Exploring the zoonotic potential of *Mycobacterium bovis* using variant calling approaches

Ruma Banerjee
Centre for Development of Advanced Computing
Muthukumar Balamurugan
Centre for Development of Advanced Computing
Rajendra Joshi (rajendra@cdac.in)
Centre for Development of Advanced Computing

Research article

Keywords: SNP, zoonosis, Mycobacterium bovis, biomarker

Posted Date: May 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-24763/v1

Abstract

Background

Mycobacterium tuberculosis var. bovis is one of the causative agents of Tuberculosis primarily known to infect cattle and is a member of the Mycobacterium Tuberculosis Complex. *M.bovis* is zoonotic and infects human and other animals resulting in immense economic losses globally. *M.bovis* is often misdiagnosed and becomes a challenge for treatment and recovery, as *M.bovis* is intrinsically resistant to certain antibiotics. Hence, the need for accurate diagnostics for zoonotic tuberculosis.

Results

In order to discern the differences which lie between *M.tuberculosis* and *M.bovis* isolates we collected a global collection of *M.bovis* isolates from NCBI and carried out variant identification studies keeping *M.tuberculosis* as the reference. Clustering approaches like Principal component analysis and distance-based UPGMA methods helped to segregate isolates into different clusters of homogeneous and heterogeneous populations, which were further analyzed for variant identification. Methods like Joint Variant Calling using population-based studies was adopted for the *M.bovis* strains, which helped to discern high confidence polymorphisms for each set of populations. Four different variant callers were used to predict SNPs present in their genomic regions. Based on the predicted SNPs, *M.bovis* samples isolated from New-Zealand were identified as a heterogeneous fast evolving population, whereas the UK samples were identified as a slow evolving homogeneous population. The core-SNP identified for each population revealed the “fixed” mutations present within the populations, whereas, the total-SNP for heterogeneous population indicated presence of clonal subpopulations. A methodology for assignment of genomic coordinates to the reference SNP cluster id of *M.tuberculosis* entries in dbSNP was also developed and implemented, which helped to identify the percentage of known variants in a population.

Conclusions

Population-specific studies for global collection of *M.bovis* samples aided in identification of SNPs responsible for zoonosis. The core-SNP set identified across global *M.bovis* isolates provide a rationale for further studies to the underlying host-tropism along with aiding in as a robust genetic biomarker for distinguishing *M.bovis* form *M.tuberculosis* in addition to the already known RDs. The variation profile reported in this study can serve as potential biomarkers for identification of *M.bovis* isolates and provide a rationale for further studies to the underlying host-tropism.

Background

The genus *Mycobacterium* is known to cause the disease tuberculosis, which infects ~ 10 million people world-wide annually, according to Global Tuberculosis Report published by World Health Organization [1]. Within this genus, *Mycobacterium tuberculosis* complex (MTBC) is known to cause tuberculosis in humans as well as animals [2]. The members of the pathogenic MTBC family are known to be host-specific. *Mycobacterium tuberculosis* var. *bovis* (*M.bovis*) infects cattle and many other species, whereas *Mycobacterium tuberculosis* H37Rv (*M.tuberculosis*) and its related substrains are mainly known to infect humans [3, 4]. *M.bovis* is also zoonotic and can spread from animals to humans, is an identified public health problem globally and is also known to have intrinsic drug resistance towards several drug molecules [5–8]. *M.bovis* infection in humans is identical with that of *M.tuberculosis* infection, differing only in being non-transmissible among immunocompetent hosts [7, 9]. Infection due to *M.bovis*, causing Bovine Tuberculosis (BTB) is recognized as a One Health problem in certain parts of the globe where cattle-oriented domestication is practised. Close contacts with livestock and consumption of unpasteurized milk, raw or uncooked meat is the key source of zoonosis [6]. According to WHO reports, although BTB cases make up only a small portion of human tuberculosis disease burden, efforts to curb global TB by 2030 seems to be hindered by incidences of zoonosis [10]. Though TB remains to be an escalating problem worldwide, it is curable (non-treatable in cases of total drug resistance) and preventable, only if all efforts to cure it are intensified to reduce mortality and morbidity [11].

Continuous efforts of controlling global TB is delayed mainly due to less sensitive diagnostic tests, lack of effective vaccines and drugs, along with a rise in multi-drug resistant (MDR), extensively-drug resistant (XDR) and total-drug resistant (TDR) strains of TB [2].

The standard method of diagnosing BTB is the regular tuberculin skin test, which cannot distinguish between *M.tuberculosis* and *M.bovis* infections, moreover its results can be affected by cross-reactivity to BCG and other environmental mycobacteria [12]. Conventional diagnosis using culture tests takes time for confirmation, hence diagnosis gets considerably delayed [13, 14]. Several non-sequence based molecular typing techniques [15–17] have been used independently or in combination for genotyping. These methods have variable specificity, and are suitable to be used only in clinics with well-equipped microbiological laboratories [13]. Hence, a robust methodology which would help identify *M.bovis* isolates likely to infect hosts other than cattle would be useful in controlling BTB infection.

M.bovis AF2122/97 genome is 99.95% similar to *M.tuberculosis* H37Rv [18, 19] differing predominantly due to Insertion-deletions(indels) and Single Nucleotide Polymorphisms (SNP)s [2, 19], which in turn attribute towards their host-specificity [20]. MTBC members are also known for their low mutation rates and limited genomic diversity [21], which makes studying their variation profile important for identification as biomarkers. These polymorphisms obtained from analyzing whole genome sequence (WGS) data are capable enough to differentiate amongst populations and predict their host-specificity [22]. SNP profiles of *M.bovis* isolates capable of causing zoonosis may give us clue towards their changing host-associations [23]. Repositories of SNP data for TB community has been generated and is on the rise [24–29], but, all predictions till date have been done using individual samples, which have their own limitations and are known to include false-positives, due to low coverage, small read lengths and sequencing errors [30]. An effort to enlist the variation profile present within a cohort is still lacking as it becomes computationally challenging to predict SNP/Indel present across samples within a cohort during multi-sample variant prediction [31].

In the light of these facts, for effective control and treatment, a sequence-based improved and targeted rapid diagnostics for BTB is desirable to provide accurate identification. With the advent of massively-parallel next generation sequencing (NGS) techniques, bacterial populations can be sequenced to study population dynamics using WGS. Although heterogeneity analysis has been a regular phenomenon with respect to (w.r.t.) viral genomics, its application in bacterial genomes seems to be tricky. Heterogeneity refers to the genetic differences present within certain isolates of a genetically similar homogeneous population. Heterogeneity w.r.t variation profile in prokaryotic populations forms the basis of survival in stressed environment like drug resistance, or change in metabolic requirements, etc. and this provides seed to microevolution [32, 33]. Microevolution can be defined as gradual acquisition of mutations within a population to give rise to variations leading to speciation [34]. The mutations arising in a prokaryotic heterogeneous or a homogeneous population needs to be studied carefully keeping in mind the gene information, annotation and their functionality. In order to identify polymorphisms responsible for causing BTB, a cross-infecting *M.bovis* population, i.e., a heterogeneous population needs to be identified and compared against a *M.bovis* cohort capable of infecting only cattle and has low divergence ratio amongst isolates, i.e. a homogeneous population. Clustering approaches capable of distinguishing such features have been implemented in this study to identify individual populations from the global *M.bovis* isolates. The initial approach uses Principal Component Analysis (PCA), which help us identify individuals with similar variation profile [35]. The second approach for clustering of isolates was performed using distance-based UPGMA and maximum-likelihood (ML) based methods for the SNP data [36]. Based on PCA clustering and distance-based clustering, a homogenous population and a heterogenous population was identified to study the variant distribution using different variant calling approaches and their significance on each population type. A study of the variant distribution of the homogeneous and heterogeneous populations of *M.bovis* gives us an insight into the SNPs under selection and their distribution for each population type.

