Genome variation in mycobacterium tuberculosis

Genome variation in *Mycobacterium tuberculosis*

N.C. Gey van Pittius[a], S.L. Sampson[b], R.M. Warren[a] and P.D. van Helden[a*]

Genome variation is the main underlying reason for phenotypic differences observed between organisms from the same species. This includes minor variations like single nucleotide polymorphisms, but also large sequence polymorphisms due to deletions and insertions. Some of these variations are of insignificant functional importance, but others may lead to an increase in viability and, in the case of microbes, possibly an increase in pathogenicity. However, the influence of variations on phenotype is not always dependent on the size of the mutated domain, and even single nucleotide polymorphisms can cause significant changes in an organism. This is of course an important area of research into studying pathogenic organisms and their epidemiological features. *Mycobacterium tuberculosis* strains also display a degree of strain variation, which has been previously suggested to be important for the outcome of disease. Certain strains are postulated to be more or less virulent, persistent, transmissible, or immunopathological. These variations do not only have important implications for the spread of the disease, but provide us with tools to characterize transmission, which have been used extensively for a number of years in tuberculosis molecular epidemiology. In this article, we discuss the different methods which are available for detecting genome variation; we look at the different mechanisms which cause this variation (including insertions, deletions, duplications and single nucleotide polymorphisms), and review the consequences and implications of this genome variation with regard to *Mycobacterium tuberculosis* pathogenicity.

Introduction

It is well known that different strains of the same species of an organism can have dramatically different phenotypic characteristics.[1,2] For pathogenic organisms, host range and virulence (or disease-causing capabilities) have become common measures of differences among strains. This trait-specific genetic diversity has been particularly relevant in the field of virology, where fast-changing genotypes play a significant role in yearly disease outbreaks with newly evolved strains.[3] *Mycobacterium tuberculosis* strain variability has also been suggested to be important for the outcome of disease after exposure, as different clinical strains can have distinct differences in phenotype and virulence.[4–6]

Since the advent of genomics, it has become clear that most, if not all, of these phenotypic differences between strains and species can be attributed to genome variation (including variation that will lead to differential expression and regulation). The use of comparative genomics to compare and analyse genomic differences between sequenced whole genomes has become increasingly popular due to the growing availability of large amounts of completed and partially completed genome sequences (e.g. ref. 7). Variation in genome make-up has thus become an important field of study in an effort to identify disease-causing virulence factors (e.g. ref. 8). Furthermore, genome variation also allows us to study disease dynamics by identifying and tracking the spread of specific epidemic strains within a community by use of markers specific for a certain strain.[9] In this article, we focus on the methods of identifying genome variation within the species *Mycobacterium tuberculosis*; we look at the specific mechanisms that could lead to phenotypic differences between strains within this species, and we discuss the implications of the translation of these genomic variations for functional differences between strains present in a community with a high incidence of tuberculosis (TB) disease. We also discuss the way in which genomic variation translates to phenotypic differences at a species level.

[a]US/MRC Centre for Molecular and Cellular Biology, Department of Medical Biochemistry, Faculty of Health Sciences, Stellenbosch University, Tygerberg 7505, South Africa.
[b]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, U.S.A.
*Author for correspondence. E-mail: pvh@sun.ac.za

Methods for detecting genomic variation

A number of common markers are in use today for the molecular differentiation of *Mycobacterium tuberculosis* isolates. These typing techniques form the basis of modern molecular epidemiology of tuberculosis and are all based on genome variation. This includes the IS*6110* insertion sequence,[10] the direct repeat (DR) region,[11,12] the polymorphic GC-rich sequences (PGRS)[13] and the variable number tandem repeats (VNTR) sequences (or mycobacterial interspersed repetitive units — MIRUs).[14] Various combinations of these markers are used in different settings in order to study the disease dynamics of TB by identifying and tracking specific strains within the community.[15]

The insertion sequence IS*6110*

The IS*6110* insertion element, identified by Thierry and co-workers,[10] is the most well known of the insertion sequences in the mycobacterial species, due to its widespread application in molecular epidemiological studies of tuberculosis. This insertion element is 1355 base pairs (bp) in length with imperfect 28 bp terminal inverted repeats.[16] It contains two transposase ORFs (ORF 1 and 2), which overlap by 1 bp. *M. tuberculosis* strains have between 0 and 25 copies of the IS*6110* element in their genomes, and variation in both the number and location of IS*6110* elements present in the genome makes it a very useful marker of strain genotype, leading to IS*6110* DNA fingerprinting being considered as the reference typing technique for *M. tuberculosis* isolates.[15,17]

Based on IS*6110* RFLP fingerprinting, the strain collection of our Cape Town study community shows over 340 different strains of *M. tuberculosis* isolated from approximately 800 TB patients, with three dominant strain families, namely F11, the W-Beijing family (F29), and F28.[18,19] The well-known W-Beijing strain family, which is disseminated throughout the world, is currently the second most dominant family in our study community, present in approximately one out of every five patients,[20] with F11 and F28 being present in 22% and 10% of patients, respectively.

