Tuberculosis xxx (2015) 1–7

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

MOLECULAR ASPECTS

Whole genome sequence analysis of *Mycobacterium suricattae*

Anzaan Dippenaar [a, 1], Sven David Charles Parsons [a, 1], Samantha Leigh Sampson [a], Ruben Gerhard van der Merwe [a], Julian Ashley Drewe [b], Abdallah Musa Abdallah [c], Kabengele Keith Siame [a], Nicolaas Claudius Gey van Pittius [a], Paul David van Helden [a], Arnab Pain [c, 2], Robin Mark Warren [a, *, 2]

[a] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research/SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
[b] Royal Veterinary College, London, UK
[c] Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form
29 September 2015
Accepted 1 October 2015

Keywords:
Meerkat
Mycobacterium suricattae
Tuberculosis
High-throughput nucleotide sequencing

SUMMARY

Tuberculosis occurs in various mammalian hosts and is caused by a range of different lineages of the *Mycobacterium tuberculosis* complex (MTBC). A recently described member, *Mycobacterium suricattae*, causes tuberculosis in meerkats (*Suricata suricatta*) in Southern Africa and preliminary genetic analysis showed this organism to be closely related to an MTBC pathogen of rock hyraxes (*Procavia capensis*), the dassie bacillus. Here we make use of whole genome sequencing to describe the evolution of the genome of *M. suricattae*, including known and novel regions of difference, SNPs and IS*6110* insertion sites. We used genome-wide phylogenetic analysis to show that *M. suricattae* clusters with the chimpanzee bacillus, previously isolated from a chimpanzee (*Pan troglodytes*) in West Africa. We propose an evolutionary scenario for the *Mycobacterium africanum* lineage 6 complex, showing the evolutionary relationship of *M. africanum* and chimpanzee bacillus, and the closely related members *M. suricattae*, dassie bacillus and *Mycobacterium mungi*.

1. Background

The *Mycobacterium tuberculosis* complex (MTBC) comprises genotypically and phenotypically distinct lineages which have evolved through clonal expansion from a common progenitor [1]. Much of this evolution has occurred through genomic loss and such genetic regions of difference (RDs) can be used to distinguish between these lineages [2]. *M. tuberculosis sensu stricto* (lineages 1–4) is exquisitely adapted to the human host and despite infecting one third of the world's human population it has not been known to become permanently established in free-living animals [3,4]. In contrast, a distinct MTBC clade, characterized by the loss of RD9, is unique in having given rise to lineages associated with both humans (*Mycobacterium africanum* lineage 5 and 6) and a wide variety of animal hosts [2]. A group of strains which display remarkable phenotypic variation are those related to lineage 6, i.e. *M. africanum* West African 2 (WA2), a lineage which circulates within human populations with apparent attenuated virulence [5]. Animal-associated strains related to WA2, based on the loss of RD7, RD8, RD9 and RD10, include the chimpanzee bacillus [1], the dassie bacillus [6], *Mycobacterium mungi* [7], and *Mycobacterium suricattae* [8], which have been isolated from a chimpanzee (*Pan troglodytes*), rock hyraxes (dassies, *Procavia capensis*), banded mongooses (*Mungos mungo*), and meerkats (*Suricata suricatta*), respectively. However, to date, the evolutionary relationship between these strains, and the genetic basis for their apparent host tropism, has not been conclusively resolved [1,2]. Here we provide analysis of the whole genome sequence of *M. suricattae*, confirm its genetic relationship to *M. africanum* and the chimpanzee bacillus and identify genetic features which may underlie host specificity.

* Corresponding author. Tel.: +27 219389073; fax: +27 219389863.
E-mail addresses: adippenaar@sun.acza (A. Dippenaar), sparsons@sun.ac.za (S.D.C. Parsons), ssampson@sun.ac.za (S.L. Sampson), rvdm@sun.ac.za (R.G. van der Merwe), jdrewe@rvc.ac.uk (J.A. Drewe), abdallah.abdallah@kaust.edu.sa (A.M. Abdallah), kksiame@sun.ac.za (K.K. Siame), ngvp@sun.ac.za (N.C. Gey van Pittius), pvh@sun.ac.za (P.D. van Helden), arnab.pain@kaust.edu.sa (A. Pain), rw1@sun.ac.za (R.M. Warren).
[1] Authors contributed equally.
[2] Authors contributed equally.

http://dx.doi.org/10.1016/j.tube.2015.10.001
1472-9792/

2. Material and methods

2.1. Whole genome sequencing (WGS)

DNA was obtained from *M. suricattae* cultures as previously described [8] and DNA from one isolate was subjected to whole genome sequencing. The Illumina HiSeq 2000 (Illumina, California, USA) platform was used for WGS with 3 different libraries, namely PCR-free paired-end, normal paired-end, and mate-pair. The DNA libraries were sequenced on the Illumina HiSeq platform for Next-Generation sequencing (NGS) according to the manufacturer's specifications. One library was sequenced on the Illumina MiSeq platform. The whole genome sequencing approach is described in detail in Supplemental Data 1. Sequence data has been deposited at the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena/data/view/PRJEB9954), which is hosted by the EBI, under accession number: PRJEB9954.

