GENOME ANNOUNCEMENT

Draft genome sequences of two super-extensively drug-resistant isolates of *Mycobacterium tuberculosis* from China

Nan Lin[1,2], Zhangyi Liu[3], Jie Zhou[4], Shihua Wang[1] & Joy Fleming[2]

[1]College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China; [2]Key Laboratory of Non-coding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China; [3]BGI-Shenzhen, Shenzhen, China; and [4]The 4th Peoples' Hospital, Foshan City, Guangdong Province, China

Correspondence: Joy Fleming, Key Laboratory of Non-coding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. Tel.: +86 1064888464; fax: +86 1064871293; e-mail: joyfleming@moon.ibp.ac.cn
and
Shihua Wang, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China. Tel./fax: +86 59187984471; e-mail: wshyyl@sina.com

Received 18 July 2013; revised 15 August 2013; accepted 19 August 2013. Final version published online 9 September 2013.

DOI: 10.1111/1574-6968.12238

Editor: Roger Buxton

Keywords
tuberculosis; whole genome shotgun sequencing; multidrug resistance; single nucleotide polymorphisms.

Abstract

The prevalence of drug-resistance in *Mycobacterium tuberculosis* is already having a negative impact on the control of tuberculosis. We report the draft genome sequences of two super-extensively drug-resistant *M. tuberculosis* isolates from China, FJ05194 (lineage 2) and GuangZ0019 (lineage 4), and compare them with the H37Rv reference strain to identify possible sources of genetic variation associated with their extensive drug resistance. Our results suggest that their extensive drug resistance probably results from the stepwise accumulation of resistances to individual drugs.

Introduction

Forty per cent of the world's notified cases of tuberculosis (TB) in 2010 arose in India and China (WHO, 2011). In China, 5.7% of new TB cases are multidrug resistant (MDR), and 8% of MDR cases are extensively drug-resistant (XDR) (Zhao *et al.*, 2012). As drug sensitivity testing of *Mycobacterium tuberculosis* strains FJ05194 and GuangZ0019, isolated from patients in Fujian and Guangdong provinces, respectively, indicated that they are both resistant to four first-line (rifampicin, isoniazid, ethambutol and streptomycin) and four second-line TB drugs (capreomycin, kanamycin, ofloxacin, and ethionamide) (WHO, 2008), we performed genome sequencing to examine the genetic basis of their drug resistance.

Genome sequencing was performed on an Illumina HiSeq 2000 sequencer (Illumina Inc.) by generating paired-end libraries. Paired-end reads were *de novo* assembled using SOAPdenovo v1.05 (Rohde *et al.*, 2011) and gaps were filled manually. Gene prediction was performed using Glimmer v3.02 (Delcher *et al.*, 2007), and rRNA and tRNA were identified using RNAmmer (Lagesen *et al.*, 2007) and tRNAscan-SE 1.3.1 (Schattner *et al.*, 2005), respectively. Sequencing-related statistics are presented in Table 1.

We performed a comparative genomic analysis of these two super-XDR (S-XDR) isolates, using *M. tuberculosis* H37Rv as a reference (GenBank accession no. NC_000962). Raw single nucleotide polymorphisms (SNPs) were identified by aligning clean reads to the H37Rv reference genome with SOAP2 (Li *et al.*, 2009). Polymorphic bases that were present in $< 70\%$ of the reads or had a quality score < 20 were filtered out and SNPs called in repetitive regions of the H37Rv genome,

FEMS MICROBIOLOGY LETTERS

Table 1. Sequencing statistics for S-XDR isolates from Fujian (FJ05194) and Guangdong (GuangZ0019) provinces, China, and their drug-resistance-associated genes and intergenic regions

	FJ05194	GuangZ0019
Sequencing statistics		
Coverage, based on H37Rv Ref (%)	99.57	99.96
Sequencing depth	110	117
Length of raw reads	528 000 000	528 000 120
Genome length (nt)	4363 305	4373 311
Scaffolds	112	97
G + C content (mol%)	65.48	65.55
Annotation*		
Protein-coding genes (CDSs)	4165	4198
tRNAs	43	43
5S rRNA genes	1	1
16S rRNA	2	1
23S rRNA	2	1
SNPs		
Total	1287	795
Coding regions	1140	709
nsSNPs	725	437
sSNPs	415	272
Intergenic regions	147	86
Lineage	2	4
Previously reported drug resistance-associated genes (TBDream database)†		
Isoniazid		
Rv1908c (*katG*)	G421S	S315N
Rv1909c (*furA*)		T40A
Rv3566c (*nat*)		G207R
Rv1592c	E60D	–
Rv2427c-Rv2428 (*proA–ahpC*)	C–52T‡	–
Rifampicin		
Rv0667 (*rpoB*)	H526Y	S531L
Streptomycin		
Rv0682 (*rpsL*)	K43R	–
Rv3919c (*gidB*)		K163X
Ethambutol		
Rv3794 (*embA*)		A366E
Rv3795 (*embB*)	E504D	–
Rv3795 (*embB*)	D1024N	–
Rv3124		C175G
Rv3793-Rv3794 (*embC–embA*)	–	C–11A‡
Ofloxacin		
Rv0005 (*gyrB*)	N538T	–
Rv0006 (*gyrA*)	D94G§	D94G
Ethionamide		
Rv3854c (*ethA*)	A89E	H166P

