PknH and RD900 locus analysis across MTBC

name

date

RD900 region was first described as a lineage specific locus in *M. africanum* GM041182 strain (Bentley et al. 2012). The locus is 3141 bp long and contains a single complete gene (MAF12860, *tbd2*) and the 3' end of another (MAF12870, *pknH2*) (Figure 1).

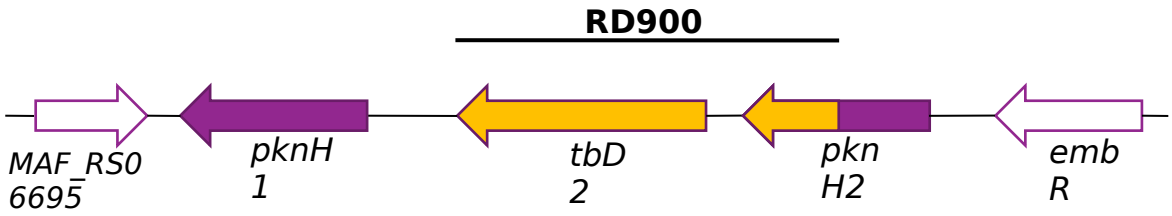


Figure 1. RD900 locus in *M. africanum* GM041182.

This region was thought to be deleted in *M. bovis* and "modern" *M. tuberculosis* lineages. RD900 region was not present in the original *M. bovis* AF2122/97 annotated genome, but was found to be actually present in the later resequencing of *M. bovis* AF2122/97 (Malone et al. 2017).

The *tbd2* gene encodes an ATPbinding cassette (ABC) transport protein that has a central ATPbinding domain and six possible membrane-spanning domains in the C-terminal portion. In addition, the N-terminal region contains Forkhead associated (FHA) domain that may confer the ability to bind DNA and thereby potentially act as a transcriptional regulator.

Tbd2 gene is flanked by similar, codirectional genes, each encoding a protein kinase (*pknH1* and *pknH2*). PknH is a transmembrane serine/threonine protein kinase (STPK). The protein has a kinase N terminal intracellular domain, followed by a proline-rich domain, a single helix predicted transmembrane region and an extracellular C terminal sensor domain (Figure 2).

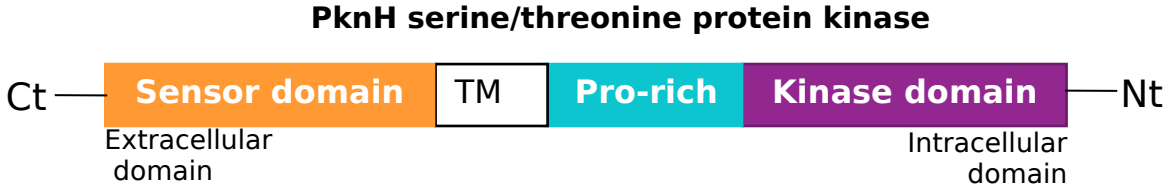


Figure 2. Representation of PknH protein domains.

Comparing the two *pknH* genes present in *M. africanum* and *M. bovis*, there is a high degree of identity between the kinase domains, except for a region of 63 bp that is deleted in *M.* africanum GM041182 pknH2. In the same region, there is a nucleotide variation in *M. bovis* AF2122 *pknH2*. As a result, a 22 amino acid deletion is present in the kinase domain of *M. africanum* GM041182 *pknH2* compared to *pknH1*, and a 22 amino acid substitution is present in the same region of *M. bovis* AF2122 *pknH2*.

The proline rich region of *M. africanum pknH2* has a deletion of about 100 bp compared to *pknH1* and to *M. tuberculosis pknH* single gene. The same deletion is present in *M. bovis* AF2122, but in this case the deletion is present in both *pknH1* and *pknH2* genes.

Finally, the sensor domain is completely different between *pknH1* and *pknH2* in *M. africanum* and *M. bovis*. Sensor domain of *M. tuberculosis* single *pknH* is homologous to the one of *pknH1* (Figure 3).

Figure 3. *PknH* protein sequence alignment. Protein sequence alignment of the single PknH from *M. tuberculosis* H37Rv (PknHTB), PknH1 and PknH2 from *M. africanum* GM041182 (PknH1MAF and PknH2MAF) and PknH1 and PknH2 from *M. bovis* AF2122 (PknH1MB and PknH2MB).

RD900 locus analysis

To understand the genomic variations in this region across the different species from the *M. tuberculosis* complex and the evolutionary meaning of these variations, we performed an analysis of this genomic region from WGS data of different isolates of *M. tuberculosis* complex species (table 1).

To avoid possible misannotations that may have been present in the assembly annotated genomes, we used the original raw reads available from the genome sequencing data (downloaded in fastq format from genbank) of different MTBC isolates. The reads were aligned to a reference sequence using bwa mem and the resulting bam files were sorted and indexed. We choose as reference sequence the *M. africanum* GM041182 RD900 locus, including *tbd2* gene, both *pknH* genes (*pknH1* and *pknH2*) and the two flanking genes *embR* and *MAF_RS06695*.

We visualized the alignments in IGV to identify the presence or absence of RD900 and the presence of the deletion in the proline-rich region. Finally, we did a variant calling analysis to identify the point mutations present in this region in the different MTBC analysed species.

RD900 presence and proline-rich deletion

We found RD900 region to be present in *M. canettii* isolates, with the same organization as *M. africanum* with the *tbd2* gene flanked by two *pknH* genes, and the deletion in the proline rich region present in *pknH2* (Figure 4). The same locus organization is present in the analysed L7 isolates. However, in *M. tuberculosis* L1 RD900 region presence is variable among the different isolates. The region is deleted in some of the isolates, where a single *pknH* gene is present with the complete proline region sequence. RD900 region is also deleted in all analysed isolates from *M. tuberculosis* “modern” lineages (L2,L3 and L4). In *M. africanum* L5 and L6, RD900 region is present but there is variability in the proline rich region sequence of *pknH1*, present in some isolates and deleted in others.

