Published in final edited form as:

Tuberculosis (Edinb). 2013 January ; 93(1): 30–39. doi:10.1016/j.tube.2012.11.002.

Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database

David Stucki^{a,b} and Sebastien Gagneux^{a,b,*}

^aSwiss Tropical and Public Health Institute, Basel, Switzerland ^bUniversity of Basel, Switzerland

Summary

Recent advances in DNA sequencing have led to the discovery of thousands of single nucleotide polymorphisms (SNPs) in clinical isolates of *Mycobacterium tuberculosis* complex (MTBC). This genetic variation has changed our understanding of the differences and phylogenetic relationships between strains. Many of these mutations can serve as phylogenetic markers for strain classification, while others cause drug resistance. Moreover, SNPs can affect the bacterial phenotype in various ways, which may have an impact on the outcome of tuberculosis (TB) infection and disease. Despite the importance of SNPs for our understanding of the diversity of MTBC populations, the research community is currently lacking a comprehensive, well-curated and user-friendly database dedicated to SNP data. First attempts to catalogue and annotate SNPs in MTBC have been made, but more work is needed. In this review, we discuss the biological and epidemiological relevance of SNPs in MTBC. We then review some of the analytical challenges involved in processing SNP data, and end with a list of features, which should be included in a new SNP database for MTBC.

Keywords

evolution; SNP; mutation; genetic diversity; drug resistance; substitution

Why are SNPs important for our understanding of TB?

The declaration of tuberculosis (TB) as a global public health emergency in 1993¹ led to renewed efforts to study the biology of the *Mycobacterium tuberculosis* complex (MTBC). For many years, the main research focus was on individual genes and proteins, but the generation of the first *M. tuberculosis* genome sequence in 1998² opened the door for more comprehensive approaches. In particular, comparative genomics studies have helped us gain a better insight into the genetic diversity and phylogenetic relationships in MTBC^{3–5}. These studies showed that the different members of MTBC primarily associated with human disease (i.e. *M. tuberculosis* sensu stricto and *M. africanum*) are more genetically diverse than previously appreciated^{6,7}. Increasingly, various “omics” approaches in TB research are being combined into what is generally known as Systems Biology⁸. Systems Biology tries to

© 2012 Elsevier Ltd. All rights reserved.

*Corresponding author. Swiss Tropical and Public Health Institute, Socinstrasse 57, CH-4002 Basel, Switzerland. Tel. +41 284 83 69. sebastien.gagneux@unibas.ch.

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

understand complex biological systems by integrating data from various disciplines; in TB for example the comprehensive data from human, animal, and computational model systems^{9,10}. There is increasing evidence that, in addition to environmental factors and human genetics, strain variation in MTBC plays a role in the outcome of TB infection and disease¹¹. Hence, there is a need to better understand the global diversity of MTBC, and determine if and how this diversity has relevance for global TB control^{12,13}. The advent of next-generation DNA sequencing (NGS) methods is likely to facilitate this task, and indeed, many genome-sequencing projects of MTBC clinical isolates are currently underway¹⁴. More than 3,800 raw genome sequences of MTBC strains have already been deposited on public sequence read archives (Figure 1), and it is safe to assume that this number will continue to grow rapidly as sequencing costs keep decreasing^{15,16}.

In contrast to the relative ease with which DNA sequencing data can be generated today, extracting useful information and compiling these in a user-friendly manner is less straightforward. In particular, thousands of genetic polymorphisms have been extracted from whole genome sequences, but the TB research community currently lacks a centralized database, which would allow accessing and handling these data more efficiently. Several TB-specific databases have been created over the last years, including genome browsers, genotyping- and drug resistance databases¹⁷, but despite these existing platforms, we lack a centralized and comprehensive repository for data on strain-specific genetic variation in MTBC. Particularly, the field would benefit greatly from a new database compiling all known single nucleotide polymorphisms (SNPs) in MTBC. Ideally, such a database should include proper annotation of these SNPs as well as all relevant metadata. Considering the increasing number of MTBC genomes becoming available, the number of MTBC SNPs identified in the coming years will surely increase by one or more orders of magnitude.

In this review, we start by summarizing the nature of SNPs in MTBC, how SNPs can be used to define phylogenetic relationships between strains, and how they might impact on the phenotype of particular MTBC variants. We then elaborate on how new SNP data can be obtained with NGS technologies, and discuss some of the analytical challenges involved. We end by advocating for a new, user-friendly SNP database for MTBC.

What are SNPs and how many do we observe?

SNPs are the most common form of genetic variation in MTBC, after insertions and deletions (InDels). A total of 9,037 SNPs were discovered by sequencing 21 clinical strains of MTBC⁵. Generally, SNPs represent single nucleotide differences between at least two DNA sequences. The term “SNP” is often used interchangeably with “mutation”, “polymorphism” or “substitution”. Strictly speaking, a change in a single base pair is generally referred to as a (point) mutation, and happens through errors during DNA replication or as a consequence of DNA damage. Such a mutation represents a relatively rare change from the “normal” base to a mutant form, and is most likely to be neutral or (slightly) deleterious; beneficial mutations can of course occur, but they are generally much less likely. Mutations that are highly deleterious will be rapidly removed by purifying selection, whereas beneficial mutations will increase because of positive selection. In addition, any mutation can increase in frequency as a consequence of random genetic drift. When an allele reaches a certain frequency in the population, we refer to it as a polymorphism (i.e. the coexistence at a specific locus of two or more alleles in a given population). The threshold for defining a new variant as a “polymorphism” as opposed to a “mutation” is arbitrary, and e.g. is set at 1% for human variants (i.e. the minority allele has to be present at a frequency of at least 1% in a given human population for the corresponding nucleotide position to be referred to as “polymorphic”). Below this threshold frequency, this new variant would be referred to as a single nucleotide “mutation”. If a new

variant becomes fixed in a given population (i.e. 100% of the members of a given population have the new variant), this variant will be referred to as a “substitution” rather than a “polymorphism”¹⁸.

In addition to the difference between single nucleotide mutation, polymorphism, and substitution, biologist often differentiate between “natural polymorphisms” and “adaptive mutations” such as those conferring drug resistance. Natural polymorphisms can be thought of pre-existing variation which defines the overall genomic diversity of a population or a particular strain background, while adaptive mutations represent *de novo* acquired changes driven by a particular selective force (e.g. exposure to antibiotics, further discussed below). However, it is not always straightforward to discriminate between these different types of mutations. For example, the intrinsic resistance of *Mycobacterium bovis* to pyrazinamide is due to an amino acid change in *pncA* (Rv2043c)¹⁹. This drug resistance-conferring mutation clearly predates the introduction of pyrazinamide and can therefore be considered a natural polymorphism. Moreover, as *all* strains of the classical *M. bovis* clade harbour this mutation, one could also refer to this mutation as a natural substitution when comparing all classical *M. bovis* to the rest of MTBC. For the sake of simplicity, we will use the term “SNP” for the remainder of this article when referring to any form of single nucleotide variant.

