Minireview

Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv

Stewart T. Cole*

Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 75724 Paris Cedex 15, France

Received 9 April 1999

Abstract ***Mycobacterium tuberculosis*, the scourge of humanity, is one of the most successful and scientifically challenging pathogens of all time. To catalyse the conception of new prophylactic and therapeutic interventions against tuberculosis, and to enhance our understanding of the biology of the tubercle bacillus, the complete genome sequence of the most widely used strain, H37Rv, has been determined. Bioinformatic analysis led to the identification of ~4000 genes in the 4.41 Mb genome sequence and provided fresh insight into the biochemistry, physiology, genetics and immunology of this much-feared bacterium. Genomic information is centralised in TubercuList (http://www.pasteur.fr/Bio/TubercuList/).**

Key words: Genomics; Tuberculosis; Repetitive DNA; Lipid metabolism

1. Introduction

More human lives have been lost to tuberculosis than to any other disease [1] and the disease is also of central importance to veterinary medicine. Since its isolation by Koch in 1882 [2], the human tubercle bacillus, *Mycobacterium tuberculosis*, has posed a formidable challenge to generations of biomedical researchers as a result of its long generation time, fastidious growth requirements and high risk of contagion. In the last decade, the ability to perform molecular genetic analysis of *M. tuberculosis* has resulted in powerful new research tools [3,4] while the availability of the complete genome sequence has provided us with a wealth of new information, knowledge and understanding of the biology of this major pathogen [5]. Comparative genomics of the agent of bovine tuberculosis, *M. bovis*, and the attenuated vaccine strain, *M. bovis* BCG, will undoubtedly shed new light on the molecular basis of pathogenicity and further the conception of new treatments and preventive therapies.

2. Integrated genome maps

An essential feature of mycobacterial genomics has been the construction of integrated genome maps. These were obtained by linking the physical map of the chromosome, obtained by pulsed field gel electrophoresis, to the contig map, comprising sets of ordered cosmid or BAC clones, via various landmarks [6,7]. Physical mapping generates an independent estimate of genome size and organisation whereas the clones that constitute the contig map represent an immortalised source of genomic DNA; this is particularly valuable for slow-growing pathogens. Large insert clones facilitate a modular approach to mycobacterial genetics and molecular biology, and constitute ideal substrates for genome sequencing [5]. Integrated maps have been generated for both *M. tuberculosis* H37Rv and *M. bovis* BCG Pasteur [6–9], both of which have a single circular chromosome which is 4.4 and 4.35 Mb in size, respectively.

3. Features of the genome sequence of *M. tuberculosis* H37Rv

Using a combined approach, in which selected BACs and cosmids were systematically analysed, together with random small insert clones, the complete genome sequence of the well-characterised H37Rv strain of *M. tuberculosis* was determined as part of a collaborative project featuring the Sanger Centre and the Institut Pasteur. The genome comprises 4 411 529 bp and has an average G+C content of 65.6% although some areas with an exceptionally high G+C content (>80%) were detected and found to correspond to a novel gene family. By means of bioinformatics, 50 genes encoding stable RNA species and 3924 genes encoding proteins were identified, and these account for >91% of the potential coding capacity [5]. This value is typical of bacterial genomes as is the gene density at one gene per 1.1 kb. In contrast to the situation in fast-growing bacteria such as *Bacillus subtilis*, the orientation of genes with respect to the direction of replication is less biased, as only 59% of them are transcribed with the same polarity as the replication forks compared to 75% in *B. subtilis* [10]. It is generally believed that higher expression levels can be obtained by coordinating directions of transcription and replication so the more even distribution seen in *M. tuberculosis* may be a reflection of its slow growth rate.

4. Genes, gene duplication and the proteome

Database comparisons were used to glean possible function information about the protein-coding genes and these were classified into 11 broad groups (Table 1). Functions were confidently attributed to ~40% of the protein-coding genes, some information or similarity was found for a further ~40% although many of these belong to the class known as conserved hypotheticals, while the remainder of the genes may well be confined to mycobacteria as they show no similarity to any other microbial sequences. About 51% of the coding sequences have arisen from gene duplication events [5],

