A Comprehensive Map of *Mycobacterium tuberculosis* Complex Regions of Difference

D. Bespiatykh,[a] J. Bespyatykh,[a] I. Mokrousov,[b] E. Shitikov[a]

[a]Federal Research and Clinical Center of Physical-Chemical Medicine, Moscow, Russia

[b]St. Petersburg Pasteur Institute, St. Petersburg, Russia

ABSTRACT *Mycobacterium tuberculosis* complex (MTBC) species are classic examples of genetically monomorphic microorganisms due to their low genetic variability. Whole-genome sequencing made it possible to describe both the main species within the complex and *M. tuberculosis* lineages and sublineages. This differentiation is based on single nucleotide polymorphisms (SNPs) and large sequence polymorphisms in the so-called regions of difference (RDs). Although a number of studies have been performed to elucidate RD localizations, their distribution among MTBC species, and their role in the bacterial life cycle, there are some inconsistencies and ambiguities in the localization of RDs in different members of the complex. To address this issue, we conducted a thorough search for all possible deletions in the WGS data collection comprising 721 samples representing the full MTBC diversity. Discovered deletions were compared with a list of all previously described RDs. As with the SNP-based analysis, we confirmed the specificities of 79 regions at the species, lineage, or sublineage level, 17 of which are described for the first time. We also present RDscan (https://github.com/dbespiatykh/RDscan), an open-source workflow, which detects deletions from short-read sequencing data and correlates the results with high-specificity RDs, curated in this study. Testing of the workflow on a collection comprising ~7,000 samples showed a high specificity of the found RDs. This study provides novel details that can contribute to a better understanding of the species differentiation within the MTBC and can help to determine how individual clusters evolve within various MTBC species.

IMPORTANCE Reductive genome evolution is one of the most important and intriguing adaptation strategies of different living organisms to their environment. *Mycobacterium* offers several notorious examples of either naturally reduced (*Mycobacterium leprae*) or laboratory-reduced (*Mycobacterium bovis* BCG) genomes. *Mycobacterium tuberculosis* complex has its phylogeny unambiguously framed by large sequence polymorphisms that present unidirectional unique event changes. In the present study, we curated all known regions of difference and analyzed both *Mycobacterium tuberculosis* and animal-adapted MTBC species. For 79 loci, we have shown a relationship with phylogenetic units, which can serve as a marker for diagnosing or studying biological effects. Moreover, intersections were found for some loci, which may indicate the nonrandomness of these processes and the involvement of these regions in the adaptation of bacteria to external conditions.

KEYWORDS MTBC, *Mycobacterium tuberculosis* complex, RD, comparative genomics, deletions, large sequence polymorphisms, regions of difference, structural variants

The *Mycobacterium tuberculosis* complex (MTBC) is a group of closely related species that can cause tuberculosis (1). The members of the complex include the following *Mycobacterium* species: *M. africanum* (i.e., MTBC lineage 5 and lineage 6) (2), *M. bovis* (3), *M. canettii* (4), *M. caprae* (5), *M. microti* (6), *M. mungi* (7), *M. orygis* (8), *M. pinnipedii* (9), *M. suricattae* (10), and *M. tuberculosis* (i.e., MTBC lineage 1 to lineage 4, lineage 7,

Citation Bespiatykh D, Bespyatykh J, Mokrousov I, Shitikov E. 2021. A comprehensive map of *Mycobacterium tuberculosis* complex regions of difference. mSphere 6:e00535-21. https://doi.org/10.1128/mSphere.00535-21.

Editor Christina L. Stallings, Washington University School of Medicine in St. Louis

Address correspondence to D. Bespiatykh, d.bespiatykh@gmail.com.

Received 10 June 2021
Accepted 1 July 2021
Published 21 July 2021

and lineage 8) (11–13). These microorganisms have no evidence of horizontal gene transfer between strains (14, 15), and more significantly, they are some of the examples of genetic homogeneity (99.9% nucleotide identity), except for *M. canettii* and other "smooth" mycobacteria (16). Due to their low diversity, classical genotyping techniques, such as pulsed-field gel electrophoresis or multilocus sequence typing, have proven to be practically inapt for accurate MTBC genotyping. Instead, IS*6110* restriction fragment length polymorphism (IS*6110*-RFLP) analysis, spoligotyping, and mycobacterial interspersed repetitive unit-variable number of tandem repeat (MIRU-VNTR) analysis were introduced for genotyping. The methods mentioned above are excellent for identifying microbial transmission routes, disease outbreaks, and new cases of reinfection. However, due to their high discriminatory power and, in some cases, effects of homoplasy, these methods are not entirely suitable for constructing a reliable phylogeny for MTBC members. Large sequence polymorphisms (LSPs) proved to be the best solution to this problem, until whole-genome sequencing (WGS) technologies developed further, and single nucleotide polymorphisms (SNPs) also became pertinent to MTBC genotyping. Together, these markers facilitated the determination of a cogent scenario for the evolution paths of members of the MTBC (11, 17).

LSPs, being unidirectional unique-event polymorphisms, were initially identified using whole-genome microarrays and bacterial artificial chromosome arrays (18, 19). Located in the so-called regions of difference (RDs), LSPs were attributed to deletions relative to the reference *M. tuberculosis* H37Rv strain, while RvDs are H37Rv-related deletions. These deletions span from several hundred base pairs to more than 10 kbp, with the largest one being 26.3 kbp long (RD$^{\text{Rio}}$) (20). In general, RDs can be divided into phylogenetically informative and noninformative regions. The latter include PE-PPE genes, prophage regions, and regions flanked by insertion sequences. These regions are often strain specific due to variability and homologous recombination. In contrast, phylogenetically informative deletions are conservative and inherited by all descendants of the strain. Moreover, these deletions are sometimes associated with the virulence or resistance of mycobacteria (17, 21, 22).

Today, next-generation sequencing technologies have made a breakthrough in mycobacterial research. Whole-genome sequencing is routinely used to investigate tuberculosis resistance, transmission dynamics, and the population structure of MTBC organisms (23, 24). Numerous bioinformatics pipelines have been developed for this purpose, making it possible to correlate genomic data and laboratory tests. For the phylogenetic study of the MTBC, various SNP-based tools have been developed, while only a few tools have been developed for the analysis of LSPs in mycobacteria. The most prominent of these tools is RD-Analyzer, which can predict species and lineages of MTBC isolates from sequenced reads based on the presence of a set comprising 31 previously defined markers (RDs). Additionally, the authors identified 6 potential RD markers for the differentiation of *M. tuberculosis* lineage 4 isolates (25).

Here, we used publicly available WGS data comprising 721 MTBC strains to search for all possible deletions in the genome. Subsequently, the found deletions were correlated with a list of 187 RDs selected from 24 studies. This allowed us to describe the specificities of 79 LSPs at the species, lineage, and sublineage levels; also, some problems that may arise when analyzing them were pointed out. In addition, we provide an RDscan workflow that was designed to find deletions and predict RDs using paired-end short-read sequencing data. Validation assessment of the workflow on a collection of ~7,000 WGS samples showed the high specificities of the identified RDs for different phylogenetic groups.

RESULTS AND DISCUSSION

Sample collection and phylogeny. Genomic analysis was performed on 9,471 SRA paired-end read sets and 367 complete *Mycobacterium tuberculosis* complex genomes. MTBC species differentiation and phylogenetic lineage confirmation were done using SNP-based SNP-IT software (26), main lineages within *M. tuberculosis* were identified based on the Coll et al. (11) classification, while the Shitikov et al. (27) and Palittapongarnpim et al. (28) classifications were used to determine more specific sublineages within lineage 2 and lineage 1, respectively. After the initial quality screening, 7,094 samples belonging to all known species and lineages met the selection criteria and were used for further analysis (see Table S1 in the supplemental material). In the final data set, *M. tuberculosis* genomes comprised most of the samples in the collection ($n = 6{,}993$) and consisted mainly of lineage 2 ($n = 2{,}365$) and lineage 4 ($n = 3{,}152$) genomes, as these are the two most globally distributed *M. tuberculosis* lineages. It should be noted that all known lineage 2 and lineage 4 sublineages were identified among the samples used in this study. Lineage 1 members were allocated to different sublineages, with support from the work of Coll et al. (11) and the more in-depth SNP schemes of Palittapongarnpim et al. (28). According to the Coll et al. typing scheme, the analysis was unable to successfully differentiate lineage 3 sequences into well-supported sublineages (only 296 of 993 samples were differentiated at the sublineage level). Other members of the complex accounted for 127 samples. Most of these members were *M. orygis* ($n = 32$) and *M. caprae* ($n = 22$) isolates, while both *M. mungi* and *M. suricattae* had only one isolate per species.

