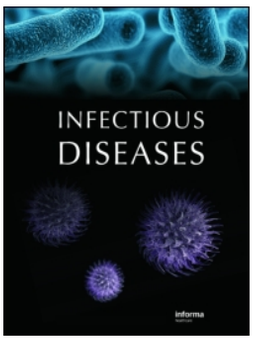


Infectious Diseases

ISSN: 2374-4235 (Print) 2374-4243 (Online) Journal homepage: http://www.tandfonline.com/loi/infd20

Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of Mycobacterium tuberculosis

Sanjay S. Gautam, Micheál Mac Aogáin , James E. Bower, Indira Basu & Ronan F. O'Toole

To cite this article: Sanjay S. Gautam, Micheál Mac Aogáin , James E. Bower, Indira Basu & Ronan F. O'Toole (2017): Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of Mycobacterium tuberculosis, Infectious Diseases, DOI: 10.1080/23744235.2017.1330553

To link to this article: http://dx.doi.org/10.1080/23744235.2017.1330553

Published online: 23 May 2017.

Taylor & Francis
Taylor & Francis Group

INFECTIOUS DISEASES,
2017; VOL. 0,
NO. 0, 1–9

https://doi.org/10.1080/23744235.2017.1330553

ORIGINAL ARTICLE

Check for updates

Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of *Mycobacterium tuberculosis*

Sanjay S. Gautam[a], Micheál Mac Aogáin[b], James E. Bower[c], Indira Basu[c] and Ronan F. O'Toole[a]

[a]School of Medicine, University of Tasmania, Hobart, Australia; [b]Department of Clinical Microbiology, Trinity Translational Medicine Institute, School of Medicine, Trinity College Dublin, St. James's Hospital, Dublin, Ireland; [c]LabPLUS, Auckland City Hospital, Auckland, New Zealand

ABSTRACT

Background: The Rangipo strain of *Mycobacterium tuberculosis* achieved notoriety in New Zealand due to its role in several tuberculosis (TB) outbreaks. Why this strain should be the source of relatively large clusters of the disease is unknown. In this work, we performed an in-depth analysis of the genome of the Rangipo strain to determine whether it offers clues to understanding its prevalence.

Methods: Next-generation sequencing was performed on nine isolates which matched the Rangipo genotypic profile. Sequence reads were assembled against the H37Rv reference genome and single-locus variants identified. Unmapped reads were compared against the genome sequences of other *M. tuberculosis* strains, in particular CDC1551, Haarlem and Erdman.

Results: Across the nine Rangipo strains, a total of 727 single-locus variants were identified with respect to H37Rv, of which 700 were common to all Rangipo strains sequenced. Within the common variants, 386 were non-synonymous, with 12 occurring in genes associated with *M. tuberculosis* virulence. Next-generation and Sanger sequencing determined the presence of three genes in the Rangipo isolates, which are absent in H37Rv, but which have been reported to be important for the pathogenicity of *M. tuberculosis*. The differentially encoded Rangipo genes consisted of transcriptional regulator EmbR2, and molybdopterin cofactor biosynthesis proteins A and B. The Rangipo strain also harbours an extended DNA helicase and an additional adenylate cyclase.

Conclusions: Our study provides new insights into the genomic content of the New Zealand Rangipo strain of *M. tuberculosis* and highlights the presence of additional virulence-related loci not found in H37Rv.

KEYWORDS
Mycobacterium tuberculosis
CDC1551
EmbR2
whole-genome sequencing

ARTICLE HISTORY
Received 27 December 2016
Revised 20 April 2017
Accepted 10 May 2017

CONTACT
Ronan F. O'Toole
ronan.otoole@utas.edu.au
School of Medicine, Faculty of Health, University of Tasmania, Medical Science 1, 17 Liverpool Street, Hobart, TAS 7000, Australia

Supplemental data for this article can be accessed here.

Introduction

Tuberculosis (TB) is the leading cause of mortality due to respiratory infection worldwide, killing ~1.5 million people in 2014 alone [1]. New Zealand has a low notification rate of TB with 6.7 cases per 100,000 population (302 cases) in 2014 [2]; however, the distribution of the disease is not uniform across the population with some districts and at-risk groups exhibiting disparate TB rates. For example, the notification rate of TB in the Asian ethnic group was 56.8 times higher (34.1 per 100,000) than in the European or other ethnic group (0.6 per 100,000) in 2014 [2]. TB rates have also been consistently higher in Māori and Pacific peoples compared to New Zealanders of European descent [2,3]. Furthermore, intermittent outbreaks of TB can significantly impact the national rate from year to year.