*Fax: (33) (1) 40 61 35 83.
E-mail: stcole@pasteur.fr

PII: S0014-5793(99)00536-0

Table 1
Broad classification of *M. tuberculosis* genes

Class	Function	Gene number	% total	Total length (kb)	% total coding
1	Lipid metabolism	225	5.7	372	9.3
2	Information pathways	207	5.2	243	6.1
3	Cell wall and cell processes	517	13.0	620	15.5
4	Stable RNAs	50	1.3	10	0.2
5	Insertion sequences and phages	137	3.4	100	2.5
6	PE and PPE protein	167	4.2	283	7.1
7	Intermediary metabolism and respiration	877	22.0	985	24.6
8	Proteins of unknown function	607	15.3	396	9.9
9	Regulatory proteins	188	4.7	162	4.0
10	Conserved hypothetical proteins	911	22.9	739	18.4
0	Virulence, detoxification, adaptation	91	2.3	95	2.4
	Non-coding sequences			434	

as is also the case for *Escherichia coli* and *B. subtilis*, bacteria with similar-sized genomes [10,11]. However, in the tubercle bacillus the degree of sequence conservation of duplicated genes is much higher and although this indicates that there must be extensive functional redundancy it is also consistent with the notion proposed by Musser and his colleagues that *M. tuberculosis* is of recent evolutionary descent [12].

The high G+C content of the genome is reflected in the amino acid composition of the proteome as amino acids such as Gly, Ala, Pro and Arg, that are encoded by G+C-rich codons, are overrepresented whereas those encoded by A+T-rich codons, such as Lys and Asn, are relatively scarce [5]. Correspondence analysis of the proteome led to the identification of two large protein families, the PE and PPE families, whose amino acid composition differs radically from that of the bulk of the proteins. Both the PE and PPE proteins are exceptionally glycine-rich while the PPE proteins also contain copious amounts of asparagine, an amino acid that is generally rare in the proteome. Curiously, asparagine is the preferred nitrogen source for *M. tuberculosis* [13], and this raises the possibility that the PPE proteins may serve as storage proteins.

Prior to completion of the genome sequence, the existence of these protein families, whose genes occupy $>7\%$ of the total coding sequence, was unknown. It was clear, however, that the genome contained two dispersed simple sequence repeats referred to as PGRS (polymorphic G+C-rich sequence) and MPTR (major polymorphic tandem repeat) [14,15] and these have since been shown to correspond to part of the 3′ ends of the PE and PPE genes. The names PE and PPE signify the presence of characteristic Pro-Glu (positions 8, 9) and Pro-Pro-Glu (positions 8–10) motifs in the well-conserved NH_2-terminal domains of the proteins and these generally precede repetitive domains of variable length. The PE_PGRS proteins contain multiple tandem repetitions of Gly-Gly-Ala (or a variant thereof), while the PPE_MPTR polypeptides are rich in repeats with the signature Asn-X-Gly-X-Gly-Asn-X-Gly. Functional information is scarce but the repetitive nature and variation in size between strains suggest that both PE and PPE proteins may represent antigens of immunological relevance. The PGRS proteins resemble EBNA, the Epstein-Barr virus nuclear antigens that block proteasome action leading to inhibition of MHC class I antigen presentation [16–18]. For further details of these unusual mycobacterial proteins consult [19].

5. Notable biological activities

For decades, the mycobacterial cell envelope, the 'waxy coat', has been the subject of intensive research as it contains a remarkable array of unusual lipids, glycolipids, mycolic acids and polyketides. It was not greatly surprising to find examples of every known lipid and polyketide biosynthetic system encoded in the genome, including enzymes usually associated with mammals and plants. More genes encoding potential lipid biosynthetic activities were uncovered than there are known metabolites thus raising the intriguing possibility that many more novel lipid and polyketide species remain to be found. A significant portion of the genome is devoted to genes involved in lipid metabolism (Table 1) and, unexpectedly, much of this encodes enzymes potentially involved in fatty acid degradation. In addition to the classical β-oxidation cycle, catalysed by the multifunctional FadA/FadB proteins, it is possible that alternative lipid oxidation pathways exist as there are >100 genes encoding enzymes that could catalyse individual steps of this cycle. These might degrade lipids present in host vacuolar or cellular membranes and hence contribute to energy metabolism [5].