In the animal adaptated species, *M. mungi*, *M. microti* and *M. pinnipedii* analysed isolates have the RD900 region present and the proline-rich deletion present in *pknH2* but not *pknH1.* In contrast, RD900 region is deleted in *M. orygis*, with one single *pknH* gene present with the complete proline-rich region sequence. Finally, in *M. caprae, M. bovis* and BCG strains RD900 presence appears to be variable in the different analysed isolates. In this case the deletion in the proline-rich region is present in both *pknH* genes in those strains with the RD900 region present but also in those with one *pknH* gene in which RD900 region is deleted. The deletion in the proline-rich region seems to be conserved in all *pknH* genes from *M. bovis* and *M. caprae* analysed isolates, except from one *M. bovis* strain (strain B2 7505) with RD900 region deleted and a single copy of *pknH* with the complete proline rich region. This strain was isolated from a human patient in Uganda (Wanzala et al., 2015) and it has been described to show *M. tuberculosis* RD patterns (Zimpel et al., 2017).

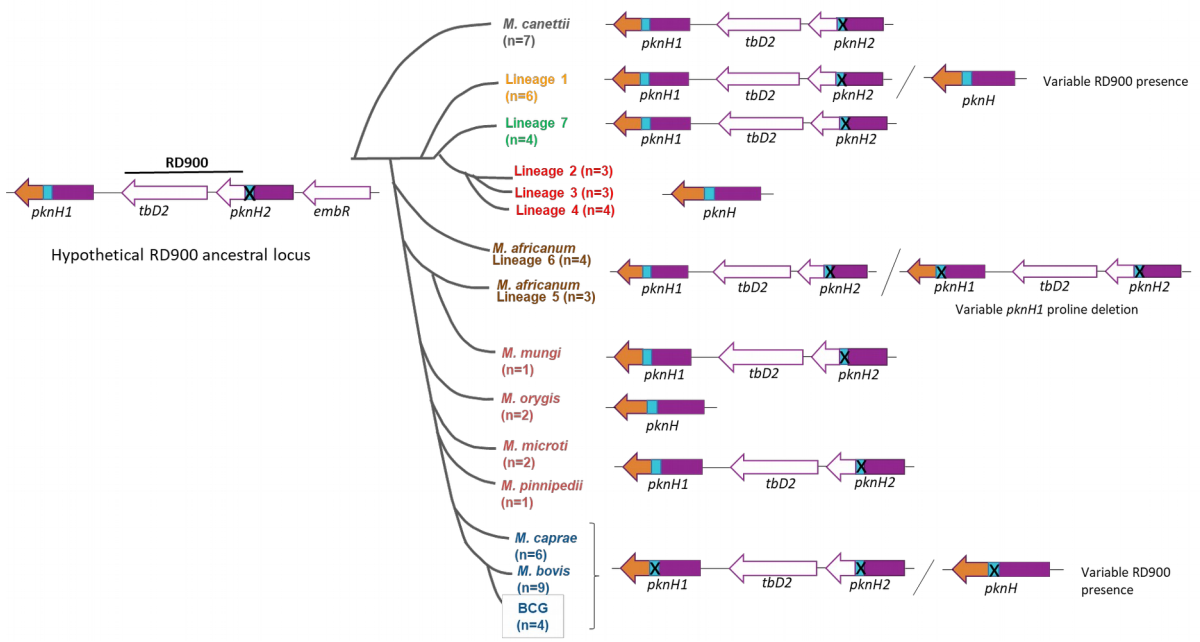


Figure 4. RD900 locus in MTBC. Representation of the RD900 locus in the different species from the MTBC. *PknH* genes are represented with the kinase domain in purple, the proline rich region in blue and the different sensor domain in orange for *pknH1* and white for *pknH2*. The deletion in the proline rich region is represented with an "X".

RD900 homologous recombination

Considering the variability of the RD900 region presence, RD900 deletion seems have happened independently in different lineages and also in different strains from the same lineage. This deletion is generated by homologous recombination between the two *pknH* flanking genes, resulting in the loss of *tbd2* gene and the generation of one *pknH* gene.

If we consider the ancestral RD900 locus as the *M. canettii* locus, based on the sequence homology of the different *pknH* domains, in the "modern" *M. tuberculosis* lineages the recombination site may be on the homologous kinase domains, resulting in an intact *pknH* complete gene composed of the kinase domain from *pknH2* and the proline-rich region, transmembrane and sensor domain from *pknH1*. In *M. bovis* and *M. caprae* strains with the complete RD900 locus, a deletion in the proline-rich domain may happened first resulting in an RD900 locus with both *pknH* genes carrying the same deletion. In *M. bovis* and *M. caprae* strains with a deleted RD900 locus the homologous recombination could happened between those two proline-rich deleted *pknH* genes resulting in one *pknH* gene with the proline-rich deletion present.