2.2. WGS data analyses

The raw sequencing data generated for the purpose of this study and 32 selected mycobacterial genomes published previously or available in public databases (Supplemental Data 1, Table S1) were analysed with widely used open source software. A pipeline was created to analyse routinely generated Illumina WGS data in order to identify genomic variation with high confidence. Briefly, low quality bases and reads were removed before aligning the reads to the *M. tuberculosis* H37Rv (GenBank NC000962.3) reference genome with three different alignment tools, namely Novoalign (Novocraft), Burrows-Wheeler Aligner (BWA) [9], and SMALT [10]. The Genome Analysis Tool Kit (GATK) was used to detect single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) [11]. Variants identified in all three of the alignments were annotated and used in subsequent analysis. The BEDTools software package was used to identify regions with zero depth of coverage; these regions were considered to be potential genomic deletions in the strain of interest, and are referred to as regions of difference (RDs) [12]. Data analysis is described in Supplemental Data 1.

2.3. PCR confirmation of genomic deletions

PCRs were done using the HotStar-Taq system (Qiagen, Venlo, Limburg, Netherlands) in 25 μl reaction volumes. The PCR mastermix consisted of 1× Q-solution, 1× reaction buffer, 2 mM $MgCl_2$, 200 μM of each deoxyribonucleotide triphospate (dNTP), 50 μM forward primer, 50 μM internal reverse primer, 50 μM reverse primer, 1.25 U Hotstar Taq polymerase, 1 μl of the DNA template (50 ng/μl). Reactions were made up to a total volume of 25 μl with nuclease free H_2O. PCRs were carried out in a thermal cycler (GeneAmp PCR System 2400, Applied Biosystems, Foster City, CA, USA) under the following thermo-cycling conditions: an initial denaturing step at 95 °C for 15 min, followed by 35 cycles of: (a) a denaturing step at 94 °C for 1 min, (b) an annealing step at the melting temperature of the primers included in the specific reaction for 1 min, (c) an extension step at 72 °C for 1 min (depending on the expected size of the PCR products, allowing 1 min per 1000 bp), and a final extension step at 72 °C for 15 min. All PCR experiments included a negative control containing no DNA template and a positive control (*M. tuberculosis* H37Rv genomic DNA). Primer sequences used to confirm newly identified genomic deletions are included in Supplemental Data 1, Table S2.

2.4. Phylogenomic analysis

Concatenated sequences containing high-confidence variable sites (coding and non-coding SNPs) with respect to the *M. tuberculosis* H37Rv reference genome, were written to a multi-fasta file (one entry for each isolate included) and were used as input in Modelgenerator to determine the optimal substitution model that fits the data structure [13,14]. The general time reversal (GTR) model scored the lowest in the hierarchical likelihood ratio test; Bayesian information criterion (BIC), and thus described the substitution pattern occurring in the dataset most accurately. The GTR model of substitution was subsequently applied to construct a maximum likelihood phylogeny of the isolates included in this analysis with MEGA6 and Randomized Axelerated Maximum Likelihood (RaxML) with 1000 bootstrap pseudo-replicates [15–17]. Positions containing gaps or missing data were not considered for the analysis. The phylogenetic trees produced by the different algorithms were compared to determine congruence in tree topology [18].

2.5. De novo assembly

The draft genome of *M. suricattae* was generated by using the 100 bp paired-end, 35–151 bp mate-pair and 35–251 bp PCR-free Illumina HiSeq 2000 reads assembled with the *de novo* assembly software package for short reads, Velvet, version 1.2.07 [19]. A script published by the Victorian Bioinformatics Consortium (velvetk.pl) was used to determine the optimal kmer length to be used. Velvet was compiled to include higher maximum kmer lengths than the standard of 31. Velvet was run with recommended parameters according to the documentation and a kmer length of 97 was used. The resulting assembly was used as input in Platanus [20] to close the gaps between the contigs making use of the above mentioned sequencing reads as well as 35–251 bp paired-end reads generated on the Illumina MiSeq platform. The final assembly was assessed and corrected with REAPR [21]. The *M. tuberculosis* H37Rv genome annotation was transferred to the final assembly using RATT [22].

2.6. Genome comparison

Alignment of the *M. suricattae* draft genome and two *Mycobacterium* spp. reference genomes (*M. tuberculosis* H37Rv, *M. africanum* lineage 6 GM041182, accession number: GCA_000253355.1) was done using progressiveMauve to observe large regions of difference [23]. Variants identified with respect to *M. tuberculosis* H37Rv in the *Mycobacterium* isolates included in this study were compared in an attempt to identify unique variant profiles of different species. Gene ontology (GO) annotation and GO enrichment was done with GOEAST on unique non-synonymous SNPs for *M. suricattae* and chimpanzee bacillus and a comparison of enriched GO terms of unique non-synonymous variants in *M. suricattae* and chimpanzee bacillus was done with Multi-GOEAST [24].

3. Results

3.1. SNPs

It has previously been shown that *M. suricattae* harbours several genetic markers which are shared with the dassie bacillus [6,8]. These markers were confirmed by the WGS data analysed in our study and includes a single nucleotide deletion (SND) in *Rv0911* (*Rv0911*^{389}), a SNP in *Rv1510* (G1129A), and a SNP in 16S rDNA (T214G). Phylogenomic SNP analysis showed *M. suricattae* to harbour 718 unique SNPs with respect to *M. tuberculosis* H37Rv, of

Please cite this article in press as: Dippenaar A, et al., Whole genome sequence analysis of *Mycobacterium suricattae*, Tuberculosis (2015), http://dx.doi.org/10.1016/j.tube.2015.10.001