The proteome is also predicted to contain 20 cytochrome P450-containing monooxygenases and their related redox partners. While it is not unusual to encounter cytochrome P450 in bacteria [20], the tubercle bacillus has far more of these enzymes than any other prokaryote examined to date. P450s catalyse mixed oxidation of hydrophobic compounds such as long-chain fatty acids, and are involved in sterol transformation and xenobiotic metabolism. The latter activity occurs in *Pseudomonas putida*, an aerobe that degrades organic matter such as camphor by means of $P450_{CAM}$, the paradigm of the bacterial P450s [21]. Saprophytes living in the soil are likely to employ this form of metabolism and the large number of P450s found in *M. tuberculosis* may indicate that its ancestor was a soil mycobacterium.

There is a remarkable lack of genetic diversity among the members of the *M. tuberculosis* complex, with single nucleotide changes being encountered at a frequency of $\sim 3\times10^{-4}$, an exceptionally low value for a bacterium. Musser and his colleagues have suggested that this may indicate that tuberculosis was disseminated recently or that *M. tuberculosis* has encountered an evolutionary bottle-neck [12]. The basis for this remarkable homogeneity is obscure but may reflect either replication machinery of exceptionally high fidelity or a very

efficient DNA repair system. Surprisingly, the genome does not appear to encode a mismatch repair system [22] but does contain three *mutT* homologues, which may purge the nucleotide pool of the oxidised guanines whose incorporation during replication causes base mismatching [23]. It has been suggested that lack of mismatch repair may lead to fixation of polymerase-mediated slippage errors resulting in sequence variation in repetitive sequences [22].

6. Repetitive DNA

There are three main sources of repetitive DNA in the tubercle bacillus: duplicated genes and gene families, IS elements, and dispersed non-coding sequences. An unusual feature of the genomes of both *Mycobacterium leprae* and *M. tuberculosis* is the existence of duplicated copies of genes, such as *cysA* and *povE* [24], that are identical in sequence. While gene duplications are common in eubacteria, as exemplified by the EF-Tu genes of *E. coli*, extensive sequence divergence generally occurs after duplication – once again the intracellular mycobacteria appear to differ radically in this respect [12]. The most prominent multigene families present in *M. tuberculosis* are those encoding the PE and PPE proteins, and several of their respective members are nearly identical in sequence suggesting that they may result from very recent gene duplication events. Of special interest are the PE and PPE proteins belonging to the PGRS and MPTR classes, respectively, as their genes are composed of multiple tandem repetitions of simple sequences such as CGGCGGCAA and GCCGGTGTTG [15]. The sequences of these genes appear to be prone to expansion or contraction [19], probably as a result of strand slippage during replication, and this is likely to be the principal source of genomic polymorphism.

The *M. tuberculosis* genome is particularly rich in IS elements and contains 56 copies belonging to at least nine different families (Table 2). The IS elements often occur in clusters and are very scarce in a segment of the genome encompassing *oriC* [25]. This suggests that insertional hot spots may exist or that, as described in *Rhizobium* sp. strain NGR234 [26], transposons preferentially target IS islands thereby preventing essential genes from inactivation [27]. With the notable exception of IS*6110*, which transposes frequently, most of the IS elements found in strain H37Rv appear to be stable, and occur in the corresponding sites of the genomes of other *M. tuberculosis* isolates, *M. bovis* and BCG. Of interest is IS*1532,* which is found in some but not all members of the *M. tuberculosis* complex [25].

The genome of H37Rv also contains seven copies of a repetitive sequence referred to collectively as the '13E12' family.

Table 2
IS elements found in *M. tuberculosis* H37Rv

IS family	Members in *M. tuberculosis*
IS*3*	IS*6110* (16), IS*1540*, IS*1604*
IS*5*	IS*1560*, IS*1560′*, IS-*like* (2)
IS*21*	IS*1532*, IS*1533*, IS*1534*
IS*30*	IS*1603*
IS*110*	IS*1547* (2), IS*1558*, IS*1558′*, IS*1607*, IS*1608′* (2)
IS*256*	IS*1081* (6), IS*1552′*, IS*1553*, IS*1554*
IS*1535*	IS*1535*, IS*1536*, IS*1537*, IS*1538*, IS*1539*, IS*1602*, IS*1605″*
IS*L3*	IS*1555*, IS*1557* (2), IS*1557′*, IS*1561′*, IS*1606″*
Unknown	IS*1556*