Depending on their position in the genome, SNPs can be either coding or non-coding. With a coding density of 90–96%²⁰, most of the SNPs in MTBC are in coding regions of the genome. Coding SNPs can be further divided into synonymous (sSNP) and non-synonymous (nSNP) depending on whether they lead to changes in the corresponding amino acid sequence. While in average nSNPs are likely to have a stronger effect on the organism’s fitness (either beneficial or deleterious), and will therefore be under a stronger selective pressure than sSNPs, the latter are not necessarily selectively neutral. Phenomena such as the codon bias in MTBC and the general mutational bias in bacteria^{20,21} suggest that sSNPs, too, can be affected by natural selection. Similarly, non-coding SNPs are often considered selectively “neutral”, but increasingly the importance of non-coding (i.e. untranslated) regions of the bacterial genome for gene regulation is becoming evident²².

SNPs are relatively rare events in MTBC. They occur approximately every 3 kb of DNA sequence⁵. Hence, there is about three times less genetic variation in MTBC than in humans²³. Together with other bacterial pathogens such as *Yersinia pestis*, *Bacillus anthracis*, *Mycobacterium leprae* or *Salmonella enterica* serovar Typhi, MTBC has been referred to as “genetically monomorphic”, even though MTBC harbours significantly more variation than other monomorphic bacteria²⁴. One interesting observation is that the large majority of SNPs in MTBC occur as singletons (variants that only occur in a single strain). No clear explanation currently exists for this phenomenon, but a possible effect of background selection has been proposed^{6,25}.

SNPs are phylogenetically informative in MTBC

The comparably low frequency of SNPs and limited ongoing horizontal gene transfer in MTBC result in low levels of homoplasy (i.e. the independent occurrence of the same SNP in phylogenetically unrelated strains)^{5,6}. Hence, SNPs represent robust markers for inferring phylogenies and for strain classification¹². SNPs can also be used to measure evolutionary distances between strains, i.e. to estimate the time of divergence of strains from their genetic distance, if a mutation rate is known²⁶.

The first step in generating SNP data is referred to as SNP discovery and usually involves comparative sequencing of multiple genes or whole genomes in two or more strains of interest. Once a set of SNPs has been identified, these can then be used to screen additional strains using one of many available SNP-typing platforms^{27,28}. Importantly, the usefulness

of a given SNP-set is dependent on the amount of effort put into the initial identification of these SNPs, in particular on the number of strains included at the discovery stage. Poor representation of the relevant strain diversity during the discovery process will result in a biased set of SNPs which can lead to erroneous phylogenetic inferences; this phenomenon is known as “phylogenetic discovery bias” and has been discussed in detail elsewhere^{24,29–31}.

In 1997, Sreevatsan and colleagues sequenced 26 drug resistance-associated genes in 842 clinical isolates and identified two nSNPs which were unrelated to drug resistance³². Based on these two SNPs, a classification scheme into three Principle Genetic Groups was developed, which has been widely used in the past. The whole genome sequence of H37Rv published in 1998² established a first reference against which other genomes could be compared. In 2002, CDC1551 was sequenced³³ allowing for a first insight into the genome-wide SNP-diversity in *M. tuberculosis*; 1,075 SNP differences were found between the two strains. The whole genome sequence of *Mycobacterium bovis* AF2122/97³⁴ and the partial genome of the “Beijing” strain 210 became available shortly thereafter, generating an increased collection of SNPs for genotyping purposes. Two studies took advantage of the availability of these four genome sequences and identified phylogenetically informative SNPs to genotype large strain collections and identify phylogenetic groups within MTBC. However, as outlined above, both of these studies suffered from a phylogenetic discovery bias due to the low number of genomes used for SNP discovery. Hence, the resulting phylogenies presented by these groups were similarly affected by this problem^{24,31}. By contrast, three other studies used *de novo* sequencing of six genes³⁵, 56 genes³⁶ and 89 genes⁶, respectively, in large strain collections and produced unbiased phylogenies which were more congruent with those inferred using other methods, i.e. genomic deletions³⁷. However, even the phylogeny by Hershberg *et al.* based on 89 whole gene sequences was unable to completely resolve all the branches within the tree. In 2010, Comas *et al.* published the first whole-genome based global phylogeny of human-associated MTBC⁵. As expected given MTBC’s largely clonal population structure, this genome-based phylogeny was highly congruent with those published previously, but all main lineages were now clearly resolved (Figure 2). This phylogeny has since then been used as a reference for phylogenetic studies including an increasing number of MTBC strains^{7,38}.

The growing number of individual gene sequences and whole genomes in MTBC has already resulted in the identification of thousands of SNPs, which in recent years have been incorporated into various SNP-typing schemes. Because MTBC is largely clonal, for each of the main lineages “diagnostic” or “canonical” SNPs can be extracted and used as markers to assign unknown isolates to a particular phylogenetic group or lineage. Various SNP-typing assays have been developed on different platforms^{32,35,39–46}, and the latest assays can interrogate multiple SNPs simultaneously in one reaction. Examples include assays designed for the MassArray and the Luminex platform^{44–46}. These methods provide the robust phylogenetic framework necessary for genotype-phenotype- and other association studies¹³. For example, there is increasing evidence for phenotypic differences between strains, and studies need to be conducted to assess if these differences are due to genetic differences between MTBC clades¹¹.

Even though a lot of progress has been made in our understanding of the global phylogenetic diversity of MTBC, much remains unknown with respect to both human- and animal-associated MTBC diversity⁴⁷. For example, in addition to the six main human-associated MTBC lineages, a seventh lineage was recently discovered in TB patients from Ethiopia³⁸. Similarly, new animal-associated lineages have been identified in several African mammal species, indicating that more MTBC diversity exists^{48,49}. Moreover, in addition to focusing on differences between the main lineages of MTBC, we also need to look deeper into the diversity within individual lineages. For example, the Beijing family of strains is a

sublineage of Lineage 2 (Figure 2) and is currently a strong focus of research because of its association with drug resistance⁵⁰, hypervirulence in animal models^{51,52}, and recent expansion in human populations^{53,54}. Moreover, even though phenotypic diversity has been associated with the different MTBC lineages (e.g. in the elicited innate immune responses), individual strains within these lineages exhibit a wide range of phenotypes⁵⁵, suggesting that in addition to strain-specific effects, sub-lineage structure should also be considered^{56,57}. Only with a full understanding of the nature and phenotypic consequence of MTBC diversity will we be able to properly evaluate new diagnostics, vaccines and treatment options¹².

To achieve a more comprehensive view of the global diversity of MTBC, we propose an iterative process, in which, first, genome sequencing of the most diverse strains is performed to identify new phylogenetically informative SNPs. These SNPs are then used to genotype large collections of strains, whereby some strains will be classified into known lineages and others identified as novel. Genome sequencing of these unclassified strains will then define their phylogenetic positions, identify new lineages and corresponding signature SNPs, which can be used in a following round of genotyping.

