Published in final edited form as:
Vaccine. 2014 October 14; 32(45): 5998–6004. doi:10.1016/j.vaccine.2014.07.113.

Impact of in vitro evolution on antigenic diversity of *Mycobacterium bovis* bacillus Calmette-Guerin (BCG)

Richard Copin[#1], **Mireia Coscollá**[#2,3], **Efstratios Efstathiadis**[4], **Sebastien Gagneux**[#2,3], and **Joel D. Ernst**[#1,5]

[1]Department of Medicine, Division of Infectious Diseases, New York University School of Medicine, 522 First Avenue, Smilow 901, New York, New York, 10016, USA [2]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland [3]University of Basel, Petersplatz 1, Basel 4003, Switzerland [4]Center for Health Informatics and Bioinformatics, New York University Langone Medical Center, New York, NY, USA [5]Departments of Microbiology and Pathology, New York University School of Medicine, 522 First Avenue, Smilow 901, New York, New York, 10016 USA

[#] These authors contributed equally to this work.

Abstract

Mycobacterium bovis bacillus Calmette-Guerin (BCG), the only vaccine currently used against tuberculosis, is an attenuated derivative of *M. bovis* that has been propagated in vitro for more than 40 years. We have previously reported that the experimentally-verified human T cell epitopes of the *M. tuberculosis* complex (MTBC) are the most conserved elements of the genome; whether immune recognition is the force driving the conservation of epitopes in the MTBC is unknown. Therefore, we sequenced the genomes of 12 BCG strains to determine whether T cell epitopes were under selection pressure during BCG in vitro evolution. We constructed a genome-wide phylogeny and refined the previously-determined BCG phylogeny. Notably, we identified a new cluster between BCG Japan and BCG Russia, and repositioned the relationships of several strains within the lineage. We also compared the sequence diversity of 1,530 experimentally verified human T cell epitopes in the BCG vaccines with those in the MTBC. We found 23% of the known T cell epitopes are absent, and that the majority (82%) of the absent epitopes in BCG are contained in 6 proteins encoded in 2 regions of difference (RD) unique to BCG strains. We also found that T cell epitope sequences in BCG are more conserved than non-epitope sequences in the same gene. Finally, we find evidence that epitope sequence variation in BCG potentially affects human T cell recognition. These findings provide new insight into sequence variation in a slow-

Corresponding authors: Joel D. Ernst, Department of Medicine, Division of Infectious Diseases, New York University School of Medicine, 522 First Avenue, Smilow 901, New York, New York, 10016 USA, Phone: 212-263-9410; joel.ernst@med.nyu.edu, Sebastien Gagneux, Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute and University of Basel, Socinstr. 57, 4051 Basel, Switzerland, Phone: +41-61-284-8369; Sebastien.Gagneux@unibas.ch.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest
All authors declare no conflict of interest.

growing bacterium closely related to the MTBC that has been subjected to prolonged passage outside of a mammalian host, and indicate little difference in the extent of variation in vivo and in vitro.

Introduction

Tuberculosis (TB) remains a major global health challenge, causing an estimated 8.8 million new cases and 1.3 million deaths in 2012 [1]. In 1921, Albert Calmette and Camille Guérin developed the only vaccine currently used against TB. This vaccine, known as Bacillus Calmette-Guerin (BCG), is an attenuated derivative of *Mycobacterium bovis*, the causative agent of bovine TB. More than 120 million doses of BCG are administered annually in more than 100 countries [2]. Although BCG is effective in preventing disseminated TB in children [3, 4], the efficacy of BCG against pulmonary TB in adults varies from 0 to 80% [5]. Various hypotheses have been put forward to explain the low and variable efficacy of BCG. These include human genetic factors, variable exposure (and immune responses) to environmental mycobacteria, and differences in the circulating *M. tuberculosis* populations [6-8]. Mismatches between the antigenic composition of BCG and virulent *M. tuberculosis* may also contribute, but have not been systematically examined. Shortly after its original development, BCG was distributed to laboratories in multiple countries in Europe, Asia, and North- and South America for local preparation of vaccine. This process led to the diversification of BCG into distinct sub-strains. Lyophilization was introduced in 1961, which allowed long-term storage of seed stocks, with each of the BCG sub-strains named after the country, city, or laboratory where it was propagated. By that time, BCG Pasteur for example, had been subjected to 1,173 passages [9]. Although it is likely that most laboratories established seed stocks shortly after lyophilization became available, detailed information is not readily available.

Several studies have compared the genomic diversity of BCG sub-strains and identified several deletions, known as regions of difference (RD), as well as tandem duplications (DU) and single nucleotide polymorphisms (SNPs)[10, 11]. This diversity has been used to construct phylogenetic trees for BCG and infer the evolutionary history of individual BCG sub-strains [12]. Whether the diversity between individual BCG strains contributes to the variable outcomes of BCG vaccine trials is unknown, but evidence from studies in humans (reviewed in [6]) and experimental animals [13] have shown that BCG strains differ in their ability to induce specific cellular immune responses.

Immunity to TB depends on T lymphocytes. Infection with *M. tuberculosis* induces antigen-specific T cell responses in humans [14] and mice [15]. Moreover, T cell-deficient humans, nonhuman primates, and mice are susceptible to rapidly-progressive disease [16-18]. Immunization with BCG induces antigen-specific T cell responses in humans [7], although a clear correlate of vaccine-induced immunity to human TB has not yet been identified. Activation of conventional T cells after infection or vaccination depends on recognition of specific peptides bound to MHC (HLA in human) molecules on antigen presenting cells. For a given peptide to be immunogenic, it must bind with sufficient affinity to one or more MHC/HLA proteins, of which there are numerous allelic variants. The peptide/MHC complex must then be recognized by clonotypic T cell antigen receptors in the repertoire of the host. Variation in the sequence of a peptide epitope can result in loss of recognition by T cells [19], indicating that a close match between the sequence of immunogenic peptides in a vaccine and in the natural pathogen may be necessary to induce optimal immunity.

