

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.

<http://hdl.handle.net/2066/183946>

Please be advised that this information was generated on 2018-04-11 and may be subject to change.

RESEARCH ARTICLE

Open Access



Large-scale genomic analysis shows association between homoplastic genetic variation in *Mycobacterium tuberculosis* genes and meningeal or pulmonary tuberculosis

Carolien Ruesen¹, Lidya Chaidir², Arjan van Laarhoven¹, Sofiati Dian³, Ahmad Rizal Ganiem³, Hanna Nebenzahl-Guimaraes^{4,5,6}, Martijn A. Huynen⁷, Bacht Alisjahbana², Bas E. Dutilh^{7,8} and Reinout van Crevel^{1*}

Abstract

Background: Meningitis is the most severe manifestation of tuberculosis. It is largely unknown why some people develop pulmonary TB (PTB) and others TB meningitis (TBM); we examined if the genetic background of infecting *M. tuberculosis* strains may be relevant.

Methods: We whole-genome sequenced *M. tuberculosis* strains isolated from 322 HIV-negative tuberculosis patients from Indonesia and compared isolates from patients with TBM ($n = 106$) and PTB ($n = 216$). Using a phylogeny-adjusted genome-wide association method to count homoplasmy events we examined phenotype-related changes at specific loci or genes in parallel branches of the phylogenetic tree. Enrichment scores for the TB phenotype were calculated on single nucleotide polymorphism (SNP), gene, and pathway level. Genetic associations were validated in an independent set of isolates.

Results: Strains belonged to the East-Asian lineage (36.0%), Euro-American lineage (61.5%), and Indo-Oceanic lineage (2.5%). We found no association between lineage and phenotype (Chi-square = 4.556; $p = 0.207$). Large genomic differences were observed between isolates; the minimum pairwise genetic distance varied from 17 to 689 SNPs. Using the phylogenetic tree, based on 28,544 common variable positions, we selected 54 TBM and 54 PTB isolates in terminal branch sets with distinct phenotypes. Genetic variation in Rv0218, and absence of Rv3343c, and *nanK* were significantly associated with disease phenotype in these terminal branch sets, and confirmed in the validation set of 214 unpaired isolates.

Conclusions: Using homoplasmy counting we identified genetic variation in three separate genes to be associated with the TB phenotype, including one (Rv0218) which encodes a secreted protein that could play a role in host-pathogen interaction by altering pathogen recognition or acting as virulence effector.

Keywords: Pulmonary tuberculosis, Tuberculous meningitis, Whole genome sequencing, Homoplasmy

* Correspondence: Reinout.vancrevel@radboudumc.nl

¹Department of Internal Medicine, Radboud university medical center, Nijmegen, The Netherlands

Full list of author information is available at the end of the article



Background

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains a major global health problem [1]. Active TB mostly affects the lungs but may also spread to other organs. TB meningitis (TBM), which represents approximately 1–5% of all TB cases, is the most severe manifestation of TB, resulting in death or neurological disability in about half of those affected [2, 3]. It is largely unknown why certain people develop pulmonary TB (PTB) and others TBM. Host immune-related factors clearly play an important role, as shown by the increased risk of TBM for patients with advanced HIV infection, and the overrepresentation of young children among TBM patients. Host genetic factors may also play a role; single studies have linked susceptibility to TBM with variation in candidate genes [4–8].

Besides the host, genetic diversity of infecting *M. tuberculosis* strains may also affect disease phenotype. Even though *M. tuberculosis* is considered a clonal organism, there is considerable genetic variation in the genomes of infecting *M. tuberculosis* isolates [9, 10]. Epidemiological studies have reported significant differences among *M. tuberculosis* lineages in terms of virulence [11, 12], transmission [9, 13, 14], progression to active disease after infection [15], and response to treatment [16, 17]. In vitro studies have supported these findings by showing *M. tuberculosis* genotype-specific differences in the human immune response [18–21].

Animal studies have shown that *M. tuberculosis* strains differ in their ability to invade the central nervous system (CNS). Five *M. tuberculosis* genes (Rv0311 (unknown function), Rv0805 (intermediary metabolism and respiration), *pknD* (protein kinase D), Rv0986 (cell wall and cell processes), and MT3280 (unknown function)) have been associated with invasion or survival in the CNS but not in lung tissues in mice [22]. Especially *M. tuberculosis pknD* was associated with invasion of brain, but not lung epithelia in guinea pigs [23], as was confirmed by another study showing that *pknD* vaccination offered significant protection against bacterial dissemination to the brain in guinea pigs [24]. Similarly, in mice, clinical isolates from TBM patients disseminated extensively to cause meningitis, whereas *M. tuberculosis* H37Rv and clinical isolates from PTB patients did not [25]. In rabbits, production of phenolic glycolipid has been linked with the increased propensity of East-Asian/Beijing strains to cause TBM [26]. Finally, four *M. tuberculosis* genes were crucial for invading an artificial blood brain barrier in an in vitro model using primary human brain microvascular endothelial cells: *PE-PGRS18* (unknown function), Rv0987 (cell wall and cell processes), *grcC2* (intermediary metabolism and respiration), and *PPE29* (unknown function) [27].