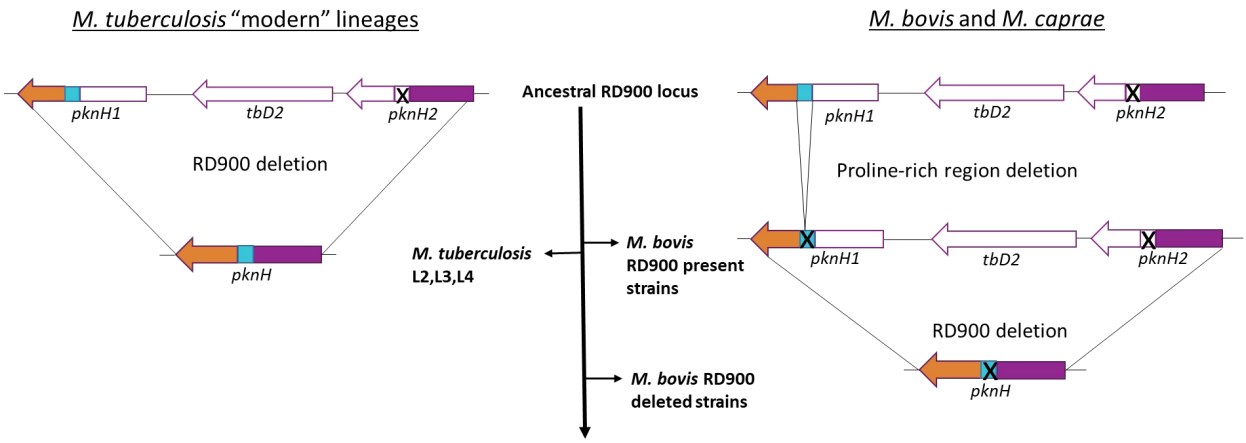


Figure 5. Hypothetical RD900 homologous recombination in *M. tuberculosis* "modern" lineages and *M. bovis* and *M. caprae*.

If we look at this genome region in other *Mycobacterium* species, in *M. marinum* we found another probable recombination pattern that could be associated with a possible ancestral locus. *M. marinum* genome has three *pknH* genes. One of this *pknH* genes (named here *pknH2b*) is separated by 12 Kb of antoher *pknH* gene (named here *pknH2a*). The identity between the protein sequences of the different domains of these genes indicate that a recombination event could happened between *pknH2a* and *pknH2b* genes resulting in the *pknH2* gene present in the MTBC species with the complete RD900 region.

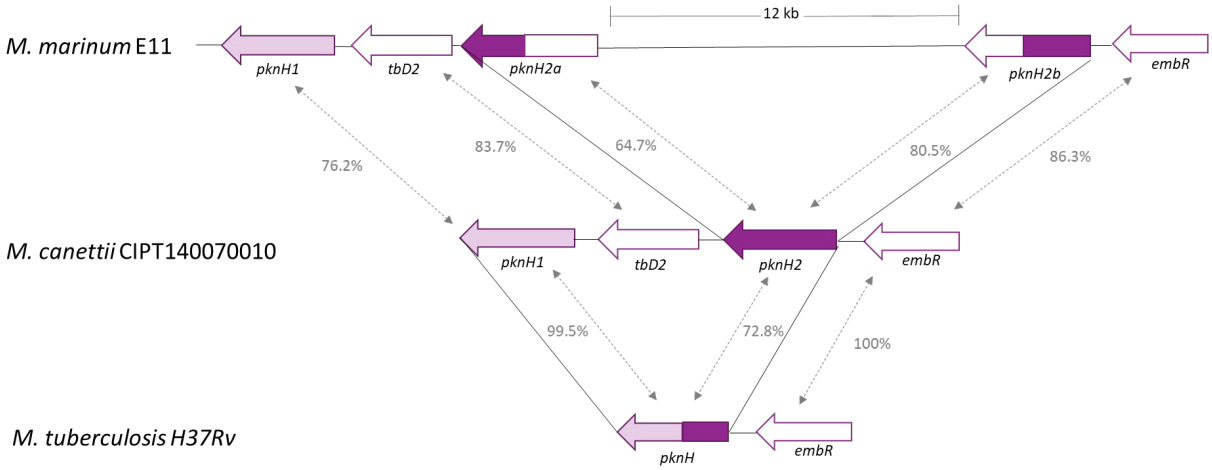


Figure 6. *M. marinum* RD900 locus and hypothetical homologous recombination events in RD900 locus. The percentage of protein sequence identity for the products of the genes present in this region between *M. marinum* E11, *M. canetti*i CIPT140070010 and *M. tuberculosis* H37Rv strains are indicated in the grey arrows.

Point mutations identification

To better understand the evolutionary process of this particular region we analysed the mutations (synonymous, missenses and frameshifts) present in this region across the different strains. To do that we used the sorted bam files of the alignments and we did the variant calling, filtered and consequence calling using freebayes and bcftools. The results for the missense and frameshifts mutations are represented in a heat map with the presence/absence data for each mutation. Mutations present in just one isolate were excluded. In the clustered heat map we can see that some of the isolates of the same lineage are clustered together whereas other isolates don't, reflecting the variability of this region across the different strains and lineages.

