Title: Whole genome sequencing provides additional insights into recurrent tuberculosis classified as endogenous reactivation by IS*6110* DNA fingerprinting

Abstract

Recurrent tuberculosis (TB) after successful TB treatment occurs due to endogenous reactivation (relapse) or exogenous reinfection. We revisited the conclusions of relapse in a high TB incidence setting that were drawn on the basis of IS*6110* restriction fragment length polymorphism (RFLP) analysis in a large retrospective cohort study in suburban Cape Town, South Africa. Using whole genome sequencing (WGS), we undertook pair-wise genome comparison of *Mycobacterium tuberculosis* strains cultured from diagnostic sputum samples collected at the index and recurrent TB episode for 25 recurrent TB cases who had been classified as relapse based on identical DNA fingerprint patterns in the earlier study. We found that paired strain genome sequences were identical or showed minimal variant differences in 22 of 25 recurrent TB cases, consistent with relapse. One showed 20 variant differences, suggestive of exogenous reinfection. Two of the 25 had mixed infections, each with the index episode strain detected as the dominant strain at recurrence in one of these patients, the minority strain harboured drug-resistance conferring mutations (*rpoB*, *katG*). In conclusion, our study highlights the additional value of WGS for investigating recurrent TB in settings with high infection pressure and closely related circulating strains, where the extent of re- and mixed infection may be underestimated.

Keywords: recurrent tuberculosis; whole genome sequencing; tuberculosis; *Mycobacterium tuberculosis*

1. Introduction

Recurrent tuberculosis (TB), defined as active TB among individuals who were apparently successfully treated for a previous TB episode, contributes significantly to the TB burden in high incidence settings[1-3]. Recurrent TB occurs either due to endogenous reactivation (relapse) or exogenous reinfection with *Mycobacterium tuberculosis* [4]. Classic molecular typing methods such as mycobacterial interspersed repetitive units variable number tandem repeat (MIRU-VNTR) typing and IS*6110* restriction fragment length polymorphism (RFLP) have been used to distinguish relapse from exogenous reinfection[5-10]. These studies defined endogenous reactivation as paired patient isolates from successive disease episodes showing identical IS*6110* fingerprints or MIRU-VNTR patterns. In contrast, exogenous reinfection was inferred if the paired patient isolates had distinct IS*6110* fingerprints or variation in more than two MIRU-VNTR alleles[8,11-13].

The increased resolution of whole genome sequencing (WGS) previously allowed the elucidation of substantial genomic diversity among *M. tuberculosis* isolates harbouring near identical MIRU-VNTR patterns and identical IS*6110* RFLP fingerprints[14,15]. The base-pair level resolution of WGS by employing next generation sequencing technologies allows for the identification of micro-evolutionary events other than IS*6110* transposon or MIRU-VNTR differences[16,17] and offers a valid differentiation between relapse and exogenous reinfection as the cause of recurrent disease, even in high incidence settings where the genomic diversity of circulating strains may be limited[14].

Traditional molecular typing tools thus have limited resolution to distinguish highly similar strains circulating in a community. Accordingly, recurrent TB due to exogenous reinfection with a closely related strain to the strain present during the first disease episode may be misclassified as relapse when traditional molecular typing tools are used. These misclassifications, in turn, may lead to the underestimation of the contribution of transmission to the TB burden in high TB incidence settings. In addition, mixed infections with potentially drug resistant strains may go undetected at first, leading to further treatment complications[18].

In this study, we revisited the conclusion of relapse as the mechanism of recurrent disease in a subset of TB cases from a high TB incidence setting in Cape Town, South Africa[6]. We anticipated that WGS could offer additional insights into whether these TB cases were correctly classified as relapse by IS*6110* RFLP. We also investigated the possibility of mixed infections and drug-resistance at recurrence.