In 2001, an analysis of an outbreak of TB in New Zealand which occurred between November 1996 and May 2000 was reported [4]. Forty-three of the 61 TB cases were confirmed as belonging to the outbreak based on IS6110-based restriction fragment length polymorphism (RFLP) typing of the isolates, with the remaining 18 cases determined by epidemiological contact tracing. One of the patients had previously served a prison sentence in the Tongariro/Rangipo Prison in 1998, and as a result, the strain subsequently became referred to as the Rangipo strain [5]. A subsequent outbreak in 2002 in the Hawke's Bay involving 19 active cases of Rangipo TB was associated with a high rate of transmission as determined by the presentation of TB disease or latent TB infection in 16.4% and 20.0%, respectively, of household or other close contacts [5]. A molecular typing study of *M. tuberculosis* complex isolates from 2003 to 2007 found that the Rangipo strain was associated with the highest number of district health boards (DHBs) in New Zealand (13 out of 18) compared to other clusters which were related to a median of 2 DHBs [6].

The reasons underlying the high infectivity and prevalence of the Rangipo strain are not known. Prison incarceration was a factor common to a number of cases [4]. The role of the prison environment in the epidemiology of TB has been well documented with active TB occurring at higher levels than among the general population [7]. This has been attributed to the fact that a disproportionate number of prisoners may exhibit TB risk factors including drug or alcohol misuse, homelessness or low socioeconomic status [8]. In addition, the location of prisoners is normally based on crime rather than public health considerations which can contribute to overcrowding, delayed diagnosis or inadequate treatment [7]. It is believed that prisons can act as reservoirs for TB transmission in the wider community [9].

It has been postulated that the Rangipo strain may be more infectious than other strains in circulation [5]; however, there are currently little published experimental data available in regard to this. We have previously shown that the Rangipo strain belongs to the Euro-American Lineage 4 of *Mycobacterium tuberculosis* complex based on earlier large sequence and single nucleotide polymorphic analyses [10]. In this work, we examined isolates of the Rangipo strain using whole-genome sequencing to shed light on potential differences in the encoded pathogenicity of this strain. In addition to single-locus variations, we discuss a series of virulence genes which distinguish the Rangipo strain from H37Rv and other reference *M. tuberculosis* strains.

Materials and methods

Bacterial strains

Mycobacterial Interspersed Repetitive Unit (MIRU) 24-loci typing was performed as previously described [11] on clinical isolates of *Mycobacterium tuberculosis* complex at LabPLUS, Auckland City Hospital. Isolates were defined as belonging to a cluster where they contained the same copy number at all loci. In this work, eight isolates with the identical 24-loci MIRU profile (233325153324 341444223362) of the Rangipo strain [5], and an additional ninth isolate, strain 356, which differed by one MIRU locus, were selected for whole-genome sequencing. The isolates analysed were from across multiple geographical locations and collected over a two-year period. The isolates were grown in Mycobacteria Growth Indicator Tube (MGIT) media and the cells lysed using glass beads for subsequent genomic DNA (gDNA) extraction.

Next-generation sequencing

gDNA of the *M. tuberculosis* isolates were purified, RNase-treated and quantified. The gDNA was tagged and amplified using a Nextera® XT DNA Library Preparation Kit, San Diego, CA and Nextera® XT Index Kit, San Diego, CA. The libraries generated were cleaned using Agencourt AMPure XP beads, normalized and then pooled. The concentration of the pooled library was determined by qPCR using a KAPA Library Quantification Kit, Wilmington, MA. 15 pM of the pooled library was loaded into the cartridge of a MiSeq Reagent Kit v2

which was run on an Illumina® MiSeq, San Diego, CA instrument and the generated FASTQ sequence files were collected.

Whole-genome sequence analysis

The FASTQ files of the nine Rangipo strains of *M. tuberculosis* were imported into the Geneious R9.0 software suite [12]. Paired-end sequence reads were trimmed (error probability limit of 0.05) and mapped (random multiple base matches) to the publicly available annotated genome of *M. tuberculosis* reference strain H37Rv (accession number NC_000962.3) [11] in the first instance. The maximum variant *p*-value was set at 10^{-6} when exceeding 65% bias. Single-locus variations (SLVs) were called at a minimum variant frequency of 95% and a minimum mean genome coverage of 20 and annotated as previously described [13]. Described repetitive regions in the H37Rv genome were masked to mitigate spurious variant calling [14]. A list of Rangipo common non-synonymous SLVs was created and classified based on the clusters of orthologous groups (COGs). Information on the known role of these genes with respect to mycobacterial virulence was obtained from the published literature.

Identification of differentially encoded genes in the Rangipo strain

To identify differentially encoded genes that are present in Rangipo, *de novo* assembly was performed on unmapped FASTQ reads from each of the Rangipo strains that did not map to the reference genome H37Rv using a trim error probability limit of 0.05. A consensus sequence was generated for each of the *de novo* assemblies and open reading frames (ORFs), with a minimum length of 150 nucleotides, standard genetic code and start codons ATG, TTG and CTG. Interior ORFs were included based on the assumption that start and/or stop codons could be located outside of the minimum ORF length. The predicted protein products of the ORFs were compared against other *M. tuberculosis* strains in the database of non-redundant protein sequences by BLASTP analysis (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [15] to identify Rangipo strain orthologues of genes present in other *M. tuberculosis* strains but absent in H37Rv. Subject sequences with $\geq$98% identity to the query sequence at the amino acid level were selected and their known or predicted functions recorded. FASTQ sequence reads were also assembled against the annotated genomes of other strains of *M. tuberculosis* Euro-American Lineage 4, in particular CDC1551 (NC_002755), Erdman (NC_020559) and Haarlem (NC_022350). Lineage numbers refer to the Gagneux lineage classification system [16].