The direct repeat region

Spoligotyping (spacer oligotyping) is a PCR-based method that exploits the unique structure of the direct repeat (DR) region in *M. tuberculosis*.[12,21] The DR region is made up of direct repeat sequences interspersed with variable spacer sequences,[11] which in combination have been termed DVRs.[12,22] This region

has been shown to be polymorphic in different clinical isolates of *M. tuberculosis* and *M. bovis,*[11,23] due to IS*6110* insertion, homologous recombination-driven deletion and strand slippage-generated duplication.[22,24–27] Different *M. tuberculosis* strains thus contain different numbers and variants of DVRs and are distinguished from each other as such. The spoligotyping method uses reverse line-blot hybridization to detect the presence and absence of the spacers of 43 unique DVR repeat sequences after being amplified from the genomic DNA of the *M. tuberculosis* isolate.[21] We have used this method extensively in our own studies for the accurate classification of strains of episodes of recurrent tuberculosis.[28]

The polymorphic GC-rich sequences

Ross and co-workers[13] identified a number of sequences in the genome of *M. tuberculosis* with a GC content of around 80%, containing characteristic expansions of a consensus repeat CGGCGGCAA, which encode multiple tandem repeats of a glycine-glycine-alanine or a glycine-glycine-asparagine motif. These polymorphic GC-rich sequences (PGRS) form the C-terminal domains of a subgroup of the PE multigene family, which is present in approximately 99 copies in the genome of *M. tuberculosis*.[29] DNA fingerprinting analysis using these repeats as probes shows extensive polymorphism in different strains of *M. tuberculosis,* making it useful for exploitation in molecular epidemiology as an informative secondary probe.[13,30–32]

The mycobacterial interspersed repetitive units

The genome of *M. tuberculosis* contains 41 MIRU loci.[14] MIRUs (mycobacterial interspersed repetitive units) are 40–100 bp minisatellite DNA elements often found as tandem repeats and dispersed in intergenic regions of the genome.[14] Twelve of these MIRU loci display polymorphisms (variations in MIRU tandem repeat copy numbers and sequence variations between repeat units) when compared in different clinical isolates of *M. tuberculosis,* and thus present an opportunity for use as powerful markers in tuberculosis molecular epidemiology, population genetics and phylogenetic studies.[33] In our study community, these markers are often utilized in the differentiation of *M. tuberculosis* strains belonging to the low IS*6110*-copy number clusters, which are difficult to separate into smaller groups and of which the genetic relationships to other strains have thus traditionally been difficult to elucidate.[34,35]

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) result from random nucleotide mutations. SNPs can occur synonymously (sSNPs) or non-synonymously (nsSNPs) in genes, resulting in either silent mutations or in amino acid changes, respectively. nsSNPs cause changes in protein sequence and are thus an agent of phenotypic evolution in an organism. Two nsSNP polymorphisms in the genes *katG* and *gyrA* have been used in a restricted fashion to assign *M. tuberculosis* strains to three principal genetic groups.[36] sSNPs, on the other hand, cause no change to protein sequence and are thus more useful targets for phylogenetic studies due to their neutrality and ease of assay.[37] A study of 230 sSNPs in 432 *M. tuberculosis* complex strains demonstrated that sSNP genotyping efficiently delineates relationships among closely-related *M. tuberculosis* strains, as well as successfully resolving the phylogenetic positions of organisms belonging to the low IS*6110*-copy number clusters,[37] thereby providing valuable new insights into phylogenetic relationships among different *M. tuberculosis* strains. Nevertheless, care should be taken when using these markers as their accuracy relies heavily on the use of properly distanced reference strains, failure of which results in distorted and re-rooted outgroups.[38]

Agents of genomic variation

In the preceding section we discussed different regions of genome variation in *M. tuberculosis* and the ways in which these could be utilized to identify polymorphic *M. tuberculosis* strains. This section will focus on ways in which these polymorphisms originate.

Insertions

Unlike the DR and MIRU markers, IS*6110* not only serves as an excellent marker to differentiate between strains, it is also an agent of genome evolution, since it acts as a highly mobile unit able to integrate (transpose) into multiple places in the genome of *M. tuberculosis*.[24] These insertions can have a number of different effects, for example an insertion event may knock out or disrupt a gene,[39,40] it can up- or down-regulate expression of downstream genes via its internal promoter,[41,42] or it may cause genomic rearrangements to take place.[26,43] These effects all have the potential to change the phenotype of the organism and influence the fitness of the strain. We have previously shown that IS*6110* insertion events in the strains present in our study community predominantly takes place into coding regions,[39] with 57 of 95 (60%) of discrete IS*6110* insertion sites occurring within potential open reading frames in the genome of *M. tuberculosis*.[44] This points to a high rate of gene disruption via IS element insertion in the genome of *M. tuberculosis* and implicates the IS*6110* element as a significant agent in genome variation between strains. A further 7% of insertions could not be mapped to the genome of *M. tuberculosis* H37Rv (implying that the percentage of gene disruption may be even higher), and 33% were found to be intergenic, which may also disrupt gene regulation. Examples of disrupted genes include: adenylate cyclases, phospholipase C (*plcD*), small membrane protein family (*mmpS1*), serine-threonine kinases (*pknJ*), and PPE gene family members.

Deletions

Deletions are another important mode of genome variation. These deletions could affect whole gene sequences or parts of genes, as well as regulatory regions up- and downstream of genes. Previously, it has been suggested that deletions mediated by the IS*6110* insertion sequences may be an important mechanism of variation in the genome of *M. tuberculosis*.[43,45] This was confirmed by work from our group, which suggested that adjacently situated, inversely orientated IS*6110* elements in clinical isolates of *M. tuberculosis* can mediate genome deletion by mechanisms that include (but are not limited to) homologous recombination.[26,27] This finding highlights the importance of the IS*6110* element not only as an instrument of genome evolution via insertion events, but also via deletion mechanisms.