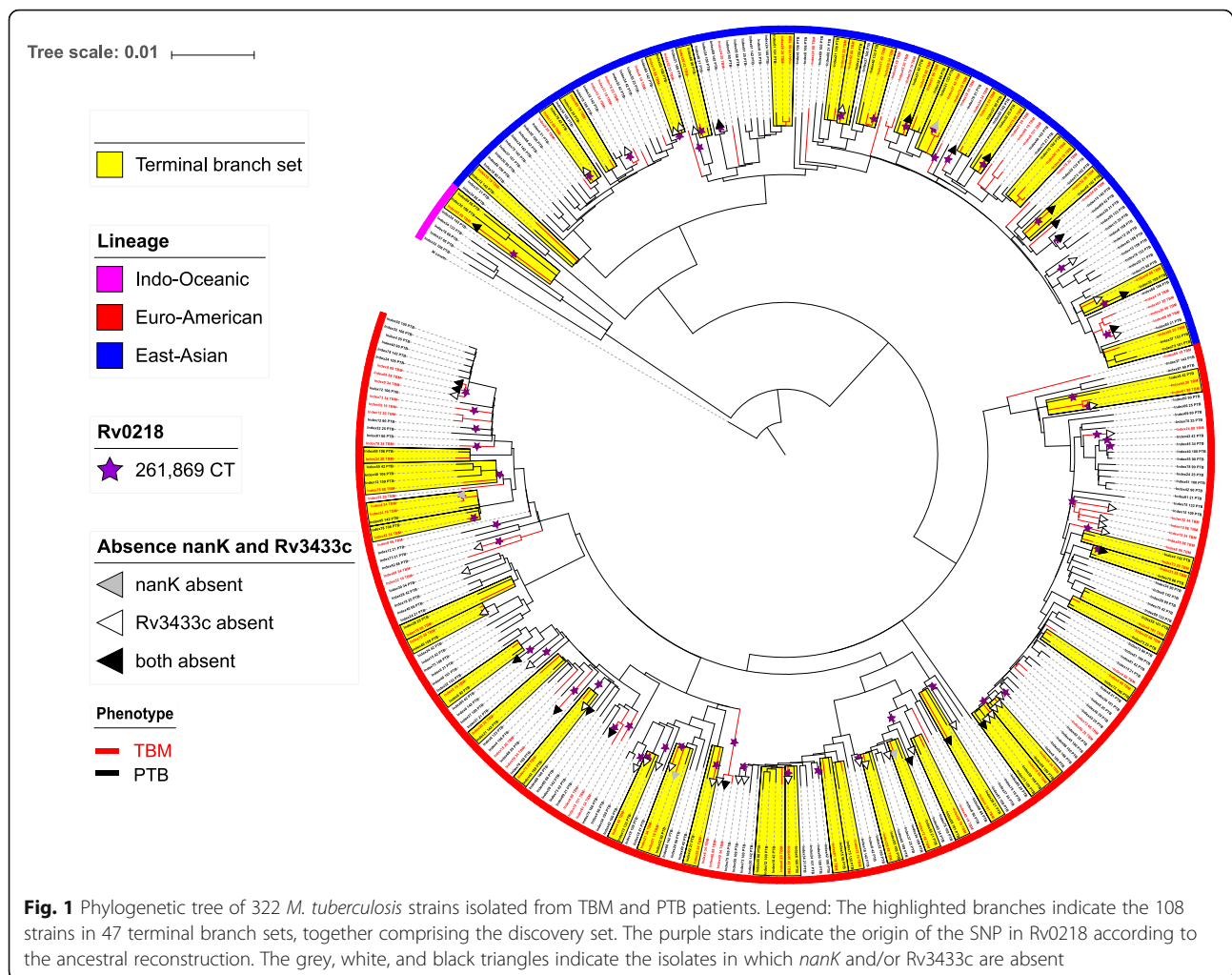
Much less is known about the role of *M. tuberculosis* genotype in TBM in humans. Most studies have examined associations of *M. tuberculosis* lineage with disease phenotype. Compared to other lineages, strains belonging to the East-Asian lineage were associated with extrapulmonary tuberculosis in one study [28], but not in another [29], while other studies found no association of *M. tuberculosis* lineage and disease localisation [30, 31]. Specifically looking at TBM, one study from Vietnam found the Euro-American lineage to be associated with PTB rather than TBM [32]. Only one study used whole genome sequencing to compare strains from TBM and PTB patients; large-scale and smaller genomic rearrangements, inversions, indels and single nucleotide polymorphisms (SNPs) in eight cerebrospinal fluid (CSF)-derived strains were not found in 69 comparison respiratory strains isolated from independent sputum samples [33]. In the current study, we used a much larger set of isolates and a novel approach to examine the effect of the *M. tuberculosis* genotype on the susceptibility to TBM. We compared *M. tuberculosis* genomes isolated from 216 PTB patients and 106 TBM patients from Indonesia, all HIV-negative, to detect homoplastic genetic variants associated with either PTB or TBM.

Results

Lineage distribution and phylogeny construction

M. tuberculosis isolates from established patient cohorts in Bandung, Indonesia were selected for whole genome sequencing. All available *M. tuberculosis* strains isolated from HIV-negative TBM patients and randomly selected strains from twice as many PTB patients from the same setting were included, one strain was selected per patient. Compared to the 216 PTB patients, the 106 TBM patients were from a similar ethnic background, but slightly younger, more often male, and more often previously treated for TB (Additional file 1: Table S1). Based on a 62-SNP barcode [34] 61.5% of the strains belonged to the Euro-American lineage (63.4% for PTB; 57.5% for TBM), 36% to the East-Asian lineage (33.3% for PTB; 41.5% for TBM), and 2.5% to the Indo-Oceanic lineage (3.2% for PTB; 0.9% for TBM). The lineage distribution did not differ significantly for strains isolated from TBM compared to PTB patients (Chi-square = 3.230; $p = 0.199$).

A phylogenetic tree was constructed based on 28,544 variable common nucleotide positions among the 322 *M. tuberculosis* isolates. The phylogeny showed that the TBM phenotype was not restricted to a certain *M. tuberculosis* lineage; instead it arose many times independently (Fig. 1). In addition, the tree showed a high degree of strain heterogeneity. On average, two strains differed by about 1000 SNPs, and this pairwise distance did not differ among PTB and TBM strains (data not shown), indicating that there was equal genetic diversity



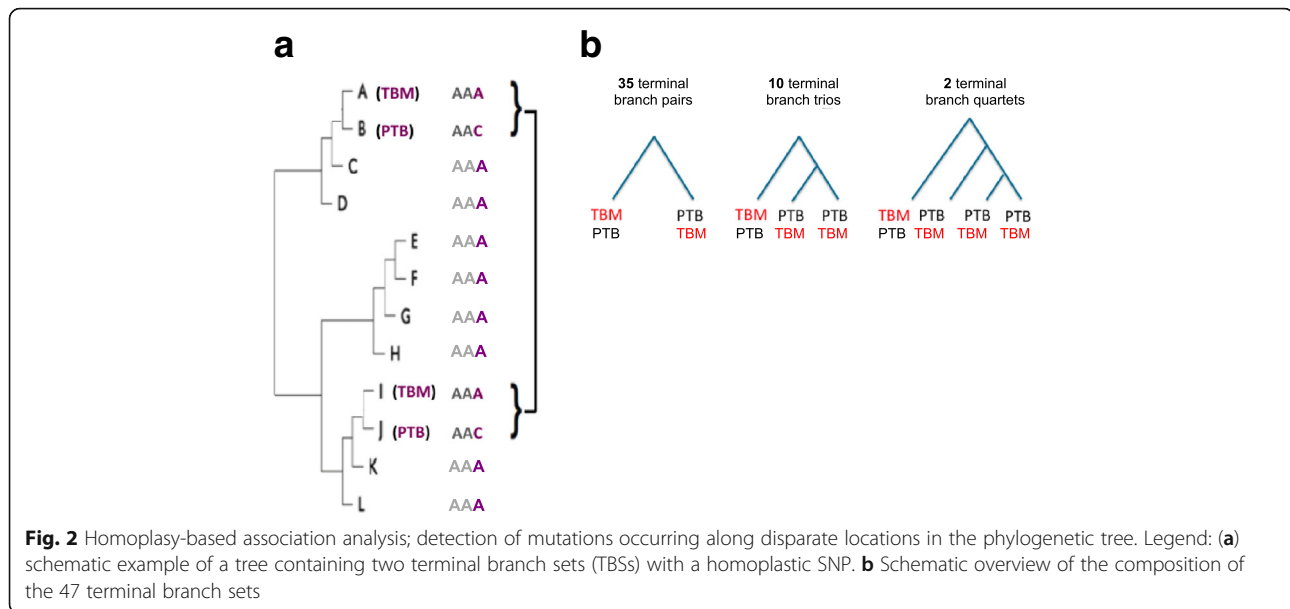
within PTB and within TBM strains. In addition, the minimum pairwise genetic distance varied from 17 to 689 SNPs (data not shown), indicating that there was no clustering of strains (≤ 12 SNPs distance [35]). For TBM strains the minimum genetic distance ranged from 17 to 1785, and for PTB strains from 31 to 803 SNPs (data not shown).

