

RESEARCH ARTICLE

Open Access



Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates

PA Black^{1†}, M. de Vos^{1†}, GE Louw¹, RG van der Merwe¹, A. Dippenaar¹, EM Streicher¹, AM Abdallah², SL Sampson¹, TC Victor¹, T. Dolby³, JA Simpson³, PD van Helden¹, RM Warren^{1*†} and A. Pain^{2†}

Abstract

Background: Whole genome sequencing has revolutionised the interrogation of mycobacterial genomes. Recent studies have reported conflicting findings on the genomic stability of *Mycobacterium tuberculosis* during the evolution of drug resistance. In an age where whole genome sequencing is increasingly relied upon for defining the structure of bacterial genomes, it is important to investigate the reliability of next generation sequencing to identify clonal variants present in a minor percentage of the population. This study aimed to define a reliable cut-off for identification of low frequency sequence variants and to subsequently investigate genetic heterogeneity and the evolution of drug resistance in *M. tuberculosis*.

Methods: Genomic DNA was isolated from single colonies from 14 rifampicin mono-resistant *M. tuberculosis* isolates, as well as the primary cultures and follow up MDR cultures from two of these patients. The whole genomes of the *M. tuberculosis* isolates were sequenced using either the Illumina MiSeq or Illumina HiSeq platforms. Sequences were analysed with an in-house pipeline.

Results: Using next-generation sequencing in combination with Sanger sequencing and statistical analysis we defined a read frequency cut-off of 30 % to identify low frequency *M. tuberculosis* variants with high confidence. Using this cut-off we demonstrated a high rate of genetic diversity between single colonies isolated from one population, showing that by using the current sequencing technology, single colonies are not a true reflection of the genetic diversity within a whole population and vice versa. We further showed that numerous heterogeneous variants emerge and then disappear during the evolution of isoniazid resistance within individual patients. Our findings allowed us to formulate a model for the selective bottleneck which occurs during the course of infection, acting as a genomic purification event.

Conclusions: Our study demonstrated true levels of genetic diversity within an *M. tuberculosis* population and showed that genetic diversity may be re-defined when a selective pressure, such as drug exposure, is imposed on *M. tuberculosis* populations during the course of infection. This suggests that the genome of *M. tuberculosis* is more dynamic than previously thought, suggesting preparedness to respond to a changing environment.

Keywords: Genetic complexity, Clinical isolates, *Mycobacterium tuberculosis*, Heterogeneity, Next generation sequencing, Relaxed variant filtering

* Correspondence: rw1@sun.ac.za

†Equal contributors

¹DST-NRF Centre of Excellence for Biomedical Tuberculosis Research/SA MRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa

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

Background

Whole genome sequencing (WGS) has revolutionised the detailed interrogation of mycobacterial genomes at base-pair resolution. Application of this technology has provided novel insights into the evolution of members of the *Mycobacterium tuberculosis* complex [1–3]. More recently, WGS has been applied to investigate the evolution of drug resistance based on the hypothesis that additional mutational events may precede or occur concurrently with known resistance conferring mutations [4–8]. A number of recent studies have assessed the genomic stability of *M. tuberculosis* during the evolution of drug resistance (Reviewed by Trauner et al., 2014) [4], producing diametrically opposed results. Some studies have demonstrated genomic stability [6, 7, 9, 10], while more recent reports have shown genomic instability with the emergence of additional genetic variants independent of those observed in drug target genes conferring resistance [5, 11–14]. Various studies have demonstrated the acquisition of known resistance conferring mutations in the *M. tuberculosis* genome during the course of infection and subsequent drug treatment. This highlights the potential of *M. tuberculosis* to diversify and adapt under selective pressure [11, 13, 15]. In other work, WGS analysis of serially collected sputum samples of Tuberculosis (TB) patients provided overwhelming evidence of the presence of drug resistant sub-populations [5]. This study emphasised the need to investigate the implications of inter- and intra-host *M. tuberculosis* genetic diversity for transmission and disease outcomes [5]. The presence of drug resistant sub-populations is a major concern when considering the variable sensitivity of standard genetic and microbiological tests for the diagnosis of drug resistance *M. tuberculosis* [16].

It is important to note that the identification of sub-populations within the context of WGS is dependent on the read frequency cut-off values used in the standard variant filtering approach. Typically, these frequency cut-off values are set at >70 %, implying that only variants fixed within a population are identified i.e. variants present at only >70 % of the sequencing reads are analysed [6, 8]. Failure to adjust these algorithms to accommodate for the presence of low-frequency sub-populations has led authors to conclude that the population structure of *M. tuberculosis* is homogeneous [6, 8, 17, 18]. By using an alternative variant calling approach (a minimum read depth of 50 and a minimum variant frequency of 4 %) a recent study demonstrated the rapid expansion and collapse of different sub-populations that evolved in parallel during the evolution of extensively drug resistant (XDR-TB) [14]. However, while WGS is increasingly being used to investigate drug resistance and evolution in *M. tuberculosis*, there is still uncertainty surrounding the error rate of sequencing, read alignment and the detection of variants. In an age where WGS is increasingly relied on for defining the structure of

bacterial genomes, it is important to investigate the reliability of next-generation sequencing reads where a variant is only present in a minor percentage of the sequencing reads. This therefore questions the threshold at which underlying populations, as indicated by the percentage of sequencing reads supporting a variant allele, can be confirmed as true variants as opposed to sequencing errors.

This study aimed to define a reliable cut-off for identification of heterogeneous variants from WGS data. Subsequently this cut-off was used to investigate: 1) heterogeneity in single colonies isolated from 13 rifampicin mono-resistant *M. tuberculosis* clinical isolates, and 2) the evolution of isoniazid resistance in rifampicin mono-resistant isolates in 2 patients.

Results

Strains were initially selected on the basis of clinical diagnostic records, then subjected to further analysis to confirm their resistance profiles and strain genotypes. Phenotypic drug resistance testing of the parental isolates ($n = 13$) and their associated single colonies ($n = 36$) confirmed the rifampicin mono-resistant profile in all cases. As expected, all isolates carried a mutation in the *rpoB* gene; genotyping of parental isolates identified one of the following mutations in the *rpoB* gene: Ser531Leu, His526Tyr or Leu533Pro (Table 1). These mutations were retained in the corresponding single colonies after selection on agar plates containing rifampicin. IS6110 genotyping and spoligotyping revealed that the rifampicin mono-resistant isolates originated from different genetic backgrounds (LCC, Haarlem, Beijing and EAI), representing the broad strain diversity circulating in the Western Cape, South Africa. Genotyping by IS6110

Table 1 Genotypic characterisation of *M. tuberculosis* clinical isolates used for the investigation into genomic heterogeneity

Isolate name	rpoB ^a	Spoligotype classification
R160	Ser531Leu	LCC
R376	Ser531Leu	Haarlem
R458	Ser531Leu	Unknown/unique
R486	Leu533Pro	Beijing
R631	His526Tyr	Unknown/unique
R637	Ser531Leu	Beijing
R641	Leu533Pro	Beijing
R721	Ser531Leu	Beijing
R912	His526Tyr	EAI
R965	Leu533Pro	Beijing
R966	His526Tyr	Beijing
R1035	Ser531Leu	LAM
R1415	His526Tyr	Beijing

LCC low copy clade, EAI East African Indian, LAM Latin American Mediterranean
^aAmino acid change according to the *Escherichia coli rpoB* gene sequence

Table 2 Validation of variants with a read frequency ranging between 20 and 100 % using targeted PCR and sanger sequencing