In a microarray analysis of 100 clinical isolates of *M. tuberculosis,* dozens of discrete deletions were found, which ranged in size from 196 bp to 11 986 bp.[46] The largest of these single deletion events caused the deletion of 217 ORFs, and the range of ORF deletion per deletion event ranged from 0–33 ORFs, with a median size of 15 ORFs.[46] This has led to the theory of reductive evolution, where it is postulated that *M. tuberculosis* undergoes changes on its way towards a minimal genome in which only the necessary genes are kept and the rest can be lost.[4] As proof of this, Small and co-workers have shown that dispensable *M. tuberculosis* genes are twice as likely to be absent, and five times as likely to be pseudogenes in the genome of *M. leprae.* Furthermore, even when present, dispensable genes are not expressed in this organism.[46]

Gene duplications – the PE and PPE gene families

The existence of two highly duplicated gene families within the genome of *M. tuberculosis* provides evidence for the presence of another means of gene disruption and phenotypic changes within this species, namely gene duplication. The PE and PPE gene families are large (99 and 68 members, respectively), dispersed throughout the genome of *M. tuberculosis*, and comprise approximately 10% of the genome.[29] The PE and PPE gene families are named after the presence of a proline-glutamic-acid (PE) motif at positions 8 and 9 in a conserved N-terminal domain of around 110 amino acids[47] for the PEs, and a highly conserved N-terminal domain of around 180 amino acids, with a proline-proline-glutamic acid (PPE) motif at positions 7–9 for the PPEs.[29] The C-terminal domains of both of these protein families are of variable size and sequence and contain repeat sequences of different copy numbers in a number of cases.[47] The mobility of the PE and PPE genes present them as agents of genome evolution as they have the potential to disrupt gene sequences with each duplication event, as well as inserting into regulatory elements in intergenic regions. These multigene families are themselves also prone to gene disruption and variation, as a number of *IS6110* insertions and disruptions were identified within the PPE gene family,[39] and PPE gene polymorphisms were detected by hybridization with IS*6110* flanking sequences (S.L. Sampson *et al.*, unpubl. results).

Although evidence has been presented that indicates that they may be cell wall associated proteins,[44,48,49] and that they may inhibit antigen processing or may be involved in antigenic variation due to the highly polymorphic nature of their C-terminal domains,[29,47,50] the function of these glycine-rich protein families are still unknown,[29] and as such the influence of disruptions within these genes are also unknown. Because they are located in a wide variety of genomic positions, where they may also be regulated differently, these multiple gene family members may have developed different functional roles through evolution. This is supported by the fact that not only are the same PE and PE-PGRS genes differentially expressed within different *M. tuberculosis* strains during *in vitro* growth,[51] but different members of the gene family are also expressed differentially under different *in vitro* and *in vivo* conditions within the same organism (S.L. Sampson *et al.*, unpubl. results; and refs 52–54). Furthermore, sequence and protein size differences were detected between the same PE and PPE orthologues in different strains of *M. tuberculosis*,[8,55] and hybridization analysis performed in our laboratory on 18 randomly selected clinical isolates with five PPE 5′ terminal probes revealed some highly variable fragments in terms of size (S.L. Sampson *et al.*, unpubl. results).

Additional gene sequence analysis revealed long tandem repeats (LTRs) in three members (Rv1753c, Rv1917c and Rv1918c) of the MPTR (major polymorphic tandem repeat) subgroup of the PPE family (S.L. Sampson *et al.*, unpubl. results; and refs 29, 44). This is the largest PPE subgroup and proteins of this subgroup contain multiple repeats of the motif AsnXGlyXGlyXAsnXGly encoded by a consensus repeat sequence GCCGGTGTTG, separated by 5-bp spacers.[56] Analysis of the sequence data from the three PPE-MPTR genes containing LTRs obtained from different clinical strains demonstrated that expansion or contraction of tandem repeat regions (by slipped strand mispairing or homologous recombination) was responsible for variations in size between the same genes from different strains (S.L. Sampson *et al.*, unpubl. results; ref. 44). These differential expression and sequence variation results contribute to a growing body of evidence supporting the hypothesis that the members of these two gene families may function as variable surface antigens. The fact that these genes encode for about 4% of the proteome of the organism (if all members are expressed), indicates that they most probably fulfil an important function or functions in the organism.

Single nucleotide polymorphisms

nsSNPs arising from random nucleotide mutations may, as described in the preceding section, be an important agent of genome variation in *M. tuberculosis*, because nucleotide changes lead to amino acid replacements in protein sequences. It has been extrapolated from data generated from 56 genes studied in several hundred *M. tuberculosis* complex strains, that about one synonymous nucleotide substitution occurs per 10 000 synonymous nucleotide sites in structural genes within the genome of *M. tuberculosis*.[36] Due to the greater propensity for random mutations to occur in nonsynonymous positions of codons, it points to an even higher rate of mutations that will likely influence protein sequence and structure/function and which will ultimately lead to evolutionary changes in the specific strain. Fraser *et al.*[57] estimated genome-wide frequency of SNPs to be about 1 in 3000 nucleotide sites, with Musser[58] estimating this to correspond to around 1500 total SNPs, with approximately 750 present in ORFs. In agreement with this assumption, Gutacker and co-workers[37] have shown that 65% of the ~900 SNPs detected in a comparison between the genomes of *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551 were nsSNPs.

Consequences of genome variation

In the preceding sections we have discussed methods of detecting genomic differences between different strains of *M. tuberculosis*, as well as the mechanisms by which these differences originate. The present section explores the influence of these genomic changes when translated to functional differences, which may lead to phenotype changes and thus be driving forces in the evolution of the organism.

In a wide-ranging study of mycobacterial growth in human macrophages, 17 different clinical isolates of *M. tuberculosis* as well as the laboratory strain H37Rv were assayed to determine the influence of interstrain genomic variation on intra-macrophage mycobacterial replication.[59] Although no consistent or marked differences were observed between the clinical isolates, they all grew consistently faster (29% on average) than H37Rv, implying greater levels of fitness when compared to the laboratory strain.[59] This could be because H37Rv has been subjected to decades of culturing and possible corresponding progressive loss in virulence.

In a follow-up study by Hoal-van Helden and co-workers,[60] diversity of *in vitro* cytokine responses by human macrophages to infection with different *M. tuberculosis* strains was investigated. Macrophages infected *in vitro* with the same 17 different clinical *M. tuberculosis* strains as well as the laboratory strain H37Rv, produced varying levels of both TNF-α and IFN-γ (cytokines released by macrophages upon activation). Once again, the laboratory strain H37Rv was markedly less effective at inducing cytokines than the clinical isolates, while efficiency of response generation also varied between different clinical isolates, probably reflecting genomic variations and subsequent differences in the host response elicited to infection.