*Gene prediction was performed using Glimmer v3.02 (Delcher *et al.*, 2007). rRNA and tRNA were identified using RNAmmer (Lagesen *et al.*, 2007) and tRNAscan-SE 1.3.1 (Schattner *et al.*, 2005), respectively.
†TBDream database (Sandgren *et al.*, 2009).
‡IGR-SNPs, position relative to the start codon of the flanking gene.
§Heterozygous call in Illumina sequencing results, but confirmed as an nsSNP by Solexa sequencing.

defined as exact repetitive sequences of ≥ 25 bp in length, identified using either BLAST, RepeatMasker or Trf (Benson, 1999; Tarailo-Graovac & Chen, 2009) were excluded. A total of 1287 and 795 SNPs were identified in FJ05194 and GuangZ0019, respectively (Table 1; Supporting Information, Table S1). The lineages of the two isolates were determined by performing a phylogenetic analysis using 22 additional drug-sensitive strains (H37Rv and 21 publicly available genomes) which represent global diversity (Comas *et al.*, 2010); FJ05194 belongs to lineage 2 and GuangZ0019 to lineage 4 (Fig. S1A). Based on the phylogenetic tree we removed SNPs present in the drug-sensitive lineage 2 and lineage 4 isolates from the FJ05194 and GuangZ0019 SNP datasets, respectively (i.e. we retained only those that were unique to FJ05194 or GuangZ0019), to better identify genetic variation related to drug resistance in these two isolates.

The presence of non-synonymous SNPs (nsSNPs) in certain genes is strongly correlated with drug resistance; for example, nsSNPs at the 516, 526 and 531 sites of *rpoB* are correlated with rifampicin resistance (Heep *et al.*, 2001). We examined the distribution of the unique nsSNPs/intergenic region SNPs (IGR-SNPs) from each isolate (FJ05194: 221; GuangZ0019: 180) in the drug resistance genes included in the TBDream database (Sandgren *et al.*, 2009); 17 nsSNPs and two IGR-SNPs previously associated with drug resistance were found in 13 of these genes and two IGRs, respectively (Table 1). We amplified the 13 nsSNP-containing drug resistance genes by PCR and resequenced them by Sanger sequencing on an ABI3730 sequencer to verify the SNPs. Results were fully consistent with Illumina sequencing results, confirming the quality of the original genome sequences.

Our results suggest that the extensive drug resistance of these isolates has probably arisen due to the accumulation of resistances to individual drugs. Resistance to six of the eight drugs tested in this study could be attributed to mutations in genes previously reported to be associated with these resistances (Table 1). For example, the H526Y and S531L mutations in *rpoB*, associated with resistance to rifampicin (Heep *et al.*, 2001), and the A89E and H166P mutations in *ethA*, associated with ethionamide resistance (DeBarber *et al.*, 2000), were present in these isolates. However, we did not find mutations in genes classically associated with resistance to capreomycin and kanamycin (*rrs*, *tlyA* and the *eis* promoter; Maus *et al.*, 2005a, b; Zaunbrecher *et al.*, 2009), suggesting that there are likely to be novel mutations in these isolates associated with these resistances. Of interest, several genes for efflux pumps thought to be involved in the transport of drugs across the membrane carry nsSNPs unique to isolate FJ05194 [Rv0933 (*pstB*) and Rv2333c (*stp*)] or

GuangZ0019 (Rv2687c) (Raman & Chandra, 2008). Further experimental investigation will be required to identify the key SNPs which confer drug resistance and to determine whether these efflux pumps are involved in resistance mechanisms. The genomic locations of all the drug-resistance-associated genes/IGRs from the TBDream database detected in these two isolates are shown in Fig. S1B.

