

The whole genome sequence of Polish vaccine strain *Mycobacterium bovis* BCG Moreau

Katarzyna Krysztopa-Grzybowska,¹ Jakub Lach,² Maciej Polak,¹ Dominik Strapagiel,² Jaroslaw Dziadek,³ Marcin Olszewski,⁴ Aleksandra A. Zasada,¹ Ania Darlińska,¹ Anna Lutyńska,⁵ Ewa Augustynowicz-Kopeć⁶

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT Currently, tuberculosis immunoprophylaxis is based solely on *Bacillus Calmette-Guérin* (BCG) vaccination, and some of the new potential tuberculosis vaccines are based on the BCG genome. Therefore, it is reasonable to analyze the genomes of individual BCG substrains. The aim of this study was the genetic characterization of the BCG-Moreau Polish (PL) strain used for the production of the BCG vaccine in Poland since 1955. Sequencing of different BCG lots showed that the strain was stable over a period of 59 years. As a result of comparison, BCG-Moreau PL with BCG-Moreau Rio de Janeiro (RDJ) 143 single nucleotide polymorphisms (SNPs) and 32 insertion/deletion mutations (INDELs) were identified. However, the verification of these mutations showed that the most significant were accumulated in the BCG-Moreau RDJ genome. The mutations unique to the Polish strain genome are 1 SNP and 2 INDEL. The strategy of combining short-read sequencing with long-read sequencing is currently the most optimal approach for sequencing bacterial genomes. With this approach, the only available genomic sequence of BCG-Moreau PL was obtained. This sequence will primarily be a reference point in the genetic control of the stability of the vaccine strain in the future. The results enrich knowledge about the microevolution and attenuation of the BCG vaccine substrains.

IMPORTANCE The whole genome sequence obtained is the only genomic sequence of the strain that has been used for vaccine production in Poland since 1955. Sequencing of different BCG lots showed that the strain was stable over a period of 59 years. The comprehensive genomic analysis performed not only enriches knowledge about the microevolution and attenuation of the BCG vaccine substrains but also enables the utilization of identified markers as a reference point in the genetic control and identity tests of the stability of the vaccine strain in the future.

KEYWORDS BCG vaccine, BCG Moreau, BCG Moreau PL, tuberculosis, next-generation sequencing, complete genomic sequence

Tuberculosis (TB), formerly known as the white plague, is an infectious disease caused by the *Mycobacterium tuberculosis* species complex. Because of its infectious nature, chronic progression, involving complex defense mechanisms of the immune system, the need for long-term and exhaustive treatment, the emergence of multidrug resistant forms, and TB-HIV coinfections, TB is a constant burden and public health challenge (1). TB has been one of the main causes of morbidity and human mortality in the world for centuries. According to World Health Organization (WHO), 10 million people worldwide fell ill with TB in 2019, with 1.4 million deaths. The disease affects people of all ages, while the greatest burden is observed among adult men (56%) compared to the group of adult women (32%) and children (12%). Almost 70% of all global TB cases occur in eight countries: India (26%), Indonesia (8.5%), China (8.4%), Philippines (6%), Pakistan

Editor Olivier Neyrolles, CNRS-University of Toulouse, Toulouse, France

Address correspondence to Katarzyna Krysztopa-Grzybowska, kkrysztopa@pzh.gov.pl.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 21 December 2023

Accepted 24 April 2024

Published 17 May 2024

Copyright © 2024 Krysztopa-Grzybowska et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

(5.7%), Nigeria (4.4%), Bangladesh (3.6%), and South Africa (3.6%) (2). It is estimated that approximately 25% of the world's population are carriers of *M. tuberculosis*, and the active form of the disease will affect 5%–10% of this population (3, 4). Approximately 8.2% of all TB cases worldwide in 2019 were people with HIV. The significant exacerbation of the epidemiological situation of TB in the world is mainly due to multidrug- or rifampicin-resistant TB (multidrug or rifampicin-resistant [MDR/RR]-TB), which caused 3.3% of new cases and 18% of previously treated patients in 2019. India (27%), China (14%), and the Russian Federation (8%) have the highest number of cases of MDR/RR-TB (2).

Mycobacterium bovis Bacillus Calmette-Guérin (BCG) is the only vaccine currently available against TB. It is a live-attenuated vaccine that is one of the oldest and most widely used vaccines in the history of medicine. About 100 million newborns around the world receive it every year. The original BCG strain was obtained by attenuation of the bovine TB pathogen *M. bovis* using passages in the years 1908–1921 (5). Since 1924, the parent strain of *M. bovis* BCG has been distributed to laboratories around the world to start vaccine production. This strain has been passaged and despite attempts to standardize this process, different conditions in individual laboratories led to the isolation of about 50 genetically separate substrains of BCG, for example, BCG-Pasteur, BCG-Tokyo, BCG-Danish, and BCG-Moreau (6, 7). Stopping further differentiation of the BCG strain became possible thanks to the development of a system of lyophilized seed lots in the 1960s and then issuing detailed recommendations by the WHO regarding the permitted number of passages in the BCG vaccine preparation process (5). The effectiveness of the BCG prophylactic vaccination estimated in many clinical and epidemiological studies varies. The greatest effectiveness of BCG vaccination was observed in North America and Northern Europe, while the lowest effectiveness was noted in tropical regions. The observed differences, as the vaccine researchers argue, may reflect differences between populations in terms of: exposure to environmental nontuberculous mycobacteria (NTM), BCG substrains and immunization schedules used, the virulence of *M. tuberculosis* and other factors, such as chronic intestinal parasitic infections, different concentrations of vitamin D, and iron of vaccinated persons (8–10).

Currently, most of the world's population is supplied with BCG vaccines purchased by UNICEF on behalf of the Global Alliance for Vaccines and Immunization. UNICEF uses only four BCG vaccine suppliers that use three BCG substrains: BCG-Danish (Statens Serum Institute in Denmark), BCG-Russia (BB-NCIPD in Bulgaria and Serum Institute in India), and BCG-Tokyo (BCG Japan Laboratory). More than 90% of all BCG vaccines are produced globally from the BCG-Danish, BCG-Russia, BCG-Tokyo, BCG-Pasteur, and BCG-Moreau RDJ (11).

The BCG-Moreau strain is derived from the original BCG strain that was sent from the Pasteur Institute in Paris to Brazil by Dr. Julio Elvie Moreau on May 22, 1925. In 1951, Prof. Arlindo de Assis, Director of the BCG Center in Brazil, submitted the BCG-Moreau strain back to Paris for a comprehensive examination. The strain was donated to the Serum and Vaccines Factory in Lublin in 1954 by Dr. Van Deirse, the head of the BCG production laboratory at the Pasteur Institute. A year later, the production of the native BCG vaccine from the BCG-Moreau substrain began in Poland. From then on, the percentage of newborns vaccinated intradermally in Poland has been steadily increasing (12). Currently, more than 90% of newborns born in Poland are vaccinated with BCG each year (13).

The *M. bovis* BCG Moreau strain (for the purposes of the study, this strain will be marked PL) has not been genetically characterized so far, although it has been used for TB vaccine production in Poland for over 65 years. This study analyzes the whole genome sequence of *M. bovis* BCG Moreau PL and compares it with the genome of the closest BCG substrain (*M. bovis* BCG Moreau RDJ) and the most distant substrain (*M. bovis* BCG Pasteur). Additionally, five lots of the vaccine strain from 1957 to 2015 were compared to track the acquisition of mutations over time and to assess the genetic stability of the Polish strain.

analysis of the *M. bovis* BCG Moreau PL and *M. bovis* BCG Moreau RDJ (AM412059.2) genomes, the presence of 175 mutations, including 143 single nucleotide polymorphism (SNP) mutations and 32 INDEL mutations was found in the Polish strain (Table S1a). Among the identified SNP mutations, 77 were located in the coding regions and caused changes in the amino acid sequences of the proteins, while three could have affected the encoded proteins due to their location in the probable promoter regions. Among non-synonymous point mutations, mutations in genes encoding Pro-Glu (PE) proteins accounted for more than half (45/77), of which 29 concern the *PE_PGRS43b* gene. Another group consisted of mutations in proteins involved in intermediary metabolism and respiration (12/77), of which seven were located in the gene encoding biotin sulfoxide reductase (*bisC*). Eight mutations occurred in genes of proteins with unidentified functionality (hypothetical proteins), three in genes of proteins involved in lipid metabolism, and three in the gene encoding the transposase. The fewest non-synonymous SNPs were identified in the genes of regulatory proteins (2/77) and genes of proteins associated with the bacterial cell wall and cell processes (2/77). No single mutation of this type was identified among the genes associated with virulence, detoxification, and adaptation. The predominant number of SNP mutations in PE protein genes is indicative of their particular high polymorphism, occurring even in very closely related strains. The functional annotation of the identified SNPs using PROVEAN (21) showed that among the identified mutations, only one was potentially significant from the point of view of protein functionality. Among the INDEL mutations identified, 17 involved coding regions. The two largest deletions were 9,758 bp and 100 bp long (Fig. 2). These closely spaced deletions involved a total of nine genes. An in-depth analysis of genes that were completely or partially deleted in the *M. bovis* BCG Moreau PL genome, based on data available in the Mycobrowser database (mycobrowser.epfl.ch.) (22), showed the presence of 10 genes in this region, including 3 of them related to the virulence of mycobacteria: *mazF6* toxin gene, *mazE6* antitoxin gene and gene encoding one of the universal stress proteins. This region also contains genes encoding two of the twelve P-type ATPases found in the mycobacterial genome (CtpG and CtpF) and an important transcription factor (CmtR).