Confirmation of the presence of differentially encoded genes by PCR amplification and sanger re-sequencing

PCR was used to confirm the presence or absence of differentially encoded genes in the Rangipo and H37Rv strains. For this, pairs of oligonucleotides were designed using Primer 3 software (version 2.3.4) to amplify the full length and internal regions of the genes assayed (Supplementary Table S1). A consensus gene sequence, generated by aligning Rangipo genomes with the genome of MTBC strain CDC1551, was used as a template to design the primers. All pairs of primers were verified using the nucleotide BLASTN tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [15] to ensure that specific complementary regions were present only within the region of interest. For amplification reactions, a total reaction volume of 25 μL was used containing 0.5 μL DNA, 0.5 μL of 200 pM primer (forward and reverse each), 0.5 μL of 200 μM dNTPs (Invitrogen, Carlsbad, CA), 0.25 μL of 2U DNA polymerase (Invitrogen, Carlsbad, CA), 5 μL of 5 × buffer (Invitrogen, Carlsbad, CA) and 17.75 μL of ultrapure distilled water (Invitrogen, Carlsbad, CA) was used. The thermocycler settings were 98 °C for 30 seconds, followed by 30 cycles of 98 °C for 10 seconds, 55 °C for 30 seconds, 72 °C for 2 minutes, followed by 72 °C for 5 minutes and hold at 4 °C. The amplified fragments were separated by agarose gel electrophoresis and visualized using a Chemidoc XRS system (Bio-Rad Universal Hood II). The PCR products were excised from the agarose and column purified and Sanger sequencing of the products was performed at the Australian Genomic Research Facility (AGRF).

Phylogenetic tree building

The Rangipo genome sequences were assembled against the *M. tuberculosis* reference genome H37Rv to generate consensus sequences for each isolate. A maximum-likelihood phylogenetic tree was then built using the generalized time reversible (GTR) substitution model in PhyML [17] and included the annotated whole-genome sequences of the *M. tuberculosis* Lineage 4 strains

CDC1551 (NC_002755), H37Rv (NC_000962.3), Haarlem (NC_022350) and Erdman (NC_020559).

Results

Single-locus variations identified in the Rangipo strain of MTBC with respect to H37Rv

Single-locus variations (SLVs) were identified after aligning each Rangipo genome to H37Rv. A total of 727 SLVs were identified of which 700 polymorphisms were common to all 9 of the Rangipo isolates sequenced with respect to H37Rv. This compares to a previous report of 747 single nucleotide polymorphisms (SNP) with respect to H37Rv that were present in ten *M. tuberculosis* Rangipo isolates following genome assembly [18]. Of the common Rangipo SLVs, a total of 386 non-synonymous SLVs were identified which consisted of 354 missense variants, 21 frameshift mutations, 4 in-frame insertions, 4 stop codon gains, 2 in-frame deletions and 1 loss of stop codon mutation. The number of differences between individual Rangipo isolates ranged from 0 to 19 variant sites. A threshold of $\leq$5 SNPs between *M. tuberculosis* isolates has previously been proposed as an indicator of recent TB transmission between patients [19,20]. A number of the pairwise distances between isolates fall within the $\leq$5 SNP threshold (Table 3), but it should be noted that SNP distances between TB isolates can be influenced by factors such as time between patient sampling, local TB incidence and homogeneity of *M. tuberculosis* strains in some regions [21,22].

Non-synonymous variations in genes of known or predicted function

The genes containing the 386 non-synonymous SLVs were analysed with respect to their functional classification in terms of COGs [23]. Of these, 156 genes belonged to a single COG, with an additional 11 genes that could be assigned to two COGs. The remaining 219 genes were not assigned to an existing COG in the KEGG database (http://www.genome.jp/kegg/) [24] (Table 1). The most abundant COG among the genes containing the non-synonymous SLVs was COG C, energy production and conversion ($n = 16$) followed by COG L, replication, recombination and repair ($n = 15$). Conversely, the lowest numbers were classified under COG F, nucleotide transport and metabolism ($n = 1$); COG N, cell motility ($n = 2$); and COG D, cell cycle control, cell division and chromosome partitioning ($n = 3$). The Rangipo strains commonly contained variations with respect to H37Rv in genes determining metabolic function (59.5%) (Category: C, E, F, G, H, P, Q, R., S, I), followed by genes determining cellular process and signalling (23.5%) (Category: D, M, N, O, T, V), information storage and processing (16.8%) (Category: J, K, L) and phage- and transposon-associated proteins (1.63%) (Category: X). Rangipo common SLVs occurred most frequently in COG C (energy production and conversion; $n = 16$), COG L (replication, recombination and repair; $n = 15$) and COG H (coenzyme transport and metabolism; $n = 14$). It is important to note that COG categories C, H and L contain a relatively high number of genes, i.e. 102, 123, 81, respectively, which will likely impact on the frequency at which variants are detected in these categories.