Manca and co-workers[61] provided early evidence for bacterial phenotype differences by showing that the *M. tuberculosis* strain CDC1551 is not more virulent than other *M. tuberculosis* strains (for example, the clinical isolates HN60 and HN878 and the

laboratory strains H37Rv and Erdman) in terms of *in vivo* and *in vitro* growth, but that it does elicit a more vigorous host immune response. Similarly, significant differences were observed in the ability of different *M. tuberculosis* clinical isolates to grow within human monocytes, with the four most rapidly growing isolates all being members of the Beijing strain family.[62] The Beijing strain 210 was also shown to have an enhanced capacity to grow in human macrophages when compared to small cluster and unique *M. tuberculosis* strains, while eliciting the production of similar amounts of cytokines.[63] Patients infected with Beijing strains more often have fever unrelated to disease severity, toxicity, or drug resistance,[64] and a study by Lopez *et al.*[65] showed that mice infected with Beijing strains had extensive pneumonia, a non-protective immune response signified by early but short-term TNF-α and iNOS expression, and high levels of early death when compared with infections with H37Rv. In the same study, it was shown that Canetti strains induced limited pneumonia, continued expression of TNF-α and iNOS, and almost complete survival, while Somali and Haarlem strains displayed varying survival rates.[65]

In a comparison of the survival rates of three different *M. tuberculosis* strains (*M. tuberculosis* Erdman, H37Rv and CDC1551) in a rabbit model of tuberculosis, Erdman strain appeared to be the most virulent, with the greatest spectrum of disease and fewer organisms required to cause tubercle formation.[66] These differences may possibly be correlated with genomic differences found in the RD6 deletion region between the Erdman and H37Rv strains.[66] Kato-Maeda and co-workers[4] previously observed that the likelihood that a certain *M. tuberculosis* strain would cause pulmonary cavitation decreased with the increase in the amount of deletions in its genome.

Another dramatic example of genomic variations translating to phenotype changes is observed in drug resistance generation by SNPs. It is well known that resistance to antimycobacterial drugs is due to a number of genomic mutations in specific genes present within the genome of *M. tuberculosis*.[67] Table 1 gives an indication of some of the main SNP positions, which are known to confer drug resistance to the organism, clearly showing the remarkable effect that even single nucleotide polymorphisms could have on the phenotype of an organism.

Table 1. Major SNP positions conferring drug resistance.

Drug	Gene (codon)
Isoniazid (INH)	*katG* (315) *inhA* (–15) *ahpC* (–10) *kasA* *Ndh* (110)
Rifampicin (RMP)	*RpoB* (531,526,526)
Streptomycin (SM)	*Rps*L (43, 88) *rrs* (513)
Ethambutol (EMB)	*emb*B (306)
Pyrazinamide (PZA)	*pnc*A (139)
Fluoroquinolones (FQ)	*Gyr*A,B (95)

Adapted from Victor *et al.*[67]

Deletions are also major agents of phenotypic change, and a growing body of evidence is pointing to deletions as the most important factor distinguishing different species of the same genus.[68–72] This evidence is becoming especially evident in the genus *Mycobacterium*, where a large number of genomes from different species have been or are in the process of being sequenced (for a complete updated list see the Gold Genomes Online Database at http://wit.integratedgenomics.com/GOLD/). An important example of a mycobacterial region that, when