Analysis of the impact of these two deletions (9,758 bp and 100 bp) showed that eight whole genes: *mazE6*, *ctpG*, *BCGM1995c*, *BCGM1996c* (*cmtR*), *BCGM1997*, *BCGM1998*, *ctpF* and *BCGM2000c*, and fragments of the *mazF6* and *BCGM2001c* genes are removed. As per the *in silico* translation (insilico.ehu.es/translate) (23), the potential presence of two hypothetical proteins, with a length of 151 aa and 59 aa, was identified. Possible protein sequences were analyzed using the Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST) (24) algorithm for the non-redundant protein sequences database. The result was obtained only for the larger protein (151 aa) – it was 98% similar to a conserved *M. tuberculosis* membrane protein of uncharacterized function. The assembly of the sequence of this protein with the *BCGM2001c* gene product in Standard Protein BLAST showed the identity of the first 145 aa. The deletions described above led to

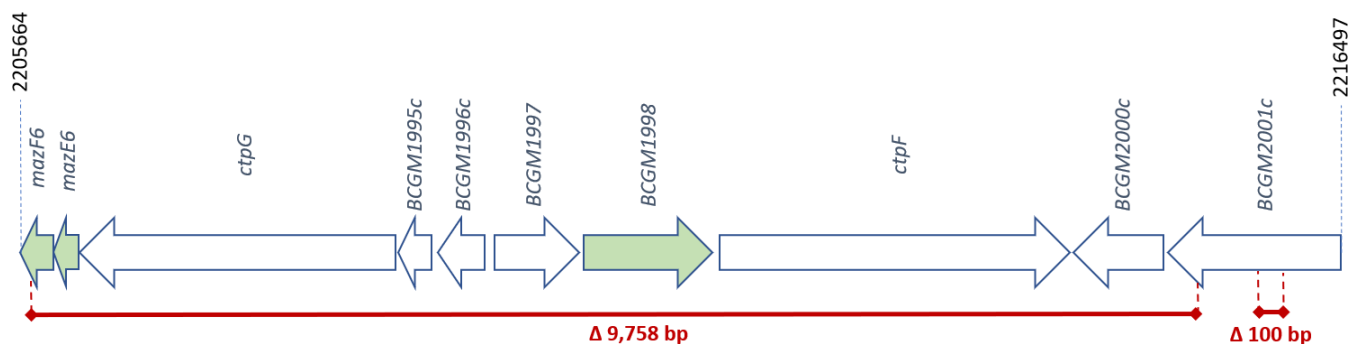


FIG 2 Schematic diagram of the two largest deletions identified in the *M. bovis* BCG Moreau PL genome relative to the *M. bovis* BCG Moreau RDJ reference genome. Genes associated with mycobacterial virulence are colored in green.

the shortening of the protein in BCG-Moreau PL, but as modeling in SWISS-MODEL (25) showed, it still has 4 of the 12 transmembrane segments. The template modeling score (TM-score) calculated in the TM-align for the truncated protein relative to the reference was 0.33046 (26).

The TM-score was calculated for 37 proteins with significant SNP and/or INDEL mutations (26). In five cases (*PE_PGRS19*, *PE_PGRS54*, *PPE24*, *BCGM2001c*, and *BCGM3906c*), the identified mutations resulted in a change in protein folding.

Genetic variations in comparison to *M. bovis* BCG Pasteur

Comparative analyses of *M. bovis* BCG Moreau PL and *M. bovis* BCG Pasteur (AM408590.1), 2 of the most genetically distant BCG substrains, identified a total of 69 significant mutations (non-synonymous SNPs and INDELS in CDSs or probable promotor regions), including 41 SNPs and 28 INDELS (Table S1b). The largest number of SNP mutations (14/41) affected genes encoding PE and Pro-Pro-Glu (PPE) proteins, of which eight were located in the *PE_PGRS28* gene. A total of 40 SNP mutations and 24 INDEL mutations common to *M. bovis* BCG Moreau PL and *M. bovis* BCG Moreau RDJ, which distinguished them from the *M. bovis* BCG Pasteur strain, were identified (Table S1a; Table 1). Many of them are important genetic markers characteristic of the BCG-Moreau substrain described in the literature (16–19). In the Polish BCG genome SNP D322G in the *phoR* gene, the lack of DU1, insertion of IS6110 in front of the *phoP* gene, N-RD18, RD14, and RD2, deletion of the C-terminal end of the *fadD2* gene and the N-end of the *ppsA* gene, DU2-I, deletion of RD16, and deletion of S-RD13 were detected (Table S1b). Analysis of significant mutations differentiating the Polish BCG strain from the RDJ strain showed that 77/80 sequences with SNPs and 15/19 sequences with INDELS were identical to sequences in the *M. bovis* BCG Pasteur genome, indicating that most of the SNPs and INDELS differentiating *M. bovis* BCG Moreau RDJ and *M. bovis* BCG Moreau PL were accumulated in BCG-Moreau RDJ after the transfer of BCG-Moreau to Poland.

Unique genome mutations of *M. bovis* strain BCG Moreau PL

Of the five mutations (one SNP and four INDELS) that distinguished the Polish strain from both BCG-Moreau RDJ and BCG Pasteur (Table 1; Table S1b), there were potentially unique markers. Comparison of these regions with all available genomic sequences of BCG substrains deposited in GenBank revealed that the G389A SNP in the gene homologous to *Rv2553c* encoding one of the conserved membrane proteins and two INDELS: $\Delta 9,758$ bp ($\Delta 2JC$) and $\Delta 100$ bp ($\Delta 3JC$) are characteristic only of the BCG-Moreau PL genome (Fig. 3). The remaining two INDELS, described in Table S1b with no. 25. and 28. have also been identified in the following genomes: BCG-Danish (CP039850.1), BCG-Tokyo (AP010918.1), BCG-Russia (CP009243.1), and BCG-Sofia (CP064405.1).

Genetic stability of the BCG-Moreau PL strain

In the study of the genetic stability of the *M. bovis* BCG Moreau PL strain, the results for five additional genomes of lots from different years obtained by next-generation sequencing (NGS) on the Illumina platform were analyzed. Genetic similarity analysis did not show any significant differences between lots. The average nucleotide identity (ANI)

TABLE 1 The number of significant mutations identified in the *M. bovis* BCG Moreau PL genome compared to *M. bovis* BCG Moreau RDJ and *M. bovis* BCG Pasteur

Type	PL vs. RDJ	PL vs. Pasteur	PL vs. RDJ & Pasteur	PL & Pasteur vs. RDJ	PL & RDJ vs. Pasteur
SNP	80 (77, 3) ^a	41 (38, 3) ^a	1	77	40
INDEL	19 (17, 2) ^a	28 (26, 2) ^a	4	15	24

^aMutations changing the amino acid sequence of a protein, mutations in the promotor region.