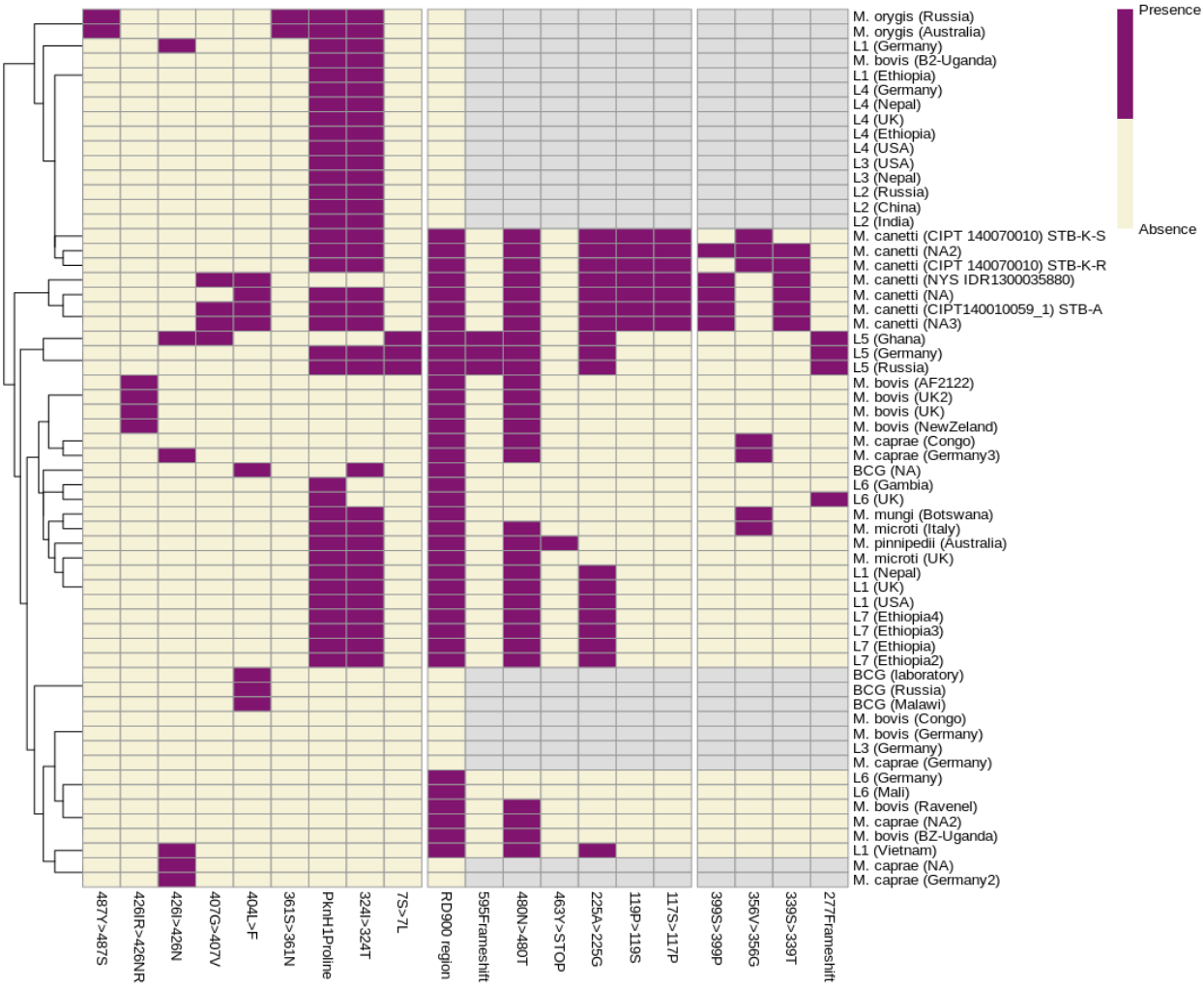


Figure 7. Presence/absence of missense and frameshift mutations in RD900 locus.

RD900 locus variability is also reflected in a dendogram constructed from a distant matrix calculated with the presence/absence data of the missense and frameshift mutations (Figure 8).

