The highly conserved bcg vaccine genome biology essay

Science, Biology

1 National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control, Beijing 102206, China2 Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, Shanghai, China3 CAS Key Lab of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China4 Key Laboratory of Medical Molecular Virology Affiliated to the Ministries of Education and Health, Shanghai Medical College; Department of Microbiology, School of Life Sciences, Fudan University, Shanghai 200433, China5 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada6 Department of Microbiology and Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China. 7 Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China# These authors contributed equally to this work.* Corresponding Authors: Kanglin Wan (wankanglin@icdc. cn), Chen Chen (chenchen@icdc. cn), Guoping Zhao (gpzhao@sibs. ac. cn), or Baoli Zhu (zhubaoli@im. ac. cn). Competing interests: The authors declare that no competing interests exist. Running Title□Diminishing of T-cell epitopes in BCG genomesKey Words□Genome; BCG; Mycobacterium tuberculosis; Epitopes; RDs;

Abstract

Background: Even though the BCG vaccine against tuberculosis (TB) has been available for more than 75 years, a third of the world's population is

still infected with Mycobacterium tuberculosis (M. tuberculosis) and about 2 million people die of TB every year. In order to reduce the immense burden of TB, a clearer understanding of the functional genes underlying BCG's action and the development of new vaccines are urgently needed. Methods and Findings: Comparative genomic analysis among 18 M. tuberculosis complex strains in this study showed that BCG strains underwent repeated human manipulation were with the higher Region of Deletions (RD) rates than those of natural M. tuberculosis, and lost several essential components that can be recognized by the immune system, such as T-cell epitopes. 188 T-cell epitopes in BCG strains were lost in various degrees. BCG Tokyo, which is non-virulent and has only lost Group 2 epitopes, is the BCG strain with the largest number of T-cell epitopes (359). We propose that the variability in protection of BCG strains is due to their different epitopes. Restoration of identified lost epitopes in BCG-Tokyo is a useful strategy for future vaccine development. We also first presented BCG as a model organism for genetics research in this study. As a new model organism, BCG strains are with very well-documented history, and now with detailed genome information. By genome comparisons, the selection process of BCG strains under human manipulation (1908~1966) was clearly presented. Conclusions: Our results revealed the cause of protection variability of BCG vaccine on genome level, and also supported that restoration of identified lost epitopes in BCG-Tokyo is a useful strategy for future vaccine development. Furthermore, the detailed genome investigation for BCG vaccines in this study would be also helpful for its usage in microbial genetics, microbial engineering and other

research fields. Key Words Genome; BCG; vaccine; Mycobacterium tuberculosis; Epitopes; RDs; SNP

Link to data: http://www. mtbgenetyping. org/mtbDB/data/download/BCG. rar

Introduction

Mycobacterium tuberculosis (M. tuberculosis) is the world's leading cause of infectious disease, Tuberculosis (TB), with enormous global impact. The report from the World Health Organization (WHO) claimed that in 2008, an estimated 11. 1 million people were newly infected with M. tuberculosis. In China alone, there were 200, 614 deaths from TB in 2007. Bacillus Calmette-Guérin (BCG), the world's most widely used vaccine, directed against tuberculosis is derived from Mycobacterium bovis (M. bovis) and was attenuated after 230 passages over a period from 1908 to 1921. Since its attenuation, the original strain of BCG has produced lots of descendant strains, which have been distributed and used in many countries/regions around the world. These strains are named based on the country or the corresponding sites, such as BCG Tokyo, Pasteur, Russia, etc. Although these BCG descendant strains share a common ancestor, they have markedly different characteristics from each other since they have been propagated for probably more than 1000 passages in different countries. In 1966, the WHO made a recommendation that vaccines should not be prepared from cultures that had undergone more than 12 passages after culturing from a defined freeze-dried seed lot. Estimates of the protection against tuberculosis imparted by BCG strains varies from nil to 80%; the greatest