Gene ^a	Sanger chromatogram result	Read frequency (%)	Sanger result
<i>Rv2316</i>	Single peak	127/127 (100.0)	True
<i>Rv1703c</i>	Single peak	152/154 (98.7)	True
<i>Rv0820</i>	Single peak	193/200 (96.5)	True
<i>Rv3220c</i>	Single peak	129/134 (96.3)	True
<i>Rv0537c</i>	Single peak	98/126 (77.8)	True
<i>Rv2692</i>	Single peak	102/142 (71.8)	True
<i>Rv1521</i>	Double peaks	157/222 (70.7)	True
<i>Rv0521</i>	Double peaks	117/159 (68.8)	True
<i>Rv1904</i>	Double peaks	119/181 (65.7)	True
<i>Rv1230c</i>	Double peaks	96/155 (61.9)	True
<i>Rv3086</i>	Double peaks	84/145 (57.9)	True
<i>Rv3083</i>	Double peaks	55/95 (57.9)	True
<i>Rv1429</i>	Double peaks	74/129 (57.4)	True
<i>Rv2577</i>	Double peaks	33/61 (54.1)	True
Intergenic (1093238)	Double peaks	89/168 (53.0)	True
<i>Rv3391</i>	Double peaks	57/113 (50.4)	True
<i>Rv2689c</i>	Double peaks	59/123 (48.0)	True
<i>Rv0970</i>	Double peaks	48/104 (46.2)	True
<i>Rv2173</i>	Double peaks	54/130 (41.5)	True
<i>Rv1929c</i>	Double peaks	52/128 (40.6)	True
<i>Rv2459</i>	Double peaks	18/45 (40.0)	True
<i>Rv2984</i>	Double peaks	34/89 (38.2)	True
<i>Rv1894c</i>	Double peaks	36/95 (37.9)	True
<i>Rv1316c</i>	Double peaks	72/198 (36.4)	True
<i>Rv1021</i>	Double peaks	42/119 (35.3)	True
<i>Rv2544</i>	Double peaks	50/151 (33.1)	True
<i>Rv3772</i>	Double peaks	47/147 (32.7)	True
<i>Rv3861</i>	Double peaks	38/119 (31.9)	True
<i>Rv3780</i>	Double peaks	30/106 (31.9)	True
<i>Rv0491</i>	Double peaks	44/138 (31.9)	True
<i>Rv2957</i>	Single peak	47/154 (30.5)	False
<i>Rv1549</i>	Double peaks	42/138 (30.4)	True
<i>Rv3703c</i>	Single peak	37/126 (29.4)	False
<i>Rv0594</i>	Double peaks	61/202 (29.2)	True
<i>Rv0372c</i>	Single peak	85/296 (28.7)	False
<i>Rv1479</i>	Single peak	34/120 (28.3)	False
<i>Rv0780</i>	Single peak	44/159 (27.7)	False
<i>Rv0092</i>	Double peaks	47/170 (27.6)	True
<i>Rv0688</i>	Double peaks	56/209 (26.8)	True
<i>Rv0282</i>	Single peak	54/205 (26.3)	False
<i>Rv0663</i>	Double peaks	41/158 (25.9)	True
<i>Rv0522</i>	Double peaks	66/263 (25.1)	True

Table 2 Validation of variants with a read frequency ranging between 20 and 100 % using targeted PCR and sanger sequencing (Continued)

<i>Rv1660</i>	Double peaks	52/239 (21.8)	True
<i>Rv1627c</i>	Single peak	34/161 (21.1)	False
<i>Rv0654</i>	Double peaks	34/167 (20.4)	True
<i>Rv2934</i>	Double peaks	47/232 (20.3)	True

^aAll variant positions and WGS results are listed in the supplementary data (Additional file 2: Table S2)

also showed that the parental isolates and single colonies were identical, confirming the absence of mixed infection.

Identifying a reliable cut-off to detect genomic heterogeneity

To define a reliable cut-off for heterogeneous variant detection, we employed a combination of WGS and statistical analyses of parental isolates and their associated single colonies. Single colonies were obtained by plating serial dilutions of the parental isolate onto 7H10 agar containing 2 µg/ml RIF. Two to three single colonies isolated from each of the 13 parental rifampicin mono-resistant *M. tuberculosis* isolates (Table 1) were selected and subjected to WGS analysis. This analysis of 36 single colonies revealed the presence of a total of 153 possible sequence variants between corresponding single colonies using *M. tuberculosis* H37Rv as the alignment standard. To determine whether a variant which was present in less than 100 % of the reads reflected the presence of either genetically distinct subpopulations or sequencing artefacts, 46 of the 153 possible variants were selected for validation by Sanger sequencing (Table 2). All 6 variants present in ≥ 71.8 % of the reads were verified with a single chromatogram peak demonstrating that only the dominant nucleotide was detected. In addition, 25 variants present in between 30.5 and 70.7 % of the reads were also determined to contain two nucleotides indicated by the presence of two chromatogram peaks. Of the 16 variants with a read frequency between 20.3 and 30.5 %, 7 were shown to be false positives as only a single chromatogram peak was observed. We investigated the frequency of the 7 false positive variants across all of the single colonies as recurrence of these variants in other genomes might suggest that the false variants were called due to mapping errors. Only one of the 7 variants (in *Rv0282*) was found to be recurrent and identified in 22 of the single colonies.

We performed a receiver operator characteristic (ROC) curve analysis to define a reliable cut-off for the identification of true heterogeneous variants. ROC curve analysis disproved the null hypothesis, with an area under the curve (AUC) of 0.835 (95 % CI: 0.722 – 0.948). A cut-off value of 30 % read frequency was defined with a true positive value of 0.795 and a false positive value of 0.143 i.e. a true positive rate of 79.5 % and a false positive rate of

14.3 % (Fig. 1). Using the conventional cut-off value of 70 % there would be a true positive rate of 17.9 % and a false positive rate of 0 %. Accordingly, we defined a cut off for the description of true heterogeneity at a defined nucleotide position as ≥ 30 % i.e. if a variant identified in at least 30 % of the Illumina sequencing reads (after filtering and GenomeView analysis), it is likely that the variant is truly present within the genome of a sub-population of bacilli. Using this analysis we also defined a variant with a read frequency of greater than 70 % to be fixed within a population.

Identifying genomic heterogeneity in individual colonies from rifampicin mono-resistant isolates

We next applied the ≥ 30 % variant frequency cut-off value to assess genomic heterogeneity within and between individual colonies. Using this cut-off value this analysis 36 of single colonies revealed the presence of a total of 114 possible sequence variants between corresponding single colonies using *M. tuberculosis* H37Rv as the alignment standard. From the 114 possible variants, 42 were found to have a read frequency of greater than 70 %, while 72 were

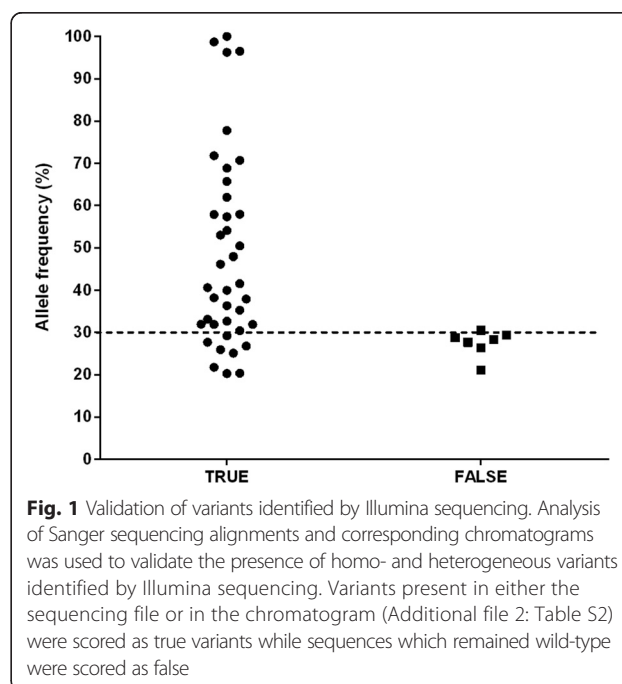


Table 3 Variants identified in corresponding single colonies derived from different clinical isolates