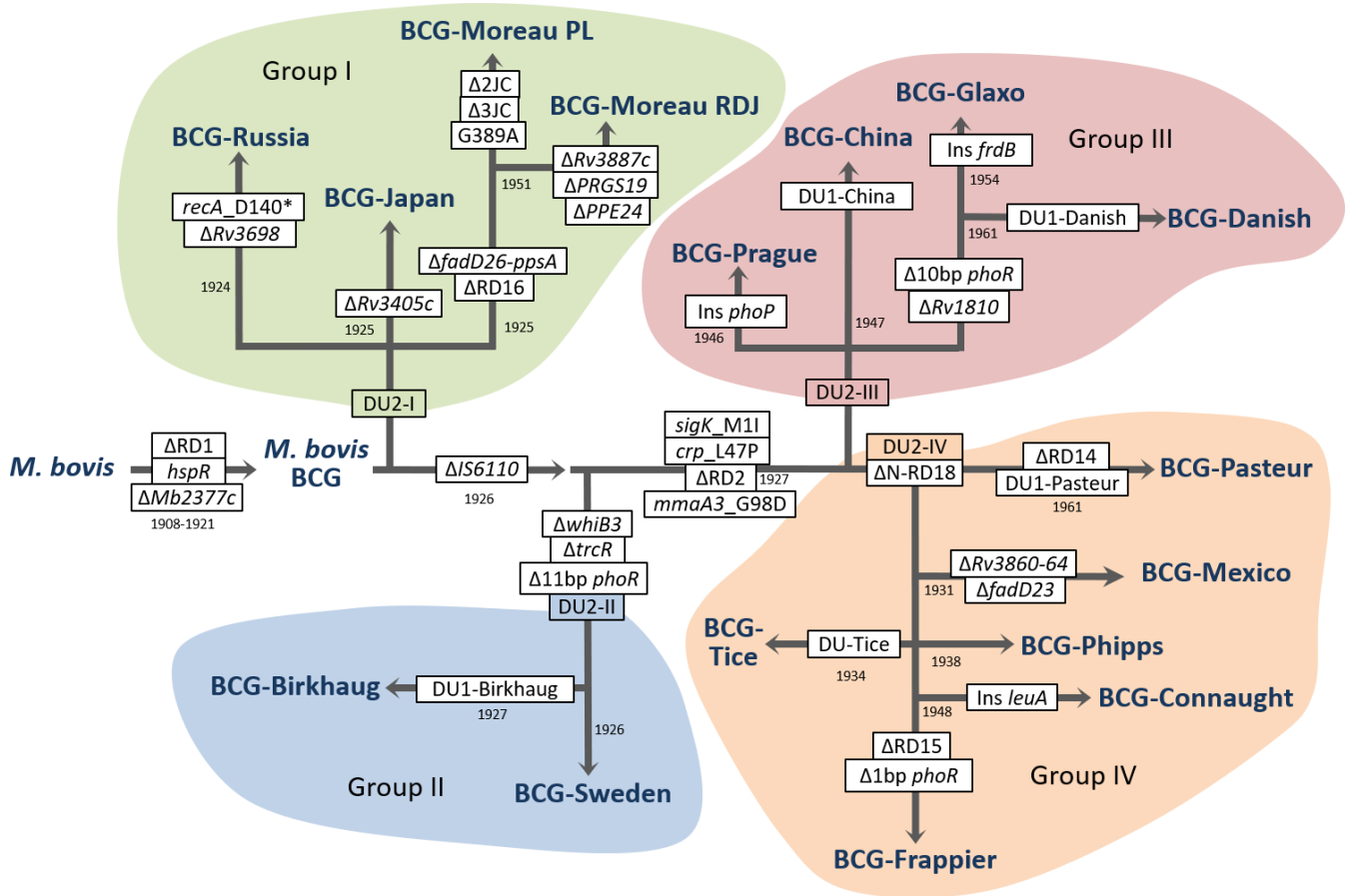


FIG 3 Diagram of the BCG microevolution, including the year a given substrain was obtained and the most important genetic changes (16–18, 27–29).

of the studied genomes was $\geq 99.988\%$, showing a very high similarity between them. Furthermore, all identified mutations were found to be common to the five lots tested.

DISCUSSION

BCG–live, attenuated vaccine is the only vaccine used in immunoprophylaxis of TB since 1921. The specific conditions of the BCG vaccine production technology in different countries over the decades have caused specific genetic changes of the original BCG vaccine strain and the emergence of genetically diverse progeny strains, which have been referred to as BCG substrains. The BCG vaccines produced from different substrains are characterized by different immunogenicity, efficacy, and safety profile (6, 7, 18, 30–32). Although a hundred years have passed since the BCG vaccine was first used, despite the enormous progress in scientific research, it has not been replaced with a new preparation (33). Therefore, it seems that further work on a better characterization of the BCG vaccine strains will allow for a more reliable assessment of the effectiveness and quality control of vaccines, as well as a complete understanding of the mechanisms of action (11).

Taking into account the evidence that the genome of mycobacteria is susceptible to rearrangements, such as deletions or duplications, the genotyping of BCG substrains plays an important role in controlling the production processes in terms of the stability of the vaccine strains (11, 34). This is important because, in addition to genetic differences between BCG substrains, variants of the same substrain have also been observed. Well-documented examples of such variants include the BCG-Tokyo, BCG-Danish, or BCG-Pasteur substrains (17, 18, 27, 35–37). Given that the genetic variations of the vaccine strains may affect the effectiveness and safety of the vaccines produced, it is

necessary to fully characterize individual BCG substrains and monitor their homogeneity over the years. Genetic stability studies have so far been mainly based on typing the multiple deletion (RD) and tandem duplication (DU) regions, that is, defined genetic markers that distinguish the substrains of BCG (20, 38–40). Methods based on the amplification of the genetic markers specific to particular BCG substrains defined in the literature are very simple, fast, and inexpensive, which makes them suitable for ongoing control of the of vaccine strain identity (41). However, these methods do not allow for the complete characterization of a strain, but only its selected regions. In recent years, the development of NGS techniques has enabled the precise identification of differences between some BCG substrains and the verification of their genetic stability (42). Asadian *et al.* (42) analyzed strains obtained from different batches of vaccines produced in the years 2010–2019 from the BCG-Pasteur substrain. Sequencing using the Illumina platform did not show batch-to-batch differences, yet only minor changes from the *M. bovis* BCG Pasteur 1173P2 reference genome were shown. The authors emphasized that it would be reasonable to support the sequencing-based technique with a short-read using a long-read technique to elucidate the sequence of regions containing repeats (42). Monitoring of spontaneous mutations in the whole genome has become possible with the development of methods based on NGS technologies and may turn out to be a new standard for the quality control of the BCG vaccine (43). Stability studies using NGS on Roche's 454 platform were performed for the BCG-Russia strain. This study compared the results obtained for three seed lots: 311 (1963), 977 (1982), and 368 (2006). Only two low-grade lesions were identified, confirming the genetic stability of the BCG Russia substrain vaccine used in production (44). The latest genomic studies of the Russian working seed lot no. 368 of the BCG-Russia strain showed divergent results, including those with respect to the sequence of the *glnD* gene. The authors concluded that these data may suggest internal polymorphism between lots produced by different manufacturers (Medgamal and Microgen) resulting from differences in the production process. The observed heterogeneity may also be partly related to different bioinformatics tools and accepted qualitative criteria used in the analysis of NGS data, which was noted out by researchers (45).

In the years 1947–1950, mass vaccinations of children and adolescents aged 1–18 were carried out in Poland, using a vaccine based on the Danish substrain. The vaccination campaign was hampered by numerous postvaccination reactions in the form of lymphadenitis in newborns vaccinated orally. Replacing the route of administration changed the location of the reactions but did not reduce them and, in fact, increased their number. A significant improvement occurred in 1954 after the vaccine substrain was changed from Danish to Brazilian (BCG-Moreau), with virtually no complications. The BCG Moreau substrain is derived from the original BCG strain sent from the Pasteur Institute in Paris to Brazil by Dr. Julio Elvie Moreau on 22 May 1925. It was transferred to Poland in 1954 to the Serum and Vaccine Factory in Lublin. A year later, the production of the native BCG vaccine from the BCG-Moreau substrain began in Poland. The freeze-drying of strains was used in the production process from the beginning (12). The first vaccines made in Lublin were oral vaccines, produced since 1957 from lot no. 1250557, which was replaced after 82 passages with lot no. 21078, initially called the experimental seed lot, and then, after implementing the principles of the seed lot system in production in 1982, it was officially the first seed lot of the BCG vaccine in Poland (Fig. 4). After its exhaustion in 1986, the second seed lot no. 70586 was introduced. Currently, in the production of the “BCG 10 Anti-TB Vaccine” in Poland, Biomed-Lublin uses two working seed lots in parallel: working seed lot no. 8112000, derived from the second seed lot no. 70586 and working seed lot no. 3122000, derived from the first seed lot no. 21078. In this study, to test the stability of the Polish vaccine strain, NGS was carried out on the Illumina platform involving five lots from different years: oral vaccine no. 1250557 (1957), two current working seed lots no. 8112000 (2000) and 3122000 (2000), and final products no. 00615 (2015) and no. 01115 (2015).

