The 409 bp tandem repeat spanning genes yxaK and yxaL is absent from the *Bacillus subtilis* chromosome

The complete genome sequence of *Bacillus subtilis* was published in 1997 (4). One of its surprising features was the presence of a nearly perfect tandem repeat of 409 bp, spanning yxaK and yxaL. In comparison, the longest tandem repeat of the *Escherichia coli* chromosome is 200 bp in length (10). The stability of tandem duplications has been studied using artificial systems in which deletion events confer a selectable phenotype, both in *E. coli* (2, 3, 5, 7, 8) and in *B. subtilis* (C. Bruand, personal communication). In the *E. coli* chromosome, for tandem repeats of 600–800 bp, the deletion rate is in the range of $10^{-4}$ to $10^{-8}$, depending on experimental systems. Deletion rates have been found in the range of $10^{-4}$ in the *B. subtilis* chromosome (C. Bruand, personal communication). Once lost, such repeats cannot re-appear; therefore the presence of the 409 bp tandem repeat in the *B. subtilis* chromosome could be the indication of a selective advantage for the strain.

We decided to investigate further this region of the chromosome. Fig. 1(a) shows a restriction map of *B. subtilis* 168 in the yxaK and yxaL region. Large black arrows indicate a nearly perfect 409 bp direct repeat spanning genes yxaK and yxaL. Whereas the first repeat spans 80% of yxaK, the second repeat contains the last 22 bp of yxaK, and the first 355 bp of yxaL. A single extra A, interrupting the homology in the second repeat (which is therefore 410 bp long), is indicated with an asterisk. According to this map, a Southern blot analysis of this region with a probe homologous to yxaK–yxaL should reveal a 1.8 kb DraI fragment and a 4 kb XmnI fragment containing the duplication. For the DraI restriction, an additional 1.1 kb fragment outside the repeat is expected. The result of such a Southern blot, performed on strain 168 as well as two different reference strains is shown in Fig. 1(b). The DraI and XmnI fragments supposed to contain the duplic-
A microbial mat composed of iron bacteria

Microbial mats can be defined as biofilms that are visible to the naked eye. Most studies of microbial mats have been of those sustained by photosynthesis (2, 3, 13), but such studies, apart from a single contribution on microbial mats in sulphurous waters in a cave (12), have been limited to photosynthetic mats. Microbial mats with iron oxide coatings were found at the edge of a hydrothermal vent field discharging water of pH 5.3–5.5 and rich in dissolved iron at a depth of about 1000 m off Hawaii (10). Scanning electron microscopy showed unbranched bacterial filaments 50–150 μm long and 1–2 μm diameter. Microbial mats with high iron content were also found associated with suspected hydrothermal venting at a depth of about 600 m in Crater Lake, Oregon, USA (4). Scanning electron microscopy demonstrated that sheaths of Leptothrix and stalks of Gallionella, both well known genera of iron bacteria, were major components of the mat. Study of the above two sites involved the use of submersibles, limiting the amount of work that could be done. Subsequently, an iron-oxidizing microbial mat was found in a more accessible environment at Aarhus, Denmark, permitting more detailed study (5, 6). The mat was located on a cement fountain and adjacent stone wall over which flowed water rich in ferrous iron (up to 250 μM). Light microscopy showed Leptothrix ochracea and Gallionella to be major components of the mat. In the course of a study on iron transport and retention in an ochre-rich watercourse (1), we have observed a similar microbial mat community at a second accessible site, as reported below.

L. ochracea (11) has been known for nearly two centuries, but pure culture has never been achieved and identification is based on morphology and habitat. The organism is restricted to near neutral waters rich in ferrous iron where prolific filamentous growth is often visible to the naked eye and the thick-walled sheaths survive long after they have been vacated by bacterial cells. In view of the habitat, iron autotrophy has been suspected but never proved. We find that the long, curved unbranched sheaths become almost invisible ‘ghosts’ and shrivel after a few minutes in 1% oxalic acid, a solvent for ferric iron. So iron, even if not used as an energy source, seems to be an essential structural component of the sheaths. Gallionella (9) has also been known for nearly two centuries, occurs in near neutral waters rich in ferrous iron and is identified on the basis of morphology. Pure culture has been achieved and iron autotrophy proved (8), but the most striking feature in samples from nature, the helical stalks of hydrated ferric oxide, are not always produced in culture (7).