protection reported in the UK (~80%) by the Medical Research Council is strikingly different from trials by the US Public Health Service in Georgia, Alabama, and Puerto Rico, which all recorded protection of less than 30%. Such variability in protection was suggested to be attributable to factors such as genetic differences in the BCG strains used for immunization, environment influences and host genetic factors. In this context, efforts to re-engineer BCG vaccines with the ability to prevent latent TB reactivation, providing long lasting protection and being devoid of collateral effects in immunosuppressed people, are urgent. A key factor among the possible scenarios attributable to the variability in protection of the vaccination is the genetic differences among BCG strains. To obtain a more comprehensive understanding of the diversity of BCG strains and identify more candidate sites for vaccine development, we determined the genome sequence of six BCG strains in this study, and used the genome sequences of 18 M. tuberculosis complex (MTBC) strains to analyze their mutation sites (RDs and SNPs), with special reference to 483 experimentally-verified human T-cell epitopes.

Methods

Sequencing and Assembly of BCG Genomes

Whole genome sequences of M. bovis (AF2122/97), six strains of M. bovis BCG (BCG China, BCG Russia, BCG Tice and BCG Danish, Tokyo 172 and Pasteur) and five strains of M. tuberculosis (H37Rv, H37Ra, F11, KZN1435 and CDC1551) were downloaded from NCBI. Detailed information about these strains is listed in Table S2. All strains of M. bovis BCG used in this

study were provided by the American Type Culture Collection (USA). We sequenced six strains of BCG (BCG-Frappier, BCG-Glaxo, BCG-Moreau, BCG-Phipps, BCG-Pragure, and BCG-Sweden) with illumine genome analyzer. The genome coverage were > 100-fold. Genomic DNA was extracted from BCG colonies on L-J medium using CTAB and 2 µg DNA from each strain was used for sequencing. Sequencing reads from the six BCG strains were assembled into draft genomes using SOAPdenovo (BGI) (Table S2).

RD, Absence genes, Lost epitopes and SNP identification

The 3945 CDS sequences from M. bovis AF2122/97 were compared one by one with the other 17 Mycobacterium strains (Table S2) using BLAT for identifying PA (Presence/Absence) genes. Absence genes were defined by sequence alignment <60%. All identified Absence genes were also checked by aligning with the original sequencing reads using SOAP and some negative Absence genes caused by assembly errors were filtered out. To further filter out false-negative Absence genes in the draft genomes, only those Absence genes not located at the ends of a contig were considered to be Absence genes. Region of Deletions (RD) which cover one or more Absence genes were identified based on the location of these identified Absence genes. The epitope was classified as the lost epitoeps if it was located in a Absence gene or in a deletion region of non-Absence gene. In addition, only these epitopes without any blast match in the genome could be left as the lost epitopes in this strain. To identify SNPs, we first obtained the 3945 gene sequences for each of the 17 strains based on BLAT results and aligned them using ClustalW. Only SNP sites with a coverage of more

than 20 and without ambiguous sites (" N") in their flanking 10 bp regions were kept.

Phylogenetic analysis

Phylogenetic analysis was first performed using SNPs from the concatenated sequences of 17 housekeeping genes (Table S1). A Neighbor-joining tree (Figure 2a) was obtained using MEGA. A further topological structure tree (Figure 2b) was based on all Absence genes and was obtained using Cluster and Treeview.

Results

Hyperconserved BCG strains

We sequenced the genomes of six BCG strains (BCG-Frappier, BCG-Glaxo, BCG-Moreau, BCG-Phipps, BCG-Pragure, and BCG-Sweden) using Illumina sequencing (Figure 1) and compared them with the available sequences of six other BCG strains (BCG-China, BCG-Tice, BCG-Russia, BCG-Danish, BCG-Tokyo, and BCG-Pasteur) and five M. tuberculosis strains (F11, H37Ra, H37Rv, KZN1435 and CDC1511) obtained from the NCBI bacterial genome database. We first determined the evolutionary relationship between these MTBC strains by constructing an NJ phylogenetic tree based on 17 housekeeping genes (Figure 2a and Table S1). Like their common ancestor, M. bovis AF2122/97, and the 5 M. tuberculosis isolates, all of the BCG strains had similar genome sizes (~4. 2M) and GC contents (~0. 65) (Table S2). It was previously shown that the genomes of M. tuberculosis strains are conserved sharing 99. 9% identity at the nucleotide sequence level. Here, we identified a total of 2157 SNPs in the 12 BCG strains, and 1444 SNPs in the 5