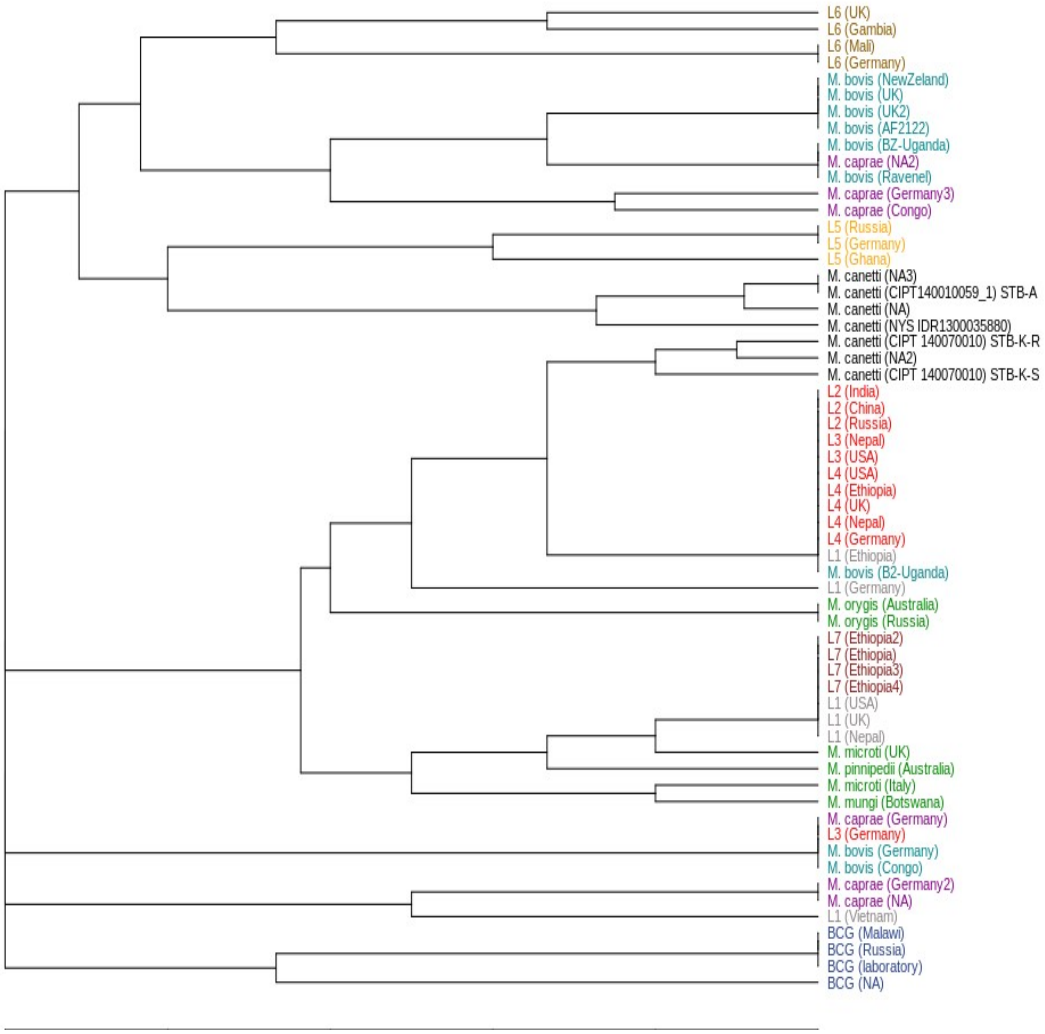


Figure 8. Dendogram constructed from presence/absence data of missense and frameshifts mutations.

In the RD900 deleted strains, independent homologous recombination events between *pknH* genes could likely result in a truncated gene, but if we look at the point mutations in *pknH* gene of the RD900 deleted strains we found an intact gene with no frameshift mutations, suggesting that different independent homologous recombination are generating the same potential functional gene. This suggest a possible selective advantage for the reduction of the number of kinase caused by RD900 deletion in these strains and the potential importance to conserve one *pknH* gene.

Limitations of the approach

In this analysis we have aligned the genome sequencing reads to a reference sequence. The limitation of this strategy is that it is not possible to detect large insertions or deletions that are not present in the reference. Another limitation on analysing this particular region is the presence of multi-mapping reads. Because of the homology between *pknH1* and *pknH2* genes, reads can map to more than one place in the reference.

Another approach to analysis this region is to do *de novo* assembly of the genome and to align the contigs from the assembly to a reference sequence. The limitation of this method is that the sequence of the RD900 locus is split into different contigs if the length of the reads is lower than 150 bp, which is the case for the most part of the analysed reads (Figure 9). This problem can also be reason why this region is likely to be misassemble due to the high sequence identity between the two *pknH* genes.

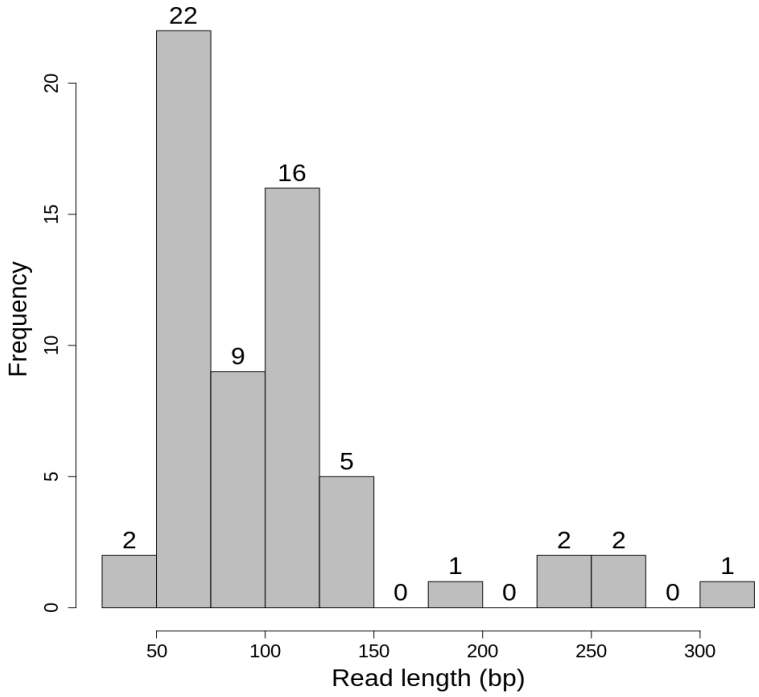


Figure 9. Read length distribution of WGS analysed data.

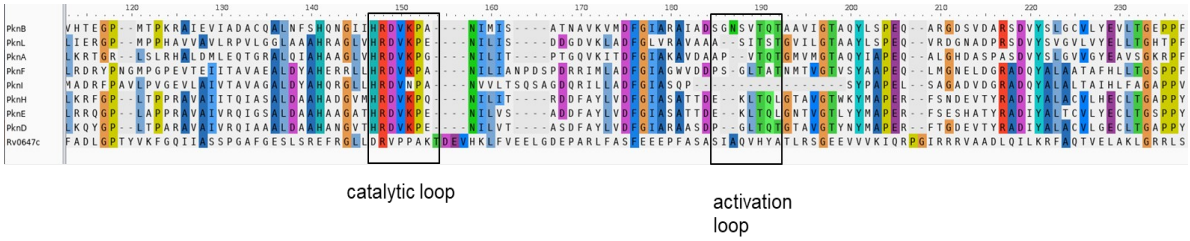
Table 1. Genome data of the 60 MTBC isolates used in this analysis.