We previously reported that 491 experimentally-verified human T cell epitopes of the *M. tuberculosis* complex (MTBC) were highly conserved [20], with 95% of the epitopes examined harboring no amino acid sequence variants in 21 genetically-diverse *M. tuberculosis* clinical strains. These findings implied that for the majority of these immune targets, antigenic variation does not contribute to immune evasion. If epitopes in *M. tuberculosis* are conserved by selection pressure exerted by human T cell recognition, we hypothesized that 40 years of in vitro evolution between the first derivation of BCG and its lyophilization would lead to accumulation of diversity. The goals of the present study were 1) to determine the sequence variation in a close relative of *M. tuberculosis* (i.e. BCG) following prolonged in vitro passage in the absence of immune selection; and 2) to define if MTBC epitopes in BCG are more diverse than in clinical MTBC isolates because of the absence of immune pressure.

Results

Genome sequences and comparative genomics

To study the genetic content, variation, and phylogenetic relationship of contemporary BCG variants, we sequenced the genomes of 12 *M. bovis* BCG strains, including Australia, Connaught, Copenhagen, Denmark, Glaxo, Japan, Mexico, Pasteur, Phipps, Prague, Tice, and Russia. On average, these strains differed from one another by 28 SNPs, with a minimum of 9 SNPs between BCG Denmark and Glaxo, and a maximum of 50 SNPs between BCG Russia and Mexico (Table S1). The majority (119 of 144, 84%) of the SNPs in the BCG strains were present in coding regions. This proportion was not significantly different from that observed between BCG strains and the inferred common ancestor of the MTBC (88%; 1,149/1,316) [20]. We used the 144 SNPs to generate a genome-based phylogeny of BCG strains (Figure 1A). This phylogeny was congruent with previously reported genealogies using RDs and SNPs [11], except that BCG Tice was genetically closer to BCG Pasteur (11 SNPs difference), and BCG Connaught was more distantly related (28 SNPs difference) to BCG Pasteur, as previously reported [21]. Additionally, the whole genome phylogeny allowed us to further resolve relations among samples. We found that BCG Russia and Japan were monophyletic, indicating that these strains originated from a common progenitor. Also in the past, it has been unclear whether or not BCG Copenhagen and BCG Denmark were the same strain [12]. Our findings show that these strains differ by 12 SNPs, with BCG Denmark being more closely related to BCG Glaxo.

To assess the impact of in vitro evolution on the substitution rate in BCG sequences, we used a Bayesian approach [22] and the historical dates of strain distribution [12] as calibration points to determine the date of origin of the most recent common ancestor (MRCA) of all BCG strains (Figure 1A); our estimate indicated an MRCA dated between 1915 and 1923. We then calculated an evolutionary rate of 1.5×10^{-7} substitutions/site/year (95% highest posterior density HPD [1.2×10^{-7}, 1.9×10^{-7}]) (Figure S1), which is similar to the mutation rate for *M. tuberculosis* in clinical settings [23], but higher than the long term substitution rate estimated for the MTBC [24].

To characterize the selection forces driving BCG sequence diversity, we determined the ratio of nonsynonymous to synonymous single nucleotide changes (dN/dS) of each BCG strain with respect to the inferred MTBC ancestor. Overall, the genome-wide dN/dS ranged from 0.51 to 0.57, indicating that BCG sequences are generally under purifying selection. Similar to what we reported for MTBC [20], we found that the dN/dS for BCG genes that correspond to the essential genes of *M. tuberculosis* was significantly lower than for the corresponding non-essential genes (Figure 2A, 0.42 vs 0.57; Mann-Whitney test p<0.00001). When we compared dN/dS in BCG strains with those reported for MTBC [20], we obtained a similar dN/dS range for essential (0.41-0.53 in BCG compared to 0.45-0.67 in MTBC) and non-essential genes (0.53-0.58 in BCG compared to 0.56-0.78 in MTBC). To evaluate the selection pressures acting during the in vitro evolution of the different BCG sub-strains (that is, after derivation from the parental *M. bovis*), we determined the dN/dS between any two pairs of BCG genomes. Median pairwise dN/dS among BCG strains was 0.37.

Comparison of M. tuberculosis antigen and epitope content and sequence variation in BCG strains

Experimentally-verified human T cell epitopes were filtered to generate a list of 1,530 nonredundant peptide T cell epitopes. For each peptide, we established the number of epitope sequences affected by amino acid substitutions or complete or partial deletions in the sequenced BCG strains. On average, 23% (358/1,530) of the human T cell epitopes identified in *M. tuberculosis* were deleted from the BCG strains; all of the deleted epitopes are encoded in previously reported RDs. Notably, RD1, RD2 and RD14, which are deletions in BCG strains relative to the remainder of the MTBC, account for 87% of the deleted T cell epitopes (Table 1). The remaining deleted epitopes were associated with other RDs also identified in the parental *M. bovis*. We thus classified BCG strains into three groups based on the deletions defined by RD1, RD2 and RD14, and calculated the number of intact T cell epitopes present in each group (Figure 1B). BCG Russia and Japan constituted Group 1 with the fewest T cell epitopes deleted (320/1,530; 20.9%). All other BCG strains with the exception of BCG Pasteur were classified into Group 2 with an intermediate number of absent T cell epitopes (376/1,530; 24.6%). BCG Pasteur had the highest number of deleted epitopes and is the sole member of Group 3 (380/1,530; 24.8%).