To equilibrate the number of samples within the groups, for further phylogenetic analysis, the data set was subsampled to contain ~10 samples per species/sublineage (Table S2). Samples were chosen so that the final data set would include the maximum variety of samples belonging to different WGS projects. The maximum-likelihood phylogenetic tree of 721 MTBC genomes was inferred using 30,166 SNPs and rooted on *M. canettii*, the phylogenetically closest relative of the MTBC (Fig. 1). The present phylogenetic analysis demonstrated that the clustering of MTBC isolates is fairly consistent with those of previously published MTBC phylogenies (11, 23), with the advantage of this assay being that it was able to combine in one tree members of different species and keep consistent sublineage differentiations.

A phylogenetic tree showed that two main evolutionary branches can be distinguished. One clade consists of *M. tuberculosis* members, among which ancient and modern lineages can be discerned. The second clade contains human-adapted lineage 5 and lineage 6 members and animal-adapted MTBC species. In agreement with the previously published study, animal-adapted species can be divided into four clades: A1 (*M. suricattae*, *M. mungi*, chimpanzee bacillus, and "Dassie" bacillus [chimpanzee and "Dassie" bacilli are not included in this study]), A2 (*M. microti* and *M. pinnipedii*), A3 (*M. orygis*), and A4 (*M. caprae* and *M. bovis*) (23).

Deletion discovery in MTBC genomes. For the detection of deletions, the same set of paired-end short-read samples belonging to all main MTBC lineages and sublineages was used ($n = 721$). In total, 14,471 deletions were found in the data set, the largest of which was 29,106 bp (this is an RD^{Rio} deletion, which has been falsely said to be 2,800 bp longer due to poor coverage in that region of a specific sample); on average, 20 (SD = 14) deletions per genome were discovered (Fig. 2a and b). The outlier peak was discovered among deletion lengths at 9,238 bp (Fig. 2c), which corresponded to deletions in the CRISPR locus.

The largest average length of deletion per sample was found in lineage 6 isolates (3,209 bp per sample), followed by animal-adapted species (3,110 bp per sample), and the highest frequency of deletions per genome was observed in animal-adapted genomes (31 deletions per sample, SD = 12) (Fig. 2a).

Fisher's exact test for enrichment estimation of 969 samples affected by deletions in genes, based on the TubercuList database annotation (http://genolist.pasteur.fr/TubercuList/), identified two overrepresented functional categories comprising 165 genes ($P < 0.05$): "PE/PPE families" ($P = 1.89E{-}39$; $n = 118$) and "Insertion sequences and phages" ($P = 0.01$; $n = 47$). The second enrichment test was performed with annotation, based on a gene's essentiality for the bacterial life cycle (29); this test showed that the only enriched category in the investigated gene set is "nonessential" ($P = 1.69E{-}24$; $n = 838$). However, two essential genes disrupted by deletions were present in the call

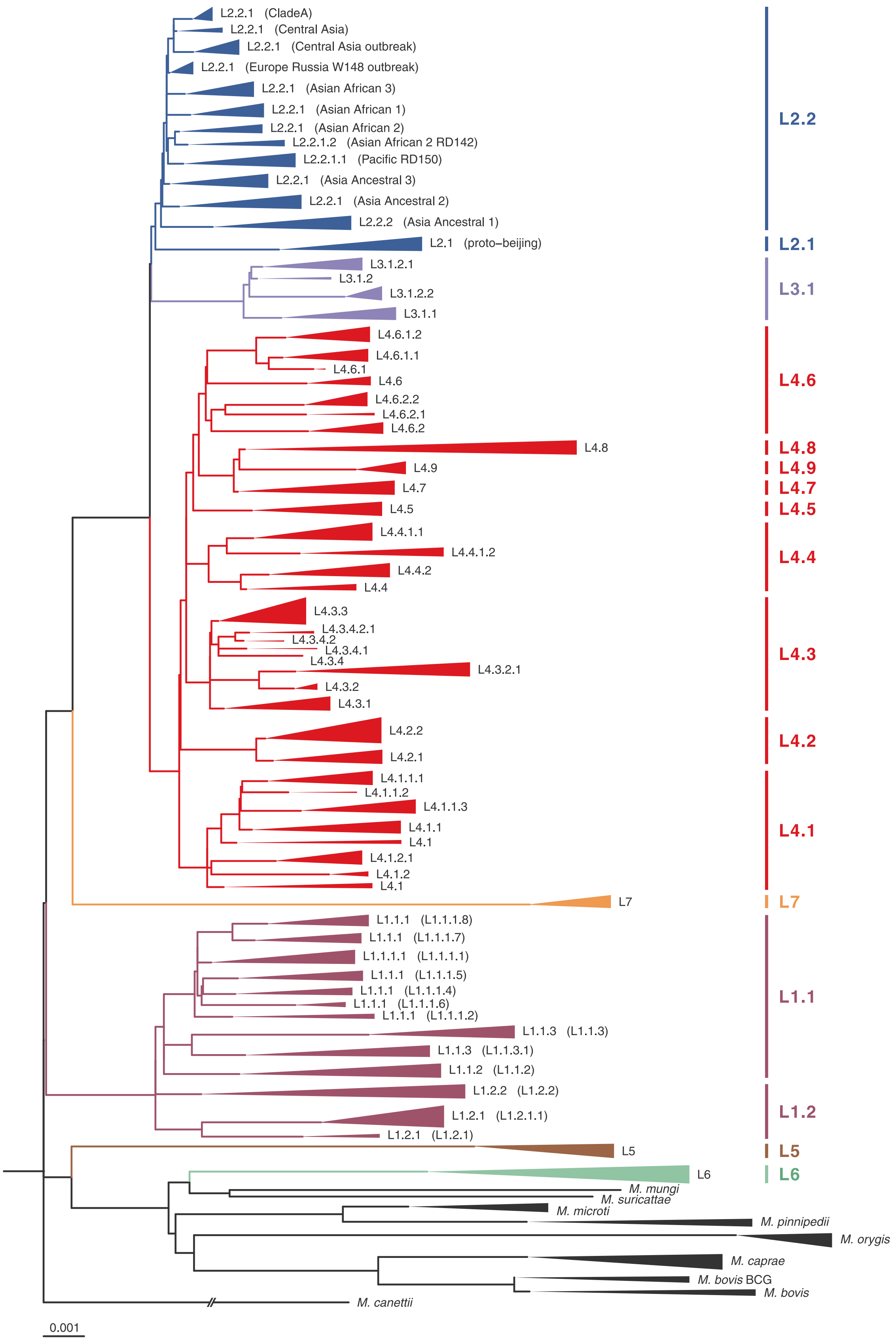

FIG 1 Maximum-likelihood phylogeny of MTBC species. Maximum-likelihood phylogenetic tree of 721 genomes, inferred using 30,166 nonrecombinant core genome SNPs. The scale bar indicates the number of nucleotide substitutions per site. The tree is rooted on *M. canettii* (branch length is omitted).