	Single colony 1			Single colony 2			Single colony 3			Total variation
	Total ^a	Fixed ^b	Hetero ^c	Total ^a	Fixed ^b	Hetero ^c	Total ^a	Fixed ^b	Hetero ^c	
R160	5	2	3	4	3	1	-	-	-	9
R376	3	1	2	8	2	6	6	1	5	17
R458	3	3	0	5	2	3	3	0	3	11
R486	0	0	0	2	0	2	-	-	-	2
R631	5	0	5	6	0	6	8	0	8	19
R637	0	0	0	0	0	0	0	0	0	0
R641	9	9	0	3	0	3	6	0	6	18
R721	0	0	0	4	0	4	0	0	0	4
R912	4	4	0	6	3	3	7	7	0	17
R965	4	4	0	1	1	0	-	-	-	5
R966	5	0	5	1	0	1	6	0	6	12
R1035	0	0	0	0	0	0	0	0	0	0
R1415	0	0	0	0	0	0	0	0	0	0

^aTotal number variants unique between the corresponding single colonies

^bFixed variants as defined as having a read frequency of $\geq 70\%$

^cHeterogeneous variants as defined as having a read frequency $< 70\%$ and $\geq 30\%$

-A third single colony was not available for the comparison

found to have a read frequency of between 30 and 70 % (Table 3, (Additional file 1)). The number of variants identified was independent of the strain background, and ranged from 0 to 19 in different clinical isolates. Importantly, genomic heterogeneity was observed both within single colonies (where a variant is only present in a proportion of the reads) and between single colonies isolated from one parent.

Comparing single colonies and the entire population isolated from sputum

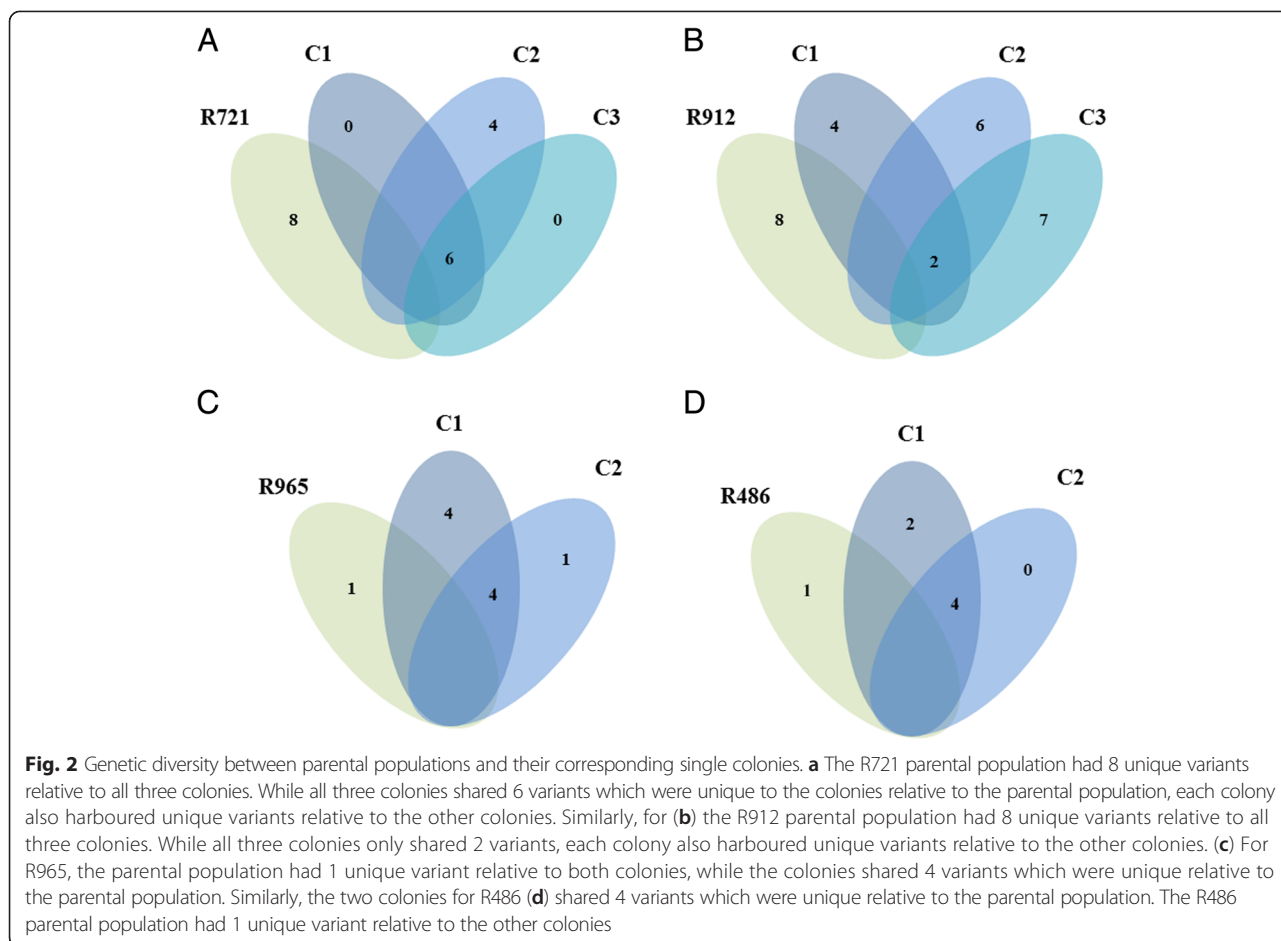
WGS data was available for four of the parental *M. tuberculosis* isolates used to make the single colonies. Therefore, to determine whether the single colonies reflected the genetic heterogeneity of the parent isolate, we compared the WGS of the single colonies from four rifampicin mono-resistant *M. tuberculosis* clinical isolates to the sequences of their corresponding parent population (R721, R912, R965 and R486). As shown in Fig. 2 this comparison identified 6, 4, 4 and 2 variants that were unique to the single colonies relative to the parental genomes for R721, R965, R486 and R912, respectively. These variants were shared between all of the single colonies cultured from their corresponding parental isolates. In addition, the individual single colonies also harboured unique variants that were absent in the corresponding parental isolates as well as absent in the other related single colonies. Conversely, the parental isolates also each harboured unique variants that were absent in all corresponding single colonies (Fig. 2). Together these results suggest that all of the genomic variants identified as unique to the

single colonies were masked by the dominant population present in the parent isolates.

Intra-patient evolution of drug resistance

Having established the utility of our analytical approach in identifying inter- and intra-isolate variation, we next went on to analyse genome heterogeneity during the evolution of drug resistance. Here, we analysed MDR isolates (R807 and R1210) collected approximately 7 and 11 months after the rifampicin mono-resistant isolates R721 and R912, respectively (Table 4). We investigated the heterogeneity of SNPs across the genome of the rifampicin mono-resistant and MDR entire populations relative to the *M. tuberculosis* H37Rv reference genome. We subsequently compared this list of SNPs between the rifampicin mono-resistant and MDR isolates to identify variants unique to each isolate as well as the evolutionary events associated with the emergence of drug (isoniazid) resistance (Table 5).