In the future, genome sequencing is likely to replace any sort of genotyping in MTBC, including SNP-typing. While SNP-typing is an ideal tool for phylogenetic strain classification in MTBC, it does not have the necessary resolution for standard molecular epidemiological investigation such as defining chains of transmission or differentiating cases of relapse from re-infection; MIRU-VNTR in combination with spoligotyping is still the gold standard for these applications⁵⁸. Genome sequencing, on the other hand, generates a complete “barcode” of a strain, including the evolutionary background, drug resistance mutations or virulence-associated polymorphisms, and at the same time provides high resolution for transmission studies^{59,60}. Yet, until large-scale genome-sequencing becomes more readily available in standard laboratories, SNP-typing in MTBC will continue to play an important role in TB research and control.

The functional consequences of SNPs

In addition to being useful phylogenetic markers, SNPs carry functional information. The best-characterized “SNPs” in MTBC are drug resistance-conferring mutations. Drug resistance in MTBC is largely caused by single nucleotide mutations^{61–64}. Many drug resistance-conferring mutations have been identified, and are publicly available in the *TBDRamDB* database⁶⁵ (currently containing information on 1447 mutations relevant for most anti-TB drugs (Table 1)). This kind of molecular information is crucial for the development of new and faster diagnostic methods to detect drug resistance. While for the first-line anti-TB drugs, the most important drug resistance-conferring mutations have been identified and incorporated into promising new diagnostic tools^{66–68}, many mutations remain unknown, including many of those causing resistance to the 2nd-line drugs. Besides the mutations causing drug resistance, other associated mutations could also be targeted in the future. For example, two recent studies have shown that compensatory mutations in the RNA polymerase of MTBC can contribute to the fitness of rifampicin-resistant strains^{69,70}. While the molecular mechanisms involved remain to be determined, other mutations associated with drug resistance (e.g. compensatory mutations) could be used for molecular diagnostics even if they are not directly responsible for the drug resistance phenotype.

Although drug resistance-conferring mutations are most important as far as TB control is concerned, most SNPs in MTBC are unrelated to drug resistance. Indeed, thousands of SNPs have already been identified across MTBC strains and lineages. Many of these might translate into cellular and/or clinically relevant phenotypic changes. This is particularly true

given the fact that up to two thirds of SNPs in MTBC are non-synonymous, which is unlike most other organisms in which sSNPs predominate⁶. The reason for the high proportion of nSNPs in MTBC is unclear but has been proposed to be the consequence of the relatively short evolutionary age of MTBC (i.e. purifying selection has not had time to remove nSNPs which on average are slightly deleterious)⁷¹. Alternatively, it might reflect a reduction in the efficacy of purifying selection as a consequence of increased genetic drift in MTBC⁶. While intragenic and sSNPs are not necessarily neutral (e.g. they can affect regulatory regions), the effect of nSNPs is likely to be more pronounced. Yet, in contrast to drug resistance-conferring mutations, the effects of these nSNPs are less evident and likely to be more subtle, rendering the study of the biological and epidemiological consequences of these SNPs difficult.

One possible way forward is to use *in silico* predictions of the effects of SNPs. There are a number of freely available tools that can be used for this purpose. A prominent example is Sorting Intolerant From Tolerant (SIFT), a tool that, based on sequence homology and amino acid properties, predicts how much a given polymorphism might affect the function of the corresponding protein⁷². Other tools include ANNOVAR⁷³ and SVA (<http://www.svaproject.org>). In MTBC, such an approach has recently been used to look at variation in mammalian cell entry (*mce*) operons⁷⁴. Ultimately however, polymorphisms predicted to effect protein function will have to be experimentally confirmed using other tools⁷⁵.

Adding another level of complexity, which so far has been largely ignored, are the possible epistatic interactions between SNPs in the same genome. A good example of this phenomenon are compensatory mutations that interact with corresponding drug resistance-conferring mutations^{69,70}. Similar effects might be occurring among other mutations⁷⁶. Hence, the phenotypic characteristics of a given strain genetic background will depend on both the individual mutations as well as on the interactions between these mutations. Another important characteristic of MTBC is the fact that horizontal gene exchange is comparably rare and as a consequence, MTBC exhibits a largely clonal population structure^{77–79}. Hence, all SNPs in a particular strain's genome are linked (i.e. they are in linkage disequilibrium), which makes attributing the phenotypic behaviour of a given strain to a particular mutation non-trivial.

In summary, thousands of SNPs are being identified in MTBC thanks to the new DNA sequencing technologies. We can use these SNPs for phylogenetic and population genetic analyses and to study drug resistance. As for the large majority of SNPs, we do not know what their functional impact might be (partially also because we do not know the functions of the relevant genes). The functional consequences of these SNPs, and their potential for driving phenotypic differences between strains should be studied in the future. In the meantime, most ongoing SNP work in MTBC largely consists of discovering and cataloguing new SNPs. Let us now discuss some of the technicalities involved in these processes.

How do we discover new SNPs in MTBC?

In the upcoming years, we expect whole genome sequencing to at least partially replace all previous genotyping methods for MTBC. So far, large-scale DNA sequencing projects have usually been performed by specialized Sequencing Centres, but new benchtop sequencing devices increasingly allow for “do-it-yourself” approaches in the standard laboratory⁸⁰. In this section, we elaborate on some of the technical aspects of NGS genome analysis, with a particular focus on the workflow during the bioinformatics analysis. The DNA sequencing technologies themselves have been covered by several recent reviews^{81–83}.

Panel A of Figure 3 shows one possible data analysis pipeline for Illumina short read data, as currently used in our laboratory^{5,69}. Starting with a FastQ file with millions of short sequence reads (50–200 bp in length), different software tools are used to align each read to a reference genome and to call variant positions^{84–88}. As a reference genome we generally use the H37Rv genome with all known variant positions replaced by the hypothetical ancestral allelic states, representing a putative reconstructed ancestor of all MTBC⁵. Each nucleotide position differing from the reference, i.e. each SNP, can be annotated with gene, amino acid change and several other features⁷³. Throughout the workflow, a number of filtering steps are used to remove SNPs showing low confidence. These include SNPs in repetitive regions of the genome, including PE/PPE genes and insertion sequences (panel B in Figure 3). A list of SNPs per strain is given as an output. Similar to single nucleotide variants, other polymorphisms such as small insertions and deletions can be analyzed, but will not be further discussed here.

The number of called SNPs is therefore a result of combining different algorithms. Panel B of Figure 3 shows schematically where true- or false-positive SNPs are lost in the filtering process (i.e. they are filtered out by the software according to the particular parameter settings). This particular approach aims at reducing the number of false-positive calls. Often, the question as to what SNPs are true (as opposed to false-positives) remains, and ideally, SNPs should be confirmed by other methods such as Sanger sequencing. The issue of data quality is key in all sequencing projects, and there is a general trade-off between data quality (i.e. the stringency of filter settings) and the number of true SNPs identified. Moving forward, there is a need for minimum requirements for SNP data generated by NGS. One possible approach could be that a particular SNP has to be confirmed by at least one independent method. However, such an approach might become increasingly difficult as the number of SNPs increases beyond a few dozen. Also, filtering thresholds will have to be defined, and difficulties resulting from ambiguous base calls or multiallelic variants discussed.