TB phenotype-associated genetic variations

Genome-wide association approaches for bacteria can broadly be categorised into allele counting and homoplasy counting methods [36]. Allele counting methods are based on the overrepresentation of an allele at the same site in cases relative to controls, introducing a risk of false-positive findings due to population stratification. Homoplasy counting on the other hand, counts repeated and independently emerging mutations that occur more often in branches of cases relative to controls. In the current study, we used a two-step approach: in the discovery phase, we used homoplasy counting by identifying terminal branch sets (TBSs) to maximize power to

identify true associations with the TB disease phenotype, uncorrected for multiple tests. In the validation phase we examined associations identified in the discovery phase using allele counting with correction for multiple testing and for phylogenetic bias to distinguish true associations from false positives, and performed ancestral reconstruction to remove possible phylogenetic bias. To divide the genomes in a discovery and a validation set, we identified isolates in terminal branch pairs, trios and quartets (i.e. separated at a terminal or near terminal branch in the phylogenetic tree) with distinct phenotypes (Fig. 2). Genetic differences between isolates within a TBS provide the strongest, homoplasy-corrected possible association with the phenotype.

The phylogenetic tree revealed a total of 47 TBSs containing 108 paired strains: 54 TBM and 54 PTB strains that make up the discovery set. The merged SNP lists consisted of 6488 variable positions with the corresponding nucleotides for the 108 strains (Additional file 2). Using the homoplasy-based association analysis, we found individual nucleotide positions, genes, or pathways where TB



disease phenotype-associated mutations repeatedly occur in different branches of the phylogenetic tree. These included 9 SNPs, 5 genes, and 1 pathway (Table 1).

We used the remaining 214 (52 TBM and 162 PTB) isolates not belonging to any of the TBSs (validation set) to verify these results. The discovery set showed a total of 6488 different non-synonymous SNPs involving 6483 dimorphic sites and 5 trimorphic sites across 2778 genes; the validation set a total of 12,211 different non-synonymous SNPs involving 12,185 dimorphic sites and 26 trimorphic sites across 3359 genes (Additional file 3). There was an overlap of 2694 non-synonymous SNPs and 1564 affected genes. Out of 9 SNPs significantly associated with either TBM or PTB in the discovery set, one was confirmed in the validation set; the mutation in Rv0218. Similarly, out of 5 genes harbouring genetic variation associated with the TB phenotype, Rv0218 was validated in the validation set (Table 1, Additional file 4: Figure S1 and Additional file 5: Figure S2). The pathway (ethylbenzene degradation) identified in the discovery set was not confirmed in the validation set.

To correct for potential phylogenetic bias in the validation set, we reconstructed the ancestral state for the SNP in Rv0218 and compared the ratio of TBM vs. PTB isolates after the occurrence of this particular SNP with the ratio of TBM vs. PTB prior to the occurrence of this SNP in the validation set (Fig. 1 and Additional file 6: Figure S3). Three branches with back mutations (2 TBM, 1 PTB branch) were excluded from the analysis. Among the 33 nodes / leaves where the SNP occurred, the average unweighted proportion of TBM isolates among the child branches was 44.7%; among 166 isolates in the validation set not harbouring this SNP, 29 (17.5%) were from TBM patients (Additional file 7: Table S2).

The Z-score for the difference between proportions was -3.83 ($p < 0.001$).

De novo genome assembly

Next, we used a reference-free *de novo* genome assembly approach to examine associations between presence/absence of genes and the TB phenotype, and to study sequences not present in the reference genome. The list of annotated coding sequences for the 108 assembled genomes in the discovery set contained 3032 distinct genes that were present in at least one, but not all of the strains (Additional file 8). The permutation analysis revealed six genes that were significantly associated with the TB phenotype. Two of these genes, Rv3433c and *nanK*, were validated in the validation set: absence of these genes was associated with TBM (Fig. 1, Table 2, Additional file 9). Together with Rv0218 they bring the total number of genes associated with the TB phenotype to three.

Effect of detected SNPs on protein function and predicted function of phenotype-associated genes

We used published algorithms to predict the effects of identified mutations on protein structure and function. The SNP in Rv0218, a protein predicted to have transmembrane helices, likely leads to a decrease of stability of the protein (Additional file 10: Table S3). For Rv3433c, and *nanK* no transmembrane helices or signaling peptides were predicted.

Discussion

To determine whether *M. tuberculosis* genetic variation is associated with the TB disease phenotype, we compared *M. tuberculosis* whole genome sequences from

Table 1 Significant SNPs, genes, and pathways identified by homoplasmy counting

		Discovery dataset (n = 108)			Validation dataset (n = 214)		
SNP-level		Strains with SNP (N)			Strains with SNP (N)		
Gene (Rv-number)	Nucleotide change	TBM (n = 54)	PTB (n = 54)	P-value	TBM (n = 52)	PTB (n = 162)	P-value [^]
Unnamed (Rv0218)	261,869CT	25	10	0.002	21	24	0.001
PPE54 (Rv3343c)	3736,628TG	22	35	0.008	27	87	0.472
PEPGRS19 (Rv1067c)	1,190,093AC	2	11	0.01	5	18	0.487
PEPGRS44 (Rv2591)	2922,848AT	0	6	0.01	2	7	0.625
PPE3 (Rv0280)	340,372TC	2	10	0.025	4	13	0.589
PEPGRS9 (Rv0746)	836,272AG	7	1	0.029	3	9	0.603
PEPGRS26 (Rv1441c)	16,18,978TC	0	5	0.032	2	3	0.369
Unnamed (Rv0064)	713,36GC	24	34	0.034	29	98	0.355
PEPGRS18 (Rv0980c)	1095,644CT	0	4	0.044	2	2	0.238
Gene-level		SNPs in gene (N)			SNPs in gene (N)		
Gene (Rv number)		TBM	PTB	P-value	TBM	PTB	P-value
PEPGRS19 (Rv1067c)		2	13	0.004	5	22	0.318
Unnamed (Rv0218)		27	11	0.007	21	32	0.001
PEPGRS26 (Rv1441c)		0	5	0.031	2	4	0.491
glmS (Rv3436c)		0	5	0.031	0	0	n.a.
Unnamed (Rv3740c)		0	4	0.05	3	0	0.011
Pathway-level		Genes with mutation (N)			Genes with mutation (N)		
Pathway name		TBM	PTB	P-value	TBM	PTB	P-value
Ethylbenzene degradation		68	54	0.032	65	224	0.071