For patient 1 numerous heterogeneous variants were identified in the rifampicin mono-resistant isolate, R721, while there were no heterogeneous variants observed in the follow-up MDR isolate, R807 (Fig. 3a, Table 4). Since the primary difference between these two isolates is the presence of a *katG* mutation (Table 5), these results suggest that the heterogeneous variants present in R721 were 'lost' during the acquisition of isoniazid resistance. Patient 2 shows contrasting results: the rifampicin mono-resistant isolate, R912, was shown to have no heterogeneous variants while its follow-up MDR isolate, R1210, harboured 5 (Fig. 3b). The variants shown to be unique to R912 were all present at a



variant frequency of greater than 70 %. Interestingly, two variants shown to be fixed within the rifampicin mono-resistant population, with a variant frequency of 100 %, were found to be present at a lower frequency within the MDR population (Fig. 3b). This finding indicates that two fixed variants in R912 may have been lost in R1210. These results revealed the acquisition and loss of numerous heterogeneous variants (Table 5), suggesting continuous genome evolution despite the evolutionary bottle neck imposed by the isoniazid selective pressure.

Finally, we compared the WGS data of MDR isolates to the single colonies generated from the rifampicin mono-resistant parental isolates. This comparison failed to identify any of the single colony specific genomic variants in the MDR isolate.

Discussion

Our study aimed to use WGS in combination with Sanger sequencing and statistical analyses to define a reliable cut-off for heterogeneous variant detection. Our data enabled

Table 4 *M. tuberculosis* clinical isolates demonstrating a number of unique variants between rifampicin mono-resistant and MDR isolates during *in vivo* evolution of isoniazid resistance

Patient	Isolate name	Phenotypic resistance	Collection date	rpoB ^c	katG	inhA promoter	Spoligotype classification	Fixed variants ^a	Heterogeneous variants ^b	Total variation
1	R721	Rifampicin mono	22/10/2003	Ser531Leu	-	-	Beijing	1	8	9
	R807	MDR	19/05/2004	Ser531Leu	Gly309Val	-	Beijing	2	0	2
2	R912	Rifampicin mono	15/09/2004	His526Tyr	-	-	East Africa Indian (EAI)	7	0	7
	R1210	MDR	12/08/2005	His526Tyr	-	-15	EAI	1	5	6

^aFixed variants as defined as having a read coverage of ≥ 70 %

^bHeterogeneous variants as defined as having a read coverage < 70 % and ≥ 30 %

^cAmino acid change according to the *Escherichia coli* rpoB gene sequence

Table 5 Isolate specific variants identified in rifampicin mono-resistant and MDR *M. tuberculosis* isolates of patient 1 and 2

Isolate	Locus	Gene	Amino acid change	Coverage of variant (%)	Gene description	Functional category ^c		
Patient 1	R721 ^a	Rv0435c	I397L	50	Putative conserved ATPase	Cell wall and cell processes		
		Rv0435c	D395Y	50	Putative conserved ATPase			
		Rv0668	<i>rpoC</i>	V1039A	30	DNA-directed RNA polymerase RpoC (RNA polymerase beta' subunit).	Information pathways	
		Rv1850	<i>ureC</i>	Q11K	48	Urease alpha subunit UreC (urea amidohydrolase)	Intermediary metabolism and respiration	
		Rv1850	<i>ureC</i>	Q11R	49	Urease alpha subunit UreC (urea amidohydrolase)		
		Rv3218		Y174H	66	Conserved protein	Conserved hypotheticals	
		Rv2004c		Ins AAG	43	Conserved protein	Conserved hypotheticals	
		Rv3563	<i>fadE32</i>	Ins AC	40	Probable acyl-CoA dehydrogenase FadE32	Lipid metabolism	
	R807 ^b	Rv3696c	<i>glpK</i>	Ins AC	73	Probable glycerol kinase GlpK (ATP:glycerol 3-phosphotransferase)	Intermediary metabolism and respiration	
		Rv1908c	<i>katG</i>	G309V	100	Catalase-peroxidase-peroxyxynitritase T KatG	Virulence, detoxification, adaptation	
	Patient 2	R912 ^a	Rv3696c	<i>glpK</i>	T91I	80	Probable glycerol kinase GlpK (glycerokinase) (GK)	Intermediary metabolism and respiration
			Rv1128c		G430S	98	Conserved hypothetical protein	Insertion sequences and phages
			Rv2236c	<i>cobD</i>	L269S	97	Probable cobalamin biosynthesis transmembrane protein CobD	Intermediary metabolism and respiration
			Rv2664		H22Q	99	Hypothetical protein	Conserved hypotheticals
Rv2772c				E149*	97	Probable conserved transmembrane protein	Cell wall and cell processes	
Rv2984			<i>ppk1</i>	P631A	96	Polyphosphate kinase PPK (polyphosphoric acid kinase)	Intermediary metabolism and respiration	
Rv3391			<i>acrA1</i>	syn (248)	99	Possible multi-functional enzyme with acyl-CoA-reductase activity AcrA1	Lipid metabolism	
Rv3537			<i>kstD</i>	syn (378)	98	Probable dehydrogenase	Intermediary metabolism and respiration	
R1210 ^b			<i>inhA promoter</i>	-15	45			
		Rv1484	<i>inhA</i>	S94A	70	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)	Lipid metabolism	
	Rv1629	<i>polA</i>	syn (146)	45	Probable DNA polymerase I PolA	Information pathways		
	Rv2935	<i>ppsE</i>	C582R	69	Phenolphthiocerol synthesis type-I polyketide synthase PpsE	Lipid metabolism		

^aRifampicin mono-resistant^bMDR^cFunctional category as classified by Tuberculist (<http://genolist.pasteur.fr/TubercuList/> and <http://tuberculist.epfl.ch/>)

us to define a read frequency cut off of 30 % for reliable Illumina sequencing variant frequency filtering i.e. a variant present in 30 % or more of the sequencing reads can be considered to be a true variant and not a sequencing error. At a variant frequency of 30 % there is a true positive of 79.5 % and a false positive rate of 14.3 %. We acknowledge that there is still a chance that true variants at a frequency lower than 30 % may be missed due to the limited resolution of Sanger sequencing. For the purpose of our analyses we regard it of greater importance to exclude false

positives from our analyses than to omit a small amount of true positives. However, we do not exclude the probability that this cut-off value may change with increased sequence read depth. Higher depth of coverage would however not limit the detection of false positive variants as the 7 false positive variants identified in this study were not associated with lower mean genome coverage. Furthermore, by defining a cut-off it allowed us to use a relaxed variant filtering approach to investigate the presence of sub-populations in *M. tuberculosis* clinical isolates.

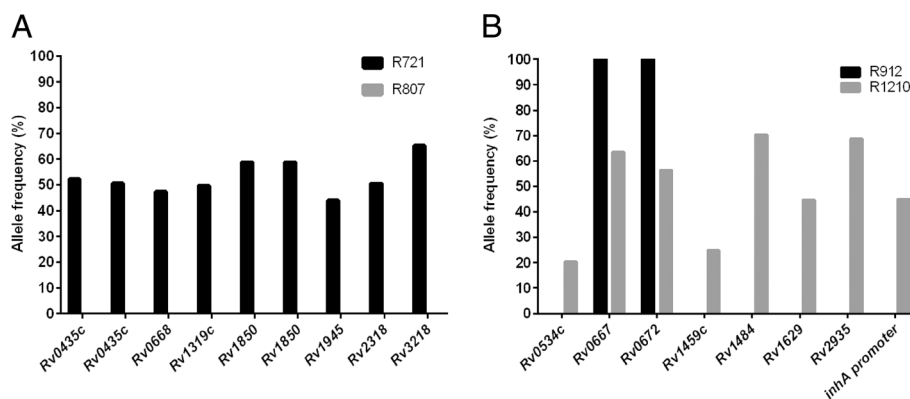


Fig. 3 Heterogeneous positions identified across the whole population genomes relative to *M. tuberculosis* H37Rv reference genome. **a**The rifampicin mono-resistant isolate (R721) for patient 1 shows numerous heterogeneous variants relative to *M. tuberculosis* H37Rv while the follow-up MDR isolate (R807) has none. **b** For patient 2 the rifampicin mono-resist isolate (R912) showed no heterogeneous variants relative to *M. tuberculosis* H37Rv, while the follow-up MDR isolate (R1210) had numerous heterogeneous variants. R912 shared 2 variants (Rv0667 and Rv0672) with R1210, where the variant was present at 100 % in R912 but was a heterogeneous variant in R1210