We observed microbial mats in the upper part of the watercourse where we studied it. Foci where the stream drains the remaining fragment of the iron-rich Sunninghill Bog, to the west of South Ascot Recreation Ground (Ordnance Survey grid reference SU 924 679). Here the stream is about 1 m wide and usually about 10 cm deep, the pH is 6.3 ± 0.2, dissolved oxygen is at about 80% full saturation, the annual temperature range is 5–14°C and the stream flows typically at 10–20 l s⁻¹. Ferrous iron seeps into the stream, resulting in concentrations of 180–270 μM. Although the amount of suspended ferric iron is commonly low, usually equivalent to 20–50 μM, the gravel bed of the stream is usually covered with ochre-rich sediments several centimetres deep. Where there is a solid substratum and flow rates are low near the stream bank, cushions of L. ochracea with diameters up to several centimetres can occur and L. ochracea may festoon sedges and other vegetation where it dips into the water. If such vegetation is wiped clear of Leptothrix, or sticks are inserted into still water, fresh growth is soon visible, often the next day in summer. The ochre-rich sediments on the stream bed may be swept away by heavy rain. When this occurs, the stream bed becomes covered by a microbial mat about 1 mm thick which may be so tough that if it is torn in an appropriate way with a stick, the current will roll it up rather than it fragmenting. If the mat is removed by raking the stream bed, it re-appears; in summer perceptible growth has been observed the following day, with a tough mat present in a few days. Light microscopy shows that the mat resembles that studied at Aarhus (5, 6), with tangled filaments of L. ochracea and stalks of Gallionella, both of which are readily identifiable from their characteristic morphology, being major structural components.

Our limited observations suggest that such microbial mats might be common and that
seeking them in ochre-rich streams could be fruitful. It is likely that the mats would need to have the stout and long filaments of *L. ochracea* as a component. Therefore, in view of the pH requirements of this organism (11), only streams near neutral would be likely to be rewarding. Visits to potential sites would need to take place within a few days of heavy rain, after removal of ochre allowed mats to develop and before further ochre deposition smothered them.

Michael J. Carlile¹ and A. William L. Dudene²

¹Department of Biology, Imperial College, Silwood Park, Ascot SL5 7PJ, UK
²T. H. Huxley School of Environment, Earth Science and Engineering, Imperial College, London SW7 2BP, UK

Author for correspondence: A. William L. Dudeny. Tel: +44 020 7594 7322. Fax: +44 020 7594 7444. e-mail: a.dudeny@ic.ac.uk


**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 convergent 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 synopsis (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 conjugal 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 processes 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
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 recombination 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

---

**Table 1. rec genes and homologous recombination proteins**

<table>
<thead>
<tr>
<th>Rec proteins and functions</th>
<th>rec genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>M. tuberculosis</td>
</tr>
<tr>
<td>RecA protein (strand transfer, ATPase)</td>
<td>recA</td>
</tr>
<tr>
<td>Exonuclease V (ATPase, helicase)</td>
<td>recB recC recD</td>
</tr>
<tr>
<td>SbcCD exonuclease</td>
<td>sbcC sbcD</td>
</tr>
<tr>
<td>ssDNA exonuclease (RecJ, ExoI)</td>
<td>recJ sbcB</td>
</tr>
<tr>
<td>RecQ helicase</td>
<td>recQ</td>
</tr>
<tr>
<td>RecE exonuclease</td>
<td>recE</td>
</tr>
<tr>
<td>RecT protein</td>
<td>recT</td>
</tr>
<tr>
<td>ssDNA-binding protein</td>
<td>sbb</td>
</tr>
<tr>
<td>RecF-RecO-RecR</td>
<td>recF recO recR</td>
</tr>
<tr>
<td>Holliday junction resolvases</td>
<td>ruwA ruwB ruwC ruwA</td>
</tr>
</tbody>
</table>

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 BLASTN 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.
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 (41 Mb) of M. tuberculosis is slightly larger than that of B. subtilis (42 Mb) and smaller than that of E. coli (46 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.

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.

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