NOTE. P-values are based on permutation analysis; bold p-values indicate validated, Bonferroni-corrected significant enrichment

[^] P-value thresholds in the validation set were Bonferroni-corrected for multiple testing by dividing them by the number of top hits in the discovery set: SNP-level: $p < 0.05/9$; gene-level: $p < 0.05/5$; pathway-level: $p < 0.05/1$

216 PTB and 106 TBM patients and searched for homoplastic mutations. We identified three genes in *M. tuberculosis* (Rv0218, Rv3433c, and *nank*) to be associated with either TBM or PTB. Previous experimental studies have assessed the importance of Rv0218. This secretome gene encodes for a protein with multiple predicted transmembrane regions and a C-terminal molybdopterin binding domain that is often found in oxidoreductases and was shown to be essential for *M. tuberculosis* in vivo growth in C57BL/6 J mouse spleen [37]. The SNP in Rv0218 is predicted to decrease the

stability of the respective protein. Secretome genes potentially influence pathogen recognition and host-pathogen interaction [38]. If mutations in these genes alter the appearance of the *M. tuberculosis* surface, this could provide a mechanism by which *M. tuberculosis* could evade the immune response and enable dissemination to extrapulmonary sites. Secretome genes are more likely to contain false-positive associations as they are under selective pressure from the immune system and phages [39]. How Rv3433c and *nank* could be related to the TB phenotype is not

Table 2 Significant genes identified by the *de novo* genome assembly analysis

		Discovery dataset (n = 108)			Validation dataset (n = 214)		
Coding sequence		Strains with CDS present (N)			Strains with CDS present (N)		
		TBM (n = 54)	PTB (n = 54)	P-value	TBM (n = 52)	PTB (n = 162)	P-value [^]
Bifunctional NAD(P)H-hydrate repair enzyme Nnr (Rv3433c)		36	47	0.011	34	144	0.002
TMAO/DMSO reductase		46	53	0.016	49	152	0.592
Antitoxin/MT2731		33	44	0.017	37	130	0.206
NPCBM-associated, NEW3 domain of alpha-galactosidase		15	6	0.022	14	19	0.063
Oxidoreductase molybdopterin binding domain protein		11	3	0.028	7	13	0.197
N-acetylmannosamine kinase (<i>nank</i>)		44	51	0.036	44	157	0.001

NOTE. P-values are based on permutation analysis; bold p-values indicate validated significant enrichment

[^] P-value thresholds in the validation set were Bonferroni-corrected for multiple testing by dividing them by the number of top hits in the discovery set: $p < 0.05/6$

obvious, although functions have been predicted based on homology detection.

To our knowledge, this is only the second attempt to relate *M. tuberculosis* genetic variation to the TB disease phenotype in humans on a genome-wide scale. The other study, by Saw et al. showed large-scale rearrangements, short translocations, inversions, indels and SNPs in eight strains cultured from CSF [33]. Non-synonymous SNPs in eight genes (*embR*, *lppD*, *PE-PGRS10*, *PE-PGRS19*, *PE-PGRS21*, *PE-PGRS49*, *PPE58*, and *Rv0278c*) were found in at least four of the eight CSF-derived strains, and in none of 69 strains isolated from sputum [33]. We did not confirm this in our set of isolates, although *PE-PGRS19* was associated with the TB phenotype in the discovery set. Moreover, we used a two-step approach, based on homoplasy counting as well as allele counting with a correction for phylogenetic bias to find mutations associated with the TB phenotype, and we performed ancestral reconstruction for the most discriminative SNP. Unlike a previous study from Vietnam [32], we found no association between *M. tuberculosis* lineage and TBM. This is no surprise given the genetic diversity even within *M. tuberculosis* lineages [9], and the observed pattern of TBM isolates scattered across the phylogenetic tree.

In concordance with previous findings [9], we found considerable genetic diversity in *M. tuberculosis* in the current study. Two isolates differed on average by 1000 SNPs, and this did not differ among PTB isolates and among TBM isolates. In addition, we did not observe any clustering, defined as two isolates differing by 12 SNPs or less [35]. The lack of clustering is probably a result of the low sampling fraction in this urban setting with thousands of incident TB cases each year.

Theoretically, two scenarios could explain the role of *M. tuberculosis* genetic variation in the development of TBM after infection with *M. tuberculosis*. First, upon infection the *M. tuberculosis* strain may carry certain mutations associated with dissemination and penetration of the blood-brain barrier. Second, a subpopulation of bacteria in the lungs of a PTB patient may develop such mutations, though it was recently shown for bacterial meningitis caused by *S. pneumonia* or *N. meningitidis* that there is no evidence for differential selection between blood and CSF, and that any mutations between these two niches is likely due to mutation hotspots or forms of diversifying selection common to both niches [40]. However, similar to the findings of Saw et al. [33], the genetic variants that we found to be associated with the TB disease phenotype were not exclusive for TBM or PTB, nor were they consistently present in all TBM or PTB strains. Therefore it seems that genetic variants may be part of a complex, multifactorial process leading to this devastating manifestation of TB, in which the human genotype or phenotype equally plays an important role [32, 41].

This is the second, and by far largest study using whole genome sequencing to link *M. tuberculosis* genotype to TBM. In this large cohort of well-characterised patients we studied strains from HIV-negative, adult patients to control for the two most important known risk factors for TBM. In addition, both patient groups were similar with regard to gender, ethnicity, and previous episodes of TB. The *de novo* assembly adds to the strengths of this study because it enabled us to examine regions of the genome that do not map to the reference genome, allowing the investigation of associations between genetic variation in these genomic regions and the TB disease phenotype. The homoplasy-based association analysis has proven to be a successful method to detect *M. tuberculosis* loci associated with a certain phenotype (e.g. transmissible vs. non-transmissible, drug-resistant vs. sensitive) [42, 43]. The major advantage is that false-positive associations due to genetic relatedness of strains with the same phenotype (i.e. 'phylogenetic bias') are filtered out, thereby increasing statistical power to find true associations. In addition, the ancestral reconstruction in the validation step ruled out the possibility that the significant association for the SNP in *Rv0218* was due to population structure.