M. tuberculosis strains. The average nucleotide diversity for pairs of any two BCG strains was only 0. 018 SNP/kb, significantly lower (P= 4. 43e-6 <0. 01, two-tailed t-test) than that for M. tuberculosis strains (0. 25 SNP/kb). These identified SNPs could be used as the new molecular mark for the BCG strain identification in future (Table S3).

Conserved, but plastic BCG genomes for its high RD rate

In this study, by aligning the 3945 genes from M. bovis AF2122/97 with the genome sequences of the 12 BCG strains, we identified 24 RDs (10 previously published and 14 new, Table S4) which cover one or more Absence genes in the genome sequences of these BCG strains. The topological structure tree based on RDs (Figure 2b) shows the relationship among the M. bovis BCG and M. tuberculosis strains more clearly than the NJ phylogenetic tree (Figure 2a). This, together with the well-documented history of BCG vaccines, enables us to accurately predict the time of the occurrence of most of these RDs (Figure 3). RD1, RD3 and Del_Mb2377c most likely occurred during the first period of attenuation (1908-1921) , while the remaining 21 RDs, which contain a total of 49 Absence genes, probably occurred during the following period of divergence (1921-1966). The presence of Absence genes in some BCG vaccines that are absent in others is likely the cause of the differences in phenotype among strains (Figure 3) and may help to explain the different levels of virulence remaining in different BCG strains or their immunological efficiency. RDs have also occurred in M. tuberculosis strains. 17 RDs covering 44 genes were found in the five strains of M. tuberculosis examined here by comparison to the M.

bovis genome (Figure 3), which are all potential molecular markers for the distinguish of Mycobacterium sp.. We compared the rate of RD occurrence in BCG and M. tuberculosis, and found that BCG has a markedly higher RD rate than that of M. tuberculosis. Differences in the historical development of BCG and M. tuberculosis strains may explain this phenomenon. Attenuation of BCG strains from M. bovis began in 1908 and was completed after 13 years (1921). From 1921-1966 BCG strains diverged due to separate culturing in different countries. Since 1966, BCG vaccines have not been prepared from cultures that have undergone more than 12 passages. In other words, since 1966, any new mutations occurring in BCG strains would not have been perpetuated. Thus, in theory, all mutations in BCG, including both SNPs and RDs, occurred during the two periods 1908~1921 and 1921~1966. The average RD number of each strain during these two periods is 3 and 2. 86, respectively. Thus the average RD rate of each BCG strain during these two periods is 0. 23/year/strain and 0. 07/year/strain, respectively (Figure 3). For M. tuberculosis was estimated to have occurred roughly 15, 000~20, 000 years ago and their average RD number is seven, the average RD rate for each M. tuberculosis strain would be 0. 00035/year/strain ~0. 00046/year/strain, less than that of BCG (0. 07/year/strain). Different with other clinical M. tuberculosis strains, H37Ra and H37Rv are both derived from their virulent parent strain H37 through a process of aging and dissociation from in vitro culture between 1905 and 1935. Genome comparison in this study represented that RD rate during this period for these two strains under human manipulation (0. 033 RD/year/strain) is clearly higher than that of other M. tuberculosis strains surviving in natural

environments (<0. 00046 RD/year/strain; Figure 3), but still lower than that of BCG strains (0. 07/year/strain). Thus, it is reasonable to speculate that BCG has a significantly higher RD rate than M. tuberculosis. These high RD rates suggest that the genome sequences of BCG strains are plastic.