Among the 700 SLVs common to the Rangipo strains sequenced, in-frame insertions were identified in four coding sequences (Rv0872c, Rv1435c, Rv2407 and Rv2823c) and deletions were detected in two coding sequences (Rv3345c and Rv0849). Gains of a stop codon were located in the genes *pstA*1 (Rv0930), *ipd*A (Rv3303), PE35 (Rv3872) and Rv0851c, and the loss of a stop codon was found in the polyketide beta-ketoacyl synthase *pks*3 gene (Rv1180) in all nine Rangipo isolates sequenced. In addition, there were 55 mutations occurring within PPE or PE_PGRS family proteins. Twelve non-

Table 1. Genes containing non-synonymous single-locus variations (nsSLVs), common to all nine Rangipo isolates sequenced, categorized with respect to clusters of orthologous groups (COGs).

COG	Functional category	Number of genes per COG
C	Energy production and conversion	16
D	Cell cycle control, cell division, chromosome partitioning	3
E	Amino acid transport and metabolism	12
F	Nucleotide transport and metabolism	1
G	Carbohydrate transport and metabolism	9
H	Coenzyme transport and metabolism	14
I	Lipid transport and metabolism	11
J	Translation, ribosomal structure and biogenesis	11
K	Transcription	4
L	Replication, recombination and repair	15
M	Cell wall/membrane/envelope biogenesis	13
N	Cell motility	2
O	Post translational modification, protein turnover, chaperones	9
P	Inorganic ion transport and metabolism	14
Q	Secondary metabolite biosynthesis, transport and catabolism	8
R	General function prediction only	9
S	Function unknown	12
T	Signal transduction mechanism	8
V	Defence mechanism	7
–	COG not designated	219
	Total	397

The genes containing the 386 non-synonymous SLVs, common to all nine Rangipo isolates sequenced, were analysed with respect to their functional classification in terms of clusters of orthologous group (COG). One hundred and fifty-six of the genes belonged to a single COG, with an additional 11 genes shared between two COGs. Two hundred and nineteen of the genes were not assigned to an existing COG in the KEGG database.

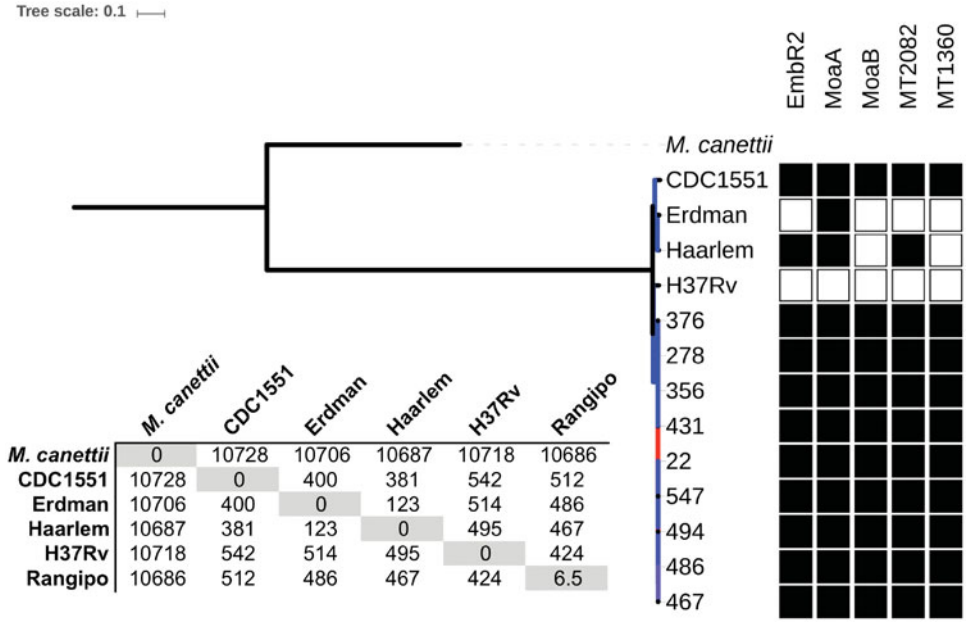
synonymous SLVs common to the Rangipo isolates were found in genes which have previously been associated with *M. tuberculosis* virulence, i.e. *aceAa*, *fadD28*, *kefB*, *mce1F*, *mycP1*, *pckA*, *pepD*, *phoR*, *pks15*, *plcB*, *pstA1*, *pstS1* [25].