Approaches like Joint Variant Calling (JVC), (concept used for the first time on prokaryotes in this study) which predict variants present in a cohort, promises to overcome the shortcomings proposed by single sample variant calling (SVC) methods, as variants are analyzed simultaneously across all samples in a population [31, 37, 38]. JVC can predict variants for low coverage data in cohorts with high sensitivity. This approach works well for both homogeneous as well as heterogeneous population, wherein, population-specific biomarkers can be identified for each cohort. JVC in diverging population can also help identify SNPs which are under selection pressure and may give rise to phenotypic variations within the population. Hence, for non-model organisms like *Mycobacterium*, polymorphisms can be detected with more confidence within populations using different variant calling approaches like, Bayesian or heuristic [37]. Variant callers capable of handling cohort data, like FreeBayes and GATK use Bayesian approach to predict variants, whereas, VarScan2 and BCFtools uses pileup results along with heuristic approaches for variant detection [39–44]. JVC was performed on a heterogenous as well as a homogenous population using a combination of various tools with different methodologies, also capable of handling prokaryotic population data for best results. Variant annotation along with functional distribution analysis of the same was done to identify the high confident variants and their contribution towards gene functionality. The TB community has a catalogue of studies related to SNPs and Indels in MTBC genomes [26], but efforts to enlist and map all to their respective chromosomal position along with their Reference SNP cluster id (rsid) is lacking. Hence, rsids were assigned to the variants mapping them to their specific chromosomal position. These polymorphisms could be used by other TB researchers for further addition and future reference of SNPs based on their unique rsids. The current study is the first report of a comparative analysis of JVC approach versus single isolate variant detection in prokaryotes using homogenous and heterogenous cohort data, namely, *M.bovis* United Kingdom, henceforth abbreviated as UK [45, 46], and New Zealand (NZ) isolates [47] respectively. JVC approach promises to improve consistency with fewer artefacts, and hence, more accurate variants were detected for homogenous as well as heterogeneous distribution of population [37, 48], that have the potential to be used as biomarkers for diagnostics and treatment purposes apart from aiding in improvising our understanding of the pathogen in each population. The SNPs identified across a heterogeneous population as compared to the homogeneous population also throws light on the differential metabolic capabilities of the isolates which may explain certain aspects of their zoonosis.

We also aim to enlist/catalogue the distinct polymorphisms present between *M.tuberculosis* and *M.bovis* by performing JVC on the global *M.bovis* population to detect SNPs which occur across all samples to identify a set of "core SNP" of *M.bovis*, which may help in understanding host-tropism in bovine hosts apart from adding onto existing list of known polymorphisms. These core-SNPs in addition to Regions of Difference (RD), may be used for lineage identification [28] in *M.bovis*, in turn aiding in identification of specific biomarkers.

Results

Population distribution and identification of SNPs

A total of 705 *M.bovis* isolates were downloaded from public domain databases and later filtered as described in the Methods section. To understand the variant distribution across *M.bovis* population, SVC was performed across all 705 *M.bovis* isolates using GATK-HC and clustering was carried out using PCA and distance-based UPGMA methods using the vcf files generated for each isolate. Clustering was also performed using ML-based methods which showed similar results like UPGMA method (Fig. 1a,b&c). Hence, results pertaining to PCA and UPGMA have been discussed further. In the PCA plot, The NZ and Mexico populations are seen to be spread across both the principal axis and form heterogeneous clusters, whereas, UK, USA and Eritrea populations are seen to form single independent clusters on the PCA plot (Fig. 1a). As is apparent from Fig. 1b&c, the *M.bovis* isolates from NZ (pink) do not seem to form a definite cluster and are found to span across the phylogenetic tree (polyphyletic). Whereas, the *M.bovis* isolates from UK (green) are found to be isolated in a single clade (monophyletic), though sharing similar variation profile with other *M.bovis* isolates from UK, USA and others. The Mexico population is also seen to co-occur with other populations like NZ and “mixed”, it was found to be almost missing in the UK clade. Hence, in terms of membership to other clusters, NZ was ranked the highest, as its presence was observed across all clades, followed by Mexico; therefore, NZ was selected for further study as a heterogeneous cluster. Isolates from USA were also found to be monophyletic and co-occurring with UK, NZ and other *M.bovis* isolates, but were not selected further for study because of their lower number of sample counts. Hence, the results obtained from both clustering methods indicate NZ to be a heterogeneous or a fast evolving population, and the UK to be a slow evolving homogeneous population. Based on the results obtained from the clustering methods, *M.bovis* isolates obtained from NZ and UK were identified as heterogeneous and homogeneous population respectively (Fig. 1a&b).

Variant Identification in M.bovis populations

In order to understand the variation between the SNP profile of homogeneous and heterogeneous populations, JVC was implemented using four variant callers, viz., BCFtools, GATK, VarScan and FreeBayes for both NZ and UK populations independently (Table 1). SNPs present in all samples predicted by individual variant caller for both populations were extracted and named as “core-UK-(tool-name)” and “core-NZ-(tool-name)” respectively. An intersection of variants predicted by all tools across samples in homogeneous as well as heterogeneous clusters, is also reported and was found to reduce false positives. These were named as “total-UK” and “total-NZ” SNPs respectively (Fig. 2a&b). ~50% of the core-SNPs in both the populations mapped onto existing rsids as compared to ~ 20% in NZ total-SNP and ~ 35% in UK total-SNP datasets. With the implementation of JVC approach for NZ population, a higher number of SNPs were reported by all tools when compared with the UK population (Table 1). Analysis of total-SNP datasets between both populations was carried out to identify common and unique SNPs. Paralogous genes corresponding to unique and common SNP datasets were identified for both populations in order to evaluate the number of spontaneous mutations occurring in single copy genes of each population.

To identify a core set of high confidence SNPs for UK and NZ populations, a consensus of all four variant callers was used, which led to 1866 SNPs (“core-UK”) in UK and 1953 SNPs (“core-NZ”) in NZ populations (Fig. 3a&b). As observed in Table 1, the total number of SNPs predicted by each tool for NZ when compared against the respective tool outcome for UK was found to be almost twice as high. This observation is in contrast to that obtained for SNPs in core-UK and core-NZ that were found to be comparable. The total-UK SNPs are similar in count to that of core-UK SNPs, whereas, in the case of heterogeneous NZ population, the total-NZ SNPs predicted are found to be high as compared to the core-NZ dataset.

Table 1
Variant prediction across UK and NZ samples

JVC												
	New Zealand population						UK population					
Tools	Total-NZ	no. of rsid mapped SNP (total-NZ)	% of rsid mapped NZ	Core-NZ	no. of mapped rsid (core-NZ)	% rsid mapping (Core-NZ)	Total-UK	no. of rsid mapped SNP (total-UK)	% of rsid mapped UK	Core-UK	no. of mapped rsid (core-UK)	% rsid mapping (Core-UK)
BCFtools	7094	1297	18.3	2093	1165	55.7	3687	1259	34.1	2312	1183	51.2
GATK-HC	6482	1279	19.7	2050	1160	56.6	3259	1252	38.4	2236	1146	51.3
VarScan2	6672	1301	18.3	1980	1141	54.5	3425	1233	36.0	1881	1011	53.7
FreeBayes	7747	1294	19.9	2097	1169	56.6	5112	1265	24.7	2328	1187	50.9
SVC												
BCFtools							4497			2297		
GATK-HC							3312			2236		
VarScan2							3897			1789		
FreeBayes							5329			2423		

A comparative analysis of the single sample variant call was also performed using all four variant callers respectively for each UK sample. As observed in Table 1, the number of merged SNPs obtained using SVC is high when compared to JVC approach for each of the individual variant callers. This may be due to the artefacts related to single sample variant prediction [38]. Strikingly, GATK showed comparable results in single sample calls as well as JVC predictions. This is attributed to the fact that GATK-HC uses a scalable variant calling approach, wherein, per sample runs are done to generate an intermediate genomic vcf (gVCF) output that is further processed for joint genotyping [41], a step absent in other variant callers.