2. Material and Methods

We re-investigated 25 recurrent TB cases who had been classified as relapse in a large retrospective cohort study in a high TB-incidence setting in suburban Cape Town, South Africa. At the time of the study, the study setting had an extremely high TB notification rate (761/100 000 for all forms of TB; 238/100 000 per year for new smear positive disease) and a low HIV prevalence[19]. The previous study aimed to investigate the temporal dynamics of TB due to relapse and reinfection in this high TB-burden setting. It was conducted among 130 smear-positive TB cases who had successfully completed their TB treatment under the directly observed treatment, short-course (DOTS) strategy between 1996 and 2007 (92% with bacteriological evidence of cure) and subsequently re-treated for smear-positive TB in the same setting. Using pairwise comparison of IS*6110* RFLP fingerprint patterns, the previous study classified 64 (49%) of 130 recurrent TB cases as relapse and the remainder as exogenous reinfections. Details of the earlier study and its setting have been previously described[6]. In the present study, we revisited the conclusion of relapse that was made for a subset of 25 recurrent TB cases with identical IS*6110* RFLP fingerprint patterns in the earlier study (Figure 1). The subset was chosen based on resource considerations and the availability of WGS-quality DNA.

M. tuberculosis isolates were cultured under biosafety level 3 conditions on Lowenstein-Jensen slants until confluent growth was observed. Cultured bacteria were heat-killed prior to phenol/chloroform DNA extraction[20]. Paired-end genomic libraries were prepared using the TruSeq DNA Sample Preparation Kits V2 (Illumina Inc, San Diego, CA, USA) according to the manufacturers' instructions. Pooled samples were sequenced on an Illumina HiSeq2000 instrument.

Raw sequence data were analysed as previously described[21,22]. Briefly, reads were trimmed with Trimmomatic[23] using a sliding window approach and an average phred quality score of 20, and aligned to the *M. tuberculosis* H37Rv reference genome (GenBank NC000962.3) with BWA[24], SMALT[25] and Novoalign (Novocraft). Genomic variants (single nucleotide variants and 1-10 base pair insertions and deletions) identified in all three alignments with SAMTools[26] and the Genome Analysis Toolkit[27] were considered with high confidence. Strain specific single nucleotide variants were used for lineage identification[28]. Qualimap was used to assess the quality of the alignments, mapping statistics are summarised in Supplementary Table S1[29,30]. For the purpose of this study a relaxed heterogeneous variant filtering approach was used to allow for the identification of underlying mixed infections. Specifically, variants identified with respect to the *M. tuberculosis* H37Rv reference genome were not filtered for allele frequency prior to comparative analysis. After pairwise comparisons, filters were applied to exclude variants found in *pe/ppe* genes, insertion sequences and phages, and repeat regions. In cases where indels were identified as unique variants, these were considered as a single discreet variant [equivalent to one single nucleotide substitution (SNP)]. Paired patient isolates with 0-5 variants between the strain from the index episode and the recurrent episode were considered relapse, a larger variant distance was taken to indicate exogenous reinfection, in accordance with the variant distance cut-offs used in previous studies[17,31,32]. Variants were considered "fixed" when supported by ≥70% of reads mapping to the variant locus, whilst variants were considered "heterogeneous" when a variant frequency between 30% and 70% were observed. Where WGS results indicated mixed infections (relatively consistent heterogeneous base frequencies indicating two distinct populations) in the second episode, the presence of phylogenetic markers (*katG*463, *gyrA*95, lineage defining regions of difference) [28,33,34] was assessed and known drug resistance conferring regions were also investigated[35].

The earlier study was approved by the Committee for Human Research, Faculty of Medicine and Health Sciences, Stellenbosch University (N09/05/144 and amendments 1 and 3)[6]. The present study used de-identified samples and data collected as part of the earlier study.

3. Results and Discussion

Of the 25 recurrent TB cases sampled, 13 (52%) were male, the median age was 29 years (inter-quartile range [IQR]: 21-39 years) (Table 1). HIV status was documented as negative in 11, positive in one, and not documented in 13 of the 25 patients. All patients had initially been diagnosed with smear-positive pulmonary TB and successfully completed their treatment according to available records, 23 of 25 had bacteriological proof of cure. All patients were subsequently re-treated for smear-positive pulmonary TB in the same study setting with a median time between the end of initial treatment and the start of the recurrent treatment of 11.7 months (IQR: 5.5-20.0 months). The IS*6110* RFLP fingerprints of the 25 TB cases showed identical genotypes between the index- and recurrent episode, and no evidence of mixed infections was reported.