Differentially encoded genes in the Rangipo strain

FASTQ sequences from the Rangipo strains, which did not map to the genome of H37Rv, were compared against the genome of other strains of *M. tuberculosis* by BLASTP analysis (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [15]. This identified a number of genes ($n = 5$) which were present in the nine Rangipo isolates sequenced in this study and which had orthologues in strain CDC1551. The presence of the above five genes was confirmed in all nine of the Rangipo isolates by PCR amplification and subsequent Sanger sequencing and BLASTX analysis of the products (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [15]. The five genes were not amplified from the H37Rv template in agreement with their absence in the annotated genome sequence of this strain as reported by Fleischmann et al. [26]. The five Rangipo gene products consisted of orthologues of transcriptional regulator protein EmbR2 (MT3428), molybdopterin cofactor biosynthesis protein A (MT3427), molybdopterin cofactor biosynthesis protein B (MT3426), an extended-length DNA helicase (MT2082) and an additional adenylate cyclase (MT1360) from strain CDC1551. The closest orthologue of the helicase MT2082 in the H37Rv genome was Rv2024c (33% query cover, 99% identity at the nucleotide level). The closest orthologue of the adenylate cyclase MT1360 in the H37Rv genome was Rv1319c (99% query cover, 85% identity at the nucleotide level). The distribution of differentially encoded gene products in the *M. tuberculosis* Rangipo isolates compared to reference genomes of strains belonging to *M. tuberculosis* Lineage 4 is shown in Figure 1.

Phylogenetic analyses of the Rangipo strain

A maximum-likelihood-based distance tree was constructed in PhyML based on the consensus whole-genome sequences of the Rangipo isolates and the whole-genome sequences of the *M. tuberculosis* Lineage 4 strains CDC1551 (NC_002755), H37Rv (NC_000962.3), Haarlem (NC_022350) and Erdman (NC_020559), using the *M. canettii* to root the tree (Figure 1). This analysis



	M. canettii	CDC1551	Erdman	Haarlem	H37Rv	Rangipo
M. canettii	0	10728	10706	10687	10718	10686
CDC1551	10728	0	400	381	542	512
Erdman	10706	400	0	123	514	486
Haarlem	10687	381	123	0	495	467
H37Rv	10718	542	514	495	0	424
Rangipo	10686	512	486	467	424	6.5

Figure 1. Differential carriage of genes in the Rangipo strain and related *Mycobacterium tuberculosis* reference genomes. A maximum-likelihood phylogenetic tree was built using the generalized time reversible (GTR) substitution model in PhyML, including reference strains of *M. tuberculosis* Lineage 4. A total of 2,944,041 sites were identified on comparing Rangipo isolates with *M. tuberculosis* Lineage 4 strains CDC1551 (SRX393042), Haarlem (SRX347319) and Erdman (SRX364193) to the H37Rv (NC_000962.3) reference genome, while using the *Mycobacterium canettii* reference genome (HE572590.1) as a rooting. Branch lengths indicate divergence distance based on 11,438 common variant sites with bootstrap values indicated as follows: blue: 88–100, mauve: 73–87, red: <25 (of 100 replicates). The presence of genes *embR2, moaA, moaB, MT2082* and *MT1360* is indicated by filled squares to the right of the tree. A pairwise SNV distance matrix indicates strain divergence in terms of median observed variant sites.

Table 2. Non-synonymous single-locus variations, common to all nine Rangipo isolates sequenced, which occur in genes that have previously been related to mycobacterial virulence.

Locus tag	Gene	Role of gene in virulence	Nucleotide change	Amino acid change
Rv0174	*mce1F*	Growth regulation in murine macrophages	1109T > C	Leu370Pro
Rv0211	*pckA*	Growth regulation in human monocyte cell line	302A > C	Asn101Thr
Rv0758	*phoR*	Regulation of cell wall hydrophobicity	515C > T	Pro172Leu
Rv0930	*pstA1*	Resistance to host immunity	913C > T	Arg305[a]
Rv0934	*pstS1*	Adhesin binding to macrophages	63_64insA	Ala22[b]
Rv0983	*pepD*	Stress response protein	1169T > C	Leu390Pro
Rv1915	*aceAa*	Growth regulation in acetate containing medium	26A > C	Glu9Ala
Rv2350c	*plcB*	Growth regulation in macrophages	1406T > C	Leu469Ser
Rv2941	*fadD28*	Virulence attenuation in BALB/c mice	545T > C	Val182Ala
Rv2947c	*pks15*	Interaction with host cells	998T > C	Val333Ala
Rv3236c	*kefB*	Bacterial persistence in Guinea pig	962G > T	Arg321Leu
Rv3883c	*mycP1*	Regulation of ESX-1	979C > T	Pro327Ser

[a]Stop gained.
[b]Frameshift variant.

indicated that isolates 467, 486 and 494 are closely related in agreement with a low number of SLVs (0–1) between these isolates (Table. 3). Of the *M. tuberculosis* reference strains included in the analysis, the genome of H37Rv clustered more closely with the Rangipo isolates (Figure 1).

Table 3. Pairwise distance matrix of single-locus variations across the nine Rangipo isolates sequenced based on reference mapping to H37Rv.

