQ DNA helicase is a suppressor of illegitimate recombination in *Escherichia coli*. *Proc Natl Acad Sci USA* **94**, 3860–3865.

9. Kowalczykowski, S. C., Dixon, D. A., Eggelston, A. K., Lauder, S. D. & Rehrauer, W. M. (1994). Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol Rev* **58**, 401–465.

10. Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

11. Kuzminov, A. (1999). Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage λ. *Microbiol Mol Biol Rev* **63**, 751–813.

12. Lloyd, R. G. & Thomas, A. (1984). A molecular model for conjugational recombination in *Escherichia coli* K-12. *Mol Gen Genet* **197**, 328–336.

13. Mizrahi, V. & Anderson, S. J. (1998). DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence? *Mol Microbiol* **29**, 1331–1339.

14. Winder, F. G. & Barber, D. S. (1973). Effects of hydroxyurea, nalidixic acid and zinc limitation on DNA polymerase and ATP-dependent deoxyribonuclease activities of *Mycobacterium smegmatis*. *J Gen Microbiol* **76**, 189–196.