The current study has several limitations. Firstly, we only focused on mutations in coding regions of the genome, as they are more likely to have functional consequences, but mutations in non-coding regions could also affect function, for instance by transcriptional and translational regulation of protein-coding sequences [44]. Secondly, the large number of genetic variants increases the risk of finding false-positive associations, although homoplasy counting enabled us to filter out many of these false-positives. We did not correct for multiple testing in the discovery set, but we used a validation set where we did correct for multiple testing for confirmation. Lastly, whether bacteria developed TBM-associated mutations before or after infecting a patient remains unclear. One way to investigate this is to compare the genomes of strains isolated from sputum and CSF from the same patient. Unfortunately we did not have the availability of paired isolates. Most TBM patients were too ill to expectorate sputum.

Conclusions

We present evidence from a homoplasy-based association analysis that three *M. tuberculosis* genes, including *Rv0218*, a cell wall-associated and/or secretome gene, are associated with the TB disease phenotype. These findings serve as an important step forward in the quest for an improved understanding of the mycobacterial determinants of TB tissue tropism. Functional validation studies are warranted to further explore the effect of mutations in these genes on protein function.

Methods

Patients and isolates

We used *M. tuberculosis* isolates from two established cohorts of Indonesian patients with confirmed TB. The first group consisted of adult patients (≥ 15 years old) with TBM admitted at Hasan Sadikin Hospital between 2006 and 2013, with *M. tuberculosis* cultured from CSF. The second group was randomly selected from a cohort of culture-positive HIV-negative PTB patients (age ≥ 15 years) from the same setting recruited between 2012 and 2015. All patients were tested for HIV, and those who were HIV-positive were excluded.

Sequencing, alignment, and variant calling

Mycobacterial DNA was extracted from cultures using cetyl trimethylammonium bromide (CTAB) or using UltraClean[®] Microbial DNA Isolation Kit (MO BIO Laboratories). A single isolate from each patient was selected for sequencing. *M. tuberculosis* DNA was sequenced on an Illumina HiSeq 2000 instrument using 2×100 bp paired-end reads at the Beijing Genome Institute in Hong Kong. After sequencing, the raw FASTQ sequence reads were filtered, including removing of adapter sequences, contamination, and low quality reads which have more than 10% N base calls, or where more than 40% of the bases have a quality score ≤ 4 . Quality control statistics are shown in Additional file 11: Table S4. Five TBM strains and four PTB strains were contaminated, based on a low GC-content, and were excluded from further analyses. Sequencing coverage was determined using the FASTQC quality control tool version 0.10.1. The proportion of bases sequenced with a sequencing error rate of 1% or less per base ranged from 93% to 97% per genome. The average coverage depth for the remaining 322 sequenced strains was 121.1, and the average percentage of bases covered by at least one read was 98.9%.

The sequence reads were aligned to reference strain *M. tuberculosis* H37Rv, accession number NC_000962.3, and variants were called using Breseq software, version 0.27.1 [45] using a minimum threshold of $30\times$ coverage. Mutations with low-quality evidence (i.e. possible mixed read alignment) were not included. The Breseq variant call output was converted to a tab-separated file for each sequence using customized Python and R scripts that are available upon request.

Phylogeny construction

A phylogeny was constructed to determine evolutionary relationships of the isolates. We extracted all 29,199 variable positions across the 322 *M. tuberculosis* sequences and concatenated them into a single alignment. Solely for the purpose of creating the phylogenetic tree, SNPs occurring in PE/PPE genes and genes related

to mobile elements (genes listed in Additional file 12: Table S5) were excluded to avoid any concern about inaccuracies in the read alignment in these parts of the genome. In addition, SNPs in an additional 40 genes previously associated with drug resistance [46] were removed to exclude the possibility that homoplasmy of drug resistance mutations would significantly affect the phylogeny [47]. After applying these filters to the initial set of 29,199 SNPs, the 28,544 remaining SNPs were used to construct the phylogenetic tree using PhyML, version 3.0 [48] using the HKY85 model with four categories for the gamma distribution, and using a hundred bootstraps.

To determine the lineage distribution of the strains and to evaluate whether an association exists between *M. tuberculosis* lineage and TB disease phenotype, we determined the lineage for each of the 322 strains using a 62-SNP barcode [34]. The resulting classification in the main *M. tuberculosis* lineages also served as a quality check for the generated Maximum Likelihood (ML)-phylogenetic tree, as it enabled us to validate that isolates belonging to the same lineage clustered together in the tree. A Chi-square test was used to statistically test the association between *M. tuberculosis* lineage and TB disease phenotype.

Homoplasmy-based association test to identify associations between *M. Tuberculosis* genotype and TB disease phenotype

We used a two-step approach: in the discovery step we aimed to maximize power by homoplasmy counting, without correction for multiple testing. In the subsequent validation step, aimed to distinguish true associations from false positives, we used allele counting with multiple testing correction, and performed ancestral reconstruction to remove possible phylogenetic bias.

To divide the genomes in a discovery and a validation set, we identified isolates in terminal branch pairs, trios and quartets (i.e. separated at a terminal or near terminal branch in the phylogenetic tree) with distinct phenotypes (Fig. 2). These terminal branch sets (TBSs) together formed the discovery set. These provide the strongest, homoplasmy-corrected possible association with the phenotype. We used the remaining genomes to validate the association. For all isolates in the TBSs, we listed the non-synonymous SNPs to create a table with all variable positions in rows, the paired isolates in columns, and the corresponding nucleotide in the cells (Additional file 2). For every SNP, an enrichment score was calculated using the following formula:

$$\log \left(\frac{(\text{number of TBM isolates with SNP} / \text{total number of TBM isolates}) + 0.001}{(\text{number of PTB isolates with SNP} / \text{total number of PTB isolates}) + 0.001} \right)$$

A permutation *p*-value for each SNP was calculated by randomising the phenotypes over the isolates 1000 times.

In parallel we grouped SNPs per gene, using the same empirical randomisation strategy to assess association, adjusted for gene length:

$$\log\left(\frac{((\text{number of SNPs in gene in TBM isolates}/\text{gene length})/\text{total number of TBM isolates}) + 0.001}{((\text{number of SNPs in gene in PTB isolates}/\text{gene length})/\text{total number of PTB isolates}) + 0.001}\right)$$

Similarly, we grouped genes with ≥ 1 SNP per *M. tuberculosis* pathway according to PATRIC [49], and calculated association using the aforementioned permutation analysis, adjusted for the number of genes in a particular pathway:

$$\log\left(\frac{((\text{number of pathway genes with } \geq 1 \text{ SNP in TBM isolates}/\text{unique gene count})/\text{total number of TBM isolates}) + 0.001}{((\text{number of pathway genes with } \geq 1 \text{ SNP in PTB isolates}/\text{unique gene count})/\text{total number of PTB isolates}) + 0.001}\right)$$

Significance of associations was determined by calculating a permutation *p*-value through randomization of the phenotypes over the isolates 1000 times. All calculations were performed with customized Perl scripts that are available upon request.