To evaluate the impact of duplication events on T cell epitope content of BCG strains, we initially inspected the available genome data for the published BCG strains for the presence and number of copies of each *M. tuberculosis* T cell epitope. We found 15 duplicated epitopes in the DU2 of BCG Pasteur, and 5 in the DU2 of BCG Prague and Danish. To extend this analysis on the unpublished BCG genomes included in this work, we used the CNV-seq computational method to detect copy number variation of epitope-encoding genes [25]. Amongst the 12 BCG genomes, a total of 77 (0.5%) T cell epitopes encoded in 37 proteins were found in more than 1 copy when compared to *M. tuberculosis* H37Rv, indicating that duplication events have a minor impact on the antigen content and diversity of BCG strains (Table S2).

Although more than 95% of the variation affecting *M. tuberculosis* T cell epitopes in BCG strains was due to deletion of genes, we also identified 9 BCG proteins containing *M. tuberculosis* T cell epitopes affected by amino acid substitutions (Table 2). Notably, two antigens used in new TB vaccines, *rv0934* (encoding PstS1) and *rv1886c* (FbpB/Ag85B), had one amino acid substitution each in all BCG strains, which affected 2 and 5 epitopes, respectively (Table 3). Using algorithms that predict peptide-HLA binding (see Methods) we observed that in PstS1, substitution of a valine at position 352 for an alanine in all BCG strains diminished the epitope binding affinity to all HLA class II alleles considered (Table 3). We also evaluated the predicted impact of SNPs on protein structure as this may relate to protein function and/or stability. We found that the amino acid substitution identified in FbpB/Ag85B common to all BCG strains was predicted to affect the protein's structure and/or stability (Table S3). FbpB/Ag85B is one of the most antigenic proteins of *M. tuberculosis*, harboring 139 experimentally confirmed human T cell epitopes indexed in the Immune Epitope Database. The mutation we identified replaces phenylalanine at position 140 with a leucine located 20 amino acids from the active site (Ser 166). This may impact the catalytic activity of FbpB/Ag85B as a mycolyl transferase and/or decrease the stability and lifespan of the protein.

Characterization of the selective forces acting on T cell epitope sequences in BCG strains

To determine whether T cell epitope sequences in BCG strains are as conserved as previously described for the MTBC [20], we compared the number of SNPs in the T cell epitope-encoding sequences found in BCG (using the MTBC MRCA as reference) and in sequences selected randomly in the remaining BCG genome (Figure 2B). T cell epitopes showed significantly fewer SNPs than random sequences (Paired t test, p=0.0012), indicating that human T cell epitopes were also conserved in the BCG genomes. To test if T cell epitopes were under purifying selection in BCG strains, we compared the dN/dS in the epitope and non-epitope regions of each antigen. Similar to MTBC [20], we found a significantly lower dN/dS in epitope regions (median 0.15) when compared to non-epitope regions (median 0.44; Mann-Whitney test, p<0.00001; Figure 2A). Notably, all SNPs found in epitope-encoding sequences were shared among all BCG strains and thus existed in the ancestor of BCG or emerged *de novo* during in vitro passage before the diversification of the individual sub-strains in different laboratories. However no additional SNP emerged in these epitope regions despite 40 years of in vitro diversification, which corresponds to 1172 passages for BCG Pasteur. To explore if the absence of an accumulation of SNPs in T cell epitope sequences during in vitro diversification might reflect the action of a particular selective pressure, we used the estimated in vitro substitution rate of BCG to predict the number of SNPs expected in the epitope regions of the 12 BCG genomes, assuming epitopes accumulate mutations at the same rate as the remainder of the genome. We estimated that between 0.12 and 0.23 SNP per genome would accumulate during 40 years (corresponding to up to 15,000 generations, assuming a generation time of 14-24 hours [27]. When combining the 12 BCG genomes, this corresponds to a maximum of 1 to 3 new SNPs in T cell epitopes. As this value is so low, these estimates suggest that even though immune selection against new mutations in epitope regions was absent during in vitro evolution of BCG, four decades of in vitro passage might be insufficient for new variation to emerge among T cell epitope sequences.

Discussion

In this study of 12 BCG genome sequences, we found that genes corresponding to the essential genes in *M. tuberculosis* were more conserved than the corresponding non-essential genes, and that T cell epitope regions in BCG strains were highly conserved, as in the rest of the MTBC [20]. The low dN/dS that we determined for T cell epitope regions in BCG genomes indicates that absence of immune recognition during in vitro passage did not affect the conservation of T cell epitope-encoding sequences in BCG. Although it is possible that structural and functional constraints that are independent of T cell recognition contribute to conservation of T cell epitope sequences as previously shown for *M canettii* [26], we determined that 40 years of in vitro evolution is insufficient to observe diversification of T cell epitope sequences.

Our comparative analysis of the genome sequences of BCG strains was undertaken to understand the impact of in vitro evolution on the genetic diversity of *M. tuberculosis* T cell epitopes. Our previous results showed that in the MTBC, T cell epitope sequences were more conserved than the rest of the genome, indicating purifying selection [20]. These findings suggested that T lymphocyte recognition may be an important factor in sequence conservation of these loci. Because BCG strains evolved for decades in the absence of T cell selection pressure, the characterization of T cell epitope diversity in their genomes is a unique opportunity to test this hypothesis. Moreover, our finding that the BCG evolutionary rate is comparable with that found for *M. tuberculosis* in clinical settings [23], facilitates the comparison of diversity and evolutionary pressures acting on BCG compared to the MTBC.