As the presence of drug-resistance mutations is often associated with a fitness cost (Muller *et al.*, 2013), it is highly likely that there are compensatory mutations present in these two S-XDR isolates. Indeed, the *proA-ahpC* IGR in FJ05194 carries an SNP that has previously been reported as a compensatory mutation (Gagneux *et al.*, 2006). Previously reported compensatory mutations in *rpoA* and *rpoC*, however, were not observed in these isolates, possibly because the S531L and H526Y mutations they carry have previously been reported to have a low fitness cost (Muller *et al.*, 2013). Further analysis of these genome sequences and comparison with the genomes of other well-characterized XDR isolates as they become available should lead to the identification of other compensatory mutations and a better understanding of the evolution and characteristics of extensive drug resistance in *M. tuberculosis*, leading potentially to more effective control measures.

Nucleotide sequence accession numbers: Sequencing reads are available at the NCBI Sequence Read Archive under the accession code SRA068167. The whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accessions ANBL00000000 (FJ05194), and ANFI00000000 (GuangZ0019).

Acknowledgement

This work was supported by the National Basic Research Program of China (Grant No.: 2012CB518700) and the Chinese Academy of Sciences (Grant No.: KSZD-EW-Z-006).

Authors contribution

N.L., Z.L. and J.Z. all contributed equally to this work.

References

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Comas I, Chakravartti J, Small PM *et al.* (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* **42**: 498–503.

DeBarber AE, Mdluli K, Bosman M, Bekker LG & Barry CE 3rd (2000) Ethionamide activation and sensitivity in multidrug-resistant *Mycobacterium tuberculosis*. *P Natl Acad Sci USA* **97**: 9677–9682.

Delcher AL, Bratke KA, Powers EC & Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**: 673–679.

Gagneux S, Burgos MV, DeRiemer K *et al.* (2006) Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*. *PLoS Pathog* **2**: e61.

Heep M, Brandstatter B, Rieger U, Lehn N, Richter E, Rusch-Gerdes S & Niemann S (2001) Frequency of *rpoB* mutations inside and outside the cluster I region in rifampin-resistant clinical *Mycobacterium tuberculosis* isolates. *J Clin Microbiol* **39**: 107–110.

Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T & Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**: 3100–3108.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K & Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966–1967.

Maus CE, Plikaytis BB & Shinnick TM (2005a) Mutation of *tlyA* confers capreomycin resistance in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **49**: 571–577.

Maus CE, Plikaytis BB & Shinnick TM (2005b) Molecular analysis of cross-resistance to capreomycin, kanamycin, amikacin, and viomycin in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **49**: 3192–3197.

Muller B, Borrell S, Rose G & Gagneux S (2013) The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. *Trends Genet* **29**: 160–169.

Raman K & Chandra N (2008) *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. *BMC Microbiol* **8**: 234.

Rohde H, Qin J, Cui Y *et al.* (2011) Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N Engl J Med* **365**: 718–724.

Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM & Murray MB (2009) Tuberculosis drug resistance mutation database. *PLoS Med* **6**: e2.

Schattner P, Brooks AN & Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**: W686–W689.

Tarailo-Graovac M & Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4: Unit 4 10.

WHO (2008) *Policy Guidance on TB Drug Susceptibility Testing (DST) of Second-Line Drugs*. WHO/HTM/TB/2008392. WHO, Geneva.

WHO (2011) *Global Tuberculosis Control 2011*. WHO/HTM/TB/2011.16. WHO, Geneva.

Zaunbrecher MA, Sikes RD Jr, Metchock B, Shinnick TM & Posey JE (2009) Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in *Mycobacterium tuberculosis*. *P Natl Acad Sci USA* **106**: 20004–20009.

Zhao Y, Xu S, Wang L *et al.* (2012) National survey of drug-resistant tuberculosis in China. *N Engl J Med* **366**: 2161–2170.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. SNPs identified in the FJ05194 and GuangZ0019 isolates.

Fig. S1. (A) Neighbour-joining phylogeny of 24 human *Mycobacterium tuberculosis* complex isolates based on 6586 high-quality SNP positions. The tree is rooted with *Mycobacterium canettii*. Branches are coloured according to the six main phylogeographical lineages of MTBC (Comas *et al.*, 2010): pink, L1 lineage (the Philippines and the Indian Ocean Rim); blue, L2 lineage (East Asia); purple, L3 lineage (India, East Africa); red, L4 lineage (Europe, America); brown, L5, *Mycobacterium africanum* (West Africa 1); green, L6, *M. africanum* (West Africa 2). (B) Distribution of SNPs and drug resistance genes/IGRs in the two S-XDR genomes. The outer circle shows SNPs unique to FJ05197 and the inner circle shows those unique to GuangZ0019. Non-synonymous SNPs are represented by red bars, synonymous SNPs by blue bars and intergenic SNPs by yellow bars. The genes and intergenic regions labelled have previously been reported to be associated with drug resistance (Sandgren *et al.*, 2009).