Loss of T-cell epitopes in plastic BCG genome

Our research about 483 experimentally-verified human T-cell epitopes (Table S5) indicated that several T-cell epitopes have been lost in BCG strains, although all of them are existed in M. bovis and 5 strains of M. tuberculosis (Figure 4). Only 295 T-cell epitopes (Table S5& S6) are presented in all of the 12 BCG strains, which were classified as Group 1 Epitopes. Our results showed that epitope sequences in M. tuberculosis and BCG are both highly conserved. Of the 483 experimentally-verified human T-cell epitopes (Table S5) , only 8 SNPs were identified in the genomes of the five M. tuberculosis strains, consistent with the previously reported observation about the highly conserved epitopes in M. tuberculosis genomes. While the BCG T-cell epitopes are also conserved with no SNP ever identified in these 295 T-cell epitopes of the 12 BCG strains examined. Besides 295 T-cell epitopes in Group 1, the other 188 T-cell epitopes in BCG strains were lost in varying degrees. The first lost of epitopes in BCG was occurred during the attenuation period (1908-1921). Just as shown in Figure 4b, 124 T-cell epitopes classified as Group 2 were lost in all BCG strains. Most (117, 94. 4%) of Group 2 T-cell epitopes are located in RD1 which encodes several essential antigens (Table S7), such as immunogenic co-regulated secreted proteins (ESAT-6 and CFP-10). Between 1926 (the dissemination time of

BCG-Sweden) and 1934 (the dissemination time of BCG-Tice), the other 28 T-cell epitopes, which are all located in RD2 and classified as Group 3, were lost during the ongoing propagation of 8 BCG strains, while the others BCG-Moreau, BCG-Russia, BCG-Tokyo and BCG-Sweden still kept these Group 3 epitopes (Figure 4b). Thus, comparing to other BCG strains, these four strains own more antigens (Table S7) recognized by the immune system, such as MPT64. Each BCG also has its unique lost epitopes (Figure 4a and Figure 4b), which are classified as Group 4. Of the 12 BCG strains examined here, BCG-Tokyo was the strain with the highest number of epitopes (359) and might be the only strain with the same epitopes number of the original vaccine strain (Figure 4a).

Discussion

Effect of human manipulation on the evolution of BCG

While M. bovis BCG, derived from M. bovis, and M. tuberculosis originated from a common ancestor, they have existed in different environments and experienced different selection pressures since segregation. BCG strains have been grown under artificial culture in labs around the world and have always been subject to human manipulation, while M. tuberculosis strains, except for the laboratory strains H37Ra and H37Rv, must survive in natural environments and are subject to the selection pressure of the human immune system. Different mutation models have arisen in BCG and M. tuberculosis as a result of these different environments and selection pressures since their segregation. Although RDs have occurred over the course of evolution in both M. tuberculosis and BCG vaccines, they have

been especially frequent in BCG. Comparison of the RD rate in BCG and M. tuberculosis shows that RDs occur and are maintained more frequently in the plastic genome of BCG than that in M. tuberculosis (Figure 3). Because they did exist in different environments and experienced different selection pressures, we hypothesize that human-manipulated BCG vaccine strains are under greater selection pressure to tolerate more RDs than natural M. tuberculosis. In other words, manipulation of strains under laboratory conditions tends to result in the loss of sequences more readily than in strains that occur in natural hosts. The relative higher RD rate between laboratory strains H37Rv and H37Ra under human manipulation (0. 033 RD/year/strain) than that in other M. tuberculosis strains surviving in natural environments (<0. 00046 RD/year/strain; Figure 3) further supports our hypothesis that human manipulation of bacterial strains results in strong positive selection of RDs. These RDs identified in several BCG strains could also explain why there are differences in the levels of virulence remaining in BCG descendant strains and their immunological efficiency. It may be possible to take advantage of these differences and manually delete or insert particular regions in BCG to develop new BCG strains with better immunological efficiency. For example, the restoration of RD1 region into BCG has been proven to improve its vaccine efficacy.