Whereas most of the T cell epitopes analyzed here were highly conserved, fifteen epitopes in nine antigens harbored amino acid changes with respect to the MRCA of the MTBC. Our results also showed that sequence diversity in epitopes can affect human T cell recognition. While the mutation we identified in PstS1 was predicted to decrease peptide binding to diverse HLA alleles without a functional impact on the protein's structure or function, the amino acid substitution in FbpB/Ag85B was predicted to impact the structural integrity and/or stability of the protein. This is consistent with results of a study comparing the culture supernatant proteome of *M. tuberculosis* and *M. bovis* BCG Copenhagen that revealed that the mature form of FbpB is absent from the supernatant of the vaccine strain [28], and is consistent with results of studies indicating increased immunogenicity when BCG is engineered to overexpress FbpB/Ag85B from *M. tuberculosis* [29].

Although SNPs are a minor cause of antigen variation in BCG strains, gene deletion has a major impact on the T cell epitope composition of BCG. Our analysis revealed that compared with the *M. tuberculosis* H37Rv reference sequence, ~120 genes were lost in BCG strains including 33 genes that encode as many as 380 *M. tuberculosis* T cell epitopes. These results confirm and extend the findings of Zhang et al. using a smaller dataset corresponding to one third of the currently known *M. tuberculosis* T cell epitopes [30]. In contrast to the findings of that study, we found that BCG Pasteur, and not BCG Phipps, has the largest number of T cell epitope sequences depleted from its genome. The vast majority of those absent sequences are encoded in only 6 proteins contained in RD1 and RD2. RD2 delimits the border between the early and the late BCG vaccines. The unique presence of the 60 RD2-encoded epitopes in BCG Russia and Japan led the same authors to propose that these strains represent superior candidates for development of a new vaccine against TB [30]. However, the results of clinical trials do not provide evidence for a correlation between the strain of vaccine used and the efficacy in preventing tuberculosis [7] [31] [32] [33]. Moreover, most of the epitopes present in RD2 are encoded in two proteins, Rv1985c (23 epitopes) and Mpt64 (24 epitopes) and it is thus likely that the large number of known epitopes in RD2 is due at least in part to the intensive epitope discovery effort targeting these two antigens.

Our findings suggest that the decreased epitope content of the BCG vaccine strains compared with *M. tuberculosis* may be one factor that contributes to the low efficacy of BCG vaccination in preventing pulmonary tuberculosis. Our findings also indicate that the remaining *M. tuberculosis* T cell epitopes are highly conserved between BCG strains; this is most likely due to a mutation rate that is insufficient to generate sequence diversity in these regions during in vitro passage.

Methods

Complete methods and any associated references are available as supplemental information.

BCG Sequencing and assembly

Detailed information about the BCG strains sequenced is listed in Table S4. Bacterial strains were cultured from single colonies. Genomic DNA was extracted as described in [20] and sequenced with the lllumina Genome Analyzer of GATC-Biotech.

Phylogenetic and evolutionary analyzes

Phylogenetic analysis was based on 144 high-confidence variable positions by specifying *M. bovis* as the outgroup. Maximum likelihood phylogenies were obtained using PhyML [34], and HKY model. Evolutionary rates were determined using BEAST 1.7.5 [35]. CODEML from PAML 3.14b [36] was used to estimate dN/dS rates.

T cell epitope variability in BCG

Human *M tuberculosis* T cell epitopes were retrieved from the Immune Epitope DataBase (http://www.iedb.org/). An epitope was considered absent from a BCG genome if its sequence identity was lower than 20% when compared to its orthologous sequence in *M tuberculosis*. For each polymorphic position in the epitopes, an independent PCR product was sequenced.

Copy Number variations involving epitopes

CNV-seq was used to detect copy number variation from data generated via next-generation sequencing using BCG Pasteur as reference [25]. Data were retrieved using R. A p value

$<10^{-4}$ was used to proclaim a gene as duplicated. The analysis focused exclusively on T cell encoding genes and other potential duplicated genes were not considered in the present work.

Epitope and amino acid tolerance prediction

NetMHCpan 2.3 and NetMHCIIpan 2.2 were used as described in [37] for predicting impact of mutations observed in T cell epitopes in BCG protein sequences on HLA class I and class II binding-affinity, respectively [38, 39]. Sift (http://sift.jcvi.org/), Polyphen-2 (http://genetics.bwh.harvard.edu/pph2/) and Provean (http://provean.jcvi.org/index.php) were used to predict the impact of amino acid changes on protein function.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Eleanor Click (Division of Tuberculosis Elimination, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, U.S. Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America) for helpful suggestions at the initial stage of this study. Supported by R01 AI090928 and HHSN266200700022C. This work was also supported by the Swiss National Science Foundation grant PP00P3_150750.