Using our cut-off of 30 % variant frequency we investigated the genomic heterogeneity within and between individual single colonies isolated from rifampicin mono-resistant *M. tuberculosis* clinical isolates. To our knowledge this is the first study that has investigated *M. tuberculosis* genomic diversity within single colonies from a clinical sample at a single time point. We observed a high rate of genetic diversity between single colonies isolated from the same parental isolate. We acknowledge that the analysis of only two to three single colonies for each patient isolate would have limited the analysis of the true heterogeneity in the total population. We also acknowledge that the selection of these single colonies on media containing rifampicin would have resulted in the loss of variants representing rifampicin susceptible colonies. This high rate of genetic diversity seen in our study is in concordance with previous studies that used relaxed filtering with regards to variant frequency [5, 12, 14]. Sun et al. showed that there can be as many as 41 (ranging between 6 and 41) variants between serial sputum samples, with 82.7 % of all the variants at frequencies lower than 20 % (based on statistical evaluation as opposed to validation by Sanger sequencing) [5]. A study by Bryant et al. showed seemingly contrasting results using the same filtering approach to investigate relapse and reinfection cases [10]. Using a minimum read depth of 4 and read frequency higher than 5 %, little genetic diversity was observed in the relapse cases. However any heterogeneous positions identified were discarded as mapping errors and no validation by Sanger sequencing was done. This may have led to an underestimation of the number of variants, a possibility highlighted by our findings that heterogeneous variants occurring between 30 and 70 % frequency are likely to be truly present within a population.

Our findings which showed the extent of diversity between parental isolates and their respective single colonies are in agreement with a recent study where *in vitro* generated mutants were compared to their drug susceptible progenitor [19]. Numerous variants only present in a proportion of the Illumina reads were identified to be unique to the parental genome. Similar to our findings the authors showed that in some daughter cells the mutant allele was lost while in others it was retained [19]. This suggests that single colonies may not be a true reflection of the genomic diversity of a clinical *M. tuberculosis* isolate. In contrast, WGS of an entire population may underestimate the extent of genetic diversity present in a clinical isolate given the complexity of the *M. tuberculosis* population structure. However, improved read depths may allow for identification of underlying populations which were undetectable in this study. The high degree of genetic diversity seen in this study is similar to that reported elsewhere [5, 12, 14] and is unlikely to have arisen as a consequence of laboratory adaptations. The number of culturing steps was kept to a minimum and previous WGS data from *in vitro* generated mutants in our laboratory showed very little genetic diversity (data not shown). This is supported by WGS of six *M. tuberculosis* H37Rv strains from multiple laboratories that showed little change after years of repeated sub-culturing, suggesting genomic stability during *in vitro* culture [20].

Our results show that single colonies may not be a true reflection of the genetic diversity present within a clinical isolate and vice versa. This finding may have important implications for genomic epidemiology since a recent study by Didelot et al. (2014) demonstrated the use of WGS to infer person to person transmission. Underlying

populations masked by dominant variants may not truly be represented in the above method and may therefore be overlooked when interpreting person to person transmission [21]. Therefore, this study highlights the importance of the methods of storage used for *M. tuberculosis* isolates. The diversity seen i) within a clinical isolate and ii) between single colonies isolated from a single *M. tuberculosis* clinical isolate needs to be taken into consideration. Storage of samples should therefore be carefully considered based on the research questions which may be asked in future studies. Numerous studies have stated that single colonies were isolated from LJ slants for storage and further use [9, 22–25], while other studies use clinical isolates as a whole representative population [5, 8, 10, 11, 13, 14]. These different approaches may impact results obtained and conclusions drawn.

Having shown genetic diversity between parental populations and their single colonies, as well as diversity between single colonies isolated from the same progenitor, we next investigated the evolution of isoniazid resistance in *M. tuberculosis* clinical isolates within two patients. Initial investigation into genetic heterogeneity in both rifampicin mono-resistant and MDR isolates relative to *M. tuberculosis* H37Rv revealed that there were numerous heterogeneous variants present within the genome. For patient 1, the rifampicin mono-resistant isolate harboured numerous heterogeneous variants, all of which were unique to this isolate when compared to its paired MDR isolate. During the acquisition of the *katG* isoniazid resistance causing mutation all of these heterogeneous variants were lost from the population. This finding suggested that the isoniazid selective pressure imposed an evolutionary bottleneck, resulting in a purification effect and the loss of heterogeneous variants. Contrasting results were observed for patient 2, as the 7 apparently fixed SNPs were lost during the acquisition of the *inhA* promoter isoniazid resistance causing mutation. Once again, this finding suggests that isoniazid selective pressure imposed an evolutionary bottleneck. The paired MDR isolate, R1210, showed a total of 8 heterogeneous variants to be unique to this isolate when compared to R912, two of which were fixed within the population of R912. This suggests that these two variants were reverting to wild type, while other variants were emerging within the population. These findings highlight the continuous genome evolution occurring after an evolutionary bottleneck is imposed on a population. The *M. tuberculosis* isolates from each patient represent two different strain lineages, namely Beijing and EAI, suggesting that genetic diversity observed during the evolution of isoniazid resistance is not limited to one lineage. The MDR isolates were collected approximately 7 and 11 months after the initial rifampicin mono-resistant samples were collected from patients' 1 and 2, respectively. Unfortunately,

clinical information such as treatment regimens and treatment adherence was not available to us at the time of the study, limiting our ability to draw conclusions regarding mutation rates and selective pressure.

The loss of an *rpoC* polymorphism from the rifampicin mono-resistant isolate from patient 1 (R721) further highlights the importance of selective pressure on defining the genetic population of *M. tuberculosis*. A proportion of 30 % of the R721 population (based on read frequency) contained an *rpoC* mutation while there was no *rpoC* mutation present in the follow-up MDR isolate (R807). While variants in the *rpoC* gene have been speculated to be putative compensatory mutations [26], this mutation may not be important for survival since the selective purification for isoniazid resistance causing mutations resulted in the loss of this putative compensatory mutation. No additional compensatory mutations were identified in the WGS data for the isolates used in this study.

Complementary to our findings are the results observed in a recent study by Eldholm et al. where it was shown that the amount of variation between serial isolates from a single patient may be higher than that observed between two patients in a transmission chain [14]. In addition they observed numerous SNPs within the mycobacterial population which occurred concurrently with drug resistance causing mutations, which were termed 'hitchhiking SNPs'. Excluding these 'hitchhiking SNPs' from their analyses, the amount of variation between sequential samples decreased, suggesting that the selective pressure of drug exposure resulted in a purifying effect. Based on these observations the authors concluded that the presence of populations with high genetic diversity (variation) facilitated the emergence of drug resistance, and that selective pressure may be a driving force in longitudinal genetic diversity [14]. Similarly, a recent study by Bergval et al. hypothesised that a selection event may result in the fixation of either a wild-type or mutant allele which was originally present in only a sub-population of *M. tuberculosis* isolates [19].