These range in size from 1355 to 1448 bp, and contain an open reading frame that, in some cases, appears to require frame shifting to produce a protein. It is unclear whether the 13E12 repeat corresponds to an IS element as its putative products show no similarity to known transposase sequences [25]. Lee et al. described a 453 bp repeat sequence that is specific for the *M. tuberculosis* complex and showed it to be present in three or four copies in various strains [28]. This 453 bp repeat corresponds to a subsequence of the 13E12 repeat. It is clear that one copy of the 13E12 repeat has inserted into the *bio* operon where it serves as the attachment site for the prophage phiRv1 [5,25]. As seven potential *att* sites for phiRv1 exist it is conceivable that the prophage will be found in other locations on the genome and hybridisation data suggest that this may be the case in the Erdman strain [28].

The *M. tuberculosis* genome also contains about 65 copies of a novel dispersed repeat referred to as the mycobacterial interspersed repetitive unit or MIRU [29] that ranges in size from 46 to 101 bp. These sequences are also found in *M. leprae* where they were termed REP1 [30]. Frothingham and coworkers [31] have described repetitive sequences called variable number tandem repeats (VNTR) and exact tandem repeats (ETR), that are probably equivalent to MIRUs. Most MIRUs occur at the 5′ end of genes and those copies situated between genes in operons often have the potential to encode peptides and appear to have inserted into sites involved in translational coupling [29]. As indicated by the terms VNTR and ETR, MIRUs can exist as tandem copies that display variable copy number and this property has been exploited for strain typing purposes [32].

7. Pathogenicity

Although *M. tuberculosis* is undoubtedly one of the most successful human pathogens ever, little is known about the molecular basis of its pathogenicity. In the 1890s, Robert Koch demonstrated that the tubercle bacillus does not produce toxins [33] and it is now clear that much, if not all, of the pathology associated with the disease results from excessive cell-mediated immune and inflammatory responses, mounted by the host in response to bacterial antigens [34]. On analysis of the genome sequence no features reminiscent of the horizontally transferred pathogenicity islands seen in enteric bacteria were found [35] and only a very limited number of genes showing similarity to known virulence genes were detected [36]. These include four copies of an operon encoding putative invasins [37] and the iron acquisition system, as well as a variety of phospholipases C and other lipases, esterases, cutinases and proteases that might act on cellular or vesicular components.

References

[1] Bloom, B.R. and Murray, C.J.L. (1992) Science 257, 1055–1064.
[2] Koch, R. (1882) Berl. Klin. Wochenschr. 19, 221–230.
[3] Jacobs Jr., W.R., Kalpana, G.V., Cirillo, J.D., Pascopella, L., Snapper, S.B., Udani, R.A., Jones, W., Barletta, R.G. and Bloom, B.R. (1991) Methods Enzymol. 204, 537–555.
[4] Pelicic, V., Jackson, M., Reyrat, J.M., Jacobs Jr., W.R., Gicquel, B. and Guilhot, C. (1997) Proc. Natl. Acad. Sci. USA 94, 10955–10960.
[5] Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth,