We used the set of 214 strains that were not in TBSs to validate candidate SNPs, genes, and pathways identified in the discovery set, using the same permutation test as described above for the discovery set. We used a *p*-value threshold of 0.05 for the discovery set. The *p*-value thresholds in the validation set were Bonferroni-corrected for multiple testing by dividing them by the number of significant (candidate) hits in the discovery set. To correct for potential phylogenetic bias in the validation set, we performed ancestral reconstruction for validated TB phenotype-associated SNPs using FASTML [50] with default parameters, and compared the proportion of TBM vs. PTB isolates prior to (i.e. older than) and after (i.e. younger than) the occurrence of the SNP in the validation set. For each node / leave where the SNP occurred, we calculated the proportion of TBM isolates among the child branches and we calculated the (unweighted) average over all of these nodes and leaves to determine the proportion of TBM isolates after the SNP (Additional file 7: Table S2). This way, every independent occurrence of the SNP contributes equally to the analysis, regardless of the number of child branches after the SNP, thus correcting for phylogenetic bias. The significance of the difference in proportion was determined by calculating the Z-score for 2 population proportions with accompanying *p*-value.

PE/PPE genes, a major challenge in the analysis of *M. tuberculosis* whole genome sequences due to the repetitive nature of these sequences, were included in the analysis. TB phenotype-associated SNPs in PE/PPE genes were manually examined to confirm that they did not fall within a repetitive region (for an example please see Additional file 13: Figure S4).

De novo genome assembly

Sequence reads were *de novo* assembled using SPAdes, version 3.6.1 [51] with default parameters. All assemblies

were evaluated, focussing on genome size, N50 length, number of contigs and scaffolds, and GC-content. The assembled genomes were annotated using Prokka, version 1.11 [52] with default parameters. Completeness and contamination of assemblies were determined with CheckM version 1.0.5 [53]. The assembly statistics are shown in Additional file 14. Presence or absence of annotated genes was determined for the 108 assembled genomes in the discovery set. An enrichment score per gene was calculated based on the frequency of occurrence in TBM vs. PTB strains using the following formula:

$$\log\left(\frac{(\text{number of TBM isolates with gene present}/\text{total number of TBM isolates}) + 0.001}{(\text{number of PTB isolates with gene present}/\text{total number of PTB isolates}) + 0.001}\right)$$

Statistical significance was again determined based on permutation by randomizing the phenotypes over the isolates 1000 times. We repeated this permutation analysis for the 214 genomes comprising the validation set, using Bonferroni-adjusted *p*-value thresholds. For the genes with a validated, significant enrichment for TBM or PTB, we confirmed their absence in the respective genomes by mapping the raw sequencing reads for these genomes back to the H37Rv reference sequence of the gene (Additional file 15: Figure S5), and visualized this with integrative genomics viewer (IGV), version 2.3.32 [54].

Prediction of mutation effects

We used two algorithms to predict the effect of the mutations on protein structure and function. I-Mutant version 2.0, which predicts the protein stability change upon single site mutation (<http://folding.biofold.org/i-mutant/i-mutant2.0.html>) [55] and PolyPhen-2, which predicts the possible impact of an amino acid substitution on the structure and function of a protein (<http://genetics.bwh.harvard.edu/pph2/>) [56] to predict the impact of the validated SNPs on protein structure and function. In addition, we used TartgetP (<http://www.cbs.dtu.dk/services/TargetP/>) [57] to predict the subcellular location of the proteins encoded by the validated genes, and TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>) [58] to predict transmembrane helices in these proteins.

Additional files

Additional file 1: Table S1. Description of baseline characteristics for PTB and TBM patients. IQR, interquartile range; SD, standard deviation. Data were missing for history of TB treatment (TBM, *n* = 7; PTB, *n* = 1); ethnicity (TBM, *n* = 58; PTB = 2). (DOCX 58 kb)

Additional file 2: Merged SNP lists consisting of variable positions for the 108 strains in the discovery set. File showing all common non-synonymous SNPs in the 108 isolates in the discovery set, including their *p*-value for association with the TB disease phenotype. (XLSX 2729 kb)

Additional file 3: Merged SNP lists consisting of variable positions for the 214 strains in the validation set. File showing all common non-synonymous SNPs in the 214 isolates in the validation set, including their *p*-value for association with the TB disease phenotype. (XLSX 8758 kb)

Additional file 4: Figure S1. *P*-values of association between SNPs and TB disease phenotype in the discovery and validation sets. Scatterplot showing the *p*-values of the SNPs found in the discovery and the validation set. *P*-values in the discovery set are shown on the x-axis; *p*-values in the validation set are shown on the y-axis. Gene names are shown for SNPs significant in both the discovery and validation set. (DOCX 82 kb)

Additional file 5: Figure S2. *P*-values of association between genes and TB disease phenotype in the discovery and validation sets. Scatterplot showing the *p*-values of the genes found in the discovery and the validation set. *P*-values in the discovery set are shown on the x-axis; *p*-values in the validation set are shown on the y-axis. Names are shown for genes significant in both the discovery and validation set. (DOCX 66 kb)

Additional file 6: Figure S3. Phylogenetic tree of 322 *M. tuberculosis* strains isolated from TBM and PTB patients. The highlighted branches indicate the 108 strains in 47 terminal branch sets, together comprising the discovery set. The purple stars indicate the origin of the SNP in Rv0218 according to the ancestral reconstruction. The nucleotide for SNP position 261,869 is indicated next to the leaf labels. (PDF 60 kb)

Additional file 7: Table S2. Ancestral reconstruction of SNP 261869TC in Rv0218. Listed are the internal nodes and leaves where the SNP in Rv0218 occurred according to the ancestral reconstruction of the SNP. (DOCX 72 kb)

Additional file 8: Annotated coding sequences present in the 108 *de novo* assembled genomes in the discovery set. Annotated coding sequences present in at least one, but not in all of the 108 *de novo* assembled genomes in the discovery set. (XLSX 1286 kb)

Additional file 9: Annotated coding sequences present in the 214 *de novo* assembled genomes in the validation set. Annotated coding sequences present in at least one, but not in all of the 214 *de novo* assembled genomes in the validation set. (XLSX 1309 kb)

Additional file 10: Table S3. Protein prediction for genomic sites associated with the TB disease phenotype. NA: No homologs of Rv0192 were found therefore protein prediction was not possible. # I-mutant predicts free energy changes of protein stability upon a point mutation under different conditions. & PolyPhen predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. ^ TargetP predicts the subcellular location of proteins based on the predicted presence of any N-terminal signal peptides. * TMHMM predicts transmembrane helices in proteins. (DOCX 57 kb)

Additional file 11: Table S4. Description of sequencing quality control parameters and statistics. Displayed are different measures of sequencing quality, used for the sequencing quality control check. (DOCX 208 kb)