A functional distribution of the total-UK and core-UK SNPs revealed maximum number of variations to be present across Intermediatory metabolism and respiration (IMR) genes (596 and 433 respectively), followed by cell wall & cellular processes (576 and 429 respectively), conserved hypothetical (528 and 387 respectively), virulence, detoxification & adaptation and insertion sequences & phages (Table 2). Largest number of SNPs were found to be varied in PE/PPE genes (187) and least variation was observed in virulence, detoxification and adaptation (11) between total-UK and core-UK datasets. The functional distribution of total-NZ and core-NZ SNPs also followed a similar trend as seen in the case of the UK population (Table 2). Notable SNP differences were observed for all functional distribution groups between total-NZ and core-NZ SNPs, with highest SNP differences belonging to IMR, conserved hypotheticals and cell wall & cellular processes. A detailed list of the functional distribution of the SNPs for total-NZ, core-NZ, total-UK and core-UK can be found in supplementary table S1(a-d).

Table 2
Functional distribution of SNPs in UK and NZ

Total-UK (2712 SNPs)						**Core-UK (1866 SNPs)**			
Functional Category	**Number of SNP (a)**	**Synonymous SNP**	**Non-synonymous SNP**	**SNP in upstream region**	**no. of stop gain/loss SNP**	**Number of SNP (b)**	**Synonymous SNP**	**Non-synonymous SNP**	**SN up reg**
Regulatory protein	143	48	76	19	NIL	109	36	55	18
Virulence,detoxification&adaptation	77	24	45	7	1	66	21	39	6
PE/PPE	298	105	160	26	7	111	26	58	20
Information & Pathways	159	59	85	13	2	117	44	63	8
Lipid Metabolism	248	89	133	24	2	173	63	94	14
Cell wall & cellular process	576	174	328	66	8	429	124	245	52
Intermediatory metabolism & Respiration	596	184	339	65	8	433	129	246	54
Insertion sequences & phages	87	29	47	11	NIL	38	8	21	9
Conserved hypothetical & Unknown	528	160	273	82	13	387	121	195	60
Total-NZ (5561 SNPs)						**Core-NZ (1953 SNPs)**			
Functional Category	**Number of SNP (a)**	**Synonymous SNP**	**Non-synonymous SNP**	**SNP in upstream region**	**no. of stop gain/loss SNP**	**Number of SNP (b)**	**Synonymous SNP**	**Non-synonymous SNP**	**SN up reg**
Regulatory protein	270	79	151	36	4	107	35	56	16
irulence,detoxification&adaptation	185	59	99	24	3	66	20	40	6
PE/PPE	614	228	331	45	13	138	40	77	15
Information & Pathways	341	118	186	34	3	116	44	63	8
Lipid Metabolism	532	174	298	53	7	189	67	103	17
Cell wall & cellular process	1091	343	625	108	15	439	132	249	50
Intermediatory metabolism & Respiration	1298	431	721	126	20	458	144	252	56
Insertion sequences & phages	161	58	78	24	1	42	16	17	9
Conserved hypothetical & Unknown	1069	320	581	144	23	404	134	204	53

It is known that in heterogeneous populations, the frequency of SNPs due to spontaneous mutations is high as compared to homogeneous populations [49]. A comparative study between the total-UK and total-NZ SNPs revealed 1617 genes having 2313 SNPs to be common. SNPs associated with 73 and 934 genes belonging to the UK and NZ population respectively were found to be unique. A comparative study of SNP distribution between these populations aided in the identification of the frequency of SNPs present in paralogs and the frequency of SNPs occurring spontaneously. Eight paralogues were found to carry SNPs in the NZ population and the other 828 genes harboured spontaneous mutations. No paralogous genes were observed to carry SNPs in the UK population. Hence, the proportion of SNPs carrying spontaneous mutation in single copy genes in the heterogeneous NZ population was found to be more as expected. A further analysis of these spontaneous mutations would help identify genes under positive selection, if any.

Identification of M.bovis **specific SNPs**

A total of 351 SNPs (known as core-SNP) were found to be common across all 705 *M.bovis* isolates, which may aid in identification of a set of SNPs for distinguishing *M.bovis* isolates. Of these, 170 were non-synonymous, 104 were synonymous, 69 were present as upstream and eight SNPs were graded as “high” effect mutations by SNPeff annotations, which hampered protein function [50]. Of these, *pstA1* and *mmpL9*, both belonging to cell wall and cellular components category, were found to be truncated genes with altered functionality. A functional distribution of these 351 SNPs revealed highest number of missense-variants belonging to the IMR category. A complete list of the functional distribution of the core-SNPs can be found in Supplementary table S2. To test the utility of 351 core SNPs, *M.bovis* isolates available in NCBI SRA were chosen randomly, which are not a part of the 705 isolates mentioned above and analysed using GATK-HC. 100% prediction accuracy for the SNPs was obtained, irrespective of their geographic origin.

Discussion

SNP shapes the population structure of prokaryotes, wherein the SNPs are responsible for important bacterial phenotypes and may aid in their survival with the ability to infect and thrive on different metabolites [7]. *M.bovis* isolates available in public domain repositories were collected and single SVC was performed to identify SNPs. Population distribution analysis of the *M.bovis* isolates using their SNP occurrences through clustering techniques identified NZ to be a more heterogeneous population as compared to the UK population. PCA clusters samples and variables with similar profiles together. In our study, the

SNP distribution of the NZ population was found to be highly varied as it co-occurred with all other populations. This indicated a rapidly evolving population, that is usually responsible for underlying subpopulations giving rise to heterogeneity within the population. This heterogeneous population is found to be polyphyletic in the distance-based tree and also has the capability of infecting hosts other than cattle as reported by Crispell et al. [47]. UK isolates although were found to be sharing SNPs with other population *M.bovis* isolates, their occurrence was sporadic and were found as a monophyletic clade in the distance-based tree. The UK isolates were present as a single cluster on the PCA plot and accordingly, were found to be a homogeneous cluster having low divergence ratio. This hypothesis was further confirmed through variant calling based on individual populations. JVC was chosen to be the best method for distinguishing SNPs for homogeneous and heterogeneous populations. This method has the ability to identify variants with high confidence value and was preferred over SVC of multiple samples, which is underpowered due to several reasons. First, JVC performs variant analysis simultaneously across all samples, whereas, in SVC, samples are analyzed individually. Secondly, JVC can clearly distinguish between homozygous reference sites and sites with missing variant data, as it emits information of the variant site present in the population. This feature is missing when SVC is performed and hence reduces the chance of false positives in case of JVC. The read quality across *M.bovis* samples varied drastically. JVC also helped to overcome this issue as it uses information from high coverage samples present within the study to fetch confident variant site information for a sample which has lower coverage at that location. JVC allows to gather variant information across all samples, thereby reducing false positives. SVC methods lack the capability of handling multiple samples at a time and hence, merging results from single sample calling would result in false discoveries. Moreover, the differences observed in the total-SNPs was found to be responsible for the overall population dynamics, which gets enlisted as high confidence variants in case of JVC as compared to SVC. Hence, variant calling using JVC method was preferred which helped us to overcome such artefacts and predict variants by joint analysis of multiple samples.