The NGS data pipeline shown in Figure 3 consists of a combination of several command-line tools and customized scripts. Even though scripting can automatize the processing of one or many genomes, running these tools requires a certain level of bioinformatic expertise. But if whole genome sequencing is to be used more broadly in clinical settings, we need good software packages with automatic SNP-calling and SNP-annotation⁸⁰. There are several (not MTBC specific) public platforms for customized and semi-automated NGS data analysis, with the most prominent and feature-rich among them known as “Galaxy”⁸⁹. Galaxy allows graphically guided analysis of NGS data. The DNA data bank of Japan (DDBJ) also features a NGS analysis platform, which is publicly available and makes automatic integration into the DDBJ-archive possible for publication (<http://p.ddbj.nig.ac.jp/pipeline/Login.do>). These platforms are designed as customized combinations of tools that can be controlled with a graphical interface. However, what is actually needed for the TB-community is a “one-click” tool to upload NGS data in FastQ format, and to get a list of polymorphisms as an output (see Box 1). Moreover, the rapidly increasing amount of SNP data generated through all the ongoing and future NGS projects should ideally be centralized in a user-friendly and well-curated database.

Box 1

“Wish list” for a new SNP database for MTBC

A future database for SNP annotation and genome analysis to serve the needs of the TB research community should include (amongst other):

- Position-independent reference numbers for SNPs for unambiguous communication between laboratories and data management. The rs# numbers of NCBI provide a suitable starting point.
- Implementation of a “one-click” genome analysis pipeline to extract SNPs and indels from uploaded fastq-files. Fastq file would be automatically processed, all SNPs given as output, compared with the existing ones in the SNP-database, and annotated. The phylogenetic position of the user-genome would be given (see below). The SNP data from the user genome could then be shared with the database to be included in a next build. Multiple genomes sequences (i.e. multiple fastq-files) can be uploaded.
- Phylogenetic context: MTBC strains harbouring a particular SNP can be shown as a list. Strains of interest (i.e. all Lineage 4 strains) can be selected and shared SNPs shown (“SNPs per selected clade”).
- Generation of a phylogenetic tree with all available genomes or SNPs, and possibility of mapping a genome uploaded by the user.
- Visualization of SNPs as an alignment of all strains piled up (as in TBDB).
- A genome map highlighting SNP positions.
- Visualization of sequence upstream- and downstream of the SNP.
- Genotyping assays for selected SNPs: either published, or automatically calculated (e.g. primer pairs for PCR amplification of the SNP-relevant genomic region).

The need for a new SNP database for MTBC

So far, most SNP data in MTBC have been computed and stored on local workstations, and only made available upon publication. For the raw NGS reads, NCBI, EBI and DDJB have created Sequence Read Archives (SRA) where these data can be deposited (<http://www.ncbi.nlm.nih.gov/sra>, <http://www.ebi.ac.uk/ena/home>, http://trace.ddbj.nig.ac.jp/dra/index_e.shtml). These archives contain the raw sequencing reads in SRA format, which can be downloaded and converted to FastQ files (<http://www.ncbi.nlm.nih.gov/books/NBK47537/>). Raw sequence reads are required whenever a re-analysis of variants (e.g. with new software algorithms) are performed.

Once the polymorphisms have been extracted from the raw NGS data, they have to be made available upon publication. With previous projects based on classical Sanger sequencing, it was often possible to present SNP data as tables or spreadsheets, but the need for an online centralized database becomes obvious when considering the large amount of SNP- and associated metadata coming out of current and future MTBC NGS projects. The number of SNPs identified in a particular NGS project will likely be between several hundreds and several tens of thousands, depending on the number of MTBC genomes analyzed. This makes the use of lists or spreadsheets problematic and error-prone. Furthermore, to make use of this SNP information, researchers need to have access to the biological context, and to be able to interlink SNP data with other metadata as well as with other databases containing e.g. information on protein structure.

At least five existing databases harbour information on polymorphisms in MTBC (Table 1). Four of them were designed to contain data on MTBC SNPs only. The most important and multifunctional among these is probably the Tuberculosis Database (TBDB)⁹⁰. It currently contains 23,795 SNPs extracted from 25 MTBC genomes (Table 1). These SNPs can be

viewed as pair-wise or multiple comparisons of genomes with a variety of display options. SNPs can also be viewed in form of sequence alignments and downloaded as text files. SNPs can be separated by coding- and non-coding regions, or clustered by functional enrichment (e.g. all polymorphisms per COG-category gene). Drug resistance mutations can be specifically sought for. Each SNP is linked to the annotated gene it falls in, which is then linked to other databases. Phylogenetically relevant SNPs, i.e. potential markers, can be extracted by generating lists of SNPs shared by members of one lineage, and excluding SNPs shared with members of other lineages.

Another database is TBDRamDB which focuses on drug resistance-conferring mutations⁶⁵. It is probably the most complete repository for drug resistance mutations in MTBC. It currently features 1447 mutations and is regularly updated with information from new publications. MGDD⁹¹ is a database to compare SNPs across 6 MTBC genomes by gene name, nucleotide position or base change. MTCID⁹² is another database comparable in function to MGDD, and has in addition geographical mapping of SNPs implemented. Both MGDD and MTCID focus on drug resistance-associated mutation, but not exclusively. However, they have limited search functions and it is unclear if these databases are still being updated. The SNP database with the largest number of MTBC SNPs is dbSNP of NCBI⁹³. Currently about 40,000 SNPs are obtained using the search query “*Mycobacterium tuberculosis*”. Unfortunately, these SNPs are not annotated, and the origin of the data is unclear. But dbSNP is likely to be the most powerful database for SNP data in other organisms, and serves as reference database for the known genetic variants in *Homo sapiens*. Several features of dbSNP established for human variation could inform the establishment of a new SNP database for MTBC.

Several large databases of human variants are available (Table 1). dbSNP comprises currently 53,558,214 human variants. Most other databases reference their SNPs to dbSNP by the respective rs# number. This is an established annotation system that functions as follows: whenever a SNP of any species is uploaded to dbSNP (from publications or directly from the user) it is assigned a unique and position-independent submitted SNP ID number (ss#) and mapped to the reference genome (position on the contig). In a next step, the ss# is linked to a unique new RefSNP ID number (rs#), or is assigned to an existing rs# if a SNP at this genomic position was found and uploaded before. An rs# can therefore contain multiple ss# numbers, which are found in a table shown in each rs# record. The rs# is also position-independent, but each record contains the genomic position of the SNP. The SNP, i.e. the rs#, is then fully annotated, and linked to other NCBI resources such as gene context or publication source. Any attribute field or database can be linked to a specific rs#. In the next “build” of the database, the SNP is then incorporated⁹⁴. Other databases containing human data can use the dbSNP data to build, link and annotate their own SNP information. A selection of other large SNP databases potentially relevant for MTBC is listed in Table 1. These databases have different contents (i.e. only a subset of the dbSNP entries or refer to restricted populations as in PASNP or JSNP), different structures and are tailored to match different requirements, such as the HapMap genome browser (<http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html>), or the topoSNP database⁹⁵, which maps SNPs to protein structures. All of these databases have specific fields, which could be included in a future MTBC SNP database. In addition to all the descriptive information of a given SNP, human dbSNP entries contain data about the frequency of the SNP in different human populations. Drawing a line to a future MTBC SNP database, frequency data could be important for drug resistance-conferring mutations, and could be calculated e.g. by lineage (e.g. the frequency of *rpoB* mutations in each lineage, calculated and updated automatically). To allow calculating frequencies of mutations, a database needs to allow for multiple uploads of the same SNP. In the final chapter of this review we discuss some other features that a new MTBC SNP database should include (Box 1).