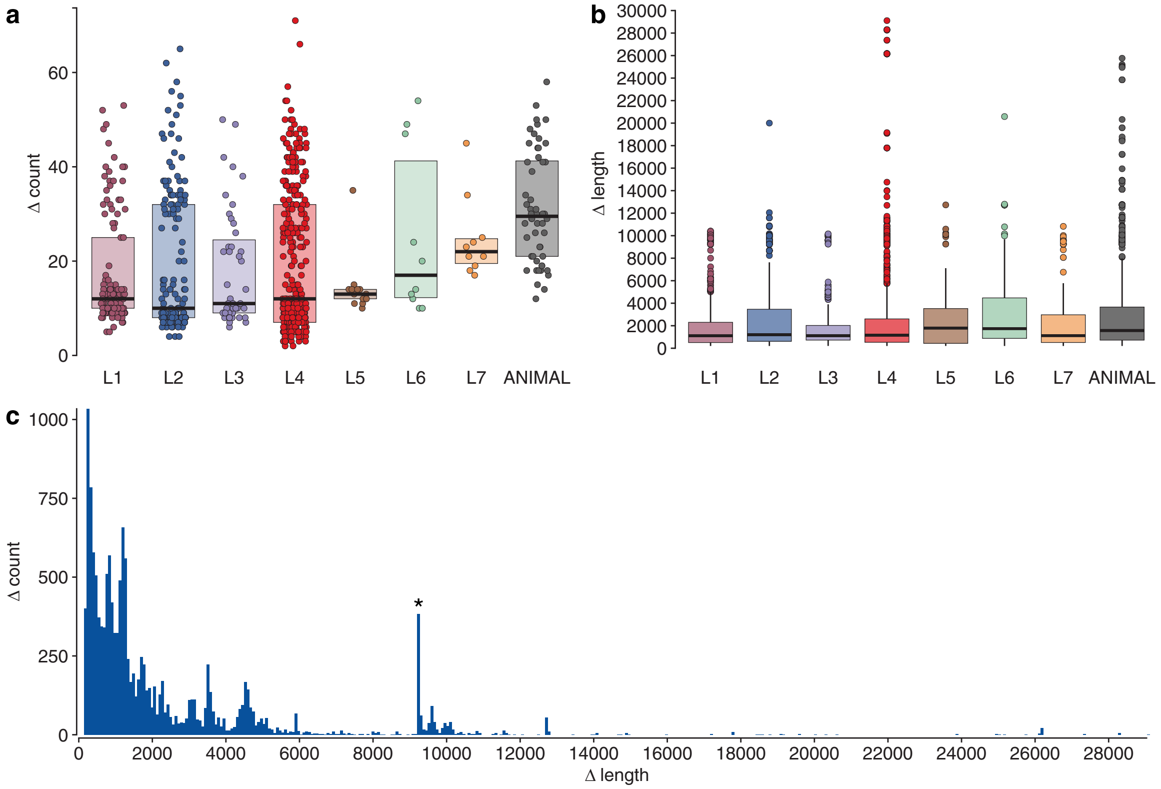

FIG 2 Characteristics of deletions in MTBC samples. (a) Deletions per genome distribution among MTBC strains. Each point represents an individual sample. The *y* axis indicates the number of deletions per sample. The box represents the interquartile range that contains 50% of the values. A line across the box indicates the median. (b) Deletion length distribution among lineages. The *y* axis shows deletion length, and points represent outliers. The boxes indicate upper and lower quartiles, and the horizontal lines mark the medians. Whiskers indicate maximum and minimum values, excluding outliers. (c) Size distribution of deletions among all samples. The outlier peak is marked with an asterisk. L1 to L7, lineages 1 to 7.

set. The Rv1122 gene (*gnd2*; 6-phosphogluconate dehydrogenase) was partially deleted in some members of lineage 4.6.2 (5 samples), while the Rv2017 gene (transcriptional regulator) was disrupted in a number of independent deletion events.

Species-, lineage-, and sublineage-specific RDs. A comprehensive list of RDs, based on previously published studies, was made to correspond to deletions from this study with already-well-described loci (Table S3). Additionally, *M. tuberculosis* H37Rv-related RvD1-5, TbD1, and RD900 deletions were used in the current study (17, 30, 31). A total of 187 RDs were included; out of these, 61 belonged to animal-adapted regions, and the remaining regions corresponded to *M. tuberculosis*.

The analysis revealed two types of deletions that agree with the results of previously published studies (17). The first type was affiliated with repeat sequences or mobile genetic elements, such as prophages and insertion sequences. The name of these deletions is collective, and their breakpoints may vary both at the species level and with strains of the same lineage level. The latter significantly reduces the putative differential capabilities of particular RDs. Some of the best-known types of such regions are RD3 (also known as DS5 and RD149) and RD11 (also known as DS10 and RD198a), related to prophage sequences phiRv1 and phiRv2, respectively. These elements are deleted among different species and lineages of the MTBC, which points to the independence of these deletions and the instability of the affected genomic regions. Other examples of these deletions are RD6, containing IS*1532*, or RD5, RvD2, RvD3, RvD4, RvD5, and RD152, related to deletions in IS*6110* flanking regions or recombination events between repeats. MID3 and MID4 also belong to this type of deletion and are associated with repetitive sequences in MTBC genomes.

The most notable RD related to this type of deletion is RD5. This deletion is often mentioned in the literature, and it has been reported to contribute to the virulence of the MTBC members (32). Initially, RD5 was described when *M. tuberculosis* H37Rv was compared with *M. bovis* BCG (31) and is often found (with various positions) among different *M. tuberculosis* strains. However, in the case of animal-adapted strains, specific deletion breakpoints are indicated (Fig. S1). *M. bovis* and *M. caprae* have the same breakpoints of deleted RDs, which can indicate the presence of this deletion in ancestor species, whereas, in the case of *M. orygis*, a wide variety of deletions was discovered with slightly different 5′-end positions, while the 3′ end corresponded to previously described RD5^{oryx} (33). For *M. microti*, the presence of RD^{mic} could not be confirmed due to the fact that the analyzed genomes, as well as the reference genome (GenBank accession no. CP010333.1), were intact at this locus. Only MTBC strains that cause tuberculosis in voles have been described to have this deletion, whereas human strains do not (6). *M. pinnipedii* as well as *M. microti*, being a member of animal-adapted clade A2, was also intact at the RD5 locus. It is not yet possible to judge the diversity of RD5 in *M. mungi*, *M. surricattae*, and "Dassie" bacillus due to the small number of publicly available strains. Nonetheless, it can be assumed that RD5^{das} (34) and RD5^{sur} (35) are identical deletions.

The second type of deletion corresponds to RDs, the flanking regions of which do not contain repeat sequences. As a result, we found 79 such deletions, which were characteristic of phylogenetic units derived from SNP analysis. It should be noted that characteristic deletions were found for all lineages, as well as for most sublineages of the complex. Out of 79 RDs, 33 were specific to *M. tuberculosis* lineages 1 to 4 and lineage 7 and correlated with well-known RDs; also, 10 new deletions that had not been described prior to this study were found (Fig. 3). Seven out of the 10 new deletions were sublineage specific, while the other 3 (RD311, RD316, and RD306) could be found across different sublineages within the lineage. The small deletion RD311 (213 bp), which leads to Rv2434c inactivation, was found among all modern Beijing strains and can serve as an additional marker for the detection of strains belonging to this group (exceptions are bmyc26 group strains, belonging to the ancient Beijing genotype family, as they do not bear a deletion in this locus [data not shown]). The RD316 (1,297 bp) deletion, resulting in the loss of the Rv3516 and Rv3517 genes, is specific to all members of lineage 3. Rv1179c and was truncated by RD306 (256 bp), which was specific to lineage 4.4.1.1 and lineage 4.4.1.2 sublineages.

For the second phylogenetic clade, comprising *M. tuberculosis* lineage 5, lineage 6, and animal-adapted MTBC species, 29 previously described and 7 novel RDs were found (Fig. 3). Newly described RD307, RD312, and RD317 were found in lineage 5 samples. RD303 (375 bp), affecting the Rv0267 (*narU*) gene, was specific to lineage 6. RD301 and RD315 were unique to *M. orygis*, while the RD305 deletion was specific to *M. caprae*.