References

1. WHO. GLOBAL TUBERCULOSIS CONTROL. World Health Organization; Geneva: 2013.
2. WHO Expert Committee on Biological Standardization. World Health Organization technical report series. 2013; (979):1–366. back cover.
3. Trunz BB, Fine P, Dye C. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. Lancet. Apr 8; 2006 367(9517):1173–80. [PubMed: 16616560]
4. Colditz GA, Berkey CS, Mosteller F, Brewer TF, Wilson ME, Burdick E, et al. The efficacy of bacillus Calmette-Guerin vaccination of newborns and infants in the prevention of tuberculosis: meta-analyses of the published literature. Pediatrics. Jul; 1995 96(1 Pt 1):29–35. [PubMed: 7596718]
5. Fine PE. Variation in protection by BCG: implications of and for heterologous immunity. Lancet. Nov 18; 1995 346(8986):1339–45. [PubMed: 7475776]
6. Ritz N, Hanekom WA, Robins-Browne R, Britton WJ, Curtis N. Influence of BCG vaccine strain on the immune response and protection against tuberculosis. Fems Microbiol Rev. Aug; 2008 32(5):821–41. [PubMed: 18616602]
7. Davids V, Hanekom WA, Mansoor N, Gamieldien H, Gelderbloem SJ, Hawkridge A, et al. The effect of bacille Calmette-Guerin vaccine strain and route of administration on induced immune responses in vaccinated infants. J Infect Dis. Feb 15; 2006 193(4):531–6. [PubMed: 16425132]
8. Evans TG, Brennan MJ, Barker L, Thole J. Preventive vaccines for tuberculosis. Vaccine. Apr 18; 2013 31(Suppl 2):B223–6. [PubMed: 23598486]
9. Gheorghiu M, Lagrange PH. Viability, heat stability and immunogenicity of four BCG vaccines prepared from four different BCG strains. Annales d'immunologie. Jan-Feb;1983 134C(1):125–47.
10. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science. May 28; 1999 284(5419):1520–3. [PubMed: 10348738]

11. Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, Valenti P, et al. Genome plasticity of BCG and impact on vaccine efficacy. Proc Natl Acad Sci USA. Mar 27; 2007 104(13):5596–601. [PubMed: 17372194]

12. Behr MA, Small PM. A historical and molecular phylogeny of BCG strains. Vaccine. Feb 26; 1999 17(7-8):915–22. [PubMed: 10067698]

13. Castillo-Rodal AI, Castanon-Arreola M, Hernandez-Pando R, Calva JJ, Sada-Diaz E, Lopez-Vidal Y. Mycobacterium bovis BCG substrains confer different levels of protection against Mycobacterium tuberculosis infection in a BALB/c model of progressive pulmonary tuberculosis. Infect Immun. Mar; 2006 74(3):1718–24. [PubMed: 16495544]

14. Havlir DV, Wallis RS, Boom WH, Daniel TM, Chervenak K, Ellner JJ. Human immune response to Mycobacterium tuberculosis antigens. Infect Immun. Feb; 1991 59(2):665–70. [PubMed: 1898911]

15. Sorensen AL, Nagai S, Houen G, Andersen P, Andersen AB. Purification and characterization of a low-molecular-mass T-cell antigen secreted by Mycobacterium tuberculosis. Infect Immun. May; 1995 63(5):1710–7. [PubMed: 7729876]

16. North R, Jung Y-J. Immunity to tuberculosis. Annu Rev Immunol. 2004; 22:599–1222. [PubMed: 15032590]

17. Lin P, Rodgers M, Smith Lk, Bigbee M, Myers A, Bigbee C, et al. Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model. Infect Immun. 2009; 77:4631–73. [PubMed: 19620341]

18. Kwan C, Ernst J. HIV and tuberculosis: a deadly human syndemic. Clin Microbiol Rev. 2011; 24:351–427. [PubMed: 21482729]

19. da Silva J. The evolutionary adaptation of HIV-1 to specific immunity. Curr HIV Res. Jul; 2003 1(3):363–71. [PubMed: 15046259]

20. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. Jun; 2010 42(6): 498–503. [PubMed: 20495566]

21. Garcia Pelayo MC, Uplekar S, Keniry A, Mendoza Lopez P, Garnier T, Nunez Garcia J, et al. A comprehensive survey of single nucleotide polymorphisms (SNPs) across Mycobacterium bovis strains and M. bovis BCG vaccine strains refines the genealogy and defines a minimal set of SNPs that separate virulent M. bovis strains and M. bovis BCG strains. Infect Immun. May; 2009 77(5): 2230–8. [PubMed: 19289514]

22. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLOS Biol. May.2006 4(5):e88. [PubMed: 16683862]

23. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis. Feb 27.2013 13(1):110. [PubMed: 23446317]

24. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. Oct; 2013 45(10):1176–82. [PubMed: 23995134]

25. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC bioinformatics. 2009; 10:80. [PubMed: 19267900]

26. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat Genet. Feb; 2013 45(2):172–9. [PubMed: 23291586]

27. James BW, Williams A, Marsh PD. The physiology and pathogenicity of Mycobacterium tuberculosis grown under controlled conditions in a defined medium. J Appl Micro. Apr; 2000 88(4):669–77.