Features of a new MTBC SNP database

Given the existing features of TBDB (discussed above), this database represents an ideal starting point for an extended SNP database for MTBC (Box 1). TBDB already includes important aspects such as the relational tables and annotations. Unique and highly valuable modules such as the phylogenetic context could be extended to deal with larger numbers of taxa (strains) and characters (SNPs). So far, SNPs in TBDB are identified based on their position in the reference genome, but with larger numbers of genome sequences and characters, as well as alternative reference genomes used by different researchers, a unique SNP ID will become necessary, similar to what has been implemented in dbSNP. This would allow for unambiguous communication across the research community, solve the problem of SNPs whose positions are not found in a particular reference genome (e.g. in regions where large deletions in H37Rv occur), and allow for specifying a position of interest on more than one genome (e.g. when referring to the recently finished genome of *Mycobacterium africanum* GM041182). Moreover, a new MTBC SNP database should be frequently updated. Regular builds are also important in case of new annotations in the reference genome(s). The database should include as many informative fields as possible, but should not store any redundant information available in other databases. Many important fields are already present in TBDB and could be adopted into an extended version (Figure 4). These fields include the “standard” data on position, nucleotide change, gene annotation, and amino acid change (synonymous/non-synonymous). Additional fields could include “essentiality” of the corresponding gene (based on experimental evidence⁹⁶), validation (methods used to confirm polymorphism and to exclude sequencing errors), source of data (publication, upload, diversity sequencing project, etc.), associations with drug resistance, phylogenetic context (e.g. “Lineage 4 marker”), and data quality scores (e.g. phred score or coverage depth from the corresponding sequencing project). Other metadata such as clinical associations (e.g. virulence, transmission, vaccine efficacy, drug treatment) could be annotated in attribute fields as they become available. Other fields could be developed based on functional predictions of SNPs, and experimental confirmation of any phenotypic effects. Here, the SIFT algorithm (see above) and other related approaches could be implemented more widely. The way by which a particular SNP was discovered has to be defined (sequencing method, confirmation by other methods), and the adjacent sequence up- and downstream of the SNP position should be shown for unambiguous identification.

Regarding the original source of the SNP data, the ideal database should harbour both the raw sequencing data and the corresponding polymorphism data. This would allow a direct link to the source data, and – if needed – the possibility to retrieve the original sequencing reads for quality assessment or application of alternative mapping algorithms or other analytical approaches. For example, FastQ files could be uploaded by users, with SNPs called automatically and put into the context of the existing SNP data (see also Box 1). Following a set of quality criteria, the new SNPs would then be automatically merged into the main SNP database. SNPs should also be linked to relevant entries in other databases. There are a variety of databases that store MTBC-specific information¹⁷, and potential links are shown in Figure 4. Some of these databases include existing and highly accessed databases such as gene annotations on TubercuList or expression data on TBDB. Others could include new functions such as visualizing the spatial location of a SNP on the corresponding protein structure. Moreover, two large Systems Biology consortia are currently working on TB (<http://www.systemtb.org/> and <http://www.broadinstitute.org/annotation/tbsysbio/home.html>⁹⁷). Linking MTBC SNP information with transcriptional, proteomic, and metabolomic data generated through these consortia should allow for a more comprehensive understanding of TB biology.

As already mentioned, frequency data on SNP distribution should be included. As the number of available whole-genome sequences increase, we will need sophisticated tools to visualize allele frequencies in a geography-dependent manner. The high clonality of MTBC and the strong genome-wide linkage disequilibrium between individual SNPs in a given genome will have to be considered in this respect. Finally, the SNP database should be easily searchable by position, reference number, keywords, or free text. All data contained in the database should also be downloadable as a bulk. In dbSNP, this is achieved by building a local copy of the database. For MBTC, as the number of SNP will be considerably smaller, this could be achieved by a bulk download in text format.

The requirements and opportunities for a new SNP database for MTBC are manifold but will not be easy to implement (Box 1). Ideally, joint efforts between NCBI, TBDB and PATRIC are needed to achieve this goal. PATRIC⁹⁸ is a platform for storage and exchange of bacterial data, featuring a variety of tools including genome browsing, BLAST, comparative pathways, and protein annotations. PATRIC has already compiled a lot of mycobacterial data, and there are plans to host bacterial SNP data as well (B. Sobral, personal communication). Thanks to the large amount of information already included, a joint venture between TBDB and PATRIC seems like an ideal way forward to establish a new SNP database for MTBC.

Conclusions

NGS studies of MTBC clinical isolates are discovering thousands of SNPs. Studying the functional effects of these SNPs and their association with phylogenetic clades should become an increasing part of the research portfolio. MTBC consist of a diverse population of strains, and this diversity should be considered when developing new tools and strategies to combat TB. A new, extended, and well-curated database is necessary to accommodate these rapidly accumulating SNP data in a user-friendly and integrated format. Ideally, dbSNP/NCBI, TBDB and PATRIC should join forces and play major roles in the development of such a database. After establishing the basic framework of such a database, specific needs and wishes of the community can be discussed and incorporated (Box 1). The aim of this review is to contribute to the discussion.

Acknowledgments

We thank Mireia Coscollá and Iñaki Comas as well as the other members of our group for the inspiring discussions and comments on the manuscript. The work in our laboratory is supported by the Swiss National Science Foundation (grant number PP0033-119205) and the National Institutes of Health (AI090928 and HSN266200700022C).

References

1. WHO. Global tuberculosis control. Geneva, Switzerland: 2011.
2. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998; 393:537–44. [PubMed: 9634230]
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 USA. 2002; 99:3684–9. [PubMed: 11891304]