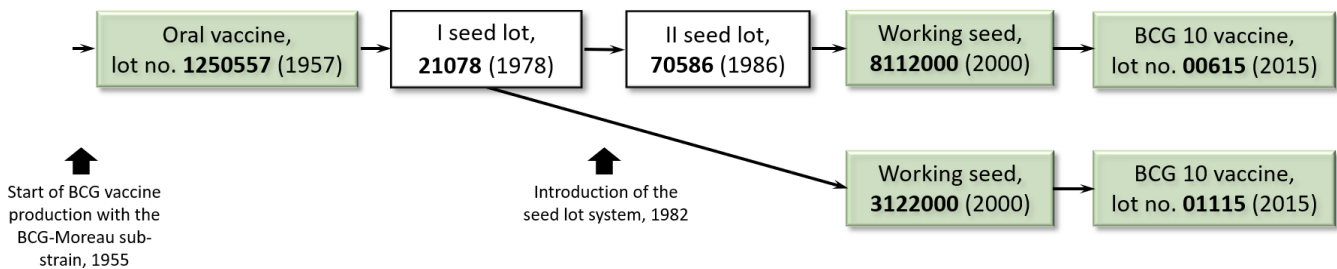


FIG 4 Diagram of the production of TB vaccines in Poland. The material used for the tests is marked in green.

This material selection was intended to identify possible genetic changes before and after the introduction of the seed lot system, differences between the working lots used for the production of the BCG vaccine and differences between a given working lot and the final product obtained from it. Analysis of genetic similarity of the five test lots of the *M. bovis* BCG Moreau PL strain did not show significant differences between lots. The average similarity of the nucleotide sequences of the working genomes obtained for each batch with the other four was $\geq 99.988\%$, which confirmed a very high similarity between them. Mutations and weak/null coverage identified for each lot were also analyzed based on Illumina sequencing data against the *M. bovis* BCG Moreau RDJ reference substrain genome. All identified changes were found to be common to the five batches tested. These results demonstrate the genetic stability of the *M. bovis* BCG Moreau PL strain over a period of almost 60 years, which is consistent with our own previous research on the genotyping of the Polish BCG lots (20). Therefore, it can be concluded that the manner, in which the vaccine strain was stored and passaged in Poland did not lead to genetic changes, and the identified mutations present in all five BCG lots occurred before 1957. However, it cannot be determined whether these changes occurred in Poland after receiving the BCG-Moreau substrain between 1954 and 1957, because we do not have the 1954 lot. The original BCG-Moreau strain that was sent from the Pasteur Institute to Brazil is also unavailable. Therefore, the data obtained in the study were compared to the genomic sequence of *M. bovis* BCG Moreau RDJ deposited in the GenBank database (19). The strain, sequenced by Gomes et al., is a strain derived from the BCG vaccine currently used in Brazil and, therefore, probably different from the original BCG-Moreau strain, imported to Brazil in 1925, about 85 years earlier. It is worth to emphasize how important it is to know the full sequence of the strain used to produce the BCG vaccine administered to the entire population. In future studies, the Polish strain will constitute a reference point for comparing the sequences of strains isolated from patients with BCG-itis after a locally produced vaccine in order to verify whether genetic changes occurred *in vivo*. The stability analysis of BCG-Moreau PL showed that the markers we identified have not changed over almost 60 years. We can confidently suspect that this will not change because knowledge about the production of the BCG vaccine and maintaining quality standards, including the use of a freeze-dried seed lots system, has achieved enormous progress since the 1960s. However, with the full genome sequence now available, stability can be easily monitored and any changes detected.

In order to obtain the complete genome sequence of the Polish vaccine strain *M. bovis* BCG Moreau PL, two new generation sequencing technologies were used – Illumina (second generation) and MinION (third generation). The strategy of combining short-read sequencing, for example, NGS Illumina, with long-read sequencing, for example, NGS MinION, provides the most optimal approach to sequencing bacterial genomes and allows the identification of both small point mutations and major genomic rearrangements. As shown in the paper, hybrid *de novo* assembly, in which long NGS MinION reads form a “scaffolding” supplemented by short NGS Illumina reads, allowed to obtain a good quality single working genome sequence that did not require many corrections. This approach is supported by recent sequencing of the BCG-Danish 1331 strain (27). The authors pointed out that using short-read sequencing for genomes

containing multiple repeat regions and large deletions and duplications can generate errors. Such sites can be effectively verified using long-read technology. That was the first published hybrid approach to genome sequencing of a BCG strain that identified inconsistencies in previous papers describing the sequence of the BCG-Danish sub-strain. The scientists also emphasized that the methodology, combining the two NGS technologies, is both time- and financially advantageous compared to the methods used in the past (27, 30).

Analysis of the study results confirmed the presence of characteristic changes of the BCG-Moreau strain described in the literature in the genome of the Polish strain BCG-Moreau PL (16–19), that is, lack of DU1 duplication, presence of DU2-I duplication, IS6110 insertion before the gene *phoP*, C-terminal deletion of the *fadD2* gene and N-terminal deletion of the *ppsA* gene, RD16 deletion, S-RD13 deletion, and lack of deletion: N-RD18, RD14, and RD2. When analyzing the sequences containing the identified SNP mutations and INDEL mutations in the *M. bovis* BCG Moreau PL substrain genome against *M. bovis* BCG Moreau RDJ, the vast majority were shown to be identical to the sequences in the *M. bovis* BCG Pasteur genome. Therefore, these mutations probably occurred in the genome of *M. bovis* BCG Moreau RDJ and not in *M. bovis* BCG Moreau PL. However, mutations that distinguish the genome of the *M. bovis* BCG Moreau PL strain from the BCG genomes available in GenBank are one SNP mutation (G389A) in the gene homologous to *Rv2553c*, encoding conserved membrane protein and two INDEL mutations ($\Delta 9,758$ bp and $\Delta 100$ bp). These deletions cover 10 genes, including among others: a toxin gene (*mazF6*), an antitoxin gene (*mazE6*), a gene encoding one of the universal stress proteins, two genes encoding ATPases (*ctpG*, *ctpF*), and a transcription factor (*cmtR*). All the genes mentioned are involved in the response to stress conditions and the survival of mycobacteria in the host tissues, although the mechanism of their action is different. The presence of free toxin MazF6 causes reprogramming transcription toward adaptation to stress conditions (46) while P-type ATPases, CtpG and CtpF, are critically important for the survival of mycobacteria because they are responsible for the detoxification of metal ions (Zn^{2+} and Ca^{2+} , respectively) inside human macrophages during infection (47, 48). These changes may constitute one of the factors contributing to the higher safety of the Polish BCG vaccine compared with BCG-Moreau RDJ described in the literature (49). However, the exact significance of the mutations unique to the Polish strain, both in terms of their impact on the vaccine safety and immunogenicity, requires more detailed analyses in the future. In the next stage of the research, analogous comparisons at the transcriptome and proteome level to reference strains were planned. It will be intriguing to verify what alternative pathways of response to stress conditions have been developed in BCG-Moreau PL strain.

In the mid-1990s, intensive research began on the genetic diversity of BCG substrains. RD, tandem duplications (DU), and SNPs were identified. On the basis of the detected genetic markers differentiating the substrains and historical knowledge of their distribution, phylogenetic trees were constructed to present the microevolution of BCG substrains (17, 18, 50). In routine BCG vaccine identity testing, multiplex PCR is used, which allows BCG to be identified accurately to the substrain (41). However, it does not allow the differentiation of variants of the same substrain as BCG-Moreau RDJ and PL. The markers we identified that are unique to the Polish substrain allowed us to distinguish it from other substrains, including RDJ. Simple PCR using the 2JC primers we designed in the study gives product only for BCG-Moreau PL (Table S2).