AC. NUMBER	MTB LINEAGE	ORIGIN	REFERENCE
ERR2642516	*M. caprae*	NA	Brites et al. 2018
ERR551023	*M. caprae*	Congo	Malm et al. 2017
ERR551704	*M. caprae*	Germany	Malm et al. 2017
ERR552526	*M. caprae*	Germany	Malm et al. 2017
ERR841382	*M. caprae*	NA	Domogalla et al. 2013
SRR650219	*M. caprae*	Germany	Research Center Borstel
ERR233356	*M. tuberculosis* L1	USA	Comas et al. 2013
ERR1200629	*M. tuberculosis* L1	Ethiopia	Comas et al. 2013
ERR234155	*M. tuberculosis* L1	Germany	Comas et al. 2013
ERR234238	*M. tuberculosis* L1	Vietnam	Comas et al. 2013
ERR234272	M. tuberculosis L1	UK	Comas et al. 2013
ERR233377	M. tuberculosis L1	Nepal	Comas et al. 2013
ERR1200603	*M. tuberculosis* L7	Ethiopia	Comas et al. 2013
ERR1200617	*M. tuberculosis* L7	Ethiopia	Comas et al. 2013
ERR1200635	*M. tuberculosis* L7	Ethiopia	Comas et al. 2013
ERR1200640	*M. tuberculosis* L7	Ethiopia	Comas et al. 2013
ERR234097	*M. tuberculosis* L5	Germany	Comas et al. 2013
ERR234199	*M. tuberculosis* L5	Ghana	Comas et al. 2013
ERR017801	*M. tuberculosis* L5	Russia	Casali et al. 2014
ERR233366	*M. tuberculosis* L6	Gambia	Comas et al. 2013
ERR234184	*M. tuberculosis* L6	Germany	Comas et al. 2013
ERR400537	*M. tuberculosis* L6	UK	Walker et al. 2015
SRR998594	*M. tuberculosis* L6	Mali	Broad Institute
ERR234098	*M. tuberculosis* L2	China	Comas et al. 2013
ERR233391	*M. tuberculosis* L3	Nepal	Comas et al. 2013
ERR233358	*M. tuberculosis* L4	USA	Comas et al. 2013
ERR125602	*M. bovis*	UK	Trewby et al. 2016
SRR5216693	*M. bovis*	New Zeland	Crispell et al. 2017
ERR551009	*M. bovis*	Germany	Walker et al. 2015
ERR552138	*M. bovis*	Republic of the Congo	Malm et al. 2017
ERR400386	*M. bovis*	UK	Walker et al. 2015

	M. bovis AF2122		UCD
SRR1173570	*M. bovis*	Uganda(BZ)	Wanzala et al. 2015
SRR1173284	*M. bovis*	Uganda(B2)	Wanzala et al. 2015
ERR027294	*M. microti*	UK	Wellcome Sanger Institute
ERR2659164	*M. microti*	Italy	Brites et al. 2018
ERR234675	*M. orygis*	Russia	Casali et al. 2014
ERR2659153	*M. orygis*	Australia	Brites et al. 2018
SRR3500411	*M. mungi*	Botswana	Alexander et al. 2016
SRR1239337	*M. pinnipedii*	Australia	Bos et al. 2014
ERR015598	*M. canetti*	NA	Wellcome Sanger Institute
SRR011186	*M. canetii*	NA	Broad Institute
ERR1336826	*M. canetii*	NA	Institut Pasteur
ERR1336823	*M. canetii*	NA	Institut Pasteur

PknH structure analysis

There are 11 serine/threonine protein kinases (STPKs) in *M. tuberculosis*, 9 of which are transmembrane proteins and are mostly conserved in different mycobacterial species. The kinase domain is highly structural conserved across all STPK and also structurally closed to the eukaryotic kinases (Figure 10). In contrast, the sensor extracellular domains are different in sequence and structure in the different mycobacterial STPKs.



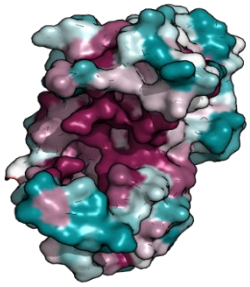


Figure 10. Kinase domain conservation in STPKs. Protein sequence alignment of the kinase domains from the different STPKs in *M. tuberculosis* and representation of the kinase residues conservation (purple residues are the most conserved).

In *M. tuberculosis* PknH, the structure of the sensor domain (corresponding to the PknH1 sensor domain in *M. bovis* or *M. africanum*) is solved. The kinase domain structure is solved in other *M. tuberculosis* STPKs that

present high identity with PknH kinase domain, as PknE (Figure 11). Based on this information, we wanted to model the structure of the whole transmembrane complex and the differences in the kinase domain of the PknH2 in *M. africanum* and *M. bovis* (22 amino acid deletion and substitution respectively).

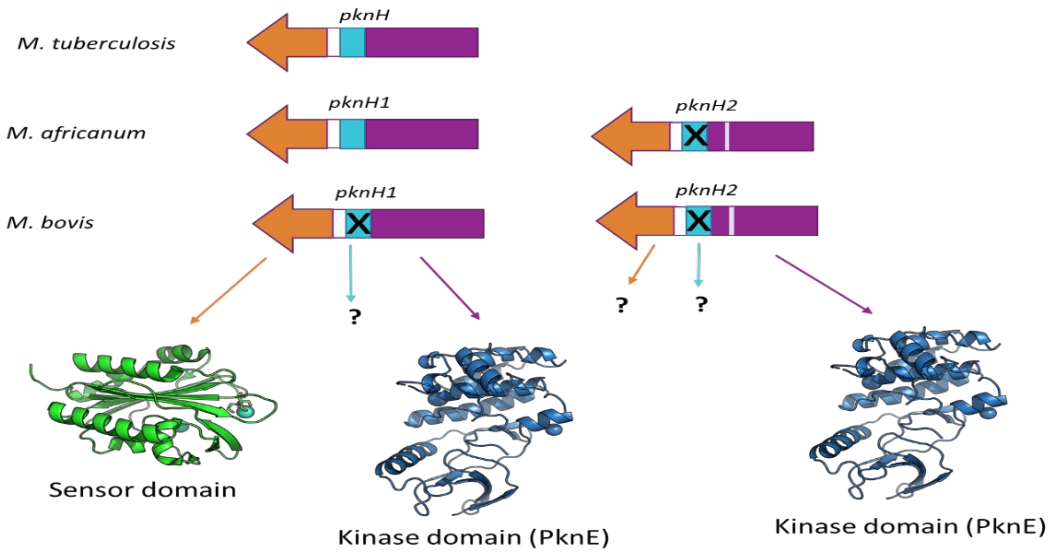


Figure 11. Known structures of PknH domains in *M. tuberculosis*.