4. Mostowy S, Cousins D, Brinkman J, Aranaz A, Behr MA. Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *J Infect Dis.* 2002; 186:74–80. [PubMed: 12089664]
5. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet.* 2010; 42:498–503. [PubMed: 20495566]
6. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* 2008; 6:e311. [PubMed: 19090620]
7. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, Thurston S, Gagneux S, Wood J, Antonio M, Quail MA, Gehre F, Adegbola RA, Parkhill J, de Jong BC. The genome of *Mycobacterium Africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. *PLoS Negl Trop Dis.* 2012; 6:e1552. [PubMed: 22389744]
8. Comas I, Gagneux S. A role for systems epidemiology in tuberculosis research. *Trends Microbiol.* 2011; 10.1016/j.tim.2011.07.002
9. Breitling R. What is Systems Biology? *Front Physiol.* 2010; 1:10.3389/fphys.2010.00009 [PubMed: 21522484]
10. Kirschner DE, Young D, Flynn JL. Tuberculosis: global approaches to a global disease. *Curr Opin Biotechnol.* 2010; 21:524–31. [PubMed: 20637596]
11. Coscolla M, Gagneux S. Does *M. tuberculosis* genomic diversity explain disease diversity? *Drug Discov Today Dis Mech.* 2010; 7:e43–e59. [PubMed: 21076640]
12. Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis.* 2007; 7:328–37. [PubMed: 17448936]
13. Comas I, Gagneux S. The past and future of tuberculosis research. *PLoS Pathog.* 2009; 5:e1000600. [PubMed: 19855821]
14. Mycobacterium - Wellcome Trust Sanger Institute. [accessed 28 Oct 2012] <http://www.sanger.ac.uk/resources/downloads/bacteria/mycobacterium.html>
15. Stein LD. The case for cloud computing in genome informatics. *Genome Biol.* 2010; 11:207. [PubMed: 20441614]
16. Wetterstrand, KA. [accessed 28 Oct 2012] DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. <http://www.genome.gov/sequencingcosts/>
17. Sharma D, Surolia A. Computational tools to study and understand the intricate biology of mycobacteria. *Tuberculosis (Edinb).* 2011; 91:273–6. [PubMed: 21398182]
18. Hartl, DL.; Clark, AG. Principles of population genetics. 4. Sinauer Associates; 2007.
19. Huard RC, Fabre M, de Haas P, Lazzarini LCO, van Soolingen D, Cousins D, Ho JL. Novel genetic polymorphisms that further delineate the phylogeny of the *Mycobacterium tuberculosis* complex. *J Bacteriol.* 2006; 188:4271–87. [PubMed: 16740934]
20. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC. After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res.* 2012; 22:721–34. [PubMed: 22377718]
21. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 2010; 6:10.1371/journal.pgen.1001115
22. Arnvig KB, Comas I, Thomson NR, Houghton J, Boshoff HI, Croucher NJ, Rose G, Perkins TT, Parkhill J, Dougan G, Young DB. Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog.* 2011; 7:10.1371/journal.ppat.1002342
23. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. *Nature.* 2008; 452:872–6. [PubMed: 18421352]
24. Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol.* 2008; 62:53–70. [PubMed: 18785837]

25. Pepperell C, Hoepfner VH, Lipatov M, Wobeser W, Schoolnik GK, Feldman MW. Bacterial genetic signatures of human social phenomena among *M. tuberculosis* from an aboriginal canadian population. *Mol Biol Evol.* 2010; 27:427–40. [PubMed: 19861642]
26. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M, Flynn JL, Fortune SM. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet.* 2011 advance online publication. 10.1038/ng.811
27. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering.* 2007; 9:289–320.
28. Wang L, Luhm R, Lei M. SNP and mutation analysis. *Adv Exp Med Biol.* 2007; 593:105–16. [PubMed: 17265721]
29. Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'Ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML, Van Ert MN, Huynh LY, Fraser CM, Keim P. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci USA.* 2004; 101:13536–41. [PubMed: 15347815]
30. Alland D, Whittam TS, Murray MB, Cave MD, Hazbon MH, Dix K, Kokoris M, Duesterhoeft A, Eisen JA, Fraser CM, Fleischmann RD. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. *J Bacteriol.* 2003; 185:3392–9. [PubMed: 12754238]
31. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol.* 2009; 7:537–44. [PubMed: 19483712]
32. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA.* 1997; 94:9869–74. [PubMed: 9275218]
33. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR Jr, Venter JC, Fraser CM. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol.* 2002; 184:5479–90. [PubMed: 12218036]
34. Garnier T, Eiglmeier K, Camus J-C, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, Simon S, Harris B, Atkin R, Doggett J, Mayes R, Keating L, Wheeler PR, Parkhill J, Barrell BG, Cole ST, Gordon SV, Hewinson RG. The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci USA.* 2003; 100:7877–82. [PubMed: 12788972]
35. Baker L, Brown T, Maiden MC, Drobniewski F. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerging Infect Dis.* 2004; 10:1568–77. [PubMed: 15498158]
36. Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V, Tonjum T, Sola C, Matic I, Gicquel B. Evolution and Diversity of Clonal Bacteria: The paradigm of *Mycobacterium tuberculosis*. *PLoS ONE.* 2008; 3.10.1371/journal.pone.0001538
37. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA.* 2006; 103:2869–73. [PubMed: 16477032]
38. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, Gadisa E, Kiros T, Habtamu M, Hussein J, Zinsstag J, Robertson B, Ameni G, Lohan A, Loftus B, Comas I, Gagneux S, Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A. Investigation of pulmonary and extrapulmonary tuberculosis in Ethiopia shows minimal zoonotic transmission of *Mycobacterium bovis* and identifies a new lineage of *Mycobacterium tuberculosis*. *Emerging Infectious Diseases.* accepted.
39. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osorio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, Alland D. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution,

- phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol.* 2006; 188:759–72. [PubMed: 16385065]
40. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis.* 2006; 193:121–8. [PubMed: 16323140]
 41. Bergval IL, Vijzelaar RNCP, Dalla Costa ER, Schuitema ARJ, Oskam L, Kritski AL, Klatser PR, Anthony RM. Development of multiplex assay for rapid characterization of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 2008; 46:689–99. [PubMed: 18077627]
 42. Abadia E, Zhang J, dos Vultos T, Ritacco V, Kremer K, Aktas E, Matsumoto T, Refregier G, van Soolingen D, Gicquel B, Sola C. Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect Genet Evol.* 2010; 10:1066–74. [PubMed: 20624486]
 43. Bouakaze C, Keyser C, de Martino SJ, Sougakoff W, Veziris N, Dabernat H, Ludes B. Identification and genotyping of *Mycobacterium tuberculosis* complex species by use of a SNaPshot Minisequencing-based assay. *J Clin Microbiol.* 2010; 48:1758–66. [PubMed: 20220173]
 44. Bouakaze C, Keyser C, Gonzalez A, Sougakoff W, Veziris N, Dabernat H, Jaulhac B, Ludes B. Matrix-assisted laser desorption ionization-time of flight mass spectrometry-based single nucleotide polymorphism genotyping assay using iPLEX gold technology for identification of *Mycobacterium tuberculosis* complex species and lineages. *J Clin Microbiol.* 2011; 49:3292–9. [PubMed: 21734028]
 45. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, Borrell S, Fenner L, Comas I, Coscollà M, Gagneux S. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. *PLoS ONE.* 2012; 7:e41253. [PubMed: 22911768]
 46. Bergval I, Sengstake S, Brankova N, Levterova V, Abadía E, Tadumaze N, Bablishvili N, Akhalaia M, Tuin K, Schuitema A, Panaiotov S, Bachiyska E, Kantardjiev T, de Zwaan R, Schürch A, van Soolingen D, van 't Hoog A, Cobelens F, Aspindzelashvili R, Sola C, Klatser P, Anthony R. Combined species identification, genotyping, and drug resistance detection of *Mycobacterium tuberculosis* cultures by MLPA on a bead-based array. *PLoS ONE.* 2012; 7:e43240. [PubMed: 22916230]
 47. Gagneux S. Host-pathogen coevolution in human tuberculosis. *Philos Trans R Soc Lond, B, Biol Sci.* 2012; 367:850–9. [PubMed: 22312052]
 48. Mostowy S, Cousins D, Behr MA. Genomic interrogation of the dassie bacillus reveals it as a unique RD1 mutant within the *Mycobacterium tuberculosis* complex. *J Bacteriol.* 2004; 186:104–9. [PubMed: 14679230]
 49. Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, Gey van Pittius NC. Novel *Mycobacterium tuberculosis* complex pathogen, *M mungi*. *Emerging Infect Dis.* 2010; 16:1296–9. [PubMed: 20678329]
 50. Borrell S, Gagneux S. Infectiousness, reproductive fitness and evolution of drug-resistant *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis.* 2009; 13:1456–66. [PubMed: 19919762]
 51. Caws M, Thwaites G, Stepniewska K, Nguyen TNL, Nguyen THD, Nguyen TP, Mai NTH, Phan MD, Tran HL, Tran THC, van Soolingen D, Kremer K, Nguyen VVC, Nguyen TC, Farrar J. Beijing genotype of *Mycobacterium tuberculosis* is significantly associated with human immunodeficiency virus infection and multidrug resistance in cases of tuberculous meningitis. *J Clin Microbiol.* 2006; 44:3934–9. [PubMed: 16971650]
 52. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis.* 2010; 10:103–11. [PubMed: 20113979]
 53. Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, Wilkinson RJ, Nicol MP. Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. *Clin Infect Dis.* 2008; 47:1252–9. [PubMed: 18834315]
 54. van der Spuy GD, Kremer K, Ndabambi SL, Beyers N, Dunbar R, Marais BJ, van Helden PD, Warren RM. Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis (Edinb).* 2009; 89:120–5. [PubMed: 19054717]

55. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog.* 2011; 7:e1001307. [PubMed: 21408618]
56. Schürch AC, Kremer K, Hendriks ACA, Freyee B, McEvoy CRE, van Crevel R, Boeree MJ, van Helden P, Warren RM, Siezen RJ, van Soolingen D. SNP/RD typing of *Mycobacterium tuberculosis* Beijing strains reveals local and worldwide disseminated clonal complexes. *PLoS ONE.* 2011; 6:e28365. [PubMed: 22162765]
57. Wada T, Iwamoto T, Hase A, Maeda S. Scanning of genetic diversity of evolutionarily sequential *Mycobacterium tuberculosis* Beijing family strains based on genome wide analysis. *Infection, Genetics and Evolution.* 2012; 12:1392–6.
58. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Loch C, van Soolingen D. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 2006; 44:4498–510. [PubMed: 17005759]
59. Schürch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, van Soolingen D. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol.* 2010; 48:3403–6. [PubMed: 20592143]
60. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodtkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med.* 2011; 364:730–9. [PubMed: 21345102]
61. Musser JM. Antimicrobial agent resistance in mycobacteria: molecular genetic insights. *Clin Microbiol Rev.* 1995; 8:496–514. [PubMed: 8665467]
62. Telenti A. Genetics of drug resistance in tuberculosis. *Clin Chest Med.* 1997; 18:55–64. [PubMed: 9098610]
63. Ramaswamy S, Musser JM. Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber Lung Dis.* 1998; 79:3–29. [PubMed: 10645439]
64. Riska PF, Jacobs WR Jr, Alland D. Molecular determinants of drug resistance in tuberculosis. *Int J Tuberc Lung Dis.* 2000; 4:S4–10. [PubMed: 10688142]
65. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis Drug Resistance Mutation Database. *PLoS Med.* 2009; 6:e1000002.
66. Hillemann D, Weizenegger M, Kubica T, Richter E, Niemann S. Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol.* 2005; 43:3699–703. [PubMed: 16081898]
67. Hain Lifescience - Mycobacteria. [accessed 28 Oct 2012] <http://www.hain-lifescience.de/en/products/microbiology/mycobacteria.html>
68. Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, Allen J, Tahirli R, Blakemore R, Rustomjee R, Milovic A, Jones M, O'Brien SM, Persing DH, Ruesch-Gerdes S, Gotuzzo E, Rodrigues C, Alland D, Perkins MD. Rapid molecular detection of tuberculosis and rifampin resistance. *New England Journal of Medicine.* 2010; 363:1005–15. [PubMed: 20825313]
69. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics.* 2011; 44:106–10. [PubMed: 22179134]
70. Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, Bentley SD, Parkhill J, Nejentsev S, Hoffner SE, Horstmann RD, Brown T, Drobniewski F. Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Research.* 2012; 20:1101/gr.128678.111
71. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 2006; 239:226–35. [PubMed: 16239014]
72. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 2009; 4:1073–81. [PubMed: 19561590]

73. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl Acids Res.* 2010; 38:e164–e164. [PubMed: 20601685]
74. Pasricha R, Chandolia A, Ponnann P, Saini N, Sharma S, Chopra M, Basil M, Brahmachari V, Bose M. Single nucleotide polymorphism in the genes of *mce1* and *mce4* operons of *Mycobacterium tuberculosis*: analysis of clinical isolates and standard reference strains. *BMC Microbiology.* 2011; 11:41. [PubMed: 21345183]
75. Lamrabet O, Drancourt M. Genetic engineering of *Mycobacterium tuberculosis*: A review. *Tuberculosis.* 2012; 92:365–76. [PubMed: 22789498]
76. Borrell S, Gagneux S. Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. *Clin Microbiol Infect.* 2011; 17:815–20. [PubMed: 21682802]
77. Supply P, Warren RM, Bañuls A-L, Lesjean S, Van Der Spuy GD, Lewis L-A, Tibayrenc M, Van Helden PD, Locht C. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Mol Microbiol.* 2003; 47:529–38. [PubMed: 12519202]
78. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci USA.* 2004; 101:4871–6. [PubMed: 15041743]
79. Liu X, Gutacker MM, Musser JM, Fu Y-X. Evidence for recombination in *Mycobacterium tuberculosis*. *J Bacteriol.* 2006; 188:8169–77. [PubMed: 16997954]
80. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med.* 2012; 366:2267–75. [PubMed: 22693998]
81. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol.* 2009; 25:195–203. [PubMed: 19429539]
82. Metzker ML. Sequencing technologies — the next generation. *Nature Reviews Genetics.* 2010; 11:31–46.
83. Lee H, Tang H. Next-generation sequencing technologies and fragment assembly algorithms. *Methods Mol Biol.* 2012; 855:155–74. [PubMed: 22407708]
84. Maq. [accessed 6 Dec 2010] <http://maq.sourceforge.net/index.shtml>
85. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754–60. [PubMed: 19451168]
86. SMALT - Wellcome Trust Sanger Institute. [accessed 18 Oct 2012] <http://www.sanger.ac.uk/resources/software/smalt/>
87. Samtools. [accessed 8 Dec 2010] <http://samtools.sourceforge.net/cns0.shtml>
88. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. *Bioinformatics.* 2011; 27:2156–8. [PubMed: 21653522]
89. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology.* 2010; 11:R86. [PubMed: 20738864]
90. Reddy TBK, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, Gellesch M, Hubble J, Jen D, Jin H, Koehrsen M, Larson L, Mao M, Nitzberg M, Sisk P, Stolte C, Weiner B, White J, Zachariah ZK, Sherlock G, Galagan JE, Ball CA, Schoolnik GK. TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res.* 2009; 37:D499–508. [PubMed: 18835847]
91. Vishnoi A, Srivastava A, Roy R, Bhattacharya A. MGDD: *Mycobacterium tuberculosis* genome divergence database. *BMC Genomics.* 2008; 9:373. [PubMed: 18681951]
92. Bharti R, Das R, Sharma P, Katoch K, Bhattacharya A. MTCID: a database of genetic polymorphisms in clinical isolates of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb).* 2012; 92:166–72. [PubMed: 22209237]
93. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–11. [PubMed: 11125122]