Additional file 12: Table S5. PE / PPE genes and drug resistance genes excluded for the phylogeny construction. Listed are the genes that were excluded from the multiple alignment used to create the phylogenetic tree. (DOCX 158 kb)

Additional file 13: Figure S4. Diagram demonstrating breseq calling a SNP in the *PE-PGRS1* gene. Displayed are 60 Illumina sequencing reads mapping to the H37Rv reference genome (shown at the top and bottom). Visual inspection of the SNP confirms that it does not occur in a region containing uniformly lower base quality scores. (DOCX 1204 kb)

Additional file 14: *De novo* genome assembly statistics. File listing different measures used to check the quality of the *de novo* genome assemblies. (XLS 73 kb)

Additional file 15: Figure S5. Read alignment demonstrating the absence of *nanK*. Displayed is the alignment of the raw sequencing reads against the H37Rv *nanK* gene. No reads are mapping to this gene, showing that it is absent in the sequenced genome. (DOCX 54 kb)

Goeman for statistical advice; Jakko van Ingen for fruitful discussions; Jordy Coolen, Maha Farhat, Daniel Garza, Robin van der Lee, and Aldert Zomer for advice on the methodology and bioinformatics; Bruno Andrade for assisting in the *de novo* assembly, and the director of the Hasan Sadikin General Hospital for accommodating the research.

Funding

This study was supported by the Royal Netherlands Academy of Arts and Sciences (KNAW). [09-PD-14 to RvC]; fellowship from the Netherlands Organization for Health Research and Development (ZonMw) and The Netherlands Foundation for Scientific Research [VIDI grant. 017.106.310 to RvC., and VIDI grant 864.14.004 to BED.]; and Radboud University fellowship [to CR]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was also supported by the TANDEM (Tuberculosis and Diabetes Mellitus) Grant of the ECFP7 (European Union's Seventh Framework Programme) under Grant Agreement no. 305279.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. Data generated or analysed during this study are included in this published article [and its supplementary information files], or are available from the corresponding author on reasonable request. The raw sequence files (FASTQ) were archived on the NCBI Sequence Read Archive and are available at: <https://www.ncbi.nlm.nih.gov/sra/SRP130118>. The individual isolates can be accessed under the following Biosample accession numbers: SAMN08376067-SAMN08376388. The Bioproject accession number is: PRJNA430531. Phylogeny data have been uploaded to TreeBASE (<http://purl.org/phylo/treebase/phylovs/study/TB2S22081>).

Authors' contributions

CR was responsible for conceptualization, data analysis, funding acquisition, and writing. LC was responsible for laboratory management, project administration, resources, and writing. AvL was responsible for conceptualization and writing. SD was responsible for inclusion of patients in the study and writing. ARG was responsible for patient management and writing. HNG was responsible for conceptualization and methodological guidance. MAH was responsible for conceptualization, methodology, resources, supervision, and writing. BA was responsible for funding acquisition, project administration, resources, and supervision. BED was responsible for conceptualization, data analysis, methodology, supervision, and writing. RvC was responsible for conceptualization, funding acquisition, project administration, supervision, and writing. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All adult patients provided written informed consent; from the age of 15, patients are no longer seen by a paediatrician [59] and parents provided informed consent for patients under 18. The consent procedure was approved by the local Institutional Review Board. The study protocols for the inclusions of patients and for bioanalysis were approved by the ethical committee of the Faculty of Medicine, Universitas Padjadjaran / Hasan Sadikin Hospital, Bandung, Indonesia under ethical registration number 0716040326.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Internal Medicine, Radboud university medical center, Nijmegen, The Netherlands. ²Health Research Unit, Faculty of Medicine, Padjadjaran University/Hasan Sadikin Hospital, Bandung, Indonesia. ³Department of Neurology, Faculty of Medicine, Padjadjaran University/Hasan Sadikin Hospital, Bandung, Indonesia. ⁴National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. ⁵Life and Health

Acknowledgements

The authors would like to thank the data management team members for data management, the residents for monitoring patients, professor Jelle

Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal. ⁶ICVS/3B's, PT Government Associate Laboratory, Braga, Portugal. ⁷Centre for Molecular and Biomolecular Informatics, Radboud university medical center, Nijmegen, The Netherlands. ⁸Theoretical Biology and Bioinformatics, Science4Life, Utrecht University, Utrecht, The Netherlands.