Raw sequence data were deposited to the European Nucleotide Archive under project accession number PRJEB32341. *M. tuberculosis* isolates were whole genome sequenced at a mean depth of coverage of 119x (±37) (Supplementary Table S1) and strains were predominantly classified as Lineage 2 and Lineage 4 (Table 1) using strain specific single nucleotide variant markers for lineage inference[28].

Pairwise genomic analysis showed that 22 of the 25 pairs of isolates showed 0-4 variants (considering fixed and hetereogeneous variants) between the index- and recurrent episode (Table 1), consistent with reactivation and in agreement with the IS*6110* RFLP-based classification of relapse. Paired isolates from patient 10 showed 20 variant differences, of which only two were heterogeneous (not supported by 100% of the reads at that position), suggesting reinfection with a closely related exogenous strain. Further interpatient comparisons (all inter-patient strain comparison data not shown) revealed that the second isolate from patient 10 differed by only one variant from the first isolate of patient 15 (Figure 2), strongly supporting this hypothesis of reinfection with a closely related circulating strain.

Patients 23 and 25 showed 757 and 833 unique heterogeneous variants (with a variant frequency less than 30%) in the isolate from the second episode, respectively, while the index isolates from each patient showed zero, and two unique variants (one of which is heterogeneous), respectively. The number of variant differences observed here are in agreement with previous studies where WGS was used to differentiate between recurrent TB due to endogenous reactivation and exogenous reinfection using a variant cut-off of five to 10[16,17]. For both pairs, the underlying (minority) strain in the second episode harboured the CTG sequence at codon 463 in *katG* in contrast to the index episode isolate, which harboured the CGG sequence at this codon confirming that the reinfecting strains belonged to a different principle genetic group when compared to the index isolate[33]. In addition, the underlying (minority) population from the second episode from patient 23 harboured a rifampicin resistance conferring mutation (*rpoB* S531L), and an isoniazid resistance conferring mutation (*katG* S315T). This finding may suggest that reinfection with an exogenous strain allowed disease progression with the strain from the index episode, resulting in a mixed infection[36]. Visual reanalysis of the original IS*6110* RFLP fingerprinting blots showed the presence of faint underlying bands in the fingerprint of the second episode of patients 23 and 25 that were not considered by the GelCompar software[6]. Similar to previous findings, no significant relationship between the time difference between two isolates and the number of variants accumulated between paired patient isolates were observed[37] (Figure 3).

4. Conclusion

In conclusion, our study highlights the additional value of WGS for investigating recurrent TB to discriminate reactivation from endogenous reinfection. The findings based on WGS largely concurred with that of IS*6110* RFLP to conclude relapse as an important underlying mechanism of recurrence in a high TB incidence setting in a large retrospective cohort study in suburban Cape Town, South Africa. Pairwise genome comparison of *M. tuberculosis* strains cultured from diagnostic sputum samples collected at the index and recurrent TB episode for 25 recurrent TB classified as relapse showed 0 – 4 variants between paired isolates in 22 cases. No drug-resistance conferring mutations were observed in the isolates collected from the recurrent episode in these 22 cases, suggesting that acquisition of drug-resistance did not drive relapse, but that the infection was not completely cleared after treatment of the first episode of TB was completed. One case showed 20 variant differences between the index and recurrent episode, which suggests exogenous reinfection with a closely related circulating strain in the community. Based on WGS, two of the 25 patients showed evidence of mixed infections in the recurrent episode, each with the index episode strain detected as the dominant strain at recurrence and the minority strain harbouring distinct phylogenetic markers indicative of reinfection with an unrelated strain. IS*6110* RFLP typing may thus have slightly underestimated the contribution of exogenous reinfection to recurrent TB in this setting. In addition, one of the patients in which a mixed infection was detected in the recurrent episode, had an underlying strain that harboured drug-resistance conferring mutations (*rpoB*, *katG*). The observation of mixed infections in recurrent TB episodes raises the question of how reinfection may influence the growth of the prior infection strain which was not completely eradicated by prior treatment. This study highlights the high resolution and additional value of WGS for investigating recurrent TB in settings where the infection pressure is high with closely related strains potentially circulating in the community, and the extent of reinfections, possibly resulting in mixed infections, may be underestimated[18].