Distribution of specific deletions across the H37Rv genome. The largest number of RDs specific for phylogenetic units ($n = 79$) was located in independent regions of the genome, while a slight symmetry in the distribution of deletions relative to the origin of replication was observed. It should also be noted that a slightly higher number of specific deletions than in other regions was found in the 1.3- to 3.0-Mbp genomic region, which is consistent with previously published findings (36).

Overlaps spanning full or partial deletion lengths were found across 29 deletions when one of the deletions intersected another or was located directly inside the largest one (Fig. 4). For *M. tuberculosis*, one such pair of overlapping deletions is RD105 and RD105ext. RD105ext is specific to the members of proto-Beijing lineage 2, while RD105 is a classic marker for all other lineage 2 members. An RD150 deletion affecting four genes (Rv1671, Rv1672c, Rv1673c, Rv1674c) was found among lineage 2.2.1.1 (Pacific RD150) isolates, and it overlaps the larger RD309 deletion, specific for lineage 4.6.2.2 (Fig. 4a).

For six regions, the intersection of RDs specific for two large evolutionary branches was found (Fig. 4b). Lineage 4.3.2.1-specific RD761 has breakpoints similar to those of lineage 5-specific RD711. RD306 (lineages 4.4.1.1 and 4.4.1.2) is almost exactly the same as RD317 (lineage 5). RD252 and RD^{bovis} are specific for lineage 4.1.1.1 and *M. bovis*, respectively, and also have similar breakpoints. Lineage 4.3.4-specific RD174 intersects with

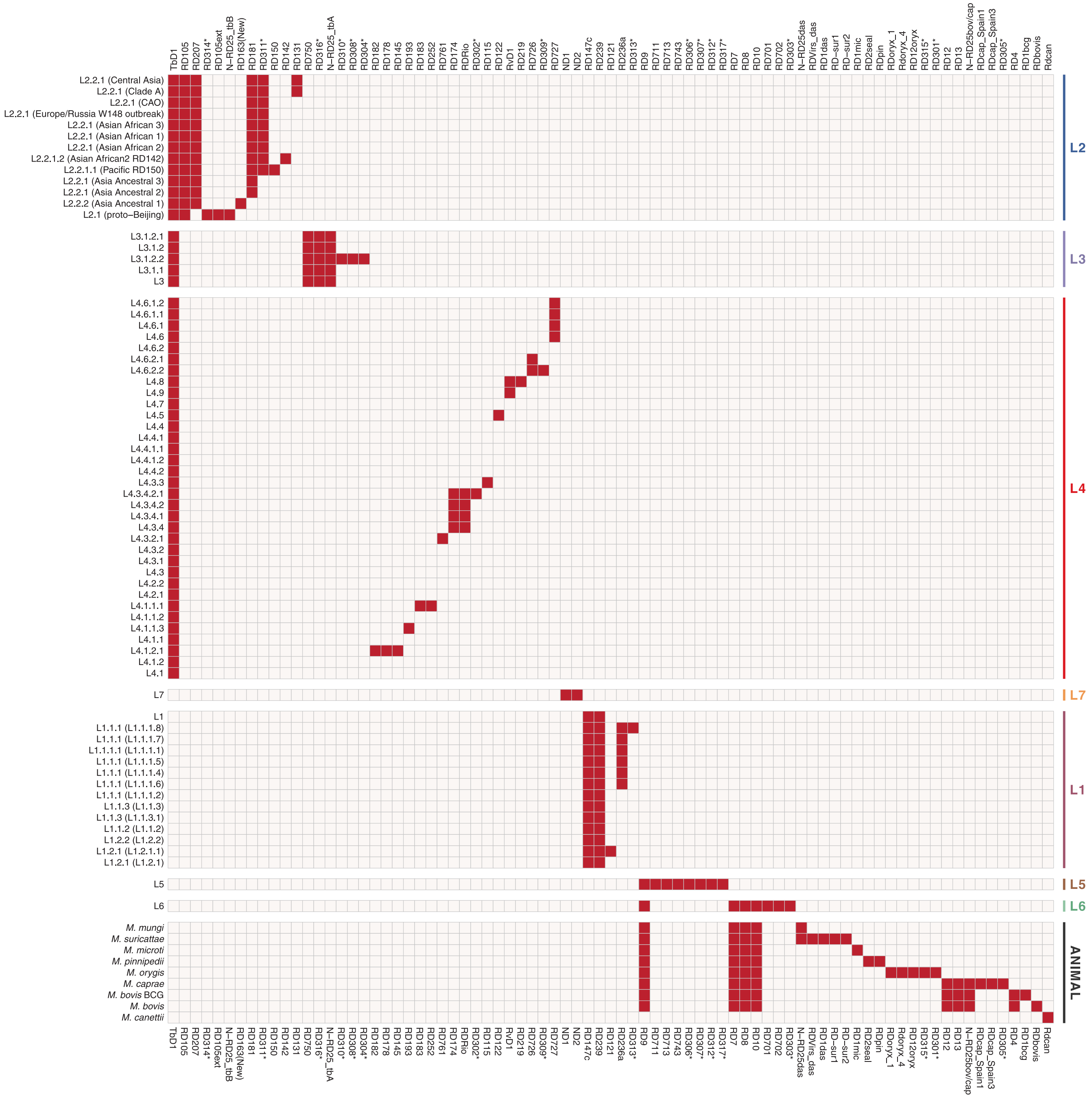

FIG 3 RD distribution across main MTBC phylogenetic units. RDs present in *M. tuberculosis* H37Rv and absent in the studied lineages are in red (RvD1 and TbD1 are exceptions). The rows represent lineages within *M. tuberculosis* or MTBC species, and each column is a specific region of difference. Lineages and species in rows are arranged according to their phylogenetic relationship based on SNP analysis. RDs found in this study are marked with asterisks.

RD743, specific to lineage 5; both deletions affect the Rv1995 and Rv1996 genes, belonging to the "Growth-Advantage" gene group (29). Another overlapping pair contains RD8 (lineage 6 animal-adapted species) and RD236a, which is specific for some lineage 1 sublineages. The N-RD25 deletion was found in different *Mycobacterium* genus members: N-RD25^{tbA} was found to be specific for many lineage 3 strains, N-RD25^{tbB} for lineage 2.1 (proto-Beijing), N-RD25$^{bovis/caprae}$ for *M. bovis/M. caprae*, and N-RD25^{das} for *M. mungi*, *M. suricattae*, and "Dassie" bacillus.

Ten overlapping RDs were identified in phylogenetic clades, including *M. tuberculosis* lineage 5, lineage 6, and animal-adapted species (Fig. 4c). One such deletion is RD12; it