Based on our findings and others' work, we propose a model to explain the effect of a selection bottleneck and random mutations on the population structure of *M. tuberculosis* clinical isolates (Fig. 4). We hypothesise that during the course of infection numerous genetic mutations arise within a mycobacterial population and in response to a selective pressure such as antibiotic exposure, clones with pre-existing drug resistant mutations are selected. Cells within the population harbouring drug resistant causing mutations survive while drug sensitive cells die. During this process numerous genetic mutations are lost from the population while mutations which occur concurrently with drug resistance causing mutations (or 'hitchhiking SNPs') remain [14]. The presence of a selective pressure therefore creates a selection bottleneck, altering the level of genetic diversity within the mycobacterial population. Subsequent

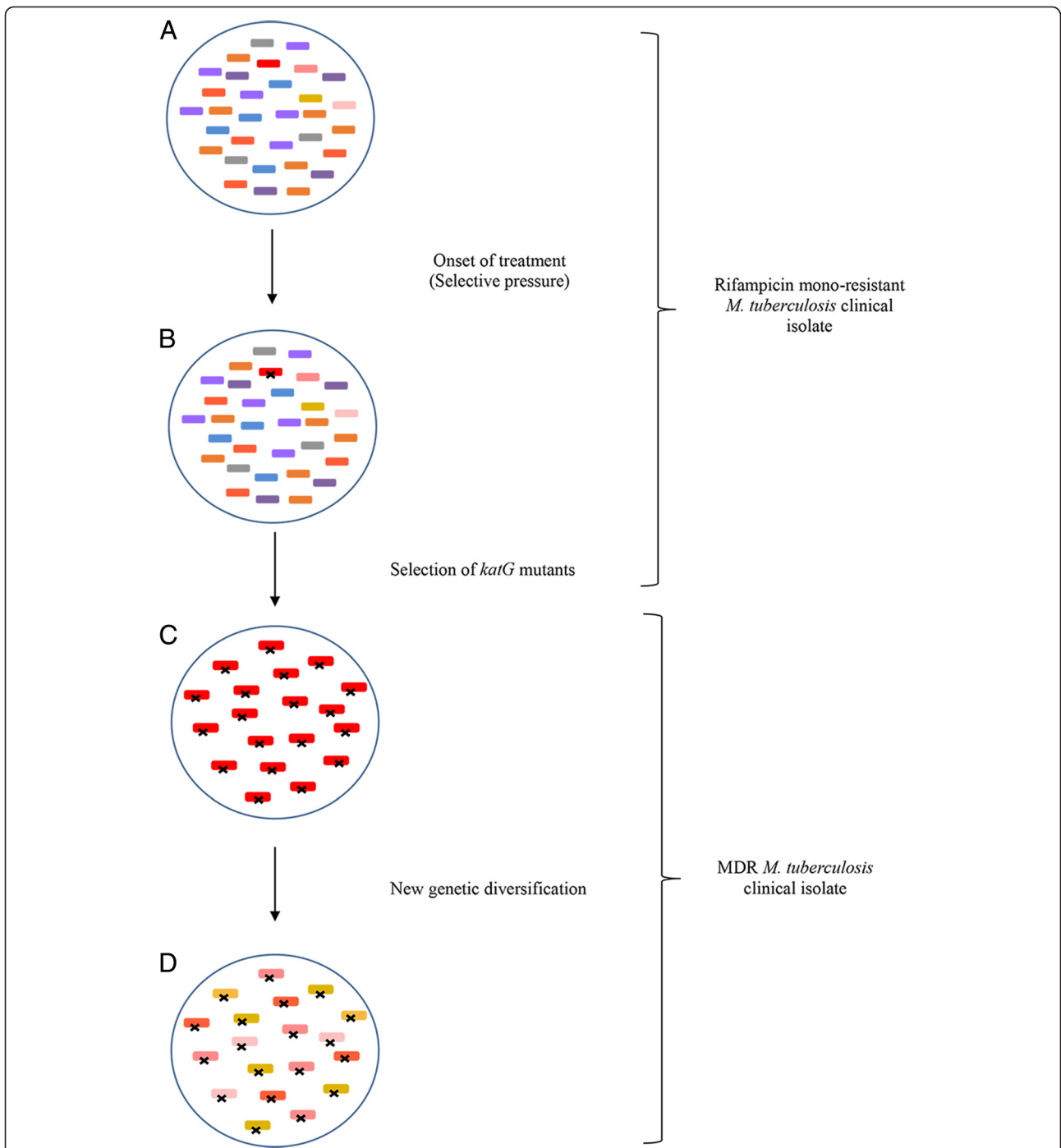


Fig. 4 Proposed model for the effect of a selection bottleneck and random mutations on the population structure of *M. tuberculosis* clinical isolates. **a** A rifampicin mono-resistant clinical *M. tuberculosis* isolate where each cell comprising the population contains an *rpoB* mutation. Numerous other genetic mutations are present thereby creating a diverse population structure. **(b)** Following the onset of treatment the genetic mutations in the population may change, and a spontaneous isoniazid resistance causing (for example *katG* gene or *inhA* promoter) mutation is selected for and becomes dominant within the population. **(c)** Selective pressure of treatment results in the emergence of an isoniazid resistant *M. tuberculosis* population where each cell contains a *katG* mutation. Numerous other genetic mutations are lost during the selection bottleneck resulting in a loss of genetic diversity. **(d)** Subsequent replication cycles and population growth results in new genetic mutations arising within the population allowing for new diversification e.g. R1210. Each cell in this MDR population retains the *rpoB* and *katG* resistance causing mutations. Key: x denotes an isoniazid resistance causing mutation (*katG*)

growth and the emergence of new SNPs allow for an increase in genetic diversity once again (Fig. 4). This subsequent increase in genetic diversity post selective pressure is substantiated by our findings that a rifampicin mono-resistant isolate and its paired MDR isolate each have numerous unique genetic differences i.e. polymorphisms are lost from the original rifampicin mono-resistant population, while different polymorphisms emerge in the MDR isolate.

Conclusions

This study investigated two key aspects involved in the use of next-generation WGS. Firstly, we investigated the confidence in the validity of variants called during bioinformatics analysis. Secondly, we investigated the difference in outputs when utilising a relaxed variant filtering approach compared to the standard filtering approach.

To our knowledge this is the first study that has investigated *M. tuberculosis* genomic diversity using single colonies isolated from clinical isolates. The surprisingly high rate of genetic diversity seen in our study is in concordance with previous studies that used a relaxed variant filtering approaches [5, 12, 14]. During the evolution of drug resistance we observed the emergence and disappearance of numerous variants within a population. Our findings allowed us to formulate a model for the selective bottleneck which occurs during the course of infection, acting as a genomic purification event. Subsequent post-bottleneck mycobacterial growth allows for new genetic diversification to occur. This proposed increase in diversity suggests that the genome is preparing to respond to a changing environment.

Methods

Strain selection and culture

This study was approved by the Health Research Ethics Committee of Stellenbosch University with the waiver of consent to retrospectively collect routine clinical isolates of *M. tuberculosis* and limited demographic and diagnostic data. To ensure patient confidentiality all patient identifiers were removed. Primary rifampicin mono-resistant clinical *M. tuberculosis* isolates were available from 13 patients that were selected from an extensive longitudinal collection of drug resistant *M. tuberculosis* isolates collected in the Western Cape, South Africa. Two of these patients had follow up isolates which were shown to be multi-drug resistant (MDR) by routine drug susceptibility testing (DST). Each of the 15 isolates were subjected to isoniazid and rifampicin DST [27], and Sanger sequencing of the *inhA* promoter and *katG* and *rpoB* genes (Additional file 2) to confirm the resistance phenotype. In addition, each isolate was further genotyped by spoligotyping and IS6110 DNA fingerprinting using internationally standardized techniques [28, 29].