Author contributions

AD, MDV, RMW, and FMM conceived this study. RMW, PDvH, AP, and SLS directed the project. AD, MDV, FMM, SAA and AP generated and/or analysed the data. AD and MDV contributed equally to the manuscript by conducting the bioinformatic analysis, interpreting the data and writing the manuscript. All authors critically reviewed the manuscript and agreed to the contents thereof.

Acknowledgments

This work was supported by the South African National Research Foundation (NRF) Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council (SAMRC) Centre for TB Research, and King Abdullah University of Science and Technology (KAUST award number BAS/1/1020-01-01). The authors acknowledge the Tuberculosis Omics Research (TORCH) Consortium headed by Prof Annelis van Rie for financial support. SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and NRF of South Africa, award number UID 86539. Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard.

Additional information

Competing financial interests

The authors declare no competing financial interests.

Accession numbers:

Whole genome sequence data have been deposited at the European Nucleotide Archive under accession number PRJEB32341.

Tables and Figures:

Table 1. Patient information and pairwise comparison of whole genome sequences from 25 recurrent TB cases previously classified as relapse.

Patient details			*M. tuberculosis* isolate, index episode					*M. tuberculosis* isolate, recurrent episode						Comparison
Patient Nr	Sex	Age at first episode (years)	Time of episode (month / year)	IS*6110* RFLP family[12-14] (cluster number)	Lineage (WGS)	Number of fixed unique variants (A)	Number of heterogeneous unique variants (B)	Time of episode (month / year)	IS*6110* RFLP family[12-14] (cluster number)	IS*6110* Evolution	Lineage (WGS)	Number of fixed unique variants (C)	Number of heterogeneous unique variants (D)	Total variation (A+B+C+D)
1	Male	29	11 / 2005	11 (1)	4.3.2	0	0	7 / 2007	11 (1)	no	4.3.2	1	0	1
2	Male	40	8 / 2002	11 (9)	4.3.2	0	0	2 / 2004	11 (9)	no	4.3.2	1	0	1
3	Female	38	5 / 2000	11 (31)	4.3.2	0	0	1 / 2002	11 (31)	no	4.3.2	0	0	0
4	Male	34	4 / 2000	11 (626)	4.3.2	1	0	6 / 2002	11 (626)	no	4.3.2	0	0	1
5	Male	41	10 / 2001	29 (208)	2.2	0	0	5 / 2002	29 (208)	no	2.2	0	0	0
6	Female	30	8 / 2002	29 (208)	2.2	0	0	10 / 2004	29 (208)	no	2.2	1	0	1
7	Male	63	10 / 2006	29 (208)	2.2	0	0	2 / 2008	29 (208)	no	2.2	0	0	0
8	Male	26	6 / 2002	29 (209)	2.2	0	0	12 / 2004	29 (209)	no	2.2	3	1	4
9	Male	16	5 / 2006	140 (330)	4.1.1.3	0	0	3 / 2007	140 (330)	no	4.1.1.3	1	0	1
10	Female	37	9 / 2003	140 (330)	4.1.1.3	10	0	7 / 2007	140 (330)	no	4.1.1.3	8	2	20
11	Female	17	7 / 2003	140 (330)	4.1.1.3	0	0	10 / 2005	140 (330)	no	4.1.1.3	0	0	0
12	Female	51	8 / 2002	140 (330)	4.1.1.3	0	0	1 / 2004	140 (330)	no	4.1.1.3	0	0	0
13	Male	28	3 / 2003	140 (330)	4.1.1.3	0	0	2 / 2004	140 (330)	no	4.1.1.3	0	0	0
14	Female	32	5 / 2004	140 (330)	4.1.1.3	1	0	2 / 2008	140 (330)	no	4.1.1.3	3	0	4
15	Female	17	9 / 2006	140 (330)	4.1.1.3	0	0	9 / 2008	140 (330)	no	4.1.1.3	2	0	2
16	Female	26	10 / 2003	29 (220)	2.2	0	0	10 / 2004	29 (220)	no	2.2	1	0	1
17	Female	23	9 / 2002	29 (670)	2.2	0	0	7 / 2004	29 (670)	no	2.2	0	0	0
18	Female	10	6 / 2002	11 (794)	4.3.2	1	0	5 / 2005	11 (794)	no	4.3.2	1	0	2
19	Female	39	11 / 1998	150 (338)	4.1.1.3	0	0	6 / 2004	150 (338)	no	4.1.1.3	0	0	0
20	Male	21	5 / 2004	29 (734)	2.2	0	0	12 / 2005	29 (734)	no	2.2	0	0	0
21	Male	19	2 / 2006	140 (381)	4.1.1.3	1	0	10 / 2007	140 (381)	no	4.1.1.3	1	0	2
22	Male	12	4 / 2004	150 (338)	4.1.1.3	0	0	8 / 2005	150 (338)	no	4.1.1.3	0	0	0
23	Female	44	1 / 2002	140 (381)	4.1.1.3	0	0	12 / 2002	140 (381)	no	4.9*	0	757	757
24	Male	45	5 / 2003	140 (330)	4.1.1.3	0	0	12 / 2004	140 (330)	no	4.1.1.3	0	0	0
25	Male	27	4 / 2005	11 (32)	4.3.2	1	1	3 / 2006	11 (32)	no	4.9*	0	833	835