	376	278	431	356	547	22	494	486	467
376		17	8	8	10	6	6	7	7
278			11	11	19	15	15	16	16
431				0	10	6	6	7	7
356					10	6	6	7	7
547						8	8	9	9
22							4	5	5
494								1	1
486									0
467									

Discussion

The association of the Rangipo genotype of *M. tuberculosis* with outbreaks of TB in New Zealand is well documented [4,5]. In this work, we investigated whether genomic data from the Rangipo genotype could inform us with respect to its inherent pathogenicity. Nine isolates of the genotype underwent whole-genome sequencing and comparative analysis was performed on the Rangipo sequences with respect to other virulent isolates of the organism.

Single-locus variant (SLV) analysis with respect to reference genome H37Rv revealed 700 polymorphisms that were common to all nine Rangipo isolates sequenced. Approximately 84.5% of these 700 polymorphisms occurred in coding regions and included 386 non-synonymous variants, 156 of which were located in genes which could be classified in one or more of the COGs (Table 1).

Twelve of the Rangipo common non-synonymous SLVs occurred in genes which have previously been reported to be important for *M. tuberculosis* pathogenesis, i.e. *aceAa*, *fadD28*, *kefB*, *mce1F*, *mycP1*, *pckA*, *pepD*, *phoR*, *pks15*, *plcB*, *pstA1*, *pstS1* (Table 2) [25]. Examples include the isocitrate lyase gene, *aceAa*, disruption of which has been associated with decreased maintenance of *M. tuberculosis* in the lungs and spleen of infected mice and reduced survival in murine macrophages and human blood monocyte-derived macrophages [27]. *M. tuberculosis* harbours four putative phospholipase C genes, *plcA-D*. Triple *plcABC* and quadruple *plcABCD* mutants are attenuated in terms of growth during the late phase of mouse infection [28]. The PhoPR two-component system is required for growth of *M. tuberculosis* under low Mg^{2+} availability *in vitro* and for growth during infection of macrophages and mice [29]. Similarly, mutation of the *pks15/1* polyketide synthase locus abrogates phenolic glycolipid production and results in decreased virulence in a rabbit model of meningeal tuberculosis [30].

While it is possible that one or more of the Rangipo common non-synonymous SLVs could affect the activity of the encoded products, experimental work with site-directed mutants would be needed to confirm the effect of specific variants on mycobacterial cell function and virulence. It is noteworthy that a single nucleotide polymorphism between *M. tuberculosis* strains CDC1551 and H37Rv causing a Leu152Pro substitution in the sensor kinase PhoR has been shown to increase cell wall hydrophobicity [31]. SLVs in the coding sequences of the *mce1* operon have been reported to be significantly high in clinical isolates of *M. tuberculosis* with *in silico* modelling predicting that a Pro359Ser substitution in Mce1A may have a diminishing effect on the stability of the protein and its biological function [32].

In addition to non-synonymous SLVs, our analysis also identified the differential carriage of genes that have been associated with the pathogenesis of *M. tuberculosis* in the Rangipo isolates with respect to reference strain H37Rv. Three virulence-related genes, which have orthologues in strain CDC1551, but which are absent in H37Rv, were identified in all nine Rangipo isolates whole-genome-sequenced in this study and were confirmed by PCR amplification and Sanger sequencing. These genes consisted of orthologues of transcriptional regulator protein EmbR2 (MT3428), molybdopterin cofactor biosynthesis protein A (MT3427) and molybdopterin cofactor biosynthesis protein B (MT3426) from strain CDC1551.

Molle et al. reported the presence in strain CDC1551 of EmbR2 (MT3428), an 1146 amino acid orthologue of the transcriptional regulator EmbR from H37Rv [33]. EmbR when phosphorylated by its cognate serine/threonine protein kinase, PknH, activates the transcription of the arabinosyltransferase genes, *embCAB*, which participate in the biosynthesis of arabinogalactans, key constituents of the mycobacterial cell wall. EmbR2 interacts with, but is not phosphorylated by, PknH [33]. Instead, EmbR2 inhibits the autokinase activity of PknH and the subsequent phosphoryl transfer to EmbR. A *pknH* mutant of *M. tuberculosis* has previously been shown to survive and replicate to a higher bacillary load in mouse lungs and spleens than its parental strain [34]. Therefore, EmbR2 through its control of the activity of the PknH/EmbR pair is believed to participate in the physiology and virulence of *M. tuberculosis* [33]. In our analysis, EmbR2 was present in the Rangipo isolates sequenced and in reference strains CDC1551 and Haarlem, but not in H37Rv or Erdman (Figure 1).

We also identified orthologues of the molybdopterin cofactor (MoCo) biosynthesis proteins MoaA (MT3427) and MoaB (MT3426) in the Rangipo isolates and the reference genome CDC1551 but not in H37Rv. MoaA participates in the conversion of guanosine triphosphate to cyclic pyranopterin monophosphate [35], while MoaB is involved in the adenylation of molybdopterin [36], key steps in MoCo biosynthesis. MoCo is an essential cofactor for redox reaction enzymes such as the *narGHI*-encoded nitrate reductase which functions in the adaptation of *M. tuberculosis* to hypoxic conditions [36], and is needed for the persistence of *M. tuberculosis* in the lungs of guinea pigs [37]. In addition, we identified an extended-length DNA helicase (MT2082, 1606 amino acids) and an additional adenylate cyclase (MT1360) in the Rangipo strain; however, their specific roles in virulence have yet to be elucidated.