Through JVC, significant differences could be observed in the distribution pattern for SNP calls between UK and NZ isolates. The core-SNP for each population revealed the “fixed” mutations present within the population, whereas the total-SNP for respective population indicated presence of clonal subpopulations. The higher the difference between the total-SNP and core-SNP for a population, the higher the heterogeneity quotient in the population. This result reveals microevolution to be predominant in NZ population, which is in cognizance to the heterogeneous SNP distribution of a population, whereas the SNPs present in the UK *M.bovis* isolates are likely to be fixed, as the total and the core-SNP were comparable in terms of numbers. This phenomenon was also reported while mapping rsids to the SNPs identified in homogeneous as well as heterogeneous populations. Rsids were mapped to the chromosomal positions of SNPs which indicate that these SNP positions have already been reported earlier. Rsids when mapped onto the core-SNP for each population, were found to be consistent, where, ~ 50% SNPs mapped onto known SNPs in each population. These SNPs may be considered as “fixed” SNPs within the population. While comparing the percentage of SNPs having rsids in total-NZ vs total-UK, NZ was found to have lesser number of rsid mapping when compared with total-UK SNPs irrespective of the variant caller. In the case of UK isolates, it is a slowly evolving homogeneous population and hence, the SNPs occurring in the population are the ones which have already been reported in the past and the total number of new SNPs is less. Whereas, in the case of heterogeneous NZ population, most of the SNPs were found to be “recent” and hence have not been reported earlier. Hence total-NZ SNPs with rsid mapping is comparatively low irrespective of the variant caller. SNPs present in the NZ population were also evaluated for their presence in paralogous genes, because, presence of SNPs in the paralogous gene groups helps the bacteria to adapt to the various changing environments [51]. Moreover, the presence of SNPs in the single copy genes due to spontaneous mutations generate heterologous bacterial populations as reported earlier by Yu [32]. Hence, in order to adapt to different hosts, NZ being a fast evolving heterogeneous population, SNPs were found to occur in the paralogous genes as well as single-copy genes, whereas, no SNPs were found to occur in the paralogs of the homogeneous UK population. However, as observed, the SNPs occurring in NZ outnumbered the SNPs occurring in the single copy genes of UK population.

Majority of SNPs associated with the IMR genes were found to be present in the NZ population. This phenomenon can be explained by the fact that Mycobacteria are faced with constant challenge of rerouting their metabolic activity with respect to their current environment, either in replicative growth or in non-replicative latent phase [52]. A lack of functional *pykA* gene in *M.bovis* due to a SNP "E220D" makes it incapable of utilizing glycerol as the sole carbon source for energy generation [53]. Searches from the UniProtKB suggests E220 to be a Magnesium (metal) binding site. Recent studies by Snášel and Pichová [54] also suggested the growth of Mycobacteria to be directly proportional to the divalent dependency of Mg^{2+} and Mn^{2+} ions for *pykA* gene. Hence, any isolate carrying a substitution E220D may not be capable of binding to Mg^{2+}, thereby making it incapable of utilizing glycerol as a carbon source. It was observed that in our study, all UK isolates had the E220D substitution in the *pykA* gene, whereas, none of the NZ isolates had the same (Supplementary table S1). This feature may be attributed to the fact that as NZ is a heterogeneous evolving population capable of cross infecting other hosts. The NZ isolates evolve themselves to utilize other sources of energy like glycerol, which may not be a case with the UK isolates. A fundamental feature of bacterial adaptation is also their ability to respire and sustain metabolism and generate ATP via oxidative phosphorylation [52]. An organism with reduced genome, such as *M.leprae* is limited in performing only anaerobic respiration by utilization of limited substrates [55]. Whereas, early studies have proved that *M.tuberculosis* grown *in vitro* were capable of utilizing exogenous substrates for respiration and survival [52, 56]. This feature is observed in a heterogeneously diversifying population, like NZ *M.bovis* isolates wherein, maximum number of SNPs are observed within the genes responsible for metabolism and respiration, which in turn supports their ability to infect hosts other than cattle, having wider availability of substrates. For instance, *Mycobacteria* have evolved highly specialized systems to extract nutrients from their hosts, in the form of lipids. One such important transporter needed for scavenging lipids from hosts is the *mce1* transporter complex [57]. A comparative study of the SNPs involved in *mce1* transporter complex in both populations revealed more number of SNPs to be present in the NZ heterogeneous population as compared to the UK homogeneous population (Supplementary Table S1). These additional SNPs may help evade the host immune responses to be able to infect more number of hosts. Hence, these observations can be attributed to the fact that *M.bovis* isolates belonging to NZ population are capable of zoonoses, which is in agreement with BTB infection in New Zealand reported by Crispell et al. [47].

The cell wall components constitute an important part of virulence and host-specificity in *Mycobacterium* [58–60]. The SNP distribution in the cell wall components contribute to pathogenesis and virulence and also forms the interface for host-pathogen interaction. The NZ isolates show a larger variation in their cell wall genes as compared to the UK isolates. This also may be attributed to the fact that this population has the capability to infect a wider range of hosts by evading immune system, from cattle to wildlife [47].

By using JVC methods, we could list the SNPs present in *M.bovis* isolates which may be responsible for zoonoses when compared against a cohort showing low divergence. A core-UK set for identifying UK isolates and a core-NZ set for identifying NZ isolates were also identified using this methodology. The mutations present in the *pstA1* and *mmpL9* genes along with other genes of the cell wall components of *M.bovis* may help accommodate the cell wall so as to increase its capability to infect multiple hosts. As reported in literature earlier, a mutant *pstsA1* gene in *M.tuberculosis* renders the bacterium hypersensitive to various stress conditions, as this gene is responsible for phosphate uptake along with inducing expression of several PE and PPE genes in the wild type. Most of the PE PPE genes are localized to the cell wall for maintaining the cell wall integrity. It was also observed by Ramakrishnan et al, that the permeability issues created by the mutant *pstsA1* gene could be replaced by deleting *pe19* gene [61]. Hence, we hypothesize that the observed mutations pertaining to the cell wall and associated genes, like the several PE-PPE genes in NZ population may together contribute in modulating the *M.bovis* envelope to resist different stress conditions while infecting multiple hosts. Another biomarker gene, *mmpL9* was found to be mutated across all *M.bovis* isolates, when compared against *M.tuberculosis*. This gene although is involved in inhibition of phagosomal maturation and oxidative stress management during infection, mutant strains were found to survive effectively in mouse infection models [62]. Hence, an integrated system-level approach to study these SNPs would help us understand the mechanism of zoonoses amongst these isolates. The core-SNP set consisting of 351 SNPs is a robust genetic marker for identification of *M.bovis* isolates in addition to already known RDs, as these were capable of identifying *M.bovis* isolates globally.

Conclusions

To conclude, this study promises to be a robust strategy for identifying traits in the form of variants present across a diverging as well as a slowly evolving population. By using strategies like PCA and distance-based UPGMA methods for clustering variants, an overall idea about the population distribution helped in identification of most evolving bacterial populations undergoing rapid mutational changes. These highly evolving divergent populations may have the potential of infecting multiple hosts or may be undergoing mutational changes under selective pressure, like the multi and extensively drug resistant strains of *Mycobacteria*. Once these populations are identified, variant calling approaches like JVC would aid in identification of specific variants in turn facilitating identification of specific biomarkers, leading to better diagnostics and cure.

Methods

Data Collection and Processing

Whole genome sequencing data of 705 *M.bovis* isolates were downloaded from NCBI Sequence Read Archive available as of March 2019 (Table 3) [45–47]. *M.bovis* samples were downloaded from NCBI SRA for cattle as host. A total of 323, 155, 139, 39, 22 and 14 samples were isolated as *M.bovis* infecting NZ, Mexico, UK, USA, Uruguay and Eritrea cattle respectively. The rest 13 *M.bovis* isolates from Guatemala, Costa Rica, Panama, South Africa, Brazil and Canada respectively were combined together as “mixed” population (Table 3). Ten additional *M.bovis* samples which are not a part of the primary dataset were randomly downloaded from NCBI SRA for further testing. Reads for each *M.bovis* isolate was checked for quality using FastQC [63]. Read length varied from 70–250 base pairs across samples. Read quality > 28 were retained. TrimGalore [64] was used to trim the reads such that > 80% of the read length were retained, as trimming of reads and retaining the good quality bases increases the accuracy of the analysis [65].