*based on filtered variants (variant frequency above 0.8), not an accurate determination of lineage in mixed infections

Figure captions:

Figure 1 – Study sample of 25 relapse cases in relation to the underlying study[6].

Figure 2. Variant differences observed among strains isolated from diagnostic samples for 2 individuals [patient 10 (blue), patient 15 (orange)] by time of the index and recurrent TB treatment episode. Variant differences are shown relative to that of the index treatment episode of Patient 10 (lower left corner). The strain isolated at the recurrent TB episode of patient 10 was more closely related to the strain isolated from another patient (15) than to the one isolated at the index episode of the same patient (10).

Figure 3. Genomic diversity of paired *M. tuberculosis* isolates indicating variant distances over the time elapsed between the index- and recurrent TB episode, excluding two reinfection cases that resulted in mixed infection in the second episode.

References:

1 Marx, F. M., Dunbar, R., Enarson, D. A. & Beyers, N. The rate of sputum smear-positive tuberculosis after treatment default in a high-burden setting: a retrospective cohort study. *PLoS One* **7**, e45724, doi:10.1371/journal.pone.0045724 (2012).

2 Glynn, J. R. *et al.* High rates of recurrence in HIV-infected and HIV-uninfected patients with tuberculosis. *J Infect Dis* **201**, 704-711, doi:10.1086/650529 (2010).

3 Middelkoop, K., Bekker, L. G., Shashkina, E., Kreiswirth, B. & Wood, R. Retreatment tuberculosis in a South African community: the role of re-infection, HIV and antiretroviral treatment. *Int J Tuberc Lung Dis* **16**, 1510-1516, doi:10.5588/ijtld.12.0049 (2012).

4 Lambert, M. L. *et al.* Recurrence in tuberculosis: relapse or reinfection? *Lancet Infect Dis* **3**, 282-287 (2003).

5 Crampin, A. C. *et al.* Recurrent TB: relapse or reinfection? The effect of HIV in a general population cohort in Malawi. *AIDS* **24**, 417-426, doi:10.1097/QAD.0b013e32832f51cf (2010).

6 Marx, F. M. *et al.* The temporal dynamics of relapse and reinfection tuberculosis after successful treatment: a retrospective cohort study. *Clin Infect Dis* **58**, 1676-1683, doi:10.1093/cid/ciu186 (2014).

7 Unis, G. *et al.* Tuberculosis recurrence in a high incidence setting for HIV and tuberculosis in Brazil. *BMC Infect Dis* **14**, 548, doi:10.1186/s12879-014-0548-6 (2014).

8 Martin, A., Herranz, M., Serrano, M. J., Bouza, E. & Garcia de Viedma, D. Rapid clonal analysis of recurrent tuberculosis by direct MIRU-VNTR typing on stored isolates. *BMC Microbiol* **7**, 73, doi:10.1186/1471-2180-7-73 (2007).