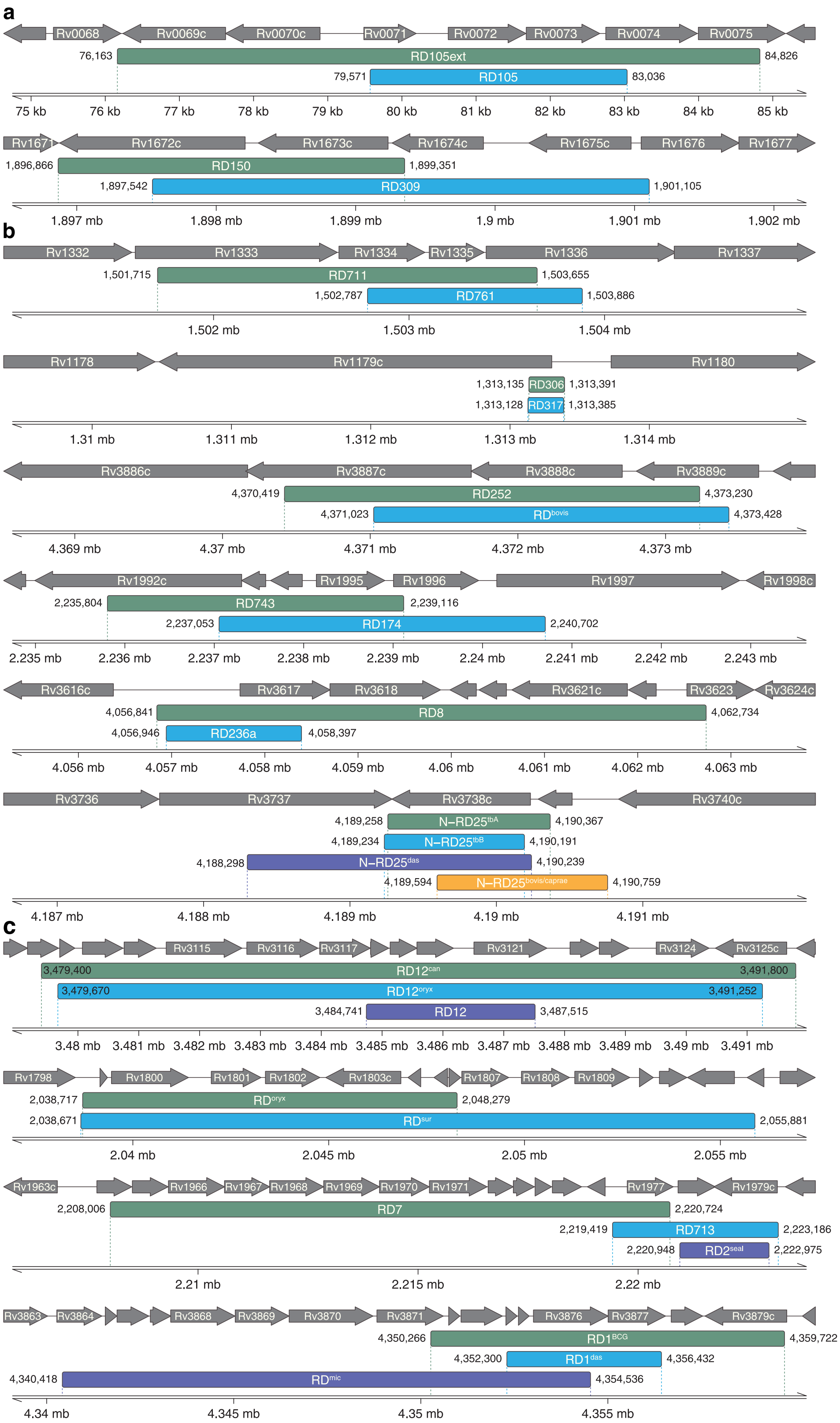

FIG 4 Overlapping RDs in different MTBC members. Deletions relative to the *M. tuberculosis* H37Rv genome are shown in green, blue, purple, and orange; gray arrows indicate genes. Overlapping RDs within *Mycobacterium tuberculosis* lineages (a), between *M. tuberculosis* (lineages 1 to 4 and lineage 7) and other MTBC members (b), and within *M. tuberculosis* lineage 5 and lineage 6 and the animal-adapted clade (c).

was originally identified as specific for *M. bovis* (31), but the deletion was also found in *M. caprae*. RD12 is overlapped by a larger RD12^{oryx} deletion specific for *M. orygis*, as well as comparably sized RD12^{can}, which was found in almost all the *M. canettii* samples included in this study. One more overlapping pair is RD^{sur1} and RD$^{oryx\_1}$, in which RD^{sur1} is much larger and affects 15 genes versus 8 genes affected by RD$^{oryx\_1}$. Consecutively, for the RD7 region (lineage 6 plus animal-adapted species), a small intersection with RD713, specific for lineage 5, was found. It was also discovered that RD713 completely overlaps RD2^{seal}, which is specific for *M. pinnipedii*. In addition, the complexity of this region lies in the fact that some BCG vaccine strains also contain a large RD2 deletion in this region (not shown in Fig. 3 and 4, since the marker is detected in only some strains [37]). The last RD related to this group was RD1, which was also originally identified in vaccine strains and is thoroughly described in previously published studies in connection with its virulence role (34).

RDscan workflow testing. To assess the performance of the RDscan pipeline, the following state-of-the-art structural variant (SV) detection tools were used: delly (v.0.8.7) (38), TIDDIT (v.2.12.1) (39), and breseq (v.0.35.7) (40). These tools were chosen because, in our practice, they produced the best results on haploid genomes detecting large deletions from mapped WGS reads.

For this benchmarking, we used *M. microti* strain OV254 (ENA database run accession no. ERR027294). Strain OV254 has been reported to harbor deletions in RD1^{mic}, RD3, RD7, RD8, RD9, RD10, MiD3, RD11 (partial), MiD1, RD5^{mic}, and MiD2 RDs (41). In addition, we manually curated the *M. microti* OV254 deletions using samplot (v.1.1.6) (https://github.com/ryanlayer/samplot) and IGV (v.2.9.4) (42). Consequently, the RD236a deletion was discovered, as well as some non-RD-specific deletions.

To compare the mentioned tools with RDscan, we used sensitivity (equation 1), precision (equation 2), and F_1 score (equation 3) as defined in the following equations:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$F_1 = 2 \times \frac{\text{sensitivity} \times \text{precision}}{\text{sensitivity} + \text{precision}} \tag{3}$$

where TP means true positive, the number of correctly identified deletions, FP means false positive, the number of nondeleted regions that were incorrectly identified, and FN means false negative, the number of deletions that were incorrectly rejected. The harmonic mean between precision and sensitivity (F_1 score) was used to determine the tool with the best balance between sensitivity and precision.

RDscan showed the highest sensitivity and F_1 score among all tools (Table 1). TIDDIT had the best precision (90%) and second-best performance (52%) but lacked in sensitivity compared to RDscan and Delly. In the case of known RDs in the genome, RDscan did not register partial deletions in RD11, Delly did not find deletions in MiD1, RD5^{mic}, and MiD2, breseq did not find deletions in RD3, MiD3, MiD1, RD5^{mic}, and MiD2, and TIDDIT did not find deletions in RD3, RD11, MiD1, RD5^{mic}, and MiD2.

This performance benchmark shows that RDscan surpasses other methods in terms of overall performance (59%) and sensitivity (92%).

TABLE 1 Comparisons of different tools for the *M. microti* OV254 genome[a]

Tool	Sensitivity	Precision	F_1
RDscan	**0.920**	0.442	**0.597**
TIDDIT	0.375	**0.900**	**0.529**
Delly	0.407	0.180	0.250
breseq	0.280	0.467	0.350

[a] Values in bold are the best results for the corresponding evaluation criteria.

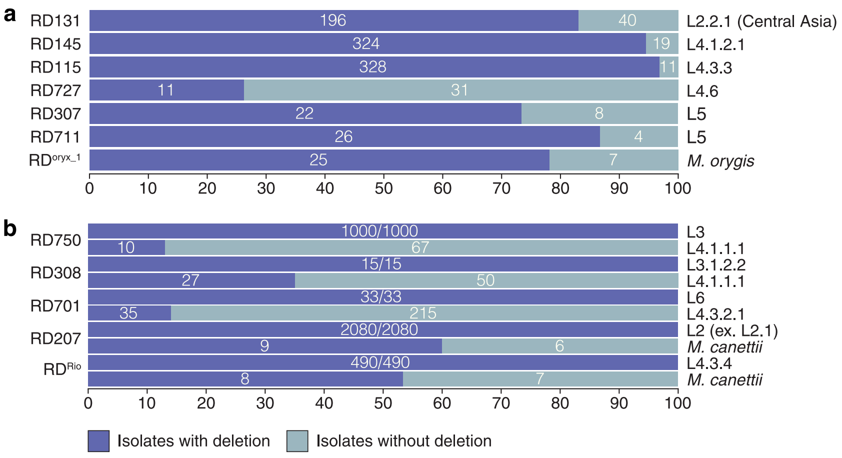

FIG 5 Ambiguous and interlineage RDs. Stacked bar plots showing the percentages of studied isolates with (blue) and without (gray) particular deletions. (a) Ambiguous RDs; (b) interlineage RDs.