Transmembrane proteins are hard to model because of the few template structures available. The PknH transmembrane and proline-rich domains are predicted as disordered regions that can't be correctly model by the modelling tools that we used.

Apart from that, the structure of the PknH1 sensor domain is already solved but there were not closely related sequences to model the PknH2 sensor domain using homology modelling. For this reason, we couldn't obtain an accurate model of the whole PknH protein.

Given the difficulties in modelling the whole protein, we focused on modelling the differences in the kinase structures of the PknH proteins in *M. africanum* and *M. bovis* PknH2 compared to PknH1. The 22 amino acid variation present in these kinases domains correspond to a region present in protein kinases called the αF-helix. In other kinases, as human PKA, this region has been described to serve as a central scaffold for active protein kinase assembly and a signal-integrating element, which connects several key areas such as the substrate-binding residues and the catalytic loop (Kornev et al.,2008).

We first used Swiss-model to model *M. africanum* PknH2 kinase domain, that has a 22 amino acid deletion in that region. In the model that we obtained the αF-helix is absent, which could compromise the correct function of the protein (Figure 12).

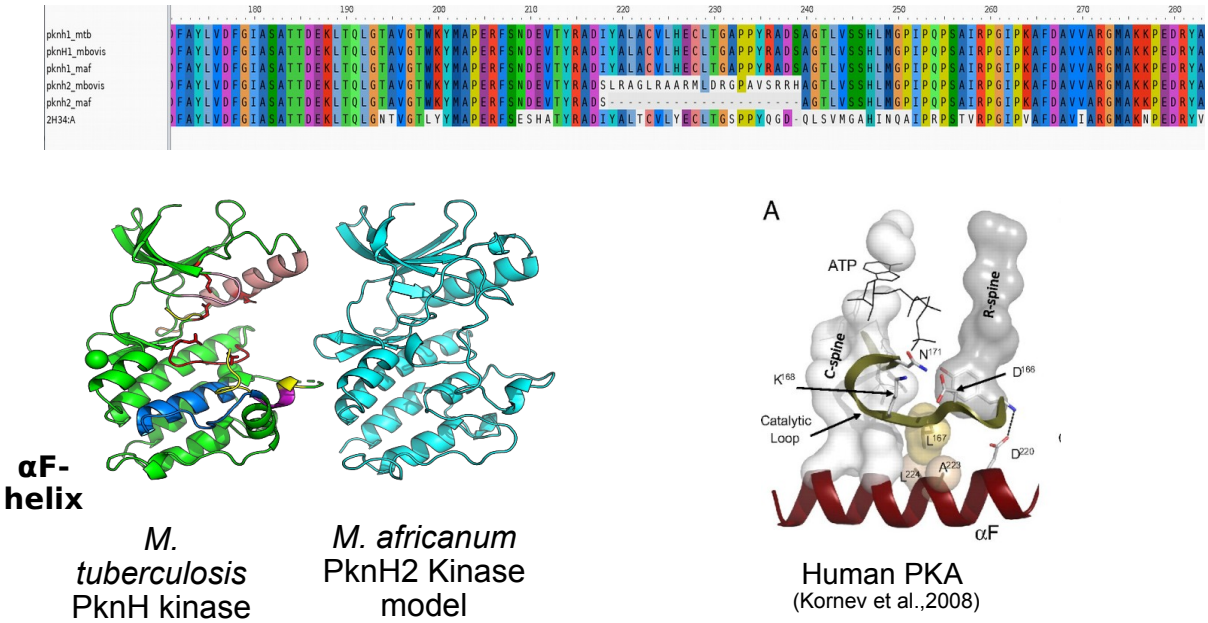


Figure 12. Protein sequence alignment of PknH kinases domains and model of *M. africanum* PknH2 kinase domain using Swiss-model. In *M. africanum* PknH2 kinase domain there is a 22 amino acid deletion compared to PknH1. This region is present in *M. bovis* AF2122 PknH2 but the amino acid sequence is different from PknH1. In the model obtained for M. africanum PknH2 the αF-helix is not present (coloured in dark blue in *M. tuberculosis* PknH).

We obtained other models from different servers that use different modelling approaches combining homology modelling with ab-initio methods, as i-TASSER. In these models the αF-helix structure seems to be retained in *M. africanum* PknH2 even in the presence of the deletion (Figure 13). The variations in PknH2 kinase αF-helix may affect the function and the substrate specificity of the kinase (Figure 14), but the information from the models is insufficient to show the real impact of these amino acid changes in the protein functionality.