9 Varghese, B., al-Omari, R., Grimshaw, C. & Al-Hajoj, S. Endogenous reactivation followed by exogenous re-infection with drug resistant strains, a new challenge for tuberculosis control in Saudi Arabia. *Tuberculosis* **93**, 246-249, doi:10.1016/j.tube.2012.12.001 (2013).

10 van der Spuy, G. D. *et al.* Use of genetic distance as a measure of ongoing transmission of Mycobacterium tuberculosis. *J Clin Microbiol* **41**, 5640-5644 (2003).

11 Warren, R. M. *et al.* Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of Mycobacterium tuberculosis. *J Clin Microbiol* **40**, 1277-1282 (2002).

12 Shen, G. *et al.* The study recurrent tuberculosis and exogenous reinfection, Shanghai, China. *Emerging infectious diseases* **12**, 1776-1778, doi:10.3201/eid1211.051207 (2006).

13 Umubyeyi, A. N. *et al.* Molecular investigation of recurrent tuberculosis in patients from Rwanda. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* **11**, 860-867 (2007).

14 Niemann, S. *et al.* Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. *PloS one* **4**, e7407, doi:10.1371/journal.pone.0007407 (2009).

15 Roetzer, A. *et al.* Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med* **10**, e1001387, doi:10.1371/journal.pmed.1001387 (2013).

16 Guerra-Assuncao, J. A. *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* **211**, 1154-1163, doi:10.1093/infdis/jiu574 (2015).

17 Bryant, J. M. *et al.* Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet Respir Med* **1**, 786-792, doi:10.1016/S2213-2600(13)70231-5 (2013).

18 van Rie, A. *et al.* Reinfection and mixed infection cause changing Mycobacterium tuberculosis drug-resistance patterns. *Am J Respir Crit Care Med* **172**, 636-642, doi:10.1164/rccm.200503-449OC (2005).

19 Verver, S. *et al.* Transmission of tuberculosis in a high incidence urban community in South Africa. *Int J Epidemiol* **33**, 351-357, doi:10.1093/ije/dyh021 (2004).

20 Warren, R. *et al.* Safe Mycobacterium tuberculosis DNA extraction method that does not compromise integrity. *J Clin Microbiol* **44**, 254-256, doi:10.1128/JCM.44.1.254-256.2006 (2006).

21 Black, P. A. *et al.* Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates. *BMC genomics* **16**, 857, doi:10.1186/s12864-015-2067-2 (2015).

22 Ates, L. S. *et al.* Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. *Nat Microbiol* **3**, 181-188, doi:10.1038/s41564-017-0090-6 (2018).

23 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

24 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

25 Ponstingl, H. & Ning, Z. SMALT - A new mapper for DNA sequencing reads. *F1000Posters* **1** (2015).

26 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

27 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

28 Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun* **5**, 4812, doi:10.1038/ncomms5812 (2014).

29 Garcia-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678-2679, doi:10.1093/bioinformatics/bts503 (2012).

30 Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292-294, doi:10.1093/bioinformatics/btv566 (2016).

31 Bryant, J. M. *et al.* Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect Dis* **13**, 110, doi:10.1186/1471-2334-13-110 (2013).

32 Walker, T. M. *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* **13**, 137-146, doi:10.1016/S1473-3099(12)70277-3 (2013).

33 Sreevatsan, S. *et al.* Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A* **94**, 9869-9874 (1997).

34 Tsolaki, A. G. *et al.* Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* **101**, 4865-4870, doi:10.1073/pnas.0305634101 (2004).

35 Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* **7**, 51, doi:10.1186/s13073-015-0164-0 (2015).

36 du Plessis, D. G., Warren, R., Richardson, M., Joubert, J. J. & van Helden, P. D. Demonstration of reinfection and reactivation in HIV-negative autopsied cases of secondary tuberculosis: multilesional genotyping of Mycobacterium tuberculosis utilizing IS 6110 and other repetitive element-based DNA fingerprinting. *Tuberculosis (Edinb)* **81**, 211-220, doi:10.1054/tube.2000.0278 (2001).

37 Korhonen, V. *et al.* Whole genome analysis of Mycobacterium tuberculosis isolates from recurrent episodes of tuberculosis, Finland, 1995-2013. *Clin Microbiol Infect* **22**, 549-554, doi:10.1016/j.cmi.2016.03.014 (2016).