BCG PA T-cell epitopes can be used to develop new vaccines

T-cell antigens consist of epitope regions of pathogens that interact with human T cells and are recognized by the immune system. Studies in pathogenic viruses, bacteria and protozoa have revealed that genes

encoding antigens and their T-cell epitope regions tend to be highly variable as a consequence of diversifying selection to evade host immunity. However, Comas et al. reported the sequencing of the genomes of 21 strains representing the global diversity and six major lineages of MTBC. Most of the 491 experimentally-confirmed human T-cell epitopes showed little sequence variation and had a lower ratio of nonsynonymous to synonymous changes than seen in essential and nonessential genes. This work confirmed that T-cell epitopes of MTBC are highly conserved and do not reflect any ongoing evolutionary arms race or immune evasion. The patterns observed might indicate that this highly successful pathogen has developed a distinct evolutionary strategy of immune subversion. Our results indicate that sequences of epitopes in BCG are conserved similar with those in M. tuberculosis. While we identified 8 SNPs among 483 of the 491 T-cell epitopes examined by Comas et al in 5 M. tuberculosis strains (sequence information was not provided for the remaining 8 epitopes), no SNPs were identified in these T-cell epitopes in the 12 BCG strains examined here. Although T-cell epitopes are essential for immune system response and are highly conserved in BCG strains, our results indicated that several epitopes had been lost in BCG strains. The 483 T-cell epitopes (Figure 4a&b) can be divided into four groups based on their distribution in the 18 MTBC strains. While 295 T-cell epitopes were present in all 18 MTBC strains (Group 1), the 188 epitopes of Groups 2, 3 and 4 were lost in some or all of the BCG strains (Figure 4b). This finding may provide insight into differences in the protective capacity of BCG strains and could be useful for the development of new TB vaccines, such as DNA vaccine, epitope vaccine or recombination vaccine. Of

the 188 epitopes lost in some or all BCG strains, 124 T-cell epitopes (Group 2) are absent in all BCG strains but have been maintained in the five strains of M. tuberculosis, and most (117, 94. 4%) are located in RD1 (Figure 4b). RD1 encodes a pair of highly immunogenic co-regulated secreted proteins (ESAT-6 and CFP-10) that contain T- and B-cell epitopes. In 2003, Pym found that the restoration of ESAT-6 in BCG improves its vaccine efficacy. Likewise, the 188 epitopes identified here could also be candidates for restoration into BCG to improve its vaccine efficacy. BCG strains have RDs that are strain-specific. The strains most commonly in use such as BCG Glaxo, Danish and Pasteur have the largest number of RDs. It has been proposed that one of the reasons behind the partial vaccine efficacy of BCG is that it has become too attenuated to successfully mimic natural MTB infection. Some empirical evidence favoring this hypothesis is provided by the finding that BCG-Japan induced greater cytotoxicity and T helper 1 responses in infants than BCG-Danish. BCG-Japan was also proven to induce higher frequencies of mycobacterial-specific polyfunctional and cytotoxic T cells and higher concentrations of Th1 cytokines than that of BCG-Russia. Our results showed that the 12 BCG strains have different numbers of T-cell epitopes. BCG Tokyo, which is non-virulent and has only lost Group 2 epitopes (Figure 4a), is the BCG strain with the largest number of T-cell epitopes that can be recognized by the immune system. It might be the only strain with the same epitopes number with the first BCG vaccine strain in 1921. We propose that BCG Tokyo is the best candidate strain for use in the development of a new and better vaccine.

Conclusion

In summary, our results showed that 188 T-cell epitopes essential for human immune system response had been lost in BCG strains in varying degrees. The higher RD rate in human-manipulated BCG strains suggests that these vaccine strains that had undergone human manipulation were under a dramatically different selection model from natural strains. Our results also suggest that BCG-Tokyo, the strain with the highest number of T cell epitopes, may be the best candidate strain for developing a better vaccine strain. Deletion or insertion of epitopes identified here that are present in M. tuberculosis but absent in some or all BCG strains, may be a useful strategy for vaccine development.

Data Access

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AKYQ00000000-AKYV00000000. Genome sequences of 6 BCG strains could be accessible (http://www. mtbgenetyping. org/mtbDB/data/download/BCG. rar) .

Acknowledgements

This study received financial support from the Transmission Mode of Tuberculosis project of the National Key Program of Mega Infectious Diseases (2013ZX10003006-002), CHINA-CANADA Joint Health Research Initiative Proposal (812111251) and the National Natural Science Funding of China (81201322)