94. [accessed 22 Oct 2012] The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation - The NCBI Handbook - NCBI Bookshelf. <http://www.ncbi.nlm.nih.gov/books/NBK21088/>
95. Stitzel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.* 2004; 32:D520–D522. [PubMed: 14681472]
96. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol.* 2003; 48:77–84. [PubMed: 12657046]
97. Aderem A, Adkins JN, Ansong C, Galagan J, Kaiser S, Korth MJ, Law GL, McDermott JG, Proll SC, Rosenberger C, Schoolnik G, Katze MG. A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *MBio.* 2011; 2:e00325–00310. [PubMed: 21285433]
98. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW. PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infect Immun.* 2011; 79:4286–98. [PubMed: 21896772]
99. Flicek P, et al. Ensembl 2012. *Nucleic Acids Research.* 2011; 40:D84–D90. [PubMed: 22086963]
100. Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, Bhak J, Liu E, Tongsima S. the HUGO Pan-Asian SNP Consortium. PanSNPdb: The Pan-Asian SNP Genotyping Database. *PLoS ONE.* 2011; 6:e21451. [PubMed: 21731755]
101. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.* 2002; 30:158–62. [PubMed: 11752280]
102. International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010; 467:52–8. [PubMed: 20811451]
103. Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, Brahmachari SK, Brookes AJ. HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.* 2009; 37:D797–802. [PubMed: 18948288]

Figure 1. Number of MTBC samples with raw genome sequences available in the Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA)
Search query was “*Mycobacterium tuberculosis* complex” in the NCBI *Biosample* database, and results were extracted from filter “Used in SRA”. Y-axis represents cumulative numbers of entries on October 15 of each year.

Figure 2. Phylogenetic tree of 22 whole genome sequences of MTBC and *M. canettii* used as the outgroup
Modified from ^{7,45}.

Figure 3. Example of a genome (re-) sequencing analysis pipeline for MTBC and reduction of the number of SNPs due to filtering

A. The whole-genome sequencing data analysis pipeline for SNP calling as currently used in our laboratory. Input data are Illumina short reads in “fastq”-format. Outputs are single nucleotide variants per strain compared to the hypothetical ancestor of MTBC. **B.** Schematic illustration of steps where (true or false-positive) SNPs remain undiscovered or are lost due to filtering.

Figure 4. A simplified scheme of a future relational SNP-database for MTBC SNPs, including links to other databases

“rs” numbers are unique IDs from dbSNP (reference cluster). This proposal for a new MTBC SNP database is independent of the platform hosting the central database.

Table 1
Relevant SNP databases for MTBC

Databases already containing SNP data of MTBC, and examples of relevant SNP-databases of human variation, which could serve as examples for a future MTBC SNP-database.

Name	Species	# of SNPs (# of genomes ⁴)	Features	URL ^{Ref}
<i>MTBC SNP-databases</i>				
dbSNP	<i>M. tuberculosis</i>	40,303 MTBC ² (3827 MTBC samples in SRA)	NCBI curated relational SNP-database for all organisms, MTBC SNPs are not annotated	http://www.ncbi.nlm.nih.gov/projects/SNP/ ⁹³
TBDB	<i>M. tuberculosis</i>	23'795 (25 MTBC ³)	Relational database with various MTBC data sets such as expression, diversity, proteins, ChIPseq, publications etc. SNPs are well annotated and interlinked with other tables, but not updated.	http://www.tbdb.org/ ⁹⁰
PATRIC	<i>M. tuberculosis</i>	0 (75 MTBC)	Extensive relational database for various bacterial pathogens linking genomic data with NIH disease, epitopes etc. SNP database in preparation.	http://www.patricbrc.org/portal/portal/patric/Home ⁹⁸
MGDD	<i>M. tuberculosis</i>	n.a. (6 MTBC)	One-by-one comparison of 6 MTBC strains; not updated since 2008	http://mirna.jnu.ac.in/mgdd/ ⁹¹
MTCID	<i>M. tuberculosis</i>	n.a.	List of mainly drug resistance conferring mutations	http://ccbb.jnu.ac.in/Tb/ ⁹²
TBDReaMDB	<i>M. tuberculosis</i>	1447 (0)	Drug resistance conferring mutations	http://www.tbdreamdb.com/ ⁶⁵
<i>Relevant SNP-databases from other organisms, that could serve as example databases</i>				
dbSNP	Various	53,558,214 human SNPs ⁴ (34'970 human samples in SRA)	NCBI curated SNP database, interlinked with many other databases; can also contain indels, IS sequences, microsatellites	http://www.ncbi.nlm.nih.gov/projects/SNP/ ⁹³
ENSEMBL	Various	synchronized with dbSNP	SNP database with extensive (graphical) links to other features (genomic context, genes, population genetics, phylogenetic	http://www.ensembl.org/index.html ⁹⁹

Name	Species	# of SNPs (# of genomes ¹)	Features	URL ^{Ref}
PASNP	<i>H. sapiens</i>	55,998 SNPs (from 1719 individuals)	context, flanking sequence, etc.) Pan-Asian SNP database with extensive graphical browsing	http://www4a.biotech.or.th/PASNP ¹⁰⁰
JSNP	<i>H. sapiens</i>	197,195 human SNPs (n.a.)	Japan SNP database with SNP data from genotyping	http://snp.ims.u-tokyo.ac.jp/ ¹⁰¹
HapMap genome browser	<i>H. sapiens</i>	1,440,616 genotyped SNPs in 1184 individuals ⁵	Haplotype database	http://hapmap.ncbi.nlm.nih.gov/ ¹⁰²
GWAS central	<i>H. sapiens</i>	62,322,744 entries (n.a.) from dbSNP build 135	Former HGVbase, links human genetic association studies with dbSNP rs#	https://www.gwascentral.org/ ¹⁰³
topoSNP	<i>H. sapiens</i>	27,417 SNPs (publication 2004)	Mapping of human non-synonymous SNPs from OMIM and dbSNP to protein structures	http://gila.bioengr.uic.edu/snp/toposnp/ ⁹⁵

¹ Complete genomes or resequenced

² Number found in dbSNP for keyword “*Mycobacterium tuberculosis*”, but rs#s are invalid links.

³ MTBC genomes under “Diversity sequencing” on tdb.org

⁴ Number of RefSNP Clusters (rs#s) in build 137 as found in http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

⁵ As in HapMap³¹⁰²