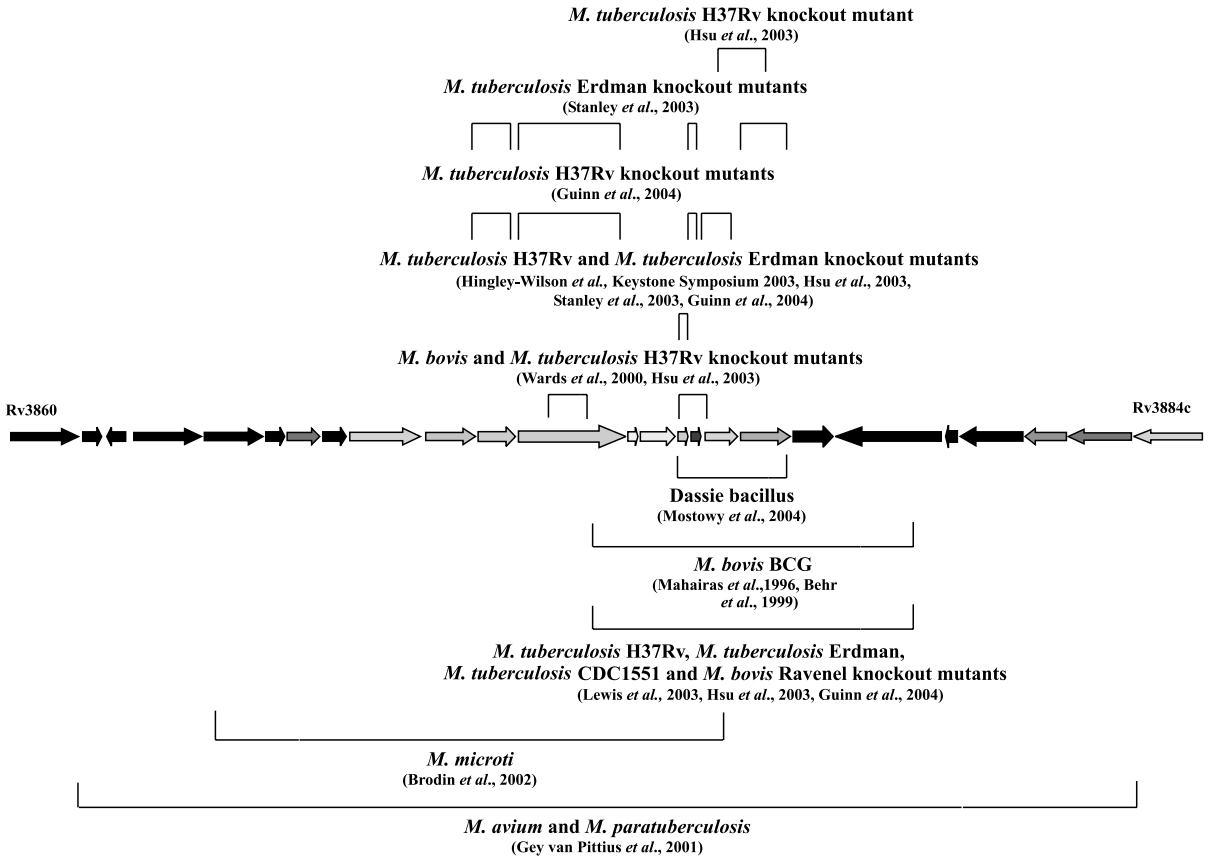

Fig. 1. Diagrammatic representation of the genes present in the RD1 region, detailing published positions of natural and knockout deletions causing attenuation of the specified species.

totally or partly deleted, causes significant phenotypic change, is that of the RD1 deletion region. This region was deleted during serial passage of *M. bovis*, generating the attenuated strain *M. bovis* BCG, which is used worldwide as a vaccine against childhood tuberculosis.[70,73–75] In the last few years, others have shown attenuation of the organism when parts of this region are deleted from *M. bovis*[76] and *M. tuberculosis* Erdman or H37Rv[74,75,77–80] (Fig. 1). Intriguingly, evidence has also become available showing deletions in this region in naturally-attenuated mycobacterial species, for example *M. microti*,[70,81] *M. avium, M. paratuberculosis*,[7] and the Dassie bacillus.[82] Furthermore, even when the whole region is present, significant down-regulation seems to exert the same effect as deletion, with the expression of the whole RD1 region being suppressed in the attenuated *M. tuberculosis* strain H37Ra.[83]

Implications of *M. tuberculosis* genome variation

The study of *M. tuberculosis* strain variation could lead to useful information that would help identify targets for drug and/or vaccine design. It also provides us with markers to study and understand the epidemic by using molecular epidemiological techniques. Molecular epidemiology attempts to answer a number of questions about the epidemic, for example: whether the disease is driven by transmission or by reactivation (which is important for understanding the efficacy of current treatment regimes), whether transmission occurs within the household or in the community, whether drug resistance is acquired or transmitted, whether relapse cases are true relapses or reinfections, whether the frequency of given strain families are the same in primary and subsequent episodes of disease, whether certain strains are better adapted to certain host conditions, whether a previous infection is able to protect against subsequent infection, and whether vaccines will thus be able to prevent disease. This will help us to understand disease dynamics, ensure efficient monitoring of any interventions (for example, chemo- or vaccine-therapy and its outcome), identify points for intervention, help identifying strains for analysis (e.g. for microarrays, deletion mutants, antigens, or virulence) and ultimately assist with the improvement (by design) of an efficient anti-tuberculosis vaccine.