Despite its imperfections, BCG has remained the only licensed vaccine against TB for over 100 years. Since 1974, it has been included in the WHO Extended Vaccination Program. The legitimacy of conducting BCG vaccinations in the world is primarily owed to the proven protection of children against the most severe forms of TB, meningitis, and miliary form. However, protection against the most common form – pulmonary TB – varies greatly between 0% and 80% (8). Randomized follow-up studies showed that no prior infection with *M. tuberculosis* or no exposure to environmental mycobacteria is associated with higher efficacy of BCG against pulmonary TB and possibly miliary

and meningeal TB (51). However, it seems that many more factors are causing such a significant variation in vaccine effectiveness. One of the most important factors may be the BCG substrain used for its production (18). Ritz et al. in 2008 provided a critical summary of the available studies comparing the efficacy of BCG vaccines in animals and humans. The analysis confirmed that the induced immune response and protection against TB differed according to the BCG substrain used to produce the vaccine. However, the results were so contradictory that it was not possible to identify a single favorite among the BCG substrains. As emphasized, identification of the BCG substrain with the best protective properties would have a huge impact on TB control at the population level (52). Two large randomized clinical trials of BCG revaccination in Brazil and Malawi failed to demonstrate efficacy against TB, although both trials did not take into account *M. tuberculosis* infection status among participants (53–55). More recent studies in the uninfected population have shown the potential of BCG revaccination to protect against latent infection. Studies verifying these results, conducted on a larger cohort, are to be completed in 2026 (55, 56). This demonstrates a potential new use for BCG. However, not only revaccination of a specific target group but also changing the route of administration (57) and genome recombinations (58) are tested, which shows that BCG can still be the subject of promising research. The Polish strain, apart from the genetic markers characteristic of the Moreau substrain described in the literature, is a natural mutant in several genes related to the response to stress and intracellular survival in host cells. Among them are genes encoding two P-type ATPases, CtpG and CtpF. As a recent research has demonstrated, their deletion disturbs detoxification systems reducing mycobacterial virulence. The *ctpF* gene shows high level of activation under different conditions, including infection, oxidative/nitrosative stress and hypoxia, promoting a non-replicating persistence (NRP) or dormancy state of bacilli. The authors suggested that this may be a good target for the attenuation of *M. tuberculosis* in terms of obtaining a new vaccine candidate against TB (47). The advantages of whole cell live vaccines include a broad antigenic repertoire, triggering a diverse immune response, mimicking immunity elicited by natural TB infection (56). However, it is necessary to develop immunological correlates of protection, in the form of a set of biomarkers (biosignature), to facilitate the development and comparison of vaccines (59).

Not a single genetic marker, but all changes in the BCG substrain genome contribute to vaccine effectiveness in a given population. Therefore, it is reasonable to track these changes and characterize BCG genomes against other substrains and genomes of virulent strains of *M. tuberculosis*. This study found that among the significant non-synonymous SNP mutations identified in the *M. bovis* BCG Moreau PL genome, in relation to the *M. bovis* BCG Moreau RDJ genome and to the *M. bovis* BCG Pasteur genome, most mutations occurred in genes encoding the PE and PPE proteins, which proves their high polymorphism within the BCG substrains. This conclusion is consistent with scientific reports showing both sequence and expression level variability of the *pe/ppe* genes among BCG substrains (18, 60). PE and PPE are two large families of proteins specific for mycobacteria, whose genes occupy approximately 7% of the coding regions in the genome of the *M. tuberculosis* complex species. Their name comes from the conserved proline and glutamic acid motifs PE and PPE at the N-terminal ends of the proteins. The *pe/ppe* genes are of great interest to scientists due to their high variability (61). Although discovered more than 20 years ago, the function of these proteins, and the polymorphic subfamilies, in particular, remains unclear. Many studies suggest that some PE and PPE proteins play an important role in the virulence of mycobacteria (61–67). Genes encoding some of the PE and PPE proteins have been shown to be overexpressed for mycobacteria isolated from lung granulomas compared to bacteria cultured *in vitro* (63, 68). Surface exposure or secretion into the extracellular environment allows the PE/PPE proteins to interact directly with host cells. Some proteins probably activate signal transduction and modulate cytokine production via the TLR2/4 receptors on the surface of macrophages, others inhibit autophagy and inflammatory responses to avoid an attack by the immune system or shift the response from Th1 to Th2, which promotes

intracellular survival of mycobacteria (69). This variation may result in different levels of protection for BCG vaccines produced from different substrains (60). In contrast, it is worth noting that BCG has a defect in the secretion of PE/PPE proteins due to the loss of the RD5 region, making it difficult to relate the results of research on virulent strains to vaccine strains in this matter (70, 71).

MATERIALS AND METHODS

Vaccine strains and DNA isolation

The genetic stability study analyzed the NGS results on the Illumina platform for five strain lots of the Polish *M. bovis* BCG Moreau strain derived from different years (Fig. 4). The oldest of the tested lots, no. 1250557, came from an oral vaccine that was freeze dried in 1957, 2 years after the start of production of the BCG vaccine in Poland from the Moreau substrain. Two more lots: no. 8112000 and no. 3122000, were produced from two different master seed lots in 2000. Both are working seed lots used currently to produce the Polish national BCG vaccine. The last two lots tested: no. 00615 and no. 01115 are the final products, that is, "BCG 10 Anti-TB Vaccine" produced in 2015, from different working seed lots 8112000 and 3122000, respectively.

The mycobacteria were cultured on Ogawa medium for 21 days at 37°C. Bacterial DNA was isolated according to the methodology of van Soolingen et al. 1991 (72) with minor modifications.

Whole genome sequencing and reference genome

The DNA of the five tested lots of the *M. bovis* BCG Moreau PL strain was sequenced on the Illumina NextSeq 500 platform in a 2 × 250 bp configuration, using the NEXtera XT kit (Illumina) for libraries preparation (240× coverage per genome).

Genome sequencing of the working seed lot no. 3122000 of the *M. bovis* BCG Moreau PL strain was additionally carried out on the MinION platform (40× coverage). The libraries have been prepared using the Ligation Sequencing Kit (Oxford Nanopore Technologies).

Genome assembly and analysis

In the first step, raw fastq files from the Illumina platform were trimmed using Trim Galore! to purify reads from the outermost sections and to remove poor quality sequences (73). Reads from the NGS Illumina platform for the five strain lots of the Polish *M. bovis* BCG Moreau substrain were aligned separately to the *M. bovis* BCG Moreau RDJ reference genome ([AM412059.2](#)) (19) using the Bowtie2 tool (74). Subsequently, consensus sequences were created for the five lots of the Polish BCG strain in BCftools (75). In fastANI, the average nucleotide sequence similarity (ANI) was calculated for each lot in relation to the other four lots (76). Sequencing results obtained from the Illumina platform were used for the identification of mutations (variants) using breseq in the genomes of the five lots of the Polish BCG strain against the reference genome *M. bovis* BCG Moreau RDJ (77). The variant designation threshold was the occurrence of changes in reads with a frequency of at least 80%. The breseq program was also used to identify sites in the reference genome that were not covered by Illumina reads, which may indicate a deletion of a given region. All identified mutations detected by breseq, differentiating the five lots of the Polish BCG strain and the uncovered regions, were verified in Integrative Genomics Viewer (IGV) (78).

Reads from the MinION platform for the working seed lot no. 3122000 were aligned to the *M. bovis* BCG Moreau RDJ reference genome using the minimap2 tool (79).

The initial sequence of the Polish vaccine strain (working seed lot no. 3122000) was generated as a result of the hybrid genome assembly of long reads obtained from the MinION platform and short reads obtained from the Illumina platform using hybridSPAdes (80). All-important mutations, that is, those lying at CDS and at probable

promotor regions (up to 50 nt upstream of the START codon), were verified by Sanger sequencing. The Sanger sequencing was performed using primers listed in Table S2.

Verifications and corrections allowed to obtain the complete sequence of the Polish strain *M. bovis* BCG Moreau (working seed lot no. 3122000), which was deposited in the GenBank database where it was automatically annotated with the “Best-placed reference protein set; GeneMarkS-2+” method according to the “NCBI PGAP” scheme and received accession number [CP085532](#). Genome visualization was performed using Proksee (v6.0.2) (15).

In order to check in which genome of BCG Moreau the identified mutations occurred, the sequence of the Polish substrain was additionally compared with the most distant sister substrain of BCG – and *M. bovis* BCG Pasteur 1173P2 ([AM408590.1](#)) (18).

Among the significant mutations (non-synonymous SNPs and INDELs in CDSs or in probable promotor regions) identified in the *M. bovis* BCG Moreau PL genome, mutations specific only to the Polish vaccine strain were searched. For this purpose, sequences containing mutations from the BCG Moreau PL genome against both BCG Moreau RDJ and BCG Pasteur were compared with genomic sequences of the BCG substrains (taxid: 33892) deposited in GenBank using BLASTN (NCBI).

The theoretical impact of identified mutations was analyzed using the [insilico.ehu.es](#) tool (genetic code: Bacterial and Plant Plasmid) (23), and the functional category of proteins, in which mutations were identified was assigned according to [myco-browser.epfl.ch](#) (22). Protein modeling was performed using SWISS-MODEL (25), and the impact of identified mutations on protein stability was verified in the TM-align (26). The functional annotation of the identified SNPs was performed using PROVEAN Protein (21).