T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, A., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S. and Barrell, B.G. (1998) Nature 393, 537–544.
[6] Brosch, R., Gordon, S.V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B.G. and Cole, S.T. (1998) Infect. Immun. 66, 2221–2229.
[7] Philipp, W.J., Poulet, S., Eiglmeier, K., Pascopella, L., Subramanian, B., Heym, B., Bergh, S., Bloom, B.R., Jacobs Jr., W.R. and Cole, S.T. (1996) Proc. Natl. Acad. Sci. USA 93, 3132–3137.
[8] Gordon, S.V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K. and Cole, S.T. (1999) Mol. Microbiol. 32, 643–656.
[9] Philipp, W.J., Nair, S., Guglielmi, G., Lagranderie, M., Gicquel, B. and Cole, S.T. (1996) Microbiology 142, 3135–3145.
[10] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F. and Danchin, A. et al. (1997) Nature 390, 249–256.
[11] Blattner, F.R., Plunkett, I.G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Wayne-Davies, N., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) Science 277, 1453–1462.
[12] 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. USA 94, 9869–9874.
[13] Grosset, J. (1993) in: Tuberculosis: A Comprehensive International Approach (Reichmann, L.B. and Hershfield, E.S., Eds.), pp. 49–74, Marcel Dekker, New York.
[14] Poulet, S. and Cole, S.T. (1995) Arch. Microbiol. 163, 87–95.
[15] Poulet, S. and Cole, S.T. (1995) Arch. Microbiol. 163, 79–86.
[16] Laal, S., Sharma, Y.D., Prasad, H.K., Murtaza, A., Singh, S., Tangri, S., Misra, R.S. and Nath, I. (1991) Proc. Natl. Acad. Sci. USA 88, 1054–1058.
[17] Levitskaya, J., Coram, M., Levitsky, V., Imreh, S., Steigerwald, M.P., Klein, G., Kurilla, M.G. and Masucci, M.G. (1995) Nature 375, 685–688.
[18] Levitskaya, J., Sharipo, A., Leonchiks, A., Ciechanover, A. and Masucci, M.G. (1997) Proc. Natl. Acad. Sci. USA 94, 12616–12621.
[19] Cole, S.T. and Barrell, B.G. (1998) in: Genetics and Tuberculosis (Novartis Foundation Symposium 217) (Chadwick, D.J. and Cardew, G., Eds.), pp. 160–172, John Wiley, Chichester.
[20] Munro, A.W. and Lindsay, J.G. (1996) Mol. Microbiol. 20, 1115–1125.
[21] Peterson, J.A. and Graham, S.E. (1998) Structure 6, 1079–1085.
[22] Mizrahi, V. and Andersen, S.J. (1998) Mol. Microbiol. 29, 1331–1339.
[23] Cole, S.T. (1998) Curr. Opin. Microbiol. 1, 567–571.
[24] Fsihi, H., De Rossi, E., Salazar, L., Cantoni, R., Labò, M., Riccardi, G., Takiff, H.E., Eiglmeier, K., Bergh, S. and Cole, S.T. (1996) Microbiology 142, 3147–3161.
[25] Gordon, S.V., Heym, B., Parkhill, J., Barrell, B. and Cole, S.T. (1999) Microbiology 145, 881–892.
[26] Freiberg, C., Fellay, R., Bairoch, A., Broughton, W.J., Rosenthal, A. and Perret, X. (1997) Nature 387, 394–401.
[27] Perret, X., Viprey, V., Freiberg, C. and Broughton, W.J. (1997) J. Bacteriol. 179, 7488–7496.
[28] Lee, T.Y., Lee, T.J., Belisle, J.T., Brennan, P.J. and Kim, S.K. (1997) Tuberc. Lung Dis. 78, 13–19.
[29] Supply, P., Magdalena, J., Himpens, S. and Locht, C. (1997) Mol. Microbiol. 26, 991–1003.
[30] Smith, D.R., Richterich, P., Rubenfield, M., Rice, P.W., Butler, C., Lee, H.-M., Kirst, S., Gundersen, K., Abendschan, K., Xu, Q., Chung, M., Deloughery, C., Aldredge, T., Maher, J., Lundstrom, R., Tulig, C., Falls, K., Imrich, J., Smyth, A., Torrey, D., Drill, S., Avruch, A., Engelstein, M., Breton, G., Madan, D., Nietupski, R., Seitz, B., Connelly, S., McDougall, S., Safer, H., Doucette-Stamm, L., Eiglmeier, K., Bergh, S., Cole, S.T., Robison, K., Jaehn, L., Gryan, G., Johnson, J., Church, G.M. and Mao, J. (1997) Genome Res. 7, 802–819.
[31] Frothingham, R. and Meeker-O'Connell, W.A. (1998) Microbiology 144, 1189–1196.
[32] Magdalena, J., Vachée, A., Supply, P. and Locht, C. (1998) J. Clin. Microbiol. 36, 937–943.
[33] Koch, R. (1891) Dtsch. Med. Wochenschr. 17, 101–102.
[34] Tascon, R.E., Colston, M.J., Ragno, S., Stavropoulos, E., Gregory, D. and Lowrie, D.B. (1996) Nature Med. 2, 888–892.
[35] Hacker, J., Blum, O.G., Muhldorfer, I. and Tschape, H. (1997) Mol. Microbiol. 23, 1089–1097.
[36] Finlay, B.B. and Falkow, S. (1997) Microbiol. Mol. Biol. Rev. 61, 136–169.
[37] Arruda, S., Bomfim, G., Knights, R., Huima-Byron, T. and Riley, L.W. (1993) Science 261, 1454–1457.