A more extensive analysis was performed to validate the efficacy of RDscan in inferring large deletions. The pipeline was run on an initial data set comprising 7,094 paired-end samples (Table S1). Putative regions of difference discovered with RDscan were compared with a database containing currently well-described RDs. As a result, the presence of all RDs discovered in a smaller data set was confirmed. In addition, for some regions, we noted some specificities that must be taken into account in further analysis.

First, for 7 out of 79 RDs, it was found that not all members of the group contained the analyzed deletion (Fig. 5a). This fact suggests that the SNP markers underlying modern typing are, in this case, phylogenetically earlier than the analyzed deletions. For example, RD115, RD145, RD131, and RD727 were not found in some samples of lineage 4.3.3, lineage 4.1.2.1, lineage 2.2.1 (Central Asia) and lineage 4.6, respectively. For RD711, specific for lineage 5, the ratio of intact strains with respect to this locus was 13.4%, which is consistent with previously published observations (43). Moreover, the RD307 deletion was found even within isolates with RD711 deleted, which was not previously described. The final notable RD is RD$^{oryx\_1}$, the deletion of which was found in 25/32 samples of *M. orygis*. In this case, the absence of the deletion should be attributed to the incorrectly rejected false-negative results of the pipeline, which were detected during manual data curation. It should be noted that part of this region is still present in these *M. orygis* samples but in a different region of the genome, which leads to false results.

Another notable group contained RDs whose specificities were reduced due to the detection of similar deletions in other populations of mycobacteria ($>$10% of the sublineage/lineage/species). In total, five such loci were identified (Fig. 5b). It should be noted that the boundaries of deletions for specific RDs never coincided with nonspecific LSPs, which once again emphasizes the independence of recombination events that have already occurred. The most prominent of these regions were RD701, RD750, and RD207, specific to lineage 6, lineage 3, and lineage 2 (except lineage 2.1), respectively. RD701 was deleted in 11% of lineage 4.3.2.1 samples, RD750 was deleted among lineage 4.1.1.1 strains, while RD207 was deleted in a significant number of *M. canettii* strains. RD^{Rio}, specific to lineage 4.3.4, is seldom detected in other mycobacteria; similar deletions were detected only among some *M. canettii* strains. However, this region is generally unstable and contains many repeats and corresponding deletions. For example, MID3 overlaps significantly with RD^{Rio} and is found in lineage 4.6.2.2, *M. pinnipedii*, *M. microti*, and *M. canettii*.

The third and most significant group consisted of overlapping RDs, which were described earlier. According to the RDscan results, only the following RD combinations can be used for reliable strain differentiation: RD105/RD105ext, RD761/RD711, RD236a/RD8, and RD1^{BCG}/RD1^{das}/RD1^{mic}.

Conclusions. Prior in-depth studies of RDs have provided a fundamental understanding of the evolutionary history of MTBC members. Furthermore, the great majority of studies may be divided into two types. The first type focuses on the investigation of phylogenetic interactions of species within the complex, where only the most significant regions are taken into account for *M. tuberculosis*. The second type concentrates on the search for deletions within *M. tuberculosis* without regard to other members of the complex.

Here, we closed the gap by collating all the previously described RDs and analyzing them on a sample set representing the complete variety of MTBC members. Although we expectedly confirmed the convergence of the main classifications at the SNP and RD levels, we also described the RDs that may overlap; in addition, we showed that some RDs are not always specific to the sublineages. It is important to note that our method has its drawbacks. Due to the high number of *M. tuberculosis* strains in the NCBI database, few other MTBC members were used in this study; this means that the found deletions, especially those that are sublineage specific, should be treated with caution. Another issue is that we were unable to classify lineage 3 samples, and most of the deletions found were characteristic of the entire lineage. The third disadvantage of this study is that all new deletions found have been identified *in silico* and require further experimental verification.

To alleviate the aforementioned disadvantages and facilitate the work with comprehensive genomics data, we have created a pipeline that can search for deletions in MTBC genomes. Its main advantage is that it is able to correlate deletions with a list that can be modified to meet the needs of a particular task. The herein-proposed RDscan pipeline can be used as a practical tool to rapidly infer known deletions in RDs and thus differentiate the strains, as well as describe the new deletions associated with the evolution of the pathogen.

MATERIALS AND METHODS

Data set. For the analysis, a collection of 9,471 draft MTBC genomes publicly available in the NCBI SRA archive (https://www.ncbi.nlm.nih.gov/sra) was used. Target SRA files were downloaded with the prefetch (v.2.10.3) tool from the SRA Toolkit (http://www.ncbi.nlm.nih.gov/books/NBK158900/). Parallel-fastq-dump (v.0.6.6) (https://github.com/rvalieris/parallel-fastq-dump) was used to extract paired-end FASTQ reads from SRA files. A quality control (QC) check on all acquired reads was done with FastQC (v.0.11.9) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In addition, 367 complete MTBC genomes were obtained from the NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/).

SNP calling and lineage typing. SNPs against the *M. tuberculosis* H37Rv (GenBank accession no. NC_000962) and *M. canettii* (NC_015848.1) genomes were inferred using the Snippy pipeline (https://github.com/tseemann/snippy). NUCmer (v.3.1) (44) was used to call SNPs from complete MTBC genomes. BCFtools (v.1.9) (https://github.com/samtools/bcftools) was used to collect statistics on called variants. Mapping quality was assessed with Qualimap (v.2.2.2) (45). MultiQC (v.1.9) (46) was used for QC report aggregation. Only samples with at least 80% of mapped reads and with a ≥50× mean coverage were used for further analysis. Lineage/sublineage typing was performed using TB-Profiler (v.2.8.12) (47), KvarQ (v.0.12.2) (48), BioHansel (v.2.4.0) (https://github.com/phac-nml/biohansel), SNP-IT (v.1.0.0) (26), and in-house python scripts used with various previously published typing schemes (11, 27, 28).

Phylogenetic analysis. A core SNP alignment was produced with snippy-core (v.4.6.0) (https://github.com/tseemann/snippy). Gubbins (v.2.4.1) (49) was used to filter out recombinant regions from the alignment. The resulting alignment was cleaned to include only core polymorphic sites with SNP sites (50). Cleaned core alignment was used to construct a phylogenetic tree via RAxML-NG (v.1.0.1) (51) using the GTR+G model and 100 bootstrap iterations; the tree was rooted on *M. canettii* (GenBank accession no. NC_015848.1). The tree was visualized with the ggtree (v.2.0.2) (52) package for R (v.4.0.2) (53).

Structural variants detection. To detect regions with structural variants (SVs), i.e., large deletions [>200 bp], regions with low coverage and with a length of ≥100 bp were extracted with covtobed (v.1.2.0) (54); further regions that were located within 1,500 bp of each other were merged with bedtools (v.2.29.2) (55). The SURVIVOR (v.1.0.7) (56) tool was used to convert the resulting .bed files with SV breakpoints to variant call format (VCF) and to further merge these files into a single multisample .vcf file. The resulting .vcf file was annotated with SnpEff (v.4.1l) (57). Identified RDs were detected by calculating median coverage in these regions with mosdepth (v.0.3.1) (58) and dividing it by median coverage of the full mapping length. A 5% threshold was used to determine whether RD regions are present in the sample. GNU parallel (v.20161222) (59) was used to speed up some parts of the analysis. Small deletions of <200 bp and deletions larger than 30,000 bp were eliminated from the analysis to reduce the number of false-positive calls. All calls were manually curated using Integrative Genomics Viewer (IGV) (v.2.8.4) (42) and samplot (v.1.0.20) (https://github.com/ryanlayer/samplot). Breakpoints were curated using *de novo*-assembled MTBC genomes; for the *de novo* assembly, genomes were cleaned of low-quality reads and adapters with fastp (v.0.20.1) (60) and assembled using Unicycler (v.0.4.8) (61). Only high-confidence deletions were kept for downstream analysis. Plots were generated within R (v.4.0.2) (53) using the ggplot2 (v.3.3.2) (62), cowplot (v.1.1.0) (https://CRAN.R-project.org/package=cowplot), Gviz (v.1.32.0) (63), lemon (v.0.4.5) (https://CRAN.R-project.org/package=lemon), and ComplexHeatmap (v.2.7.6.1004) (64) packages.