28. Mattow J, Schaible UE, Schmidt F, Hagens K, Siejak F, Brestrich G, et al. Comparative proteome analysis of culture supernatant proteins from virulent Mycobacterium tuberculosis H37Rv and attenuated M. bovis BCG Copenhagen. Electrophoresis. Oct; 2003 24(19-20):3405–20. [PubMed: 14595687]

29. Horwitz MA, Harth G, Dillon BJ, Maslesa-Galic S. Recombinant bacillus calmette-guerin (BCG) vaccines expressing the Mycobacterium tuberculosis 30-kDa major secretory protein induce

greater protective immunity against tuberculosis than conventional BCG vaccines in a highly susceptible animal model. Proc Natl Acad Sci USA. Dec 5; 2000 97(25):13853–8. [PubMed: 11095745]

30. Zhang W, Zhang Y, Zheng H, Pan Y, Liu H, Du P, et al. Genome sequencing and analysis of BCG vaccine strains. PloS one. 2013; 8(8):e71243. [PubMed: 23977002]

31. Wu B, Huang C, Garcia L, Ponce de Leon A, Osornio JS, Bobadilla-del-Valle M, et al. Unique gene expression profiles in infants vaccinated with different strains of Mycobacterium bovis bacille Calmette-Guerin. Infect Immun. Jul; 2007 75(7):3658–64. [PubMed: 17502394]

32. Gorak-Stolinska P, Weir RE, Floyd S, Lalor MK, Stenson S, Branson K, et al. Immunogenicity of Danish-SSI 1331 BCG vaccine in the UK: comparison with Glaxo-Evans 1077 BCG vaccine. Vaccine. Jul 17; 2006 24(29-30):5726–33. [PubMed: 16723176]

33. Vijayalakshmi V, Murthy KJ, Kumar S, Kiran AL. Comparison of the immune responses in children vaccinated with three strains of BCG vaccine. Indian Pediatr. Sep; 1995 32(9):979–82. [PubMed: 8935260]

34. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. Jul 1; 2005 33(Web Server issue):W557–9. [PubMed: 15980534]

35. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. Aug; 2012 29(8):1969–73. [PubMed: 22367748]

36. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS. Oct; 1997 13(5):555–6. [PubMed: 9367129]

37. Copin R, Coscolla M, Seiffert SN, Bothamley G, Sutherland J, Mbayo G, et al. Sequence Diversity in the pe_pgrs Genes of Mycobacterium tuberculosis Is Independent of Human T Cell Recognition. mBio. 2014; 5(1)

38. Hoof I, Peters B, Sidney J, Pedersen L, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics. 2009; 61:1–14. [PubMed: 19002680]

39. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus. NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. Immunome Res. 2010; 6:9. [PubMed: 21073747]

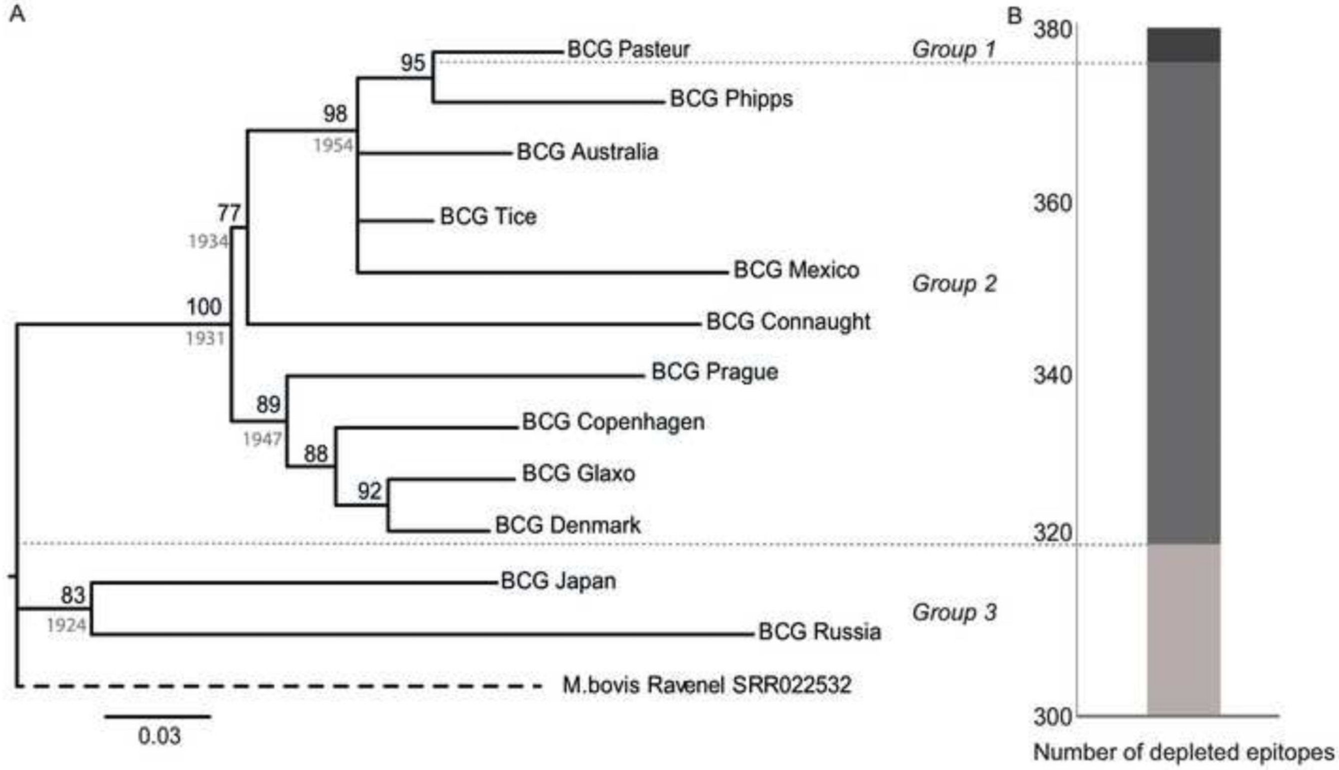

Figure 1.