Received: 15 June 2017 Accepted: 28 January 2018

Published online: 05 February 2018

References

- WHO. Global Tuberculosis Report 2015. World Health Organization; 2015. http://www.who.int/tb/publications/global_report/gtbr15_main_text.pdf.
- Ganiem AR, Parwati I, Wisaksana R, van der Zanden A, van de Beek D, Sturm P, et al. The effect of HIV infection on adult meningitis in Indonesia: a prospective cohort study. *AIDS*. 2009;23(17):2309–16.
- van Laarhoven A, Dian S, Ruesen C, Hayati E, Damen MSMA, Annisa J, Chaidir L, Netea MG, Alisjahbana B, Ganiem AR, van Crevel R. Clinical parameters, routine inflammatory markers and LTA4H genotype as predictors for mortality among 608 tuberculous meningitis patients in Indonesia. *J Infect Dis*. 2017;215:1029.
- Graustein AD, Horne DJ, Arentz M, Bang ND, Chau TT, Thwaites GE, et al. TLR9 gene region polymorphisms and susceptibility to tuberculosis in Vietnam. *Tuberculosis (Edinb)*. 2015;95(2):190–6.
- Campo M, Randhawa AK, Dunstan S, Farrar J, Caws M, Bang ND, et al. Common polymorphisms in the CD43 gene region are associated with tuberculosis disease and mortality. *Am J Respir Cell Mol Biol*. 2015;52(3):342–8.
- Hawn TR, Dunstan SJ, Thwaites GE, Simmons CP, Thuong NT, Lan NT, et al. A polymorphism in toll-interleukin 1 receptor domain containing adaptor protein is associated with susceptibility to meningeal tuberculosis. *J Infect Dis*. 2006;194(8):1127–34.
- Thuong NT, Hawn TR, Thwaites GE, Chau TT, Lan NT, Quy HT, et al. A polymorphism in human TLR2 is associated with increased susceptibility to tuberculous meningitis. *Genes Immun*. 2007;8(5):422–8.
- Hoal-Van Helden EG, Epstein J, Victor TC, Hon D, Lewis LA, Beyers N, et al. Mannose-binding protein B allele confers protection against tuberculous meningitis. *Pediatr Res*. 1999;45(4 Pt 1):459–64.
- Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. *Semin Immunol*. 2014;26(6):431–44.
- Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, et al. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in mycobacterium tuberculosis isolates. *BMC Genomics*. 2015;16(1):857.
- Guerra-Assuncao JA, Houben RM, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to relapse or reinfection with mycobacterium tuberculosis: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *J Infect Dis*. 2015;211(7):1154–63.
- Reed MB, Domenech P, Manca C, Su H, Barczak AK, Kreiswirth BN, et al. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature*. 2004;431(7004):84–7.
- Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*. 2006;103(8):2869–73.
- Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife*. 2015;4:1–17.
- de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by mycobacterium tuberculosis lineage in the Gambia. *J Infect Dis*. 2008;198(7):1037–43.
- van Crevel R, Nelwan RH, de Lenne W, Veeraragu Y, van der Zanden AG, Amin Z, et al. Mycobacterium tuberculosis Beijing genotype strains associated with febrile response to treatment. *Emerg Infect Dis*. 2001;7(5):880–3.
- Parwati I, Alisjahbana B, Apriani L, Soetikno RD, Ottenhoff TH, van der Zanden AG, et al. Mycobacterium tuberculosis Beijing genotype is an independent risk factor for tuberculosis treatment failure in Indonesia. *J Infect Dis*. 2010;201(4):553–7.
- Rakotosamimanana N, Raharimanga V, Andriamandimby SF, Soares JL, Doherty TM, Ratsitorahina M, et al. Variation in gamma interferon responses to different infecting strains of mycobacterium tuberculosis in acid-fast bacillus smear-positive patients and household contacts in Antananarivo, Madagascar. *Clin Vaccine Immunol*. 2010;17(7):1094–103.
- van Laarhoven A, Mandemakers JJ, Kleinnijenhuis J, Enaimi M, Lachmandas E, Joosten LA, et al. Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the mycobacterium tuberculosis Beijing genotype. *Infect Immun*. 2013;81(10):3750–6.
- 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(3):e1001307.
- Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. Modern lineages of mycobacterium tuberculosis exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. *PLoS One*. 2012;7(8):e43170.
- Be NA, Lamichhane G, Grosset J, Tyagi S, Cheng QJ, Kim KS, et al. Murine model to study the invasion and survival of mycobacterium tuberculosis in the central nervous system. *J Infect Dis*. 2008;198(10):1520–8.
- Be NA, Bishai WR, Jain SK. Role of mycobacterium tuberculosis pknD in the pathogenesis of central nervous system tuberculosis. *BMC Microbiol*. 2012;12:7.
- Skerry C, Pokkali S, Pinn M, Be NA, Harper J, Karakousis PC, et al. Vaccination with recombinant mycobacterium tuberculosis PknD attenuates bacterial dissemination to the brain in guinea pigs. *PLoS One*. 2013;8(6):e66310.
- Hernandez Pando R, Aguilar D, Cohen I, Guerrero M, Ribon W, Acosta P, et al. Specific bacterial genotypes of mycobacterium tuberculosis cause extensive dissemination and brain infection in an experimental model. *Tuberculosis (Edinb)*. 2010;90(4):268–77.
- Tsenova L, Ellison E, Harbacheuski R, Moreira AL, Kurepina N, Reed MB, et al. Virulence of selected mycobacterium tuberculosis clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. *J Infect Dis*. 2005;192(1):98–106.
- Jain SK, Paul-Satyaseela M, Lamichhane G, Kim KS, Bishai WR. Mycobacterium tuberculosis invasion and traversal across an in vitro human blood-brain barrier as a pathogenic mechanism for central nervous system tuberculosis. *J Infect Dis*. 2006;193(9):1287–95.
- Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship between mycobacterium tuberculosis phylogenetic lineage and clinical site of tuberculosis. *Clin Infect Dis*. 2012;54(2):211–9.
- Pareek M, Evans J, Innes J, Smith G, Hingley-Wilson S, Lougheed KE, et al. Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype. *Thorax*. 2013;68(3):221–9.
- Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerg Infect Dis*. 2013;19(3):460–3.
- Nicol MP, Sola C, February B, Rastogi N, Steyn L, Wilkinson RJ. Distribution of strain families of mycobacterium tuberculosis causing pulmonary and extrapulmonary disease in hospitalized children in cape town, South Africa. *J Clin Microbiol*. 2005;43(11):5779–81.
- Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NT, Thuong NT, et al. The influence of host and bacterial genotype on the development of disseminated disease with mycobacterium tuberculosis. *PLoS Pathog*. 2008;4(3):e1000034.
- Saw SH, Tan JL, Chan XY, Chan KG, Ngeow YF. Chromosomal rearrangements and protein globularity changes in mycobacterium tuberculosis isolates from cerebrospinal fluid. *Peer J*. 2016;4:e2484.
- Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigo J, et al. PolyTB: a genomic variation map for mycobacterium tuberculosis. *Tuberculosis (Edinb)*. 2014;94(3):346–54.
- Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dediccoat MJ, et al. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis*. 2013;13(2):137–46.
- Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol*. 2015;25:17–24.
- Sasseti CM, Rubin EJ. Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A*. 2003;100(22):12989–94.
- Zheng J, Ren X, Wei C, Yang J, Hu Y, Liu L, et al. Analysis of the secretome and identification of novel constituents from culture filtrate of bacillus Calmette-Guerin using high-resolution mass spectrometry. *Mol Cell Proteomics*. 2013;12(8):2081–95.

39. Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EP. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr Biol*. 2009;19(20):1683–91.
40. Lees JA, Kremer PH, Manso AS, Croucher NJ, Ferwerda B, Seron MV, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom*. 2017;3(1):e000103.
41. Gagneux S. Host-pathogen coevolution in human tuberculosis. *Philos Trans R Soc Lond Ser B Biol Sci*. 2012;367(1590):850–9.
42. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis. *Nat Genet*. 2013;45(10):1183–9.
43. Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, Koeken VA, Mandemakers JJ, Zomer A, et al. Transmissible Mycobacterium tuberculosis Strains Share Genetic Markers and Immune Phenotypes. *Am J Respir Crit Care Med*. 2017;195(11):1519–152.
44. Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*. 2005;21(7):399–404.
45. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol*. 2014;1151:165–88.
46. Coll F, McNERney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med*. 2015;7(1):51.
47. Farhat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome Med*. 2014;6(11):101.
48. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307–21.
49. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 2014;42(Database issue):D581–91.
50. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res*. 2012;40(Web Server issue):W580–4.
51. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In: Deng M, Jiang R, Sun F, Zhang X. (eds) *Research in Computational Molecular Biology*. RECOMB 2013. Lecture Notes in Computer Science, vol 7821. Springer, Berlin, Heidelberg.
52. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–9.
53. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043–55.
54. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–6.
55. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 2005;33(Web Server issue):W306–10.
56. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248–9.
57. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*. 2000;300(4):1005–16.
58. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305(3):567–80.
59. Ministry of Health Indonesia. Directorate General of Disease Control and Environmental Health. Petunjuk Teknis Manajemen TB Anak. 2013. <http://www.spiritia.or.id/Dok/juknisTBAnak2013.pdf>.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