BACTEC™ Mycobacterial Growth Indicator tubes (MGIT™ 960) supplemented with Oleic acid-Albumin-Dextrose-Catalase (OADC) were inoculated with each isolate and incubated in the BACTEC™ MGIT™ 960 instrument at 37 °C. Following an indication of growth positivity i.e. when a growth unit of 400 was reached, each MGIT was incubated at 37 °C for an additional 5 days to allow for optimal mycobacterial growth. A volume of 500 µl of positive culture was then used to inoculate a starter culture of 10 ml of 7H9 Middlebrook medium (Becton, Dickinson Microbiology system, Sparks, USA), supplemented with 10 % albumin-dextrose-catalase (ADC), 0.2 % (v/v) glycerol (Merck Laboratories, Saarchem, Gauteng, SA) and 0.1 % Tween80 (Becton, Microbiology systems, Sparks, USA). Subsequently, the starter cultures were grown in filtered screw cap tissue culture flasks (Greiner Bio-one, Maybach Street, Germany) without shaking at 37 °C until an optical density (OD₆₀₀) of 0.6–0.8 was reached. Contamination was assessed by Ziehl-Neelsen (ZN) staining and the plating of cultures onto blood agar plates. A 100 µl aliquot of each starter culture for the rifampicin mono-resistant ($n = 13$) and paired MDR *M. tuberculosis* isolates ($n = 2$) was plated on 7H10 solid media supplemented with OADC for DNA extraction. Serial dilutions were prepared and plated on 7H10 solid media supplemented with OADC and 2 µg/ml rifampicin for selection of single colonies.

Selection of single colonies

Single colonies were randomly selected from solid media (7H10 Middlebrook media supplemented with OADC) containing 2 µg/ml into Middlebrook 7H9 media supplemented with ADC and 0.1 % Tween80, and statically cultured to an OD₆₀₀ of above 0.8. Each single colony culture was then sub-cultured on solid media (7H10 Middlebrook media supplemented with OADC) for DNA extraction. Unnecessary sub-culturing steps were avoided to minimize the appearance of *in vitro* adaptive mutations.

DNA extraction and whole genome sequencing

Genomic DNA was isolated from single colonies for each rifampicin mono-resistant *M. tuberculosis* isolate according to standard protocols [30]. In addition, DNA was isolated from the primary cultures of the paired *M. tuberculosis* isolates demonstrating intra-patient evolution of isoniazid resistance i.e. the parental rifampicin mono-resistant isolate and a follow-up MDR isolate. Sequencing libraries for all isolates were constructed using the standard genomic DNA sample preparation kits from Illumina (Illumina, Inc, San Diego, CA), according to manufacturer's instructions. The whole genomes of the *M. tuberculosis* isolates were sequenced using either the Illumina MiSeq or Illumina HiSeq platforms.

Mapping and variant detection

An in-house automated pipeline for *M. tuberculosis* next generation sequencing (NGS) analysis was adapted to allow analysis of the sequencing data (Van der Merwe et al., manuscript in preparation). The steps involved in the pipeline are described below.

Quality assessment of the sequencing data (in FASTQ format) was done using FASTQC [31], followed by trimming of adapters and low-quality bases with a Phred quality score of less than 20 and filtering for a minimum read length of 36 using Trimmomatic [32]. A minimum read length of 36 base pairs was used for subsequent mapping. Reads were then mapped to the *M. tuberculosis* H37Rv genome (Genbank: AL123456) using three different mappers namely the Burrows-Wheeler Alignment Tool (BWA) [33], Novoalign [34] and SMALT [35]. For all libraries sequenced, over 98 % of the reference genome was covered by at least one read and an average depth of coverage of 137× was achieved. The alignment files were subjected to local realignment and de-duplication using the Genome Analysis Toolkit (GATK) [36] and Picard [37]. Variants (Single nucleotide polymorphisms (SNPs) and Insertion/Deletions (In/Dels)) in coding as well as non-coding regions were then called from each alignment file using GATK [36], and the overlap of variants identified from all three alignment files was used for further analysis. Variants were annotated using annotation data from TubercuList [38]. Variants in repetitive regions, such as *pe/ppe* and *pe_pgrs* gene families, were removed from subsequent analysis.

In this study we made use of a relaxed variant filtering approach to allow identification of variants occurring at varying read frequencies i.e. variants present at varying proportions within the population. Variants detected by GATK were filtered for a minimum read depth of 50. An initial read frequency was not defined to allow for the validation of varying frequencies with Sanger sequencing.

Identified variants were compared between the appropriate single colonies selected from the same clinical isolate. Variants present at a frequency ranging from 20–100 % were selected for validation with Sanger sequencing. Variants selected had a minimum read depth of 50 and a mapping quality score of above 50. All variants identified in this study were manually inspected using GenomeView [39].

Confirmation of variants with Sanger sequencing

A subset of 46 randomly selected genomic variants with read frequencies ranging from 20 to 100 % were validated by targeted PCR and Sanger sequencing. Oligonucleotide primers (Additional file 2) were designed using Primer3 [40] to amplify 300–600 base pairs regions flanking the variant of interest. Briefly, an aliquot of genomic DNA was added to the following reaction mix containing: 1×Q

buffer, 1× PCR buffer, 2 mM MgCl₂, 0.4 mM dNTPs, 50 μM of each primer (Additional file 2) and 1.25U Hot Star Taq polymerase (Qiagen, San Diego, CA, USA). Amplification was done under the following thermocycling conditions: 15 min (min) denaturation at 95 °C followed by 40 amplification cycles (each cycle: 94 °C for 1 min, 62 °C for 1 min, 1 min extension at 72 °C) and an elongation step of 10 min at 72 °C. PCR products were purified and sequenced with the ABI PRISM DNA Sequencer model 377, Perkin Elmer. Sequence polymorphisms were identified by comparing the consensus sequence of each isolate to the corresponding gene sequence of the *M. tuberculosis* H37Rv genome using BioEdit (v7.1.3) [41].

Sequencing results were first inspected for the presence of true variants using the ClustalW multiple alignment tool in the BioEdit software. Furthermore, the chromatograms of each sequencing file were inspected for the presence of both a wild-type and mutant peak to identify heterogeneous variants present at a low percentage within the population. Chromatograms were visually inspected for the presence of mixed peaks at the variant position identified by WGS.

To define a cut-off for the reliability of WGS results for heterogeneous variants a receiver operating characteristic (ROC) analysis was performed using IBM SPSS Statistics version 22 (IBM Corp 2013) [42].

Availability of supporting data

The data sets supporting the results of this article are available in the European Nucleotide Archive with the following accession number: PRJEB9976 and are available at: <http://www.ebi.ac.uk/ena/data/view/PRJEB9976>.

The data sets supporting the results for this article are included within the article (and its additional files). Additional file 2(.xls) (Primers used for PCR amplification and Sanger sequencing) is a table listing all primers used in the study to validate variants. Additional file 1 (.xls) (Unique variants identified in the single colony comparison analysis) contains tables of variants identified in the comparison analysis of the genomes of the single colonies isolated from clinical specimens.