PCR analysis and Sanger sequencing

PCR (Platinum SuperFi II PCR Master Mix, Invitrogen) was performed on gDNA using primers listed in Table S2. The PCR products were run on a 1.5% agarose gel, stained with Midori Green and visualized under ultraviolet light. To confirm SNPs and INDELs, the PCR products were purified with Illustra ExoProStar 1-Step (Cytiva) and Sanger sequenced.

Conclusions

The need for better TB control remains a global health priority. The live, attenuated BCG vaccine is already 100 years old and, despite being controversial, it is still the only vaccine available in TB immunoprophylaxis. The selection of the BCG substrain used for vaccine production remains an important issue. At present, it is difficult to determine which substrains will provide maximum efficacy and safety; therefore, more detailed genetic analyses of the BCG substrains are needed. The organization of the archival record of BCG distribution around the world and comparative studies of sister substrains are hampered by unclear nomenclature used initially, the use of synonyms, name changes due to a change in manufacturer or place of manufacture, or a change of substrain by the manufacturer. Despite that, over the 40 years of molecular studies of the BCG substrains, many markers have been identified that distinguish individual sister substrains – from SNP-type changes to large genomic rearrangements.

The whole genome sequence obtained is the first and only genomic sequence of the strain included in the TB vaccine production in Poland. It has been shown that it differs genetically from its closest sister strain BCG, derived from the same parent strain BCG-Moreau – *M. bovis* BCG Moreau RDJ strain. The results supplement the knowledge of the microevolution and attenuation of BCG vaccine substrains. The identified markers can be used to develop more specific identity and genetic stability tests of vaccine strains and taxonomic comparisons. The molecular characterization of BCG substrains opens the door to finding reliable correlates in the clinical assessment of the protective efficacy and safety of the BCG vaccine in humans, which is particularly important in the development of new TB vaccines and immunization strategies.

These results show the genetic stability of the Polish BCG strain over almost 60 years. Therefore, it can be concluded that the way, in which the vaccine strain had been stored

and passaged in Poland did not lead to genetic changes, and the identified mutations occurred before 1957 in the BCG-Moreau genome.

ACKNOWLEDGMENTS

The project was funded by the National Science Centre (No. 2014/15/N/NZ7/03793).

AUTHOR AFFILIATIONS

¹Department of Sera and Vaccines Evaluation, National Institute of Public Health NIH – National Research Institute, Warsaw, Poland

²Biobank Lab, Department of Oncobiology and Epigenetics, Faculty of Biology and Environmental Protection, University of Lodz, Lodz, Poland

³Mycobacterium Genetics and Physiology Unit, Institute of Medical Biology, Polish Academy of Sciences, Lodz, Poland

⁴Chair of Drug and Cosmetics Biotechnology, Faculty of Chemistry, Warsaw University of Technology, Warsaw, Poland

⁵Department of Medical Biology, National Institute of Cardiology, Warsaw, Poland

⁶Department of Microbiology, National Tuberculosis and Lung Diseases Research Institute, Warsaw, Poland

AUTHOR ORCID*s*

Katarzyna Krysztopa-Grzybowska  <http://orcid.org/0000-0002-7332-5290>

Jarosław Dziadek  <http://orcid.org/0000-0003-1427-884X>

FUNDING

Funder	Grant(s)	Author(s)
Narodowe Centrum Nauki (NCN)	2014/15/N/NZ7/03793	Katarzyna Krysztopa-Grzybowska Maciej Polak

DATA AVAILABILITY

All information on the sequence of Mycobacterium tuberculosis variant bovis BCG strain Moreau PL is available in the NCBI database: [CP085532.1](https://.ncbi.nlm.nih.gov/nucl/CP085532.1). Reference sequences used for comparisons are available in the NCBI repository with the following identifiers: [AM412059.2](https://.ncbi.nlm.nih.gov/nucl/AM412059.2) [*M. bovis* BCG Moreau RDJ (19)] and [AM408590.1](https://.ncbi.nlm.nih.gov/nucl/AM408590.1) [*M. bovis* BCG Pasteur 1173P2 (18)]. Data supporting the findings of this study are available on request from the corresponding author K.K.-G.

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Table S1 (Table S1.xlsx). Identified differences in the nucleotide sequence.

Table S2 (Table S2.docx). Characterization of primers.

REFERENCES

- Barberis I, Bragazzi NL, Galluzzo L, Martini M. 2017. The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. *J Prev Med Hyg* 58:E9–E12.
- World Health Organization. 2020. Global tuberculosis report 2020. World Health Organization, Geneva.
- Fatima S, Kumari A, Das G, Dwivedi VP. 2020. Tuberculosis vaccine: a journey from BCG to present. *Life Sci* 252:117594. <https://doi.org/10.1016/j.lfs.2020.117594>
- Fellag M, Loukil A, Drancourt M. 2021. The puzzle of the evolutionary natural history of tuberculosis. *New Microbes New Infect* 41:100712. <https://doi.org/10.1016/j.nmni.2020.100712>
- Lugosi L. 1992. Theoretical and methodological aspects of BCG vaccine from the discovery of Calmette and Guérin to molecular biology. A review. *Tuber Lung Dis* 73:252–261. [https://doi.org/10.1016/0962-8479\(92\)90129-8](https://doi.org/10.1016/0962-8479(92)90129-8)