1. Wilske B., Barbour A.G., Bergstrom S., Burman N., Restrepo B.I., Rosa P.A., Schwan T., Soutschek E. and Wallich R. (1992). Antigenic variation and strain heterogeneity in *Borrelia* spp. *Res.Microbiol.* **143**, 583–596.
2. Peeters M., Piot P. and Groen G. van der. (1991). Variability among HIV and SIV strains of African origin. *AIDS* **5**, Suppl 1: S29–S36.
3. de Jong J.C., Rimmelzwaan G.F., Fouchier R.A. and. Osterhaus A.D. (2000). Influenza virus: a master of metamorphosis. *J. Infect.* **40**, 218–228.
4. Kato-Maeda M., Rhee J.T., Gingeras T.R., Salamon H., Drenkow J., Smittipat N. and Small P.M. (2001). Comparing genomes within the species *Mycobacterium tuberculosis. Genome Res.* **11**, 547–554.
5. Kato-Maeda M., Bifani P.J., Kreiswirth B.N., and Small P.M. (2001). The nature and consequence of genetic variability within *Mycobacterium tuberculosis. J. Clin. Invest.* **107**, 533–537.
6. Kaplan G. (2003). The role of MTB lipids in immunity and virulence. *Keystone Symposium — Tuberculosis: Integrating Host and Pathogen Biology.*
7. Gey van Pittius N.C., Gamieldien J., Hide W,. Brown G.D, Siezen R.J., and Beyers A.D. (2001). The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. *Genome Biol.* **2**(10), research 0044.1–0044.18.
8. Fleischmann R.D., Alland D., Eisen J.A., Carpenter L., White O., Peterson J., DeBoy R., Dodson, Gwinn M., Haft D., Hickey E., Kolonay J.F., Nelson W.C., Umayam L.A., Ermolaeva R.M., Salzberg S.L., Delcher A., Utterback T., Weidman J., Khouri H., Gill J., Mikula A., Bishai W., Jacobs W.R. Jr., Venter J.C. and Fraser C.M. (2002). Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**(19), 5479–5490.
9. Barnes P.F. and Cave M.D. (2003). Molecular epidemiology of tuberculosis. *N. Engl. J. Med.* **349**, 1149–1156.
10. Thierry D., Brisson-Noel A., Vincent-Levy-Frebault V., Nguyen S., Guesdon J.L. and Gicquel B. (1990). Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis. *J. Clin. Microbiol.* **28**(12), 2668–2673.
11. Hermans P.W., van Soolingen D., Bik E.M., de Haas P.E., Dale J.W. and van Embden J.D. (1991). Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect. Immun.* **59**(8), 2695–2705.
12. Groenen P.M., Bunschoten A.E., van Soolingen D. and van Embden J.D. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol. Microbiol.* **10**, 1057–1065.
13. Ross B.C., Raios K., Jackson K. and Dwyer B. (1992). Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. *J. Clin. Microbiol.* **30**, 942–946.
14. Supply P., Mazars E., Lesjean S., Vincent V., Gicquel B. and Locht C. (2000). Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol. Microbiol.* **36**, 762–771.
15. Kremer K., van Soolingen D., Frothingham R., Haas W.H., Hermans P.W., Martin C., Palittapongarnpim P., Plikaytis B.B., Riley L.W., Yakrus M.A., Musser J.M. and van Embden J.D. (1999). Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J. Clin. Microbiol.* **37**(8), 2607–2618.
16. Dale J.W. (1995). Mobile genetic elements in mycobacteria. *Eur. Respir. J.* Suppl. **20**, 633s–648s.
17. van Embden J.D., Cave M.D., Crawford J.T., Dale J.W., Eisenach K.D., Gicquel B., Hermans P., Martin C., McAdam R. and Shinnick T.M. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409.
18. Warren R., Richardson M., van der Spuy G., Victor T., Sampson S., Beyers N. and van Helden P. (1999). DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. *Electrophoresis* **2**0, 1807–1812.
19. Warren R.M., Sampson S.L., Richardson M., van der Spuy G.D., Lombard C.J., Victor T.C., and van Helden P.D. (2000). Mapping of IS6110 flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol. Microbiol.* **37**, 1405–1416.
20. van Helden P.D., Warren R.M., Victor T.C., van der Spuy G., Richardson M. and Hoal-van Helden E. (2002). Strain families of *Mycobacterium tuberculosis. Trends Microbiol.* **10**, 167–168.
21. Kamerbeek J., Schouls L., Kolk A., van Agterveld M., van Soolingen D., Kuijper S., Bunschoten A., Molhuizen H., Shaw R., Goyal M. and van Embden J. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914.
22. van Embden J.D., van Gorkom T., Kremer K., Jansen R., Der Zeijst B.A. and Schouls L.M. (2000). Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J. Bacteriol.* **182**(9), 2393–2401.
23. van Soolingen D., de Haas P.E., Haagsma J., Eger T., Hermans P.W., Ritacco V., Alito A. and van Embden J.D. (1994). Use of various genetic markers in differentiation of *Mycobacterium bovis* strains from animals and humans and for studying epidemiology of bovine tuberculosis. *J. Clin. Microbiol.* **32**(10), 2425–2433.
24. Fang Z., Morrison N., Watt B., Doig C. and Forbes K.J. (1998). IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J. Bacteriol.* **180**(8), 2102–2109.
25. Warren R.M., Streicher E.M., Sampson S.L., van der Spuy G.D., Richardson M., Nguyen D., Behr M.A., Victor T.C. and van Helden P.D. (2002). Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. *J. Clin. Microbiol.* **40**(12), 4457–4465.
26. Sampson S.L., Warren R.M., Richardson M., Victor T.C., Jordaan A.M., van der Spuy G.D. and van Helden P.D. (2003). IS6110-mediated deletion polymorphism in the direct repeat region of clinical isolates of *Mycobacterium tuberculosis. J. Bacteriol.* **185**(9), 2856–2866.
27. Sampson S.L., Richardson M., van Helden P.D. and Warren R.M. (in press). An IS6110-mediated deletion polymorphism in isogenic strains of *Mycobacterium tuberculosis. J. Clin. Microbiol.* **42**(2), 895–898.
28. Warren R.M., Streicher E.M., Charalambous S., Churchyard G., van der Spuy G.D., Grant A.D., van Helden P.D. and Victor T.C. (2002). Use of spoligotyping for accurate classification of recurrent tuberculosis. *J. Clin. Microbiol.* **40**(12), 3851–3853.
29. Cole S.T., Brosch R., Parkhill J., Garnier T., Churcher C., Harris D., Gordon S.V., Eiglmeier K., Gas S., Barry C.E. III, 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. and Barrell B.G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.
30. Warren R., Richardson M., Sampson S., Hauman J.H., Beyers N., Donald P.R. and van Helden P.D. (1996). Genotyping of *Mycobacterium tuberculosis* with additional markers enhances accuracy in epidemiological studies. *J. Clin. Microbiol.* **34**(9), 2219–2224.
31. Braden C.R., Templeton G.L., Cave M.D., Valway S., Onorato I.M., Castro K.G., Moers D., Yang Z., Stead W.W. and Bates J.H. (1997). Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. *J. Infect. Dis.* **175**, 1446–1452.
32. Yang Z.H., Bates J.H., Eisenach K.D. and Cave M.D. (2001). Secondary typing of *Mycobacterium tuberculosis* isolates with matching IS6110 fingerprints from different geographic regions of the United States. *J. Clin. Microbiol.* **39**, 1691–1695.
33. Mazars E., Lesjean S., Banuls A.L., Gilbert M., Vincent V., Gicquel B., Tibayrenc M., Locht C. and Supply P. (2001). High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc. Natl Acad. Sci. USA* **98**, 1901–1906.