A. Phylogenetic relationships among BCG strains. A. Maximum likelihood phylogeny based on 144 variable common nucleotide positions across 12 *Mycobacterium bovis* BCG genome sequences. The tree is rooted with *M. bovis* Ravenel SRR022532. Node support after 1000 bootstrap replications is indicated in black. Prior dates used to calibrate the phylogeny for Bayesian coalescent analysis are indicated in grey. B. The stacked bars show the proportion of missing *M tuberculosis* T cell epitopes in the 3 BCG groups identified in panel A and delimited by the dashed lines. Group 1: BCG Russia and Japan. Group 2: BCG Australia, Connaught, Copenhagen, Denmark, Glaxo, Mexico, Phipps, Prague and Tice. Group 3: BCG Pasteur

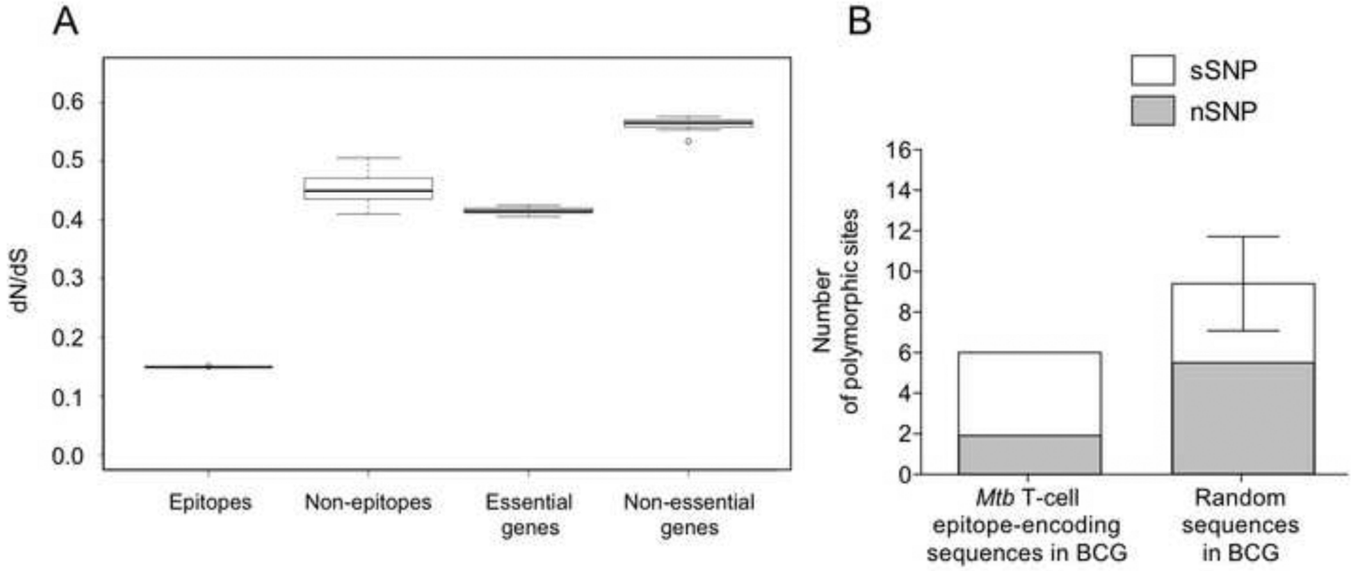

Figure 2.

Conservation of *M. tuberculosis* T cell epitopes in BCG strains. A. dN/dS in various gene classes of BCG strains. The calculation was done after comparing each of the 12 BCG genomes to the inferred most recent common ancestor of MTBC. This shows the dN/dS in 1) non-redundant epitope regions, 2) non-epitope regions of antigens, 3) *M. tuberculosis* essential genes, 4) *M. tuberculosis* non-essential genes. B. Comparison between the number of synonymous and non-synonymous SNPs found in *M. tuberculosis* non-overlapping T cell epitope regions in BCG and random sequences of same size selected from the rest of BCG genomes. The graph shows that T cell epitope regions in BCG are less affected by SNPs than expected by chance (95% confidence interval in the random replicates is indicated as error bars).

TABLE 1

Classification and characteristics of *M tuberculosis* T cell antigens absent from BCG strains.