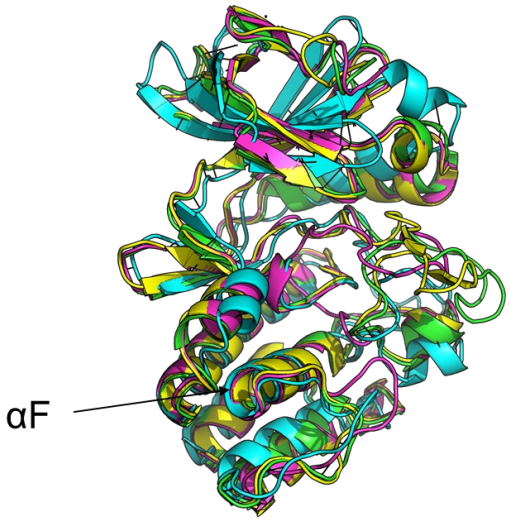
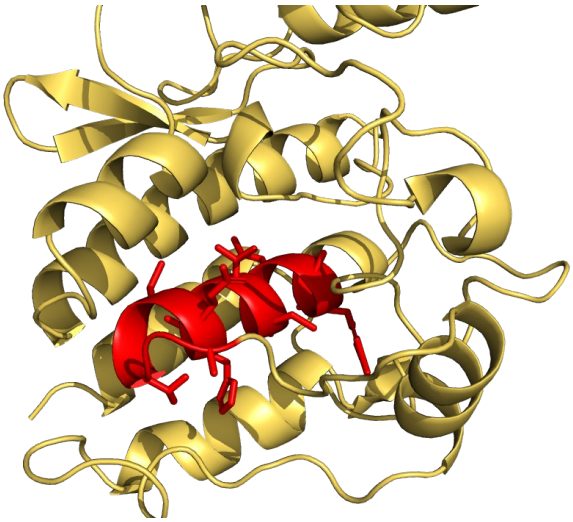
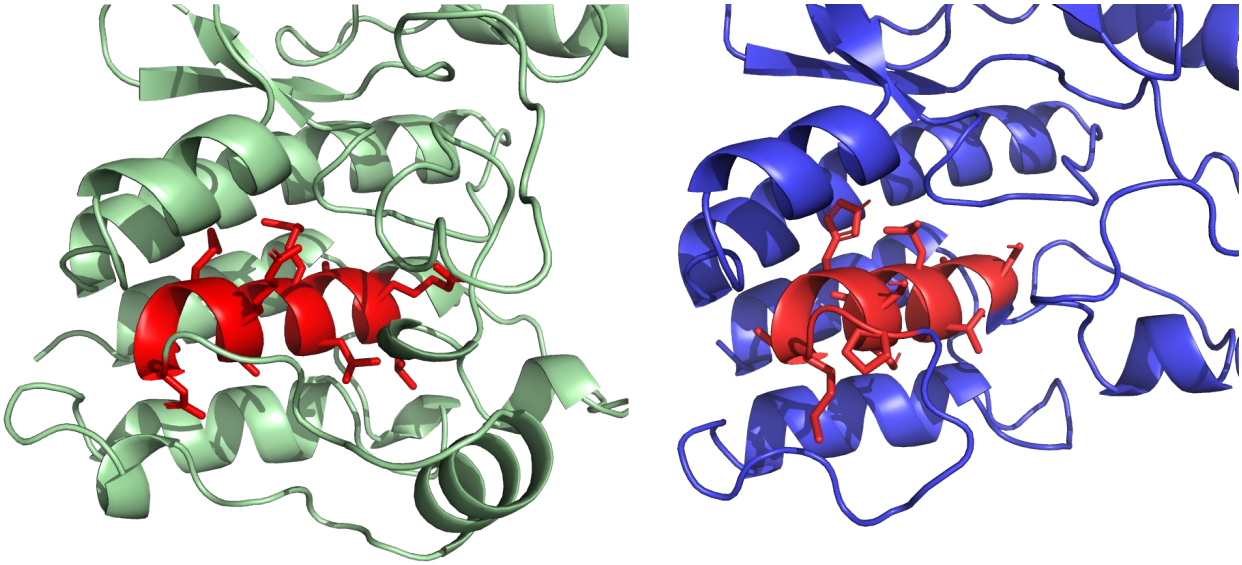


Figure 13. Superposition of 5 models of PknH2 kinase domain from *M. africanum* obtained with i-TASSER. The F helix is retained intact though perhaps the change alters substrate specificity.



(a)



(b)

Figure 14. (a) MTB pknh1 kinase structure (modelled in swissmodel from template 2h34). (b) pknH2 models. Alterations in the αF-helix sequence changes the side chains exposed to the substrate in both mbovis (left) and Africanum (right). Note that the remaining helix is not present in the africanum model because of the 22 amino acid deletion.

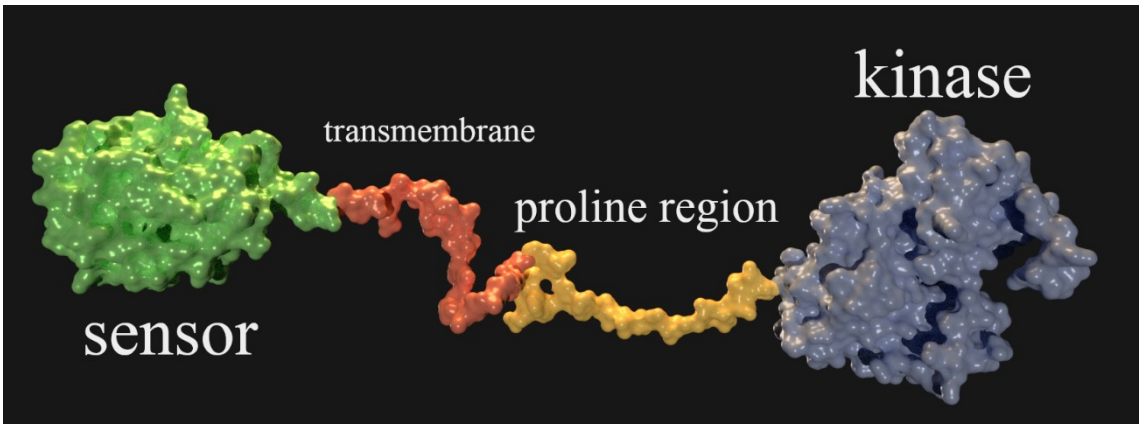


Figure 15. Putative structure of the entire pknh1 protein. This is not an accurate structure as the conformations of the TM and proline rich domains are not solvable by homology modelling. The proline region is longer than shown here. It is likely that the proline region is very flexible and may be involved in the dimerization process.