Table 3
Details of *M.bovis* isolates

Population name	Country of isolation	SRP study	no. of samples
New Zealand	New Zealand	SRP090533	192
	New Zealand	ERP005265	131
UK	UK	ERP001418	26
	UK	ERP010079	113
USA	USA	SRP053287	39
Mexico	Mexico	SRP108793	155
Uruguay	Uruguay	SRP102621	22
Eritrea	Eritrea	SRP105418	14
Mixed	Guatamala	SRP050038	3
	Costa Rica	SRP050038	3
	Panama	SRP050038	1
	South Africa	SRP020616	2
	Brazil	SRP132657	1
	Canada	SRP129014	3

Population distribution and JVC of *M.bovis*

Reference mapping for all *M.bovis* isolates were carried out using bwa-mem (0.7.17-r1188) as the aligner and *M.tuberculosis* H37Rv (Refseq id: NC_000962.3) genome as reference [66, 67]. Samtools/bcftools was used for sorting, indexing and generating pileup formats for each sample [44]. Picard was used to

identify and tag duplicate reads using the mark duplicates tool [68]. Lineage identification and validation was performed using RD-Analyzer [69]. Single sample variant identification for all 705 *M.bovis* isolates was performed using GATK HaplotypeCaller (GATK-HC) (version 3.8-0-ge9d806836), whereas, for 139 UK samples, BCFTools (1.9-172-g00011a5+), VarScan2 (v2.3.8) and FreeBayes (v1.1.0-44-gd784cf8-dirty) were also used. SNPs present in samples were merged to study the total number of variants called by each tool using BCFmerge [44]. The common SNPs across all isolate for each tool was inferred using GATK CombineVariants [41] (Fig. 4).

JVC was performed for the global collection of 705 *M.bovis* isolates using four variant callers, viz., GATK-HC [41, 40, 43], FreeBayes [70], VarScan2 [39] and Samtools/BCftools [44]. A proposed consensus of the variants predicted by above tools across all isolates is hypothesized to serve as a biomarker for global *M.bovis* isolates. Studies were carried out accordingly, but the results could not be reproduced when tested across other randomly chosen *M.bovis* isolates which were not a part of the present study (data not shown). Hence, an assessment of the population structure of *M.bovis* isolates was made by reconstructing a distance tree based on the UPGMA algorithm with 100 bootstrap replicates [36] using the vcf files obtained for all 705 samples. A maximum-likelihood tree of the 705 samples was also reconstructed using RAxML software with 100 bootstrap replicates [71]. PCA was carried out across all vcf files using the genotype-by-sequencing (GBS) tool for population genetic analysis in R [72]. Identification of homogeneous and heterogeneous population cluster was performed based on the results obtained from PCA, Maximim-likelihood and UPGMA methods. The distance-based tree was visualized using iTOL v4 tree viewer [73].

JVC was performed for both homogeneous as well as heterogeneous clusters with four different variant callers capable of handling haploid organism [37]. Heuristic tools, viz., VarScan2, BCFtools, Bayesian based tools viz., FreeBayes, Haplotype-based tools like GATK-HC were used for JVC, with ploidy "1". Multithreading option for GATK-HC enabled rapid and timely prediction of SNPs. BLAST package (version 2.6.0) was used to predict paralogues and assign rsids to the variants with dbSNP entries for *Mycobacterium* [74, 75]. Paralogs and their corresponding SNPs in each population were detected using BLASTn with a threshold of 80% “query coverage” and e-value of $<e^{-20}$. In-house custom scripts were used for data processing and data analysis. All variant calling tools were highly compute intensive and required long stretches of computational runs.

JVC was also performed for all 705 *M.bovis* isolates using GATK-HC to identify the common set of SNPs present across global *M.bovis* isolates used for this study. These SNPs were further tested for their presence in randomly selected *M.bovis* samples, which were not a part of the original 705 *M.bovis* isolates with accession numbers SRR1791699, SRR1791716, SRR1791718, SRR1791784, SRR1791808, SRR1791830, SRR1791840, SRR1791854, SRR1791885, SRR1791888. All predicted variations were annotated using SNPeff and SNPsift [50] packages. Variants were classified according to their functional category as reported in Tuberculist [66].

Assignment of rsid

Rsid accession number are available for *Mycobacterium* polymorphic sites in NCBI dbSNP available at (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/archive/mycobacterium_tuberculosis_1773/). Currently NCBI has stopped support for SNPs of non-human organisms, but data can be accessed through the above link (personal communication). Rsids for non-model organisms, such as *M.tuberculosis* lack genomic coordinates. Hence, a strategy to map rsids to specific chromosomal position for *M.tuberculosis* was implemented (Fig. 5). *M.tuberculosis* SNP data entries (37388) along with flanking sequences were downloaded from NCBI dbSNP.

Sequence alignment of these entries using standalone BLAST version version 2.6.0 [76] with *M.tuberculosis* H37Rv gene sequences as database, resulted in 45938 matches. As the number of matches were more than the query sequences, filtering was required for entries with more than one match. The BLAST output was filtered for mismatches more or less than 1 along with removal of entries with unequal subject and query length. Entries with unique and repeated rsids were analysed further. This resulted in 33565 unique rsids, along with 114 repeated rsids. It was found that the query sequences had flanking regions of varying lengths(60, 120 and 255). Hence, high-scoring Segment Pairs matching with repeated rsids were further filtered based on length of flanking sequence being 255. After filtration, 42 repeated rsids were retained and rsids corresponding to 33607 chromosomal positions were mapped. It was found that by changing the order of filtering, the number of genes being assigned rsids reduced. Two distinct rsids were assigned to the same chromosomal position in 282 entries, and hence were discarded and only unique rsid mapping onto single chromosomal position were retained. In total, we could assign 33325 unique rsids to corresponding positions of the *M.tuberculosis* genome and the vcf file has been provided as Supplementary file S3.

Declarations

Ethics statement:

Not applicable

Consent for publication:

Not applicable

Competing interests

The authors declare that they have no competing interests

Authors' contributions

RB conceived the project, carried out data analysis and wrote the manuscript, MB carried out the experiments and contributed to data analysis, RJ conceived the project and reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the BRAF facility of C-DAC for HPC requirements. The authors would also like to deeply thank Ms. K.Sunitha Manjari for her support and guidance throughout the project. This research work was funded by the National Supercomputing Mission (NSM) under Government of India.

Contributor information:

Ruma Banerjee, Email: rumas@cdac.in

Muthukumar Balamurugan, Email: muthukumarb@cdac.in

Rajendra Joshi, Email: rajendra@cdac.in

References

1. World Health Organization. Global Tuberculosis Programme: Global tuberculosis control: WHO report. Geneva: Global Tuberculosis Programme; 2019. v.
2. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol. 2018;16(4).
3. 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. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A. 2002;99(6):3684–9.
4. Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N, Hazbón MH. Phylogenomic analysis of the species of the Mycobacterium tuberculosis complex demonstrates that Mycobacterium africanum, Mycobacterium bovis, Mycobacterium caprae, Mycobacterium microti and Mycobacterium pinnipedii are later heterotypic synonyms of Mycobacterium tuberculosis. Int J Syst Evol Microbiol. 2018;68(1):324–32.
5. Kaneene JB, Miller R, Steele JH, Thoen CO. Preventing and controlling zoonotic tuberculosis: a One Health approach. Vet Ital. 2014;50(1):7–22.
6. Thoen CO, Kaplan B, Thoen TC, Gilsdorf MJ, Shere JA. Zoonotic tuberculosis. A comprehensive ONE HEALTH approach. Medicina. 2016;76(3):159–65.
7. Xiong X, Wang R, Deng D, et al. Comparative Genomics of a Bovine Mycobacterium tuberculosis Isolate and Other Strains Reveals Its Potential Mechanism of Bovine Adaptation. Front Microbiol. 2017;8:2500.
8. Kanji A, Hasan R, Ali A, et al. Single nucleotide polymorphisms in efflux pumps genes in extensively drug resistant Mycobacterium tuberculosis isolates from Pakistan. Tuberculosis (Edinb). 2017;107:20–30.
9. Malone KM, Rue-Albrecht K, Magee DA, et al. Comparative 'omics analyses differentiate Mycobacterium tuberculosis and Mycobacterium bovis and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. Microb Genom. 2018;4(3):e000163.
10. The challenges of preventing bovine tuberculosis. Bull World Health Organ. 2018;96(2):82–3.
11. Goletti D, Petruccioli E, Joosten SA, Ottenhoff TH. Tuberculosis Biomarkers: From Diagnosis to Protection. Infect Dis Rep. 2016;8(2):6568.
12. Ottenhoff TH. Overcoming the global crisis: "yes, we can", but also for TB. … Eur J Immunol. 2009;39(8):2014–20.
13. Druszczynska M, Wawrocki S, Szewczyk R, Rudnicka W. Mycobacteria-derived biomarkers for tuberculosis diagnosis. Indian J Med Res. 2017;146(6):700–7.
14. Schito M, Migliori GB, Fletcher HA, et al. Perspectives on Advances in Tuberculosis Diagnostics, Drugs, and Vaccines. Clin Infect Dis. 2015;61Suppl(3(Suppl 3):102–18.
15. Ravansalar H, Tadayon K, Ghazvini K. Molecular typing methods used in studies of *Mycobacterium tuberculosis* in Iran: a systematic review. Iran J Microbiol. 2016;8(5):338–46.
16. Moström P, Gordon M, Sola C, Ridell M, Rastogi N. Methods used in the molecular epidemiology of tuberculosis. Clin Microbiol Infect. 2002;8(11):694–704.
17. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev. 2006;19(4):658–85.
18. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence [published correction appears in Nature 1998 Nov 12;396(6707):190]. *Nature*. 1998;393(6685):537–544.
19. Garnier T, Eiglmeier K, Camus JC, et al. The complete genome sequence of Mycobacterium bovis. Proc Natl Acad Sci U S A. 2003;100(13):7877–82.
20. Garcia-Betancur JC, Menendez MC, Del Portillo P, Garcia MJ. Alignment of multiple complete genomes suggests that gene rearrangements may contribute towards the speciation of Mycobacteria. Infect Genet Evol. 2012;12(4):819–26.
21. Bryant JM, Schürch AC, van Deutekom H, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis. 2013;13:110.
22. Ford C, Yusim K, Ioerger T, et al. Mycobacterium tuberculosis–heterogeneity revealed through whole genome sequencing. Tuberculosis (Edinb). 2012;92(3):194–201.
23. Joshi D, Harris NB, Waters R, et al. Single nucleotide polymorphisms in the Mycobacterium bovis genome resolve phylogenetic relationships. J Clin Microbiol. 2012;50(12):3853–61.
24. Coll F, Preston M, Guerra-Assunção JA, et al. PolyTB: a genomic variation map for Mycobacterium tuberculosis. Tuberculosis (Edinb). 2014;94(3):346–54.
25. Coll F, McNerney R, Guerra-Assunção JA, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4812.
26. Stucki D, Gagneux S. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. Tuberculosis (Edinb). 2013;93(1):30–9.

27. Mikheecheva NE, Zaychikova MV, Melerzanov AV, Danilenko VN. A Nonsynonymous SNP Catalog of Mycobacterium tuberculosis Virulence Genes and Its Use for Detecting New Potentially Virulent Sublineages. Genome Biol Evol. 2017;9(4):887–99.
28. Dou HY, Lin CH, Chen YY, et al. Lineage-specific SNPs for genotyping of Mycobacterium tuberculosis clinical isolates. Sci Rep. 2017;7(1):1425.
29. Garcia Pelayo MC, Uplekar S, Keniry A, 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. 2009;77(5):2230–8.
30. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.
31. Huang Z, Rustagi N, Veeraraghavan N, et al. A hybrid computational strategy to address WGS variant analysis in > 5000 samples. BMC Bioinformatics. 2016;17(1):361.
32. Yu G. GenHtr: a tool for comparative assessment of genetic heterogeneity in microbial genomes generated by massive short-read sequencing. BMC Bioinformatics. 2010;11:508.
33. Navarro Y, Romero B, Bouza E, Domínguez L, de Juan L, García-de-Viedma D. Detailed chronological analysis of microevolution events in herds infected persistently by Mycobacterium bovis. Vet Microbiol. 2016;183:97–102.
34. Krysztopa-Grzybowska K, Lutyńska A. Microevolution of BCG substrains. Postepy Hig Med Dosw (Online). 2016;70(0):1259–66.
35. Sainani KL. Introduction to principal components analysis. PM R. 2014;6(3):275–8.
36. Powell JF. Dental evidence for the peopling of the New World: some methodological considerations. Hum Biol. 1993;65(5):799–819.
37. Zojer M, Schuster LN, Schulz F, Pfundner A, Horn M, Rattei T. Variant profiling of evolving prokaryotic populations. PeerJ. 2017;5:e2997.
38. Nho K, West JD, Li H, et al. Comparison of Multi-Sample Variant Calling Methods for Whole Genome Sequencing. IEEE Int Conf Systems Biol. 2014;2014:59–62.
39. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
40. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012;arXiv preprint arXiv:1207.3907.
41. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
42. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
43. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43(1110):11.10.1–11.10.33.
44. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
45. Otchere ID, van Tonder AJ, Asante-Poku A, et al. Molecular epidemiology and whole genome sequencing analysis of clinical Mycobacterium bovis from Ghana. PLoS One. 2019;14(3):e0209395.
46. Biek R, O'Hare A, Wright D, et al. Whole genome sequencing reveals local transmission patterns of Mycobacterium bovis in sympatric cattle and badger populations. PLoS Pathog. 2012;8(11):e1003008.
47. Crispell J, Zadoks RN, Harris SR, et al. Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. BMC Genom. 2017;18(1):180.
48. Olson ND, Lund SP, Colman RE, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015;6:235.
49. Schroeder JW, Yeesin P, Simmons LA, Wang JD. Sources of spontaneous mutagenesis in bacteria. Crit Rev Biochem Mol Biol. 2018;53(1):29–48.
50. Cingolani P, Platts A, Wang le. L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.
51. Tsuru T, Kobayashi I. Multiple genome comparison within a bacterial species reveals a unit of evolution spanning two adjacent genes in a tandem paralog cluster. Mol Biol Evol. 2008;25(11):2457–73.
52. Cook GM, Hards K, Vilchèze C, Hartman T, Berney M. Energetics of Respiration and Oxidative Phosphorylation in Mycobacteria. Microbiol Spectr. 2014;2(3):10.1128/microbiolspec.MGM2-0015-2013.
53. Keating LA, Wheeler PR, Mansoor H, et al. The pyruvate requirement of some members of the Mycobacterium tuberculosis complex is due to an inactive pyruvate kinase: implications for in vivo growth. Mol Microbiol. 2005;56(1):163–74.
54. Snášel J, Pichová I. Allosteric regulation of pyruvate kinase from Mycobacterium tuberculosis by metabolites. Biochim Biophys Acta Proteins Proteom. 2019;1867(2):125–39.
55. Eiglmeier K, Parkhill J, Honoré N, et al. The decaying genome of Mycobacterium leprae. Lepr Rev. 2001;72(4):387–98.
56. Bloch H, Segal W. Biochemical differentiation of Mycobacterium tuberculosis grown in vivo and in vitro. J Bacteriol. 1956 Aug;72(2):132–41.
57. Nazarova EV, Montague CR, La T, et al. Rv3723/LucA coordinates fatty acid and cholesterol uptake in *Mycobacterium tuberculosis*. Elife. 2017;6:e26969.
58. Marri PR, Bannantine JP, Golding GB. Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol Rev. 2006;30(6):906–25.
59. Banerjee R, Vats P, Dahale S, Kasibhatla SM, Joshi R. Comparative genomics of cell envelope components in mycobacteria. PLoS One. 2011;6(5):e19280.