Azhikina et al. previously reported the presence of orthologues of the five CDC1551 genes MT2082, MT3426, MT3427, MT3428 and MT1360 in clinical isolates of *M. tuberculosis* from Russia [38]. In addition, they also reported the presence of a truncated *plcD* gene which corresponded to the first 843 nucleotides of the 1545 base pair MT1799 gene of CDC1551. Interestingly, the Rangipo strain sequenced also exhibited a truncated *plcD* orthologue (905 nucleotides) with respect to MT1799.

Mycobacterium tuberculosis strain CDC1551 was first described in 1998 following a large TB outbreak in Tennessee and Kentucky between 1994 and 1996 [39]. The strain was notable for its high level of transmissibility from three patients to casual and close contacts in a community with a normally low risk for TB. Of 429 contacts, 311 (72.5%) produced positive tuberculin skin tests [39]. CDC1551 has been found to grow at a similar rate (doubling time of 25 h) as H37Rv (doubling time of 28 h) in the lungs of aerosol-infected mice during the first 14 days of infection [40]. But between day 14 and day 21, CDC1551 grew more slowly (doubling time of 105 h) than the H37Rv isolate used (doubling time of 36 h) [40]. Investigations in rabbits found that CDC1551 produced smaller granulomas containing fewer bacilli, which is indicative of lower virulence, as compared to H37Rv [41]. In addition, mice challenged with CDC1551 exhibited higher mRNA levels of inflammatory mediators TNF-α, IL-6, IL-10, IL-12 and IFN-γ, and a longer mean survival time of $>$250 days versus 185 days compared to H37Rv [40]. This has led to the suggestion that the original mini-epidemic caused by CDC1551 was due to higher transmissibility associated with the strain rather than increased in-host pathogenesis [41].

Conclusions

In summary, the Rangipo strain has been associated with high levels of transmissibility during community TB outbreaks. Our study identified key differences between the Rangipo strain and the standard *M. tuberculosis* reference strain, H37Rv that is commonly used for the assembly of whole-genome sequence reads. In addition to the presence of non-synonymous SLVs in genes linked to mycobacterial virulence, the Rangipo strain also harbours a number of virulence genes that are absent in H37Rv but present in TB outbreak strain, CDC1551. In vivo studies are needed to specifically correlate individual loci in the

Rangipo strain to its reported heightened transmissibility, and to understand the host response to infection by this strain.

Disclosure statement

The authors report no conflicts of interest.

Funding

RFOT is a recipient of a School of Medicine Research Development Grant, University of Tasmania and Royal Hobart Hospital Research Foundation grant (17-104). SSG is a recipient of a School of Medicine/Faculty of Health PhD Scholarship. Micheál Mac Aogáin was funded by Irish Research Council (EPSPD/2015/32).

ORCID

Micheál Mac Aogáin http://orcid.org/0000-0002-1726-7700
Ronan F. O'Toole http://orcid.org/0000-0002-4579-4479