RDscan workflow. An RDscan workflow was designed for deletion discovery in MTBC species using paired-end short-read FASTQ files. RDscan is implemented as a custom Snakemake (65) workflow. The workflow can be divided into two blocks; the first block finds all putative deletions, while the second scans whether already-known RDs are present in the sample. Concisely, reads are mapped to the *M. tuberculosis* H37Rv (GenBank accession no. NC_000962) reference genome using BWA-MEM (66). After the mapping step is finished, .bam files are indexed with SAMtools (67). Next BEDTools and SAMtools are used to generate .bed files with per-sample breakpoints by searching for regions with low coverage. Then SURVIVOR, GATK (68), and BCFtools are used to convert .bed files with putative deletions to .vcf files and prepare them for further steps. Duphold (69) is then used to calculate fold change for the deletion depth relative to flanking regions; the resulting .vcf files are filtered with BCFtools by minimum and maximum lengths (200 bp < deletion < 30,000 bp) and a duphold flank fold change (DHFFC) of <0.1. . vcf files from multiple files are then merged into a single call set using SURVIVOR and annotated with SnpEff; the resulting cohort .vcf file containing all deletion calls is transformed into a table using GATK, and putative RD regions are annotated. The second block starts with coverage computation using mosdepth in 79 specific RD regions identified and curated in this study. Lastly, the ratio of read depth in RD regions to full reference length depth is calculated, and results are merged into a single data frame; human-readable tables are then generated.

Functional enrichment analysis. To determine the significance of genes affected by deletions, functional categories from the TubercuList database (http://genolist.pasteur.fr/TubercuList/) and a custom database based on *M. tuberculosis* gene essentiality (29) were used. Enrichment scores of functional categories were obtained using Fisher's exact test in R (v.4.0.2).

Data availability. RDscan is an open-source software available in the GitHub repository at https://github.com/dbespiatykh/RDscan.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, EPS file, 0.5 MB.

TABLE S1, XLSX file, 0.01 MB.

TABLE S2, XLSX file, 0.03 MB.

TABLE S3, XLSX file, 0.02 MB.

ACKNOWLEDGMENT

This work was supported by the RFBR (research project no. 20-04-00686).

REFERENCES

1. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. 2009. Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol 7:537–544. https://doi.org/10.1038/nrmicro2165.
2. Vasconcellos SEG, Huard RC, Niemann S, Kremer K, Santos AR, Suffys PN, Ho JL. 2010. Distinct genotypic profiles of the two major clades of Mycobacterium africanum. BMC Infect Dis 10:80. https://doi.org/10.1186/1471-2334-10-80.
3. Garnier T, Eiglmeier K, Camus JC, 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. 2003. The complete genome sequence of Mycobacterium bovis. Proc Natl Acad Sci U S A 100:7877–7882. https://doi.org/10.1073/pnas.1130426100.
4. van Soolingen D, Hoogenboezem T, de Haas PEW, Hermans PWM, Koedam MA, Teppema KS, Brennan PJ, Besra GS, Portaels F, Top J, Schouls LM, van Embden JDA. 1997. A novel pathogenic taxon of the Mycobacterium tuberculosis complex, Canetti: characterization of an exceptional isolate from Africa. Int J Syst Bacteriol 47:1236–1245. https://doi.org/10.1099/00207713-47-4-1236.
5. Niemann S, Richter E, Rüsch-Gerdes S. 2002. Biochemical and genetic evidence for the transfer of Mycobacterium tuberculosis subsp. caprae Aranaz et al. 1999 to the species Mycobacterium bovis Karlson and Lessel 1970 (approved list 1980) as Mycobacterium bovis subsp. caprae comb. nov. Int J Syst Evol Microbiol 52:433–436. https://doi.org/10.1099/00207713-52-2-433.
6. Brodin P, Eiglmeier K, Marmiesse M, Billault A, Garnier T, Niemann S, Cole ST, Brosch R. 2002. Bacterial artificial chromosome-based comparative genomic analysis identifies Mycobacterium microti as a natural ESAT-6 deletion mutant. Infect Immun 70:5568–5578. https://doi.org/10.1128/IAI.70.10.5568-5578.2002.
7. Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, van Pittius NCG. 2010. Novel mycobacterium tuberculosis complex pathogen, M. mungi. Emerg Infect Dis 16:1296–1299. https://doi.org/10.3201/eid1608.100314.
8. van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, van Soolingen D. 2012. Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies. Emerg Infect Dis 18:653–655. https://doi.org/10.3201/eid1804.110888.
9. Cousins D. v, Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, Duignan P, Murray A, Dupont C, Ahmed N, Collins DM, Butler WR, Dawson D, Rodríguez D, Loureiro J, Romano MI, Alito A, Zumarraga M, Bernardelli A. 2003. Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. Int J Syst Evol Microbiol 53:1305–1314. https://doi.org/10.1099/ijs.0.02401-0.
10. Parsons SDC, Drewe JA, van Pittius NCG, Warren RM, van Helden PD. 2013. Novel cause of tuberculosis in meerkats, South Africa. Emerg Infect Dis 19:2004–2007. https://doi.org/10.3201/eid1912.130268.
11. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun 5:4812. https://doi.org/10.1038/ncomms5812.
12. 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. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 396:190. https://doi.org/10.1038/24206.

13. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, Antoine R, Niyigena EB, Mulders W, Fissette K, Diels M, Gaudin C, Duthoy S, Ssengooba W, André E, Kaswa MK, Habimana YM, Brites D, Affolabi D, Mazarati JB, de Jong BC, Rigouts L, Gagneux S, Meehan CJ, Supply P. 2020. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat Commun 11:2917. https://doi.org/10.1038/s41467-020-16626-6.
14. Supply P, Warren RM, Bañuls A-L, Lesjean S, van der Spuy GD, Lewis L-A, Tibayrenc M, van Helden PD, Locht C. 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. Mol Microbiol 47:529–538. https://doi.org/10.1046/j.1365-2958.2003.03315.x.
15. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. 2004. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci U S A 101:4871–4876. https://doi.org/10.1073/pnas.0305627101.
16. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. 1997. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 94:9869–9874. https://doi.org/10.1073/pnas.94.18.9869.
17. Brosch R, Gordon S. v, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99:3684–3689. https://doi.org/10.1073/pnas.052548299.
18. Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, Desmond EP, Kato-Maeda M, Behr M. 2004. Genomic analysis distinguishes Mycobacterium africanum. J Clin Microbiol 42:3594–3599. https://doi.org/10.1128/JCM.42.8.3594-3599.2004.
19. Tsolaki AG, Gagneux S, Pym AS, Goguet De La Salmoniere YOL, Kreiswirth BN, van Soolingen D, Small PM. 2005. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. J Clin Microbiol 43:3185–3191. https://doi.org/10.1128/JCM.43.7.3185-3191.2005.
20. Lazzarini LCO, Huard RC, Boechat NL, Gomes HM, Oelemann MC, Kurepina N, Shashkina E, Mello FCQ, Gibson AL, Virginio MJ, Marsico AG, Butler WR, Kreiswirth BN, Suffys PN, Silva JRL, Ho JL. 2007. Discovery of a novel Mycobacterium tuberculosis lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. J Clin Microbiol 45:3891–3902. https://doi.org/10.1128/JCM.01394-07.
21. Qin L, Wang J, Lu J, Yang H, Zheng R, Liu Z, Huang X, Feng Y, Hu Z, Ge B. 2019. A deletion in the RD105 region confers resistance to multiple drugs in Mycobacterium tuberculosis. BMC Biol 17:7–12. https://doi.org/10.1186/s12915-019-0628-6.
22. Ru H, Liu X, Lin C, Yang J, Chen F, Sun R, Zhang L, Liu J. 2017. The impact of genome region of difference 4 (RD4) on mycobacterial virulence and BCG efficacy. Front Cell Infect Microbiol 7:239. https://doi.org/10.3389/fcimb.2017.00239.
23. Brites D, Loiseau C, Menardo F, Borrell S, Boniotti MB, Warren R, Dippenaar A, Parsons SDC, Beisel C, Behr MA, Fyfe JA, Coscolla M, Gagneux S. 2018. A new phylogenetic framework for the animal-adapted mycobacterium tuberculosis complex. Front Microbiol 9:2820. https://doi.org/10.3389/fmicb.2018.02820.
24. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V, Supply P, Suresh A, Utpatel C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, de Jong BC, de Vos M, Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC, Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM, Niemann S, Comas I, van Rie A. 2019. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. Nat Rev Microbiol 17:533–545. https://doi.org/10.1038/s41579-019-0214-5.
25. Faksri K, Xia E, Tan JH, Teo YY, Ong RTH. 2016. In silico region of difference (RD) analysis of Mycobacterium tuberculosis complex from sequence reads using RD-Analyzer. BMC Genomics 17:847. https://doi.org/10.1186/s12864-016-3213-1.
26. Lipworth S, Jajou R, de Neeling A, Bradley P, van der Hoek W, Maphalala G, Bonnet M, Sanchez-Padilla E, Diel R, Niemann S, Iqbal Z, Smith G, Peto T, Crook D, Walker T, van Soolingen D. 2019. SNP-IT tool for identifying subspecies and associated lineages of Mycobacterium tuberculosis complex. Emerg Infect Dis 25:482–488. https://doi.org/10.3201/eid2503.180894.
27. Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, Ilina E, Govorun V. 2017. Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis. Sci Rep 7:9227. https://doi.org/10.1038/s41598-017-10018-5.
28. Palittapongarnpim P, Ajawatanawong P, Viratyosin W, Smittipat N, Disratthakit A, Mahasirimongkol S, Yanai H, Yamada N, Nedsuwan S, Imasanguan W, Kantipong P, Chaiyasirinroje B, Wongyai J, Toyo-Oka L, Phelan J, Parkhill J, Clark TG, Hibberd ML, Ruengchai W, Palittapongarnpim P, Juthayothin T, Tongsima S, Tokunaga K. 2018. Evidence for host-bacterial co-evolution via genome sequence analysis of 480 Thai Mycobacterium tuberculosis lineage 1 isolates. Sci Rep 8:11597. https://doi.org/10.1038/s41598-018-29986-3.
29. Dejesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, Rubin EJ, Schnappinger D, Ehrt S, Fortune SM, Sassetti CM, Ioerger TR. 2017. Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. mBio 8:e02133-16. https://doi.org/10.1128/mBio.02133-16.
30. 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. 2012. 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 6:e1552. https://doi.org/10.1371/journal.pntd.0001552.
31. Gordon S. v, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. 1999. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol 32:643–655. https://doi.org/10.1046/j.1365-2958.1999.01383.x.
32. Ates LS, Sayes F, Frigui W, Ummels R, Damen MPM, Bottai D, Behr MA, van Heijst JWJ, Bitter W, Majlessi L, Brosch R. 2018. RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in strong reduction of antigenic repertoire but little impact on protection. PLoS Pathog 14:e1007139. https://doi.org/10.1371/journal.ppat.1007139.
33. Mostowy S, Inwald J, Gordon S, Martin C, Warren R, Kremer K, Cousins D, Behr MA. 2005. Revisiting the evolution of Mycobacterium bovis. J Bacteriol 187:6386–6395. https://doi.org/10.1128/JB.187.18.6386-6395.2005.
34. Mostowy S, Cousins D, Behr MA. 2004. Genomic interrogation of the Dassie Bacillus reveals it as a unique RD1 mutant within the Mycobacterium tuberculosis complex. J Bacteriol 186:104–109. https://doi.org/10.1128/JB.186.1.104-109.2003.
35. Dippenaar A, Parsons SDC, Sampson SL, van der Merwe RG, Drewe JA, Abdallah AM, Siame KK, Gey Van Pittius NC, van Helden PD, Pain A, Warren RM. 2015. Whole genome sequence analysis of Mycobacterium suricattae. Tuberculosis (Edinb) 95:682–688. https://doi.org/10.1016/j.tube.2015.10.001.
36. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM. 2001. Comparing genomes within the species Mycobacterium tuberculosis. Genome Res 11:547–554. https://doi.org/10.1101/gr.166401.
37. Abdallah AM, Hill-Cawthorne GA, Otto TD, Coll F, Guerra-Assunção JA, Gao G, Naeem R, Ansari H, Malas TB, Adroub SA, Verboom T, Ummels R, Zhang H, Panigrahi AK, McNerney R, Brosch R, Clark TG, Behr MA, Bitter W, Pain A. 2015. Genomic expression catalogue of a global collection of BCG vaccine strains show evidence for highly diverged metabolic and cell-wall adaptations. Sci Rep 5:15443. https://doi.org/10.1038/srep15443.
38. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339. https://doi.org/10.1093/bioinformatics/bts378.
39. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. 2017. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Res 6:664. https://doi.org/10.12688/f1000research.11168.2.
40. Deatherage DE, Traverse CC, Wolf LN, Barrick JE. 2014. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. Front Genet 5:468. https://doi.org/10.3389/fgene.2014.00468.
41. Orgeur M, Frigui W, Pawlik A, Clark S, Williams A, Ates LS, Ma L, Bouchier C, Parkhill J, Brodin P, Brosch R. 2021. Pathogenomic analyses of mycobacterium microti, an esx-1-deleted member of the Mycobacterium tuberculosis complex causing disease in various hosts. Microb Genom 7:000505. https://doi.org/10.1099/mgen.0.000505.
42. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754.

43. Huard RC, Fabre M, de Haas P, Lazzarini LCO, van Soolingen D, Cousins D, Ho JL. 2006. Novel genetic polymorphisms that further delineate the phylogeny of the Mycobacterium tuberculosis complex. J Bacteriol 188:4271–4287. https://doi.org/10.1128/JB.01783-05.
44. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
45. Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292–294. https://doi.org/10.1093/bioinformatics/btv566.
46. Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/bioinformatics/btw354.
47. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, O'Grady J, McNerney R, Hibberd ML, Viveiros M, Huggett JF, Clark TG. 2019. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med 11:41. https://doi.org/10.1186/s13073-019-0650-x.
48. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. 2014. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics 15:881. https://doi.org/10.1186/1471-2164-15-881.
49. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196.
50. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mgen.0.000056.
51. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453–4455. https://doi.org/10.1093/bioinformatics/btz305.
52. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics 69:e96. https://doi.org/10.1002/cpbi.96.
53. R Core Team. 2021. R: a language and environment for statistical computing. R Foundation, Vienna, Austria.
54. Birolo G, Telatin A. 2020. covtobed: a simple and fast tool to extract coverage tracks from BAM files. J Open Source Softw 5:2119. https://doi.org/10.21105/joss.02119.
55. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
56. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun 8:14061. https://doi.org/10.1038/ncomms14061.
57. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. 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 (Austin) 6:80–92. https://doi.org/10.4161/fly.19695.
58. Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34:867–868. https://doi.org/10.1093/bioinformatics/btx699.
59. Tange O. 2020. GNU parallel 20201122 ('Biden'). Zenodo. https://doi.org/10.5281/zenodo.4284075.
60. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
61. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
62. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY. https://doi.org/10.1007/978-0-387-98141-3.
63. Hahne F, Ivanek R. 2016. Visualizing genomic data using Gviz and Bioconductor. Methods Mol Biol 1418:335–351. https://doi.org/10.1007/978-1-4939-3578-9_16.
64. Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847–2849. https://doi.org/10.1093/bioinformatics/btw313.
65. Koster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480.
66. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv https://arxiv.org/abs/1303.3997.
67. Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509.
68. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–501. https://doi.org/10.1038/ng.806.
69. Pedersen BS, Quinlan AR. 2019. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. Gigascience 8:giz040. https://doi.org/10.1093/gigascience/giz040.