60. Jackson M. The mycobacterial cell envelope-lipids. Cold Spring Harb Perspect Med. 2014;4(10):a021105.
61. Ramakrishnan P, Aagesen AM, McKinney JD, Tischler AD. Mycobacterium tuberculosis Resists Stress by Regulating PE19 Expression. Infect Immun. 2015;84(3):735–46.
62. Melly G, Purdy GE. MmpL Proteins in Physiology and Pathogenesis of M. tuberculosis. Microorganisms. 2019;7(3):70.
63. Andrews S FastQC: a quality control tool for high throughput sequence data. 2018. http://www.bioinformatics.babraham.ac.uk/projects/fastqc Accessed: 16 April, 2020.
64. Babraham Bioinformatics. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. URL https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 16 April 2020.
65. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12):e85024.
66. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;arXiv preprint arXiv:1303.3997.
67. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList–10 years after. Tuberculosis (Edinb). 2011;91(1):1–7.
68. Broad Institute. Picard Toolkit, GitHub Repository. http://broadinstitute.github.io/picard/ Accessed 16 April, 2020.
69. Faksri K, Xia E, Tan JH, Teo YY, Ong RT. In silico region of difference (RD) analysis of Mycobacterium tuberculosis complex from sequence reads using RD-Analyzer. BMC Genom. 2016;17(1):847.
70. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
71. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
72. Kagale S, Koh C, Clarke WE, Bollina V, Parkin IA, Sharpe AG. Analysis of Genotyping-by-Sequencing (GBS) Data. Methods Mol Biol. 2016;1374:269–84.
73. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256–9.
74. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
75. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
76. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

Additional Files

Additional file 1

Supplementary Table S1a - A detailed list of the functional distribution of the total SNPs present in NZ population (total-NZ) (XLSX 694 kb)

Additional file 2

Supplementary Table S1b - A detailed list of the functional distribution of the SNPs present across all isolates in NZ population (core-NZ) (XLSX 742 kb)

Additional file 3

Supplementary Table S1c - A detailed list of the functional distribution of the total SNPs present in UK population (total-UK) (XLSX 346 kb)

Additional file 4

Supplementary Table S1d - A detailed list of the functional distribution of the SNPs present across all isolates in UK population (core-UK) (XLSX 219 kb)

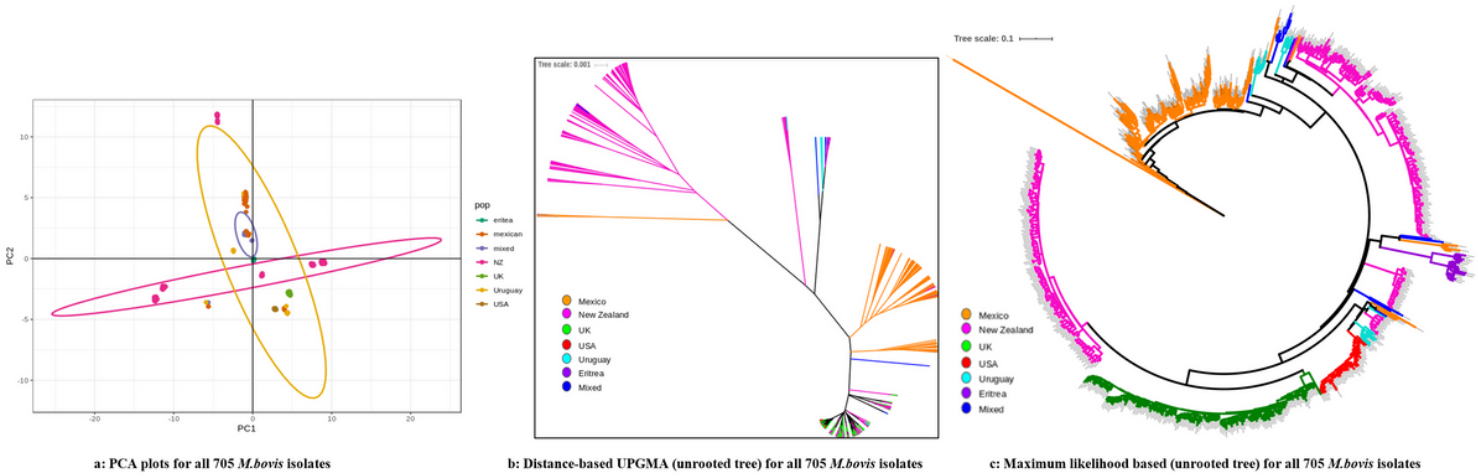
Additional file 5

Supplementary Table S2 - A detailed list of the functional distribution of the SNPs present across all isolates (core-SNP) (XLSX 137 kb)

Additional file 6

Supplementary file S3 - vcf file containing the list mapped unique rsids for Mycobacterium tuberculosis (vcf file 3 mb)

Figures



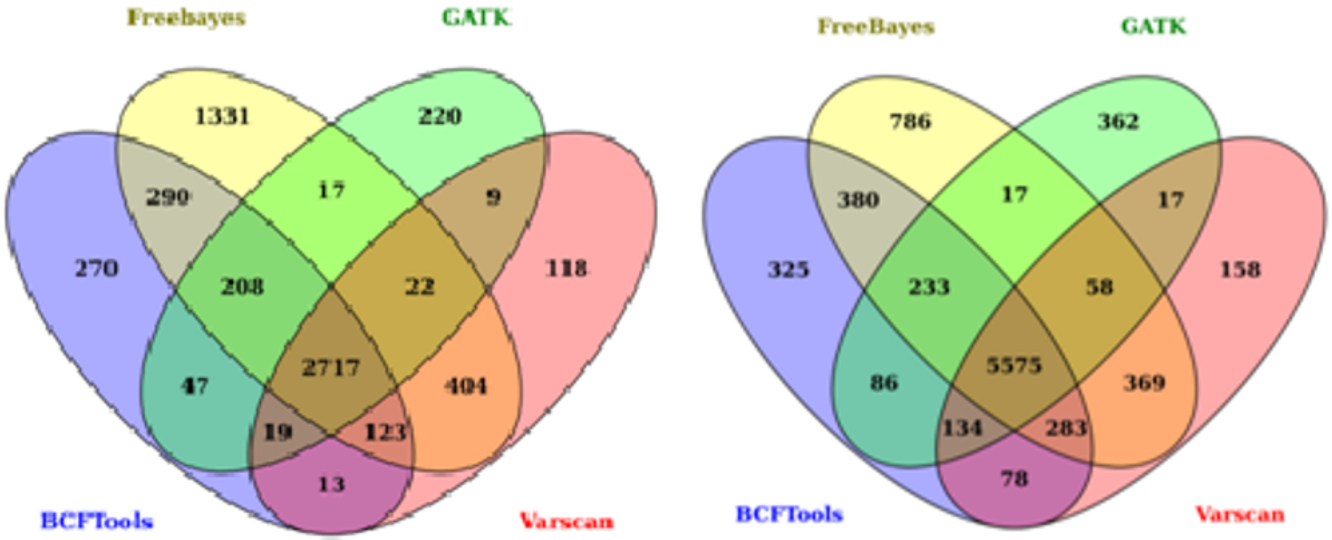
a: PCA plots for all 705 *M.bovis* isolates

b: Distance-based UPGMA (unrooted tree) for all 705 *M.bovis* isolates

c: Maximum likelihood based (unrooted tree) for all 705 *M.bovis* isolates

Figure 1

a: PCA plots for all 705 M.bovis isolates; b: Distance-based UPGMA (unrooted tree) for all 705 M.bovis isolates; Maximum likelihood based (unrooted tree) for all 705 M.bovis isolates