References

[1] World Health Organization. Global tuberculosis report. 2015; Switzerland: World Health Organization.
[2] Institute of Environmental Science and Research Ltd (ESR). Tuberculosis in New Zealand: annual report 2014. 2015; New Zealand: Institute of Environmental Science and Research Ltd (ESR).
[3] Bissielo A, Lim E, Heffernan H. Tuberculosis in New Zealand: annual report 2010. New Zealand: Institute of Environmental Science and Research Ltd.
[4] De Zoysa R, Shoemack R, Vaughan R, et al. A prolonged outbreak of tuberculosis in the North Island. New Zealand Public Health Report. 2001.
[5] McElnay C, Thornley C, Armstrong R. A community and workplace outbreak of tuberculosis in Hawke's Bay in 2002. N Z Med J. 2004;117:U1019.
[6] Sexton K, Perera S, Pandey S. Five years of molecular typing of Mycobacterium tuberculosis isolates In New Zealand, 2003 to 2007. New Zealand: Institute of Environmental Science and Research Limited, 2007.
[7] Baussano I, Williams BG, Nunn P, et al. Tuberculosis incidence in prisons: a systematic review. PLoS Med. 2010;7:e1000381.
[8] Vinkeles Melchers NV, van Elsland SL, Lange JM, et al. State of affairs of tuberculosis in prison facilities: a systematic review of screening practices and recommendations for best TB control. PLoS One. 2013;8:e53644.
[9] Sacchi FP, Praca RM, Tatara MB, et al. Prisons as reservoir for community transmission of tuberculosis, Brazil. Emerging Infect Dis. 2015;21:452–455.
[10] Yen S, Bower JE, Freeman JT, et al. Phylogenetic lineages of tuberculosis isolates in New Zealand and their association with patient demographics. Int J Tuberc Lung Dis. 2013;17:892–897.
[11] Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998;393:537–544.
[12] Kearse M, Moir R, Wilson A, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–1649.
[13] Mac Aogain M, Gautam SS, Bower JE, et al. Draft Genome Sequence of a New Zealand Rangipo Strain of *Mycobacterium tuberculosis*. Genome Announc. 2016;4:e00657-16.
[14] Sekizuka T, Yamashita A, Murase Y, et al. TGS-TB: total genotyping solution for *Mycobacterium tuberculosis* using short-read whole-genome sequencing. PLoS One. 2015;10:e0142951.
[15] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215:403–410.
[16] Gagneux S, DeRiemer K, Van T, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA. 2006;103:2869–2873.
[17] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52:696–704.
[18] Colangeli R, Arcus VL, Cursons RT, et al. Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. PLoS One 2014;9:e91024.
[19] Walker TM, Ip CLC, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13:137–146.
[20] Nikolayevskyy V, Kranzer K, Niemann S, et al. Whole genome sequencing of *Mycobacterium tuberculosis* for detection of recent transmission and tracing outbreaks: a systematic review. Tuberculosis. 2016;98:77–85.
[21] Hatherell H-A, Colijn C, Stagg HR, et al. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med. 2016;14:21.
[22] Pouseele H, Supply P. Accurate whole-genome sequencing-based epidemiological surveillance of *Mycobacterium tuberculosis*. In: Sails A, Tang Y-W, editors. Methods in Microbiology. 42. USA: Academic Press; 2015.
[23] Galperin MY, Makarova KS, Wolf YI, et al. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43:D261–D2D9.
[24] Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–D361.
[25] Forrellad MA, Klepp LI, Gioffré A, et al. Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence. 2013;4:3–66.
[26] Fleischmann RD, Alland D, Eisen JA, et al. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol. 2002;184:5479–5490.
[27] Munoz-Elias EJ, McKinney JD. *Mycobacterium tuberculosis* isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence. Nat Med. 2005;11:638–644.

[28] Raynaud C, Guilhot C, Rauzier J, et al. Phospholipases C are involved in the virulence of *Mycobacterium tuberculosis*. Mol Microbiol. 2002;45:203–217.

[29] Walters SB, Dubnau E, Kolesnikova I, et al. The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. Mol Microbiol. 2006;60:312–330.

[30] Tsenova L, Ellison E, Harbacheuski R, 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: 98–106.

[31] Schreuder LJ, Carroll P, Muwanguzi-Karugaba J, et al. *Mycobacterium tuberculosis* H37Rv has a single nucleotide polymorphism in PhoR which affects cell wall hydrophobicity and gene expression. Microbiology. 2015;161:765–773.

[32] Pasricha R, Chandolia A, Ponnan P, et al. Single nucleotide polymorphism in the genes of mce1 and mce4 operons of *Mycobacterium tuberculosis*: analysis of clinical isolates and standard reference strains. BMC Microbiol. 2011;11:41.

[33] Molle V, Reynolds Robert C, Alderwick LJ, et al. EmbR2, a structural homologue of EmbR, inhibits the *Mycobacterium tuberculosis* kinase/substrate pair PknH/EmbR. Biochem J. 2008;410:309–317.

[34] Papavinasasundaram KG, Chan B, Chung J-H, et al. Deletion of the *Mycobacterium tuberculosis* pknH gene confers a higher bacillary load during the chronic phase of infection in BALB/c Mice. J Bacteriol. 2005;187:5751–5760.

[35] Williams MJ, Kana BD, Mizrahi V. Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in *Mycobacterium tuberculosis*. J Bacteriol. 2011;193:98–106.

[36] Williams M, Mizrahi V, Kana BD. Molybdenum cofactor: a key component of *Mycobacterium tuberculosis* pathogenesis? Crit Rev Microbiol. 2014;40:18–29.

[37] Williams MJ, Shanley CA, Zilavy A, et al. bis-Molybdopterin guanine dinucleotide is required for persistence of *Mycobacterium tuberculosis* in guinea pigs. Infect Immun. 2015;83:544–550.

[38] Azhikina T, Gvozdevsky N, Botvinnik A, et al. A genome-wide sequence-independent comparative analysis of insertion-deletion polymorphisms in multiple *Mycobacterium tuberculosis* strains. Res Microbiol. 2006;157:282–290.

[39] Valway SE, Sanchez MPC, Shinnick TF, et al. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. N Engl J Med. 1998;338: 633–639.

[40] Manca C, Tsenova L, Barry CE, 3rd, et al. *Mycobacterium tuberculosis* CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. J Immunol. 1999;162:6740–6746.

[41] Bishai WR, Dannenberg AM, Jr., Parrish N, et al. Virulence of *Mycobacterium tuberculosis* CDC1551 and H37Rv in rabbits evaluated by Lurie's pulmonary tubercle count method. Infect Immun. 1999;67:4931–4934.