34. Supply P., Lesjean S., Savine E., Kremer K., van Soolingen D. and Locht C. (2001). Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J. Clin. Microbiol.* **39**(10), 3563–3571.
35. Savine E., Warren R.M., van der Spuy G.D., Beyers N., van Helden P.D., Locht C. and Supply P. (2002). Stability of variable-number tandem repeats of mycobacterial interspersed repetitive units from 12 loci in serial isolates of *Mycobacterium tuberculosis. J. Clin. Microbiol.* **40**(12), 4561–4566.
36. Sreevatsan S., Pan X., Stockbauer K.E., Connell N.D., Kreiswirth B.N., Whittam T.S. and Musser J.M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc. Natl Acad. Sci. USA* **94**(18), 9869–9874.
37. Gutacker M.M., Smoot J.C., Migliaccio C.A., Ricklefs S.M., Hua S., Cousins D.V., Graviss E.A., Shashkina E., Kreiswirth B.N. and Musser J.M. 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162**, 1533–1543.
38. Alland D., Whittam T.S., Murray M.B., Cave M.D., Hazbon M.H., Dix K., Kokoris M., Duesterhoeft A., Eisen J.A., Fraser C.M. and Fleischmann R.D. (2003). Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. *J. Bacteriol.* **185**(11), 3392–3399.
39. Sampson S.L., Warren R.M., Richardson M., van der Spuy G.D. and van Helden P.D. (1999). Disruption of coding regions by IS6110 insertion in *Mycobacterium tuberculosis. Tuber. Lung Dis.* **79**, 349–359.
40. Sampson S.L., Lukey P., Warren R.M., van Helden P.D., Richardson M. and Everett M.J. (2001). Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. *Tuberculosis* (*Edinb.*) **81**, 305–317.
41. Beggs M.L., Eisenach K.D. and Cave M.D. (2000). Mapping of IS6110 insertion sites in two epidemic strains of *Mycobacterium tuberculosis. J. Clin. Microbiol.* **38**, 2923–2928.
42. Soto C.Y., Menendez M.C., Perez E., Samper S., Gomez A.B., Garcia M.J. and Martin C. (2004). IS6110 mediates increased transcription of the phoP virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *J. Clin. Microbiol.* **42**, 212–219.
43. Fang Z., Doig C., Kenna D.T., Smittipat N., Palittapongarnpim P., Watt B. and Forbes K.J. (1999). IS6110-mediated deletions of wild-type chromosomes of Mycobacterium tuberculosis. *J. Bacteriol.* **181**, 1014–1020.
44. Sampson S., Warren R., Richardson M., van der Spuy G. and van Helden P. (2001). IS6110 insertions in *Mycobacterium tuberculosis*: predominantly into coding regions. *J. Clin. Microbiol.* **39**, 3423–3424.
45. Brosch R., Philipp W.J., Stavropoulos E., Colston M.J., Cole S.T., and Gordon S.V. (1999). Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated M. tuberculosis H37Ra strain. *Infect. Immun.* **67**(11), 5768–5774.
46. Small P.M. (2003). Comparing genomes within the species *M. tuberculosis. Keystone Symposium — Tuberculosis: Integrating Host and Pathogen Biology.*
47. Gordon S.V., Eiglmeier K., Brosch R., Garnier T., Honore N., Barrell B. and Cole S.T. (1999). Genomics of *Mycobacterium tuberculosis* and *Mycobacterium leprae.* In *Mycobacteria: molecular biology and virulence*, eds C. Ratledge and J. Dale. Blackwell Science, Oxford.
48. Doran T.J., Hodgson A.L., Davies J.K. and Radford A.J. (1992). Characterisation of a novel repetitive DNA sequence from *Mycobacterium bovis. FEMS Microbiol. Lett.* **75**, 179–185.
49. Brennan M.J., Delogu G., Chen Y., Bardarov S., Kriakov J., Alavi M. and Jacobs W.R. Jr. (2001). Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect. Immun.* **69**(12), 7326–7333.
50. Cole S.T. (1999). Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett.* **452**, 7–10.
51. Flores J. and Espitia C. (2003). Differential expression of PE and PE_PGRS genes in *Mycobacterium tuberculosis* strains. *Gene* **318**, 75–81.
52. Ramakrishnan L., Federspiel N.A. and Falkow S. (2000). Granuloma-specific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. *Science* **288**, 1436–1439.
53. Rodriguez G.M., Voskuil M.I., Gold B., Schoolnik G.K. and Smith I. (2002). ideR, an essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect. Immun.* **70**(7), 3371–3381.
54. Betts J.C., Lukey P.T., Robb L.C., McAdam R.A. and Duncan K. (2002). Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.* **43**, 717–731.
55. Banu S., Honore N., Saint-Joanis B., Philpott D., Prevost M.C. and Cole S.T. (2002). Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Mol. Microbiol.* **44**, 9–19.
56. Cole S.T. and Barrell B.G. (1998). Analysis of the genome of Mycobacterium tuberculosis H37Rv. *Novartis Found. Symp.* **217**, 160–172.
57. Fraser C.M., Eisen J., Fleischmann R.D., Ketchum K.A. and Peterson S. (2000). Comparative genomics and understanding of microbial biology. *Emerg. Infect. Dis.* **6**, 505–512.
58. Musser J.M. (2001). Single nucleotide polymorphisms in *Mycobacterium tuberculosis* structural genes. *Emerg. Infect. Dis.* **7**, 486–488.
59. Hoal-van Helden, E.G., Stanton L.A., Warren R., Richardson M. and van Helden P.D. (2001). Diversity of in vitro cytokine responses by human macrophages to infection by *Mycobacterium tuberculosis* strains. *Cell Biol. Int.* **25**, 83–90.
60. Hoal-van Helden, E.G., Hon D., Lewis L.A., Beyers N. and van Helden P.D. (2001). Mycobacterial growth in human macrophages: variation according to donor, inoculum and bacterial strain. *Cell Biol. Int.* **25**, 71–81.