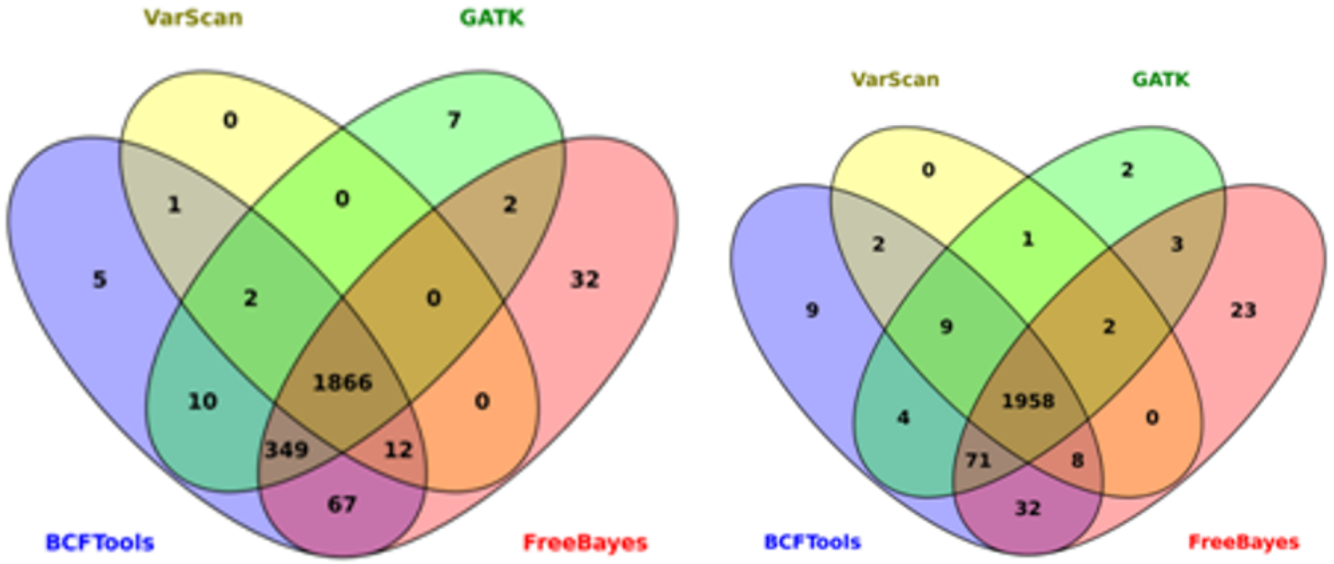


a: Total-UK SNPs predicted by al tools

b: Total-NZ SNPs predicted by all tools

Figure 2

a: Total-UK SNPs predicted by al tools; b: Total-NZ SNPs predicted by all tools



a: Core-UK SNPs predicted by all tools **b: Core-NZ SNPs predicted by all tools**

Figure 3

a: Core-UK SNPs predicted by all tools; b: Core-NZ SNPs predicted by all tools

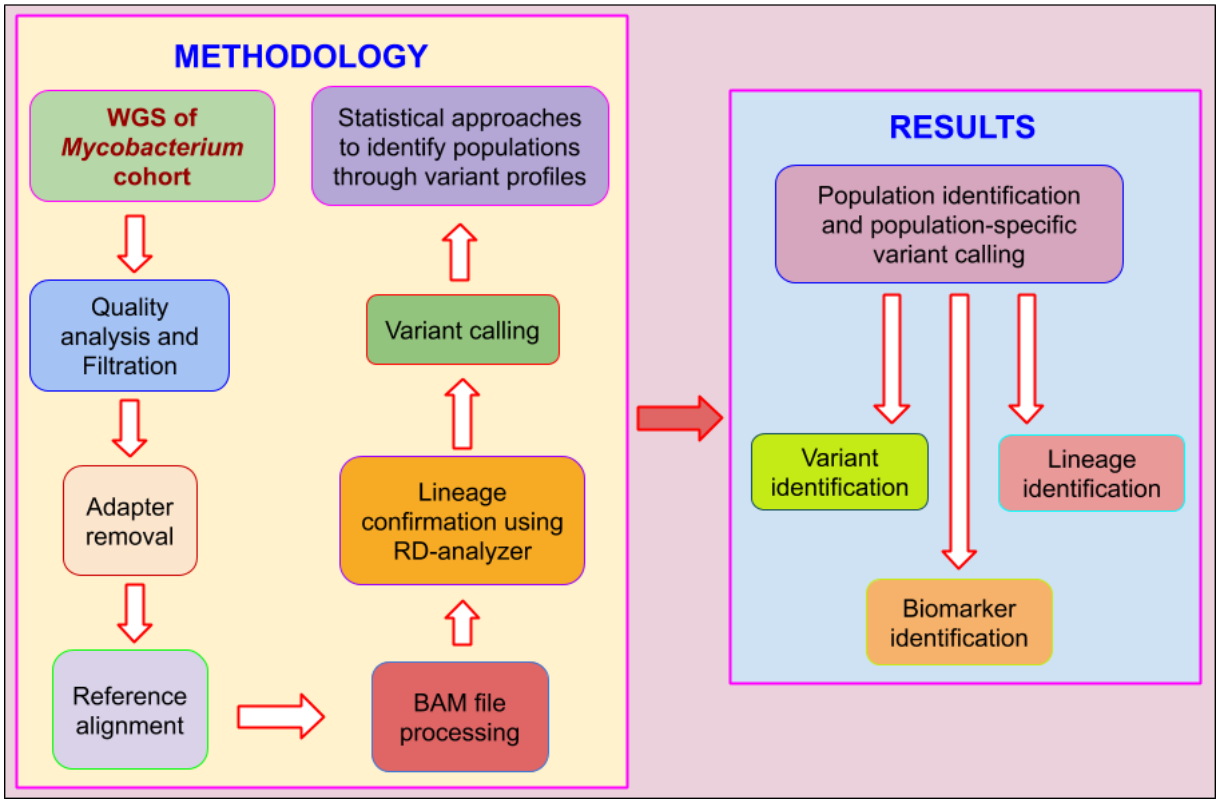


Figure 4

Methodology to identify population specific SNPs

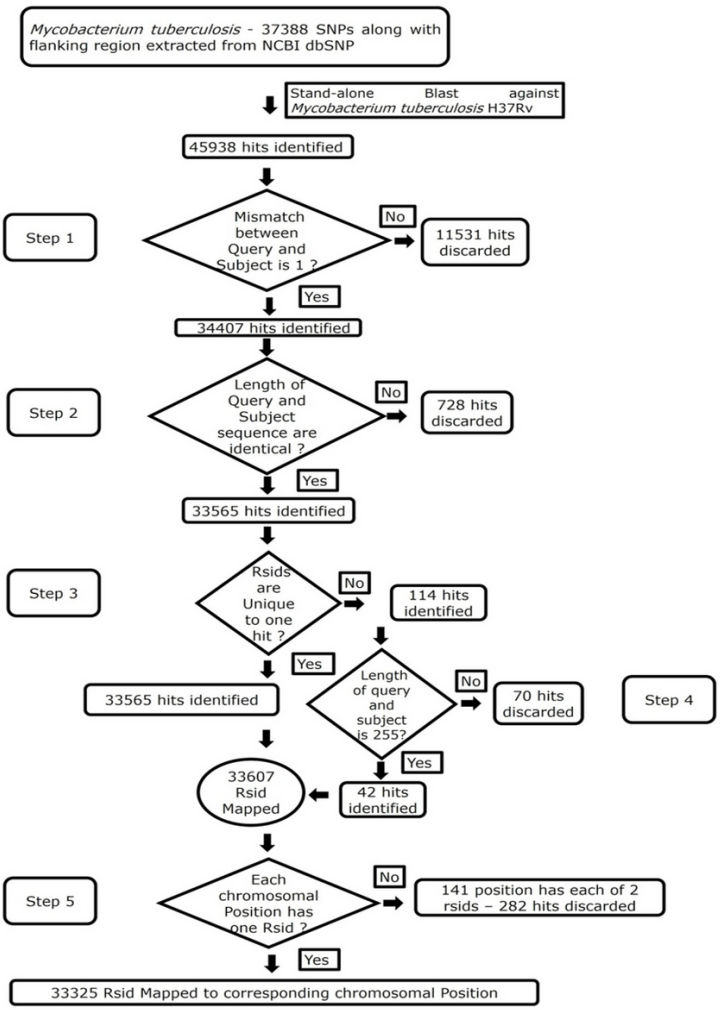


Figure 5

Flowchart for assignment of rsids

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarytableS1atotalNZ.xlsx
- SupplementaryTableS2.xls
- SupplementarytableS1ctotalUK.xlsx
- SupplementarytableS1dcoreUK.xlsx
- SupplementaryfileS3.vcf
- SupplementarytableS1bcoreNZ.xls