6. Liu J, Tran V, Leung AS, Alexander DC, Zhu B. 2009. BCG vaccines: their mechanisms of attenuation and impact on safety and protective efficacy. *Hum Vaccin* 5:70–78. <https://doi.org/10.4161/hv.5.2.7210>
7. Behr MA. 2002. BCG--different strains, different vaccines? *Lancet Infect Dis* 2:86–92. [https://doi.org/10.1016/s1473-3099\(02\)00182-2](https://doi.org/10.1016/s1473-3099(02)00182-2)
8. Ottenhoff THM, Kaufmann SHE. 2012. Vaccines against tuberculosis: where are we and where do we need to go? *PLoS Pathog* 8:e1002607. <https://doi.org/10.1371/journal.ppat.1002607>
9. Rowland R, McShane H. 2011. Tuberculosis vaccines in clinical trials. *Expert Rev Vaccines* 10:645–658. <https://doi.org/10.1586/erv.11.28>
10. Lange C, Aaby P, Behr MA, Donald PR, Kaufmann SHE, Netea MG, Mandalakas AM. 2022. 100 years of *Mycobacterium bovis* bacille Calmette-Guérin. *Lancet Infect Dis* 22:e2–e12. [https://doi.org/10.1016/S1473-3099\(21\)00403-5](https://doi.org/10.1016/S1473-3099(21)00403-5)
11. Stefanova T. 2014. Quality control and safety assessment of BCG vaccines in the post-genomic era. *Biotechnol Biotechnol Equip* 28:387–391. <https://doi.org/10.1080/13102818.2014.927200>
12. Moskwa Z. 1983. Historia Szczepień Przeciwgruźliczych I Powstanie Szczepionki BCG. *Wiadomości Lekarskie* (Warsaw, Poland: 1960). Vol. 36.
13. Korzeniewska-Koseła M. 2021. Gruźlica i choroby układu oddechowego w Polsce w 2020 roku. *Instytut Gruźlicy i Chorób Płuc*
14. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2014. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. <https://doi.org/10.1101/gr.186072.114>
15. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C-Y, Graham M, Van Domselaar G, Stothard P. 2023. Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Res* 51:W484–W492. <https://doi.org/10.1093/nar/gkad326>
16. Leung AS, Tran V, Wu Z, Yu X, Alexander DC, Gao GF, Zhu B, Liu J. 2008. Novel genome polymorphisms in BCG vaccine strains and impact on efficacy. *BMC Genomics* 9:413. <https://doi.org/10.1186/1471-2164-9-413>
17. Bedwell J, Kairo SK, Behr MA, Bygraves JA. 2001. Identification of substrains of BCG vaccine using multiplex PCR. *Vaccine* 19:2146–2151. [https://doi.org/10.1016/s0264-410x\(00\)00369-8](https://doi.org/10.1016/s0264-410x(00)00369-8)
18. Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, Valenti P, Dos Santos S, Duthoy S, Lacroix C, Garcia-Pelayo C, Inwald JK, Golby P, Garcia JN, Hewinson RG, Behr MA, Quail MA, Churcher C, Barrell BG, Parkhill J, Cole ST. 2007. Genome plasticity of BCG and impact on vaccine efficacy. *Proc Natl Acad Sci U S A* 104:5596–5601. <https://doi.org/10.1073/pnas.0700869104>
19. Gomes LHF, Otto TD, Vasconcellos EA, Ferrão PM, Maia RM, Moreira AS, Ferreira MA, Castello-Branco LRR, Degraive WM, Mendonça-Lima L. 2011. Genome sequence of *Mycobacterium bovis* BCG Moreau, the Brazilian vaccine strain against tuberculosis. *J Bacteriol* 193:5600–5601. <https://doi.org/10.1128/JB.05827-11>
20. Krysztopa-Grzybowska K, Brzezińska S, Augustynowicz-Kopeć E, Polak M, Augustynowicz E, Lutyńska A. 2012. Descendant of daughter Brazilian BCG Moreau substrain in Poland. *Vaccine* 30:5512–5518. <https://doi.org/10.1016/j.vaccine.2012.06.056>
21. Choi Y. 2012. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein, p 414–417. In *Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine*. ACM, Orlando Florida.
22. Kapopoulou A, Lew JM, Cole ST. 2011. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)* 91:8–13. <https://doi.org/10.1016/j.tube.2010.09.006>
23. Bikandi J, Millán RS, Rementeria A, Garaizar J. 2004. *In silico* analysis of complete bacterial genomes: PCR, AFLP-PCR and endonuclease restriction. *Bioinformatics* 20:798–799. <https://doi.org/10.1093/bioinformatics/btg491>
24. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. 2012. Domain enhanced lookup time accelerated BLAST. *Biol Direct* 7:12. <https://doi.org/10.1186/1745-6150-7-12>
25. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46:W296–W303. <https://doi.org/10.1093/nar/gky427>
26. Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33:2302–2309. <https://doi.org/10.1093/nar/gki524>
27. Borgers K, Ou J-Y, Zheng P-X, Tiels P, Van Hecke A, Plets E, Michielsens G, Festjens N, Callewaert N, Lin Y-C. 2019. Reference genome and comparative genome analysis for the WHO reference strain for *Mycobacterium bovis* BCG Danish, the present tuberculosis vaccine. *BMC Genomics* 20:561. <https://doi.org/10.1186/s12864-019-5909-5>
28. Krysztopa-Grzybowska K, Lutyńska A. 2016. Mikroewolucja podzespółów BCG. *Postepy Hig Med Dosw* 70:1259–1266. <https://doi.org/10.5604/17322693.1226692>
29. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM. 1999. Comparative genomics of BCG vaccines by whole genome DNA microarray. *Science* 284:1520–1523. <https://doi.org/10.1126/science.284.5419.1520>
30. Abdallah AM, Hill-Cawthorne GA, Otto TD, Coll F, Guerra-Assunção JA, Gao G, Naem R, Ansari H, Malas TB, Adroub SA, Verboom T, Ummels R, Zhang H, Panigrahi AK, McEnerney R, Brosch R, Clark TG, Behr MA, Bitter W, Pain A. 2015. Genomic expression catalogue of a global collection of BCG vaccine strains show evidence for highly diverged metabolic and cell-wall adaptations. *Sci Rep* 5:15443. <https://doi.org/10.1038/srep15443>
31. Chen JM, Islam ST, Ren H, Liu J. 2007. Differential productions of lipid virulence factors among BCG vaccine strains and implications on BCG safety. *Vaccine* 25:8114–8122. <https://doi.org/10.1016/j.vaccine.2007.09.041>
32. Ritz N, Dutta B, Donath S, Casalaz D, Connell TG, Tebruegge M, Robins-Browne R, Hanekom WA, Britton WJ, Curtis N. 2012. The influence of bacille Calmette-Guérin vaccine strain on the immune response against tuberculosis. *Am J Respir Crit Care Med* 185:213–222. <https://doi.org/10.1164/rccm.201104-0714OC>
33. Frick M. 2022. Tuberculosis vaccines pipeline report 2022 18
34. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A* 99:3684–3689. <https://doi.org/10.1073/pnas.052548299>
35. Naka T, Maeda S, Niki M, Ohara N, Yamamoto S, Yano I, Maeyama J-I, Ogura H, Kobayashi K, Fujiwara N. 2011. Lipid phenotype of two distinct subpopulations of *Mycobacterium bovis* Bacillus Calmette-Guérin Tokyo 172 substrain. *J Biol Chem* 286:44153–44161. <https://doi.org/10.1074/jbc.M111.310037>
36. Honda I, Seki M, Ikeda N, Yamamoto S, Yano I, Koyama A, Toida I. 2006. Identification of two subpopulations of Bacillus Calmette-Guérin (BCG) Tokyo172 substrain with different RD16 regions. *Vaccine* 24:4969–4974. <https://doi.org/10.1016/j.vaccine.2006.03.055>
37. Asadian M, Hassanzadeh SM, Safarchi A, Douraghi M. 2022. Genomic characteristics of two most widely used BCG vaccine strains: Danish 1331 and Pasteur 1173P2. *BMC Genomics* 23:609. <https://doi.org/10.1186/s12864-022-08826-9>
38. Seki M, Udagawa T, Sugawara I, Iwama K, Honda I, Fujita I, Hashimoto A, Yano I, Yamamoto S. 2012. The effect of passages during Japanese BCG vaccine production on genetic stability and protective efficacy. *Vaccine* 30:1460–1464. <https://doi.org/10.1016/j.vaccine.2011.12.101>
39. Fallah F, Karimi A, Abdolghafoorian H, Mohammadzadeh A, Zahraei SM, Goudarzi H, Fallah M. 2016. Molecular analysis of the bacille Calmette-Guérin vaccine strain currently being used in Iran. *Int J Tuberc Lung Dis* 20:223–227. <https://doi.org/10.5588/ijtld.15.0434>
40. Trovero A, Argüelles C, Cataldi A. 2010. Preparation of a working seed lot of BCG and quality control by PCR genotyping. *Rev Argent Microbiol* 42:4–10. <https://doi.org/10.1590/S0325-75412010000100002>
41. Markey K, Ho MM, Choudhury B, Seki M, Ju L, Castello-Branco LRR, Gairola S, Zhao A, Shibayama K, Andre M, Corbel MJ. 2010. Report of an international collaborative study to evaluate the suitability of multiplex PCR as an identity assay for different sub-strains of BCG vaccine. *Vaccine* 28:6964–6969. <https://doi.org/10.1016/j.vaccine.2010.08.045>
42. Asadian M, Safarchi A, Hassanzadeh SM, Yaseri M, Douraghi M. 2022. Genomic evidence for stability of the Bacillus Calmette-Guérin (BCG) vaccine strain (Pasteur 1173P2) from different batches in Iran. *J Appl Microbiol* 132:2379–2388. <https://doi.org/10.1111/jam.15365>