H37Rv locus tag	Gene Name	Protein name	Protein family	Region of difference	Number of *Mtb* T cell epitopes depleted from BCG strains
Rv3875	*esxA*	6 kDa early secretory antigenic target	Cell wall and Cell processes	RD1	98
Rv3874	*esxB*	10 kDa culture filtrate antigen	Cell wall and Cell processes	RD1	86
Rv3873	*PPE68*	PPE protein 68	PE/PPE	RD1	65
Rv1980c	*mpt64*	immunogenic protein MPT64	Cell wall and Cell processes	RD2	24
Rv1985c	*rv1985c*	Probable transcriptional regulatory protein	Regulatory proteins	RD2	23
Rv3878	*espJ*	hypothetical protein	Cell wall and Cell processes	RD1	17
Rv2653c	*rv2653c*	phiRv2 prophage protein	Insertion seqs and phages	RD11	12
Rv2654c	*rv2654c*	phiRv2 prophage protein	Insertion seqs and phages	RD11	6
Rv1979c	*rv1979c*	Possible conserved permease	Cell wall and Cell processes	RD2	6
Rv1769	*rv1769*	hypothetical protein	Conserved hyptheticals	RD14	4
Rv1986	*rv1986*	hypothetical protein	Cell wall and Cell processes	RD2	4
Rv0222	*echA1*	enoyl-CoA hydratase	Lipid metabolism	RD10	3
Rv1977	*rv1977*	hypothetical protein	Conserved hyptheticals	RD7	3
Rv1582c	*phiRv1*	PhiRv1 phage protein	Insertion seqs and phages	RD3	2
Rv3876	*espI*	ESX-1 secretion-associated protein EspI	Cell wall and Cell processes	RD1	1
Rv3879c	*espK*	ESX-1 secretion-associated protein EspK	Cell wall and Cell processes	RD1	1
Rv2658c	*rv2658c*	prophage protein	Insertion seqs and phages	RD11	1
Rv1256c	*cyp130*	cytochrome P450 130	Intermediary metabolism and Respiration	RD13	1
Rv1965	*yrbE3B*	integral membrane protein YrbE3b	Virulence, Detoxification, Adaptation	RD7	1
Rv1973	*rv1973*	hypothetical protein	Cell wall and Cell processes	RD7	1
Rv1983	*PE-PGRS35*	PE-PGRS protein 35	PE/PPE	RD2	1
Rv1984c	*cfp21*	cutinase precursor CFP21	Cell wall and Cell processes	RD2	1
Rv1987	*rv1987*	Possible chitinase	Cell wall and Cell processes	RD2	1
Rv1513	*rv1513*	hypothetical protein	Conserved hyptheticals	RD4	1
Rv2347c	*esxP*	putative ESAT-6 like protein ESXP	Cell wall and Cell processes	RD5	1
Rv2350c	*plcB*	membrane-associated phospholipase C2	Intermediary metabolism and Respiration	RD5	1
Rv2351c	*plcA*	membrane-associated phospholipase C1	Intermediary metabolism and Respiration	RD5	7
Rv3427c	*rv3427c*	transposase	Insertion seqs and phages	RD6	1

H37Rv locus tag	Gene Name	Protein name	Protein family	Region of difference	Number of *Mtb* T cell epitopes depleted from BCG strains
Rv3428c	*rv3428c*	transposase	Insertion seqs and phages	RD6	1
Rv3617	*ephA*	epoxide hydrolase EphA	Virulence, Detoxification, Adaptation	RD8	2
Rv3620c	*esxW*	putative ESAT-6 like protein ESXW	Cell wall and Cell processes	RD8	2
Rv3621c	*PPE65*	PPE protein 65	PE/PPE	RD8	1
Rv2074	*rv2074*	hypothetical protein	Intermediary metabolism and Respiration	RD9	1

TABLE 2

Classification and characteristics of *M tuberculosis* proteins with epitope sequence variants in BCG strains.

H37Rv locus tag	Gene name	Protein name	Protein family	Number of altered *Mtb* T cell epitopes in BCG strains
Rv0589	*mce2A*	MCE-family protein 2A	Virulence, Detoxification, Adaptation	1
Rv0934	*pstS1*	PstS1	Cell wall and Cell processes	2
Rv1300	*hemK*	HemK	Intermediary metabolism and Respiration	1
Rv1733c	*rv1733c*	Conserved hypothetical protein	Cell wall and Cell processes	2
Rv1886c	*fbpB*	Fibronectin-binding protein B/antigen 85B	Lipid metabolism	5
Rv2628	*rv2628*	Conserved hypothetical protein	Conserved hyptheticals	1
Rv3497c	*mce4C*	MCE-family protein 4C	Virulence, Detoxification, Adaptation	1
Rv3616c	*espA*	ESX-1 secretion-associated protein EspA	Cell wall and Cell processes	1
Rv3823c	*mmpL8*	Integral membrane transport protein Mmpl8	Cell wall and Cell processes	1

TABLE 3

Predicted impact of amino acid substitutions present in T cell epitope-encoding sequences of BCG strains on HLA binding affinity and protein stability or function.

H37Rv locus tag	Gene name	Epitope consensus sequences	Epitope variant sequences	HLA binding prediction	Amino acid tolerance prediction		
					SIFT	Provean	Polyphen-2
Rv0934	*pstS1*	DQVHFQPLPPAVVKLSDALI*	DQAHFQPLPPAVVKLSDALI*	Reduction of the binding affinity to all tested HLA-DRB1 molecules	Intolerant	Neutral	Neutral
Rv1733c	*rv1733c*	TVSLLTIPFAAAAGTAVQDSRSHVYAHQAQ*	TVSLLTIPFAAAAGTAVHDSRSHVYAHQAQ*	No impact	Neutral	Neutral	Neutral
Rv1886c	*fbpB*	DWYSPACGKAGCQTYKWETFLTSELPQWLSANRAVKP*	DWYSPACGKAGCQTYKWETLLTSELPQWLSANRAVKP*	No impact	Intolerant	Intolerant	Intolerant
Rv2628	*rv2628*	KVQSATIYQVTDRSH	KVQSATIYQVTDRLH	No impact	Neutral	Neutral	Neutral
Rv3497c	*mce4C*	GKTYDAYFTDAGGITPG	GKPYDAYFTDAGGITPG	No impact	Neutral	Neutral	Neutral
Rv3616c	*espA*	IISDVADIIKGTLGE	IISDVADIIKGILGE	No impact	Neutral	Neutral	Neutral
Rv3823c	*mmpL8*	AITILLLVILLIIYG	AITILLLVILLIIYR	No impact	Neutral	Neutral	Neutral
Rv0589	*mce2A*	VAFRAGLVMEAGSKVT	VASRAGLVM EAGS KVT	No impact	Neutral	Neutral	Neutral
Rv1300	*hemK*	ELVRADVTTPRLLPE	ELVRADVTTPCLLPE	Reduction of the binding affinity to HLA-DRB1*03:01 molecule	Intolerant	Intolerant	Intolerant

*non-overlapping epitope sequence.