Additional files

Additional file 1: Unique variants identified in the single colony comparison analysis. (XLSX 45 kb)

Additional file 2: Primers used for PCR amplification and Sanger sequencing. (XLSX 15 kb)

Abbreviations

ADC: Albumin Dextrose Catalase; AUC: Area Under the Curve; BWA: Burrow Wheels Aligner; DNA: Deoxyribonucleic Acid; dNTPs: Deoxyribonucleotide Triphosphate; DST: Drug Susceptibility Testing; EAI: East Africa Indian; GATK: Genome Analysis ToolKit; In/Dels: Insertions/Deletions; LAM: Latin American Mediterranean; LCC: Low Copy Clade; MgCl₂: Magnesium Chloride; MGIT: Mycobacterial Growth Indicator Units; MDR: Multi-drug Resistant;

NGS: Next generation sequencing; OADC: Oleic acid albumin dextrose catalase; OD: Optical Density; PCR: Polymerase Chain Reaction; ROC: Receiver operating characteristic; SNPs: Single Nucleotide Polymorphisms; TB: Tuberculosis; WGS: Whole Genome Sequencing; XDR: Extensively Drug Resistant; ZN: Ziehl-Neelsen.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PAB and MdV participated in the planning and design of this study, carried out molecular genetic experiments, performed bioinformatics analysis (data analysis) and were involved in manuscript preparation and editing. GEL participated in the planning and design of this study, carried out molecular genetic experiments and was involved in manuscript preparation and editing. RGvdM and AD performed bioinformatics analysis (data analysis) and was involved in manuscript preparation and editing. EMS and SLS was involved in manuscript preparation and editing. TCV contributed to the reagents, materials and analytical tools utilised in this study. AMA carried out molecular genetic experiments, contributed to the reagents, materials and analytical tools utilised in this study. TD and JAS contributed the *M. tuberculosis* isolates and phenotypic data used in this study. PDvH and AP contributed to the reagents, materials and analytical tools utilised in this study and was involved in manuscript preparation and editing. RMW participated in the planning and design of this study and was involved in manuscript preparation and editing. All authors read and approved the final manuscript.

Acknowledgments

We thank Marianna De Kock and Ruzayda van Aarde for their technical support. We thank Moleen Zunza at the Centre for Evidence Based Health Care for statistical support.

This work was supported by the South African National Research Foundation (NRF), Harry Crossley Foundation, South Africa Medical Research Council (MRC), Claude Leon Foundation, the Department of Biomedical Sciences, Stellenbosch University and King Abdullah University of Science and Technology (KAUST). SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa, award number UID 86539. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NRF.

Author details

¹DST-NRF Centre of Excellence for Biomedical Tuberculosis Research/SA MRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa. ²Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. ³National Health Laboratory Services, Green Point, Cape Town, South Africa.

Received: 25 May 2015 Accepted: 13 October 2015

Published online: 24 October 2015

References

- Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet.* 2010;42(6):498–503.
- Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, et al. Novel *Mycobacterium tuberculosis* complex isolate from a wild chimpanzee. *Emerg Infect Dis.* 2013;19(6):969–76.
- Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet.* 2013;45(10):1176–82.
- Trauner A, Borrell S, Reither K, Gagneux S. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. *Drugs.* 2014;74(10):1063–72.
- Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, et al. Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J Infect Dis.* 2012;206(11):1724–33.
- Sandegren L, Groenheit R, Koivula T, Ghebremichael S, Advani A, Castro E, et al. Genomic stability over 9 years of an isoniazid resistant *Mycobacterium tuberculosis* outbreak strain in Sweden. *PLoS One.* 2011;6(1):e16647.
- Saunders NJ, Trivedi UH, Thomson ML, Doig C, Laurenson IF, Blaxter ML. Deep resequencing of serial sputum isolates of *Mycobacterium tuberculosis* during therapeutic failure due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. *J Infect.* 2011;62(3):212–7.
- Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet.* 2014;46(3):279–86.
- Schurch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, et al. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. *Infect Genet Evol.* 2010;10(1):108–14.
- Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med.* 2013;1(10):786–92.
- Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, Rusch-Gerdes S, et al. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. *PLoS One.* 2013;8(12):e82551.
- Perez-Lago L, Comas I, Navarro Y, Gonzalez-Candelas F, Herranz M, Bouza E, et al. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. *J Infect Dis.* 2014;209(1):98–108.
- Mariam SH, Werngren J, Aronsson J, Hoffner S, Andersson DI. Dynamics of antibiotic resistant *Mycobacterium tuberculosis* during long-term infection and antibiotic treatment. *PLoS One.* 2011;6(6):e21147.
- Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA, et al. Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biol.* 2014;15(11):490.
- Meacci F, Orru G, Iona E, Giannoni F, Piersimoni C, Pozzi G, et al. Drug resistance evolution of a *Mycobacterium tuberculosis* strain from a noncompliant patient. *J Clin Microbiol.* 2005;43(7):3114–20.
- Fortune SM. The surprising diversity of *Mycobacterium tuberculosis*: change you can believe in. *J Infect Dis.* 2012;206(11):1642–4.
- Mehaffy C, Guthrie JL, Alexander DC, Stuart R, Rea E, Jamieson FB. Marked microevolution of a unique *Mycobacterium tuberculosis* strain in 17 years of ongoing transmission in a high risk population. *PLoS One.* 2014;9(11):e112928.
- Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med.* 2014;2(4):285–92.
- Bergval I, Coll F, Schuitema A, de Ronde H, Mallard K, Pain A, et al. A proportion of mutations fixed in the genomes of in vitro selected isogenic drug-resistant *Mycobacterium tuberculosis* mutants can be detected as minority variants in the parent culture. *FEMS Microbiol Lett.* 2015;362(2):1–7.
- loerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, et al. Variation among genome sequences of H37Rv strains of *Mycobacterium tuberculosis* from multiple laboratories. *J Bacteriol.* 2010;192(14):3645–53.
- Didelot X, Gady J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol.* 2014;31(7):1869–79.
- Liu F, Hu Y, Wang Q, Li HM, Gao GF, Liu CH, et al. Comparative genomic analysis of *Mycobacterium tuberculosis* clinical isolates. *BMC Genomics.* 2014;15:469.
- Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. *PLoS One.* 2014;9(3):e91024.
- Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog.* 2011;7(3):e1001307.
- Portevin D, Sukumar S, Coscolla M, Shui G, Li B, Guan XL, et al. Lipidomics and genomics of *Mycobacterium tuberculosis* reveal lineage-specific trends in mycolic acid biosynthesis. *Microbiol Open.* 2014;3(6):823–35.
- Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet.* 2012;44(1):106–10.
- Siddiqi S, Ahmed A, Asif S, Behera D, Javaid M, Jani J, et al. Direct drug susceptibility testing of *Mycobacterium tuberculosis* for rapid detection of

- multidrug resistance using the Bactec MGIT 960 system: a multicenter study. *J Clin Microbiol.* 2012;50(2):435–40.
28. Warren RM, van Helden PD, van Pittius NC G. Insertion element IS6110-based restriction fragment length polymorphism genotyping of *Mycobacterium tuberculosis*. *Methods Mol Biol.* 2009;465:353–70.
 29. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol.* 1997;35(4):907–14.
 30. Warren R, de Kock M, Engelke E, Myburgh R, Gey Van Pittius N, Victor T, et al. Safe *Mycobacterium tuberculosis* DNA extraction method that does not compromise integrity. *J Clin Microbiol.* 2006;44(1):254–6.
 31. FastQC [<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>]. Accessed 20 June 2014.
 32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20.
 33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
 34. Novoalign [<http://www.novocraft.com/main/page.php?s=novoalign>]. Accessed 20 June 2014.
 35. Pongstingl H, Ning Z: SMALT - A new mapper for DNA sequencing reads. F1000Posters 2015, 1. <http://www.sanger.ac.uk/resources/software/smalt/>.
 36. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
 37. Picard [<http://broadinstitute.github.io/picard/>]. Accessed 20 June 2014.
 38. Lew JM, Kapopoulou A, Jones LM, Cole ST. TuberculList–10 years after. *Tuberculosis.* 2011;91(1):1–7.
 39. Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y. GenomeView: a next-generation genome browser. *Nucleic Acids Res.* 2012;40(2):e12.
 40. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* 2012;40(15):e115.
 41. BioEdit: Biological sequence alignment editor for Win95/98/NT/2 K/XP/7 [<http://www.mbio.ncsu.edu/bioedit/bioedit.html>]. Accessed 20 June 2014.
 42. IBM Corp. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. Support documentation for the software can be found at the following web address. Released 2013. <http://www-01.ibm.com/support/docview.wss?uid=swg21646821>.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