43. Wada T, Maruyama F, Iwamoto T, Maeda S, Yamamoto T, Nakagawa I, Yamamoto S, Ohara N. 2015. Deep sequencing analysis of the heterogeneity of seed and commercial lots of the Bacillus Calmette-Guérin (BCG) tuberculosis vaccine substrain Tokyo-172. *Sci Rep* 5:17827. <https://doi.org/10.1038/srep17827>
44. Voronina OL, Kunda MS, Aksenova EI, Semenov AN, Ryzhova NN, Lunin VG, Gintsburg AL. 2016. Mosaic structure of *Mycobacterium bovis* BCG genomes as a representation of phage sequences' mobility. *BMC Genomics* 17:1009. <https://doi.org/10.1186/s12864-016-3355-1>
45. Narvskaya O, Starkova D, Levi D, Alexandrova N, Molchanov V, Chernyaeva E, Vyazovaya A, Mushkin A, Zhuravlev V, Solovieva N, Vishnevskiy B, Mokrousov I. 2020. First insight into the whole genome sequence variations in *Mycobacterium bovis* BCG-1 (Russia) vaccine seed lots and their progeny clinical isolates from children with BCG-induced adverse events. *BMC Genomics* 21:567. <https://doi.org/10.1186/s12864-020-06973-5>
46. Chattopadhyay G, Bhasin M, Ahmed S, Gosain TP, Ganesan S, Das S, Thakur C, Chandra N, Singh R, Varadarajan R. 2022. Functional and biochemical characterization of the MazEF6 toxin-antitoxin system of *Mycobacterium tuberculosis*. *J Bacteriol* 204:e0005822. <https://doi.org/10.1128/jb.00058-22>
47. Maya-Hoyos M, Mata-Espinosa D, López-Torres MO, Tovar-Vázquez B, Barrios-Payán J, León-Contreras JC, Ocampo M, Hernández-Pando R, Soto CY. 2022. The *ctpF* gene encoding a calcium P-type ATPase of the plasma membrane contributes to full virulence of *Mycobacterium tuberculosis*. *Int J Mol Sci* 23:6015. <https://doi.org/10.3390/ijms23116015>
48. Chen L, Li X, Xu P, He Z-G. 2022. A novel zinc exporter CtpG enhances resistance to zinc toxicity and survival in *Mycobacterium bovis*. *Microbiol Spectr* 10:e0145621. <https://doi.org/10.1128/spectrum.01456-21>
49. Bernatowska E, Skomska-Pawliszak M, Wolska-Kuśniercz B, Pac M, Heropolitanska-Pliszka E, Pietrucha B, Bernat-Sitarz K, Dąbrowska-Leonik N, Bohynikova N, Piątosza B, Lutyńska A, Augustynowicz E, Augustynowicz-Kopec E, Korzeniewska-Koseła M, Krasieńska M, Krzysztopa-Grzybowska K, Wieteska-Klimczak A, Książek J, Jackowska T, van den Burg M, van Dongen JJM, Casanova J-L, Picard C, Mikołuc B. 2020. BCG Moreau vaccine safety profile and NK cells-double protection against disseminated BCG infection in retrospective study of BCG vaccination in 52 Polish children with severe combined immunodeficiency. *J Clin Immunol* 40:138–146. <https://doi.org/10.1007/s10875-019-00709-1>
50. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J Bacteriol* 178:1274–1282. <https://doi.org/10.1128/jb.178.5.1274-1282.1996>
51. Mangtani P, Abubakar I, Ariti C, Beynon R, Pimpin L, Fine PEM, Rodrigues LC, Smith PG, Lipman B, Whiting PF, Sterne JA. 2014. Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. *Clin Infect Dis* 58:470–480. <https://doi.org/10.1093/cid/cit790>
52. Ritz N, Hanekom WA, Robins-Browne R, Britton WJ, Curtis N. 2008. Influence of BCG vaccine strain on the immune response and protection against tuberculosis. *FEMS Microbiol Rev* 32:821–841. <https://doi.org/10.1111/j.1574-6976.2008.00118.x>
53. Karonga Prevention Trial Group. 1996. Randomised controlled trial of single BCG, repeated BCG, or combined BCG and killed *Mycobacterium leprae* vaccine for prevention of leprosy and tuberculosis in Malawi. *Lancet Lond Engl* 348:17–24. [https://doi.org/10.1016/S0140-6736\(96\)02166-6](https://doi.org/10.1016/S0140-6736(96)02166-6)
54. Rodrigues LC, Pereira SM, Cunha SS, Genser B, Ichihara MY, de Brito SC, Hijjar MA, Cruz AA, Sant'Anna C, Bierenbach AL, Barreto ML, Dourado I. 2005. Effect of BCG revaccination on incidence of tuberculosis in school-aged children in Brazil: the BCG-REVAC cluster-randomised trial. *Lancet* 366:1290–1295. [https://doi.org/10.1016/S0140-6736\(05\)67145-0](https://doi.org/10.1016/S0140-6736(05)67145-0)
55. Nemes E, Geldenhuis H, Rozot V, Rutkowski KT, Ratangee F, Bilek N, Mabwe S, Makhethe L, Erasmus M, Toefy A, et al. 2018. Prevention of *M. tuberculosis* infection with H4:IC31 vaccine or BCG revaccination. *N Engl J Med* 379:138–149. <https://doi.org/10.1056/NEJMoa1714021>
56. Lai R, Ogunzola AF, Rakib T, Behar SM. 2023. Key advances in vaccine development for tuberculosis—success and challenges. *NPJ Vaccines* 8:158. <https://doi.org/10.1038/s41541-023-00750-7>
57. Darrah PA, Zeppa JJ, Maiello P, Hackney JA, Wadsworth MH, Hughes TK, Pokkali S, Swanson PA, Grant NL, Rodgers MA, Kamath M, Causgrove CM, Laddy DJ, Bonavia A, Casimiro D, Lin PL, Klein E, White AG, Scanga CA, Shalek AK, Roederer M, Flynn JL, Seder RA. 2020. Prevention of tuberculosis in macaques after intravenous BCG immunization. *Nature* 577:95–102. <https://doi.org/10.1038/s41586-019-1817-8>
58. Grode L, Seiler P, Baumann S, Hess J, Brinkmann V, Nasser Eddine A, Mann P, Goosmann C, Bandermann S, Smith D, Bancroft GJ, Reyrat J-M, van Soolingen D, Raupach B, Kaufmann SHE. 2005. Increased vaccine efficacy against tuberculosis of recombinant *Mycobacterium bovis* bacille Calmette-Guérin mutants that secrete listeriolysin. *J Clin Invest* 115:2472–2479. <https://doi.org/10.1172/JCI24617>
59. Sable SB, Posey JE, Scriba TJ. 2019. Tuberculosis vaccine development: progress in clinical evaluation. *Clin Microbiol Rev* 33:e00100-19. <https://doi.org/10.1128/CMR.00100-19>
60. Seki M, Honda I, Fujita I, Yano I, Yamamoto S, Koyama A. 2009. Whole genome sequence analysis of *Mycobacterium bovis* bacillus Calmette-Guérin (BCG) Tokyo 172: a comparative study of BCG vaccine substrains. *Vaccine* 27:1710–1716. <https://doi.org/10.1016/j.vaccine.2009.01.034>
61. Fishbein S, van Wyk N, Warren RM, Sampson SL. 2015. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol Microbiol* 96:901–916. <https://doi.org/10.1111/mmi.12981>
62. 2017. Strain variation in the *Mycobacterium tuberculosis* complex: its role in biology, epidemiology and control. Springer International Publishing. <https://www.springer.com/gp/book/9783319643694>.
63. Sampson SL. 2011. Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clin Dev Immunol* 2011:497203. <https://doi.org/10.1155/2011/497203>
64. Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD, Warren RM. 2006. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (*esx*) gene cluster regions. *BMC Evol Biol* 6:95. <https://doi.org/10.1186/1471-2148-6-95>
65. Ly A, Liu J. 2020. Mycobacterial virulence factors: surface-exposed lipids and secreted proteins. *Int J Mol Sci* 21:3985. <https://doi.org/10.3390/ijms21113985>
66. Echeverria-Valencia G, Flores-Villalva S, Espitia C. 2017. Virulence factors and pathogenicity of *Mycobacterium*. In *Mycobacterium - research and development*. IntechOpen.
67. Brennan MJ. 2017. The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis vaccination. *Infect Immun* 85:e00969-16. <https://doi.org/10.1128/IAI.00969-16>
68. Rachman H, Strong M, Ulrichs T, Grode L, Schuchhardt J, Mollenkopf H, Kosmiadi GA, Eisenberg D, Kaufmann SHE. 2006. Unique transcriptome signature of *Mycobacterium tuberculosis* in pulmonary tuberculosis. *Infect Immun* 74:1233–1242. <https://doi.org/10.1128/IAI.74.2.1233-1242.2006>
69. Qian J, Chen R, Wang H, Zhang X. 2020. Role of the PE/PPE family in host-pathogen interactions and prospects for anti-tuberculosis vaccine and diagnostic tool design. *Front Cell Infect Microbiol* 10:594288. <https://doi.org/10.3389/fcimb.2020.594288>
70. Ates LS. 2020. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. *Mol Microbiol* 113:4–21. <https://doi.org/10.1111/mmi.14409>
71. Ates LS, Sayes F, Frigui W, Ummels R, Damen MPM, Bottai D, Behr MA, van Heijst JWJ, Bitter W, Majlessi L, Brosch R. 2018. RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in strong reduction of antigenic repertoire but little impact on protection. *PLoS Pathog* 14:e1007139. <https://doi.org/10.1371/journal.ppat.1007139>
72. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. 1991. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol* 29:2578–2586. <https://doi.org/10.1128/jcm.29.11.2578-2586.1991>
73. Krueger F, James F, EwelsP, AfyounianE, WeinsteinM, Schuster-Boeckler B, HulsemansG, sclamons. 2021. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Zenodo.
74. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>

75. Danecek P, McCarthy SA. 2017. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* 33:2037–2039. <https://doi.org/10.1093/bioinformatics/btx100>
76. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9:5114. <https://doi.org/10.1038/s41467-018-07641-9>
77. Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq, p 165–188. In Sun L, Shou W (ed), *Engineering and analyzing multicellular systems: methods and protocols*. Springer, New York, NY.
78. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* 29:24–26. <https://doi.org/10.1038/nbt.1754>
79. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
80. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32:1009–1015. <https://doi.org/10.1093/bioinformatics/btv688>