61. Manca C., Tsenova L., Barry C.E. III, Bergtold A., Freeman S., Haslett P.A., Musser J.M., Freedman V.H. and Kaplan G. (1999). *Mycobacterium tuberculosis* CDC1551 induces a more vigorous host response *in vivo* and *in vitro,* but is not more virulent than other clinical isolates. *J. Immunol.* **162**(11), 6740–6746.
62. Li Q., Whalen C.C., Albert J.M., Larkin R., Zukowski L., Cave M.D. and Silver R.F. (2002). Differences in rate and variability of intracellular growth of a panel of *Mycobacterium tuberculosis* clinical isolates within a human monocyte model. *Infect. Immun.* **70**(11), 6489–6493.
63. Zhang M., Gong J., Yang Z., Samten B., Cave M.D. and Barnes P.F. (1999). Enhanced capacity of a widespread strain of *Mycobacterium tuberculosis* to grow in human macrophages. *J. Infect. Dis.* **179**(5), 1213–1217.
64. van Crevel R., Nelwan R.H., de Lenne W., Veeraragu Y., van der Zanden A.G., Amin Z., van der Meer J.W. and van Soolingen D. (2001). *Mycobacterium tuberculosis* Beijing genotype strains associated with febrile response to treatment. *Emerg. Infect. Dis.* **7**, 880–883.
65. Lopez B., Aguilar D., Orozco H., Burger M., Espitia C., Ritacco V., Barrera L., Kremer K., Hernandez-Pando R., Huygen K. and van Soolingen D. (2003). A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin. Exp. Immunol.* **133**, 30–37.
66. Manabe Y.C., Dannenberg A.M. Jr, Tyagi S.K., Hatem C.L., Yoder M., Woolwine S.C., Zook B.C., Pitt M.L. and Bishai W.R. (2003). Different strains of *Mycobacterium tuberculosis* cause various spectrums of disease in the rabbit model of tuberculosis. *Infect. Immun.* **71**(10), 6004–6011.
67. Victor T.C., van Helden P.D. and Warren R. (2002). Prediction of drug resistance in *M. tuberculosis*: molecular mechanisms, tools, and applications. *IUBMB. Life* **53**, 231–237.
68. Behr M.A., Wilson M.A., Gill W.P., Salamon H., Schoolnik G.K., Rane S. and Small P.M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523.
69. Gordon S.V., Brosch R., Billault A., Garnier T., Eiglmeier K. and Cole S.T. (1999). Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.* **32**, 643–655.
70. Pym A.S., Brodin P., Brosch R., Huerre M. and Cole S.T. (2002). Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti. Mol. Microbiol.* **46**, 709–717.
71. Brosch R., Gordon S.V., Marmiesse M., Brodin P., Buchrieser C., Eiglmeier K, Garnier T., Gutierrez C., Hewinson G., Kremer K., Parsons L.M., Pym A.S., Samper S., van Soolingen D. and Cole S.T. (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl Acad. Sci. USA* **99**(6), 3684–3689.
72. Huard R.C., Oliveira Lazzarini L.C., Butler W.R., van Soolingen D. and Ho J.L. (2003). PCR-based method to differentiate the subspecies of the *Mycobacterium tuberculosis* complex on the basis of genomic deletions. *J. Clin. Microbiol.* **41**, 1637–1650.
73. Mahairas G.G., Sabo P.J., Hickey M.J., Singh D.C. and Stover C.K. (1996). Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis. J. Bacteriol.* **178**, 1274–1282.
74. Lewis K.N., Liao R., Guinn K.M., Hickey M.J., Smith S., Behr M.A. and Sherman D.R. (2003). Deletion of RD1 from *Mycobacterium tuberculosis* mimics bacille Calmette-Guerin attenuation. *J. Infect. Dis.* **187**, 117–123.
75. Hsu T., Hingley-Wilson S.M., Chen B., Chen M., Dai A.Z., Morin P.M., Marks C.B., Padiyar J., Goulding C., Gingery M., Eisenberg D., Russell R.G., Derrick S.C., Collins F.M., Morris S.L., King C.H. and Jacobs W.R. Jr. (2003). The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proc. Natl Acad. Sci. USA* **100**(21), 12420–12425.
76. Wards B.J., de Lisle G.W. and Collins D.M. (2000). An *esat6* knockout mutant of *Mycobacterium bovis* produced by homologous recombination will contribute to the development of a live tuberculosis vaccine. *Tuber. Lung Dis.* **80**, 185–189.
77. Hingley-Wilson S., Hsu T., Chen B., Dao D., Jacobs W.R. Jr and King C.H. (2003). The Identification and characterization of *Mycobacterium tuberculosis* Erdman mutants exhibiting reduced necrosis of lung epithelial cells. *Keystone Symposium —Tuberculosis: Intergrating Host and Pathogen Biology.*
78. Stanley S.A., Raghavan S., Hwang W.W. and Cox J.S. (2003). Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proc. Natl Acad. Sci. USA* **100**(22), 13001–13006.
79. Pym A.S., Brodin P., Majlessi L., Brosch R., Demangel C., Williams A., Griffiths K.E., Marchal G., Leclerc C. and Cole S.T. (2003). Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nature Med.* **9**, 533–539.
80. Guinn K.M., Hickey M.J., Mathur S.K., Zakel K.L., Grotzke J.E., Lewinsohn D.M., Smith S. and Sherman D.R. (2004). Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of *Mycobacterium tuberculosis. Mol. Microbiol.* **51**, 359–370
81. Brodin P., Eiglmeier K., Marmiesse M., Billault A., Garnier T., Niemann S., Cole S.T. and Brosch R. (2002). Bacterial artificial chromosome-based comparative genomic analysis identifies *Mycobacterium microti* as a natural ESAT-6 deletion mutant. *Infect. Immun.* **70**(10), 5568–5578.
82. Mostowy S., Cousins D. and Behr M.A. (2004). Genomic interrogation of the dassie bacillus reveals it as a unique RD1 mutant within the *Mycobacterium tuberculosis* complex. *J. Bacteriol.* **186**, 104–109.
83. Mostowy S., Cleto C., Sherman D., and Behr M.A. (2003). An avirulent transcriptome for members of the *Mycobacterium tuberculosis* complex? *Keystone Symposium — Tuberculosis: Integrating Host and Pathogen Biology.*