DE NOVO ASSEMBLY AND BIOINFORMATICS ANALYSIS OF *M.TUBERCULOSIS* MDR STRAINS SEQUENCED ON NEXT-GENERATION SEQUENCING PLATFORM

A.Daniyarov[1], U.Kairov[1], A.Molkenov[1], U.Kozhamkulov[2], A.Akhmetova[2], A.Akilzhanova[2], Zh.Zhumadilov[2].

[1] Laboratory of Bioinformatics and Systems Biology, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan

[2] Laboratory of Genomic and Personalized Medicine, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan

INTRODUCTION

Tuberculosis remains one of the major problems in public health (1). During the last 10 years the tuberculosis incidence and mortality rate in Kazakhstan decreased by 2.4 times and 6 times, respectively. Despite the decreasing rate of incidences in Kazakhstan last years, the rate of multidrug-resistant (MDR) forms is increasing.

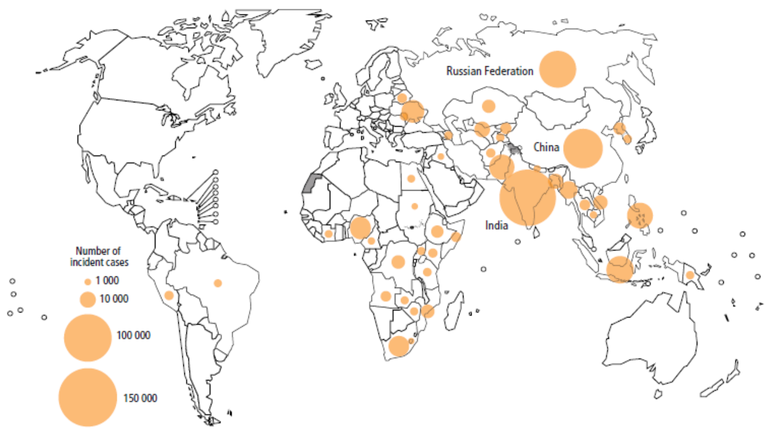


Figure 1. Estimated incidence of MDR/RR-TB in 2016, for countries with at least 1000 incident cases

According to WHO European Region report, Kazakhstan is 1 of 18 countries with a high rate of MDR tuberculosis (2). Some previous studies have been performed in Kazakhstan using genotyping and sequencing methods (3-5).

It is extremely important to examine sensitive and resistant strains with different mutations in genes encoding drug metabolism among M.tuberculosis strains from the different geographic regions.

MATERIALS AND METHODS

- Prepared genomic libraries of 7 MDR strains were sequenced on Roche 454 GS FLX Titanium NGS platform at the Centre for Life Sciences, NLA-NU.
- FastQC was applied to analyze reads quality and adapters were trimmed using Trimmomatic v.0.38 to truncate low quality reads.
- De novo assembling has been performed by Velvet v.1.2.10.
- Gene prediction and annotation were carried out with Genome Annotation Service PATRIC, using RAST toolkit.
- High quality reads were then mapped to the *M.tuberculosis H37Rv* genome (GenBank AL123456.3) using GS Reference Mapper v.2.8. The alignments files were subjected to local realignment and de-duplication using GS Reference Mapper v.2.8 and MUMmer v.3.23. Identified genomic variants were annotated using in-house prepared Python script.
- Comparative phylogenetic analysis has been performed by Maximum-Likelihood method based on Tamura-Nei model in MEGA X.

CONCLUSION

We performed comprehensive bioinformatics analysis of whole-genome data from 7 MDR isolates. We identified genomic variants (SNPs and InDels) in *de novo* assembled and annotated whole-genomes and specific/novel variants in drug-resistant genes of MDR strains circulated in Kazakhstan.

These findings may provide additional source for justification of drug sensitivity before clinical treatment as well as the basis for expansion of the current reference MTB database with genetic variability among different drug sensitivity isolates. Moreover, these results may provide supplementary information in further fundamental investigation of virulence and transmissibility patterns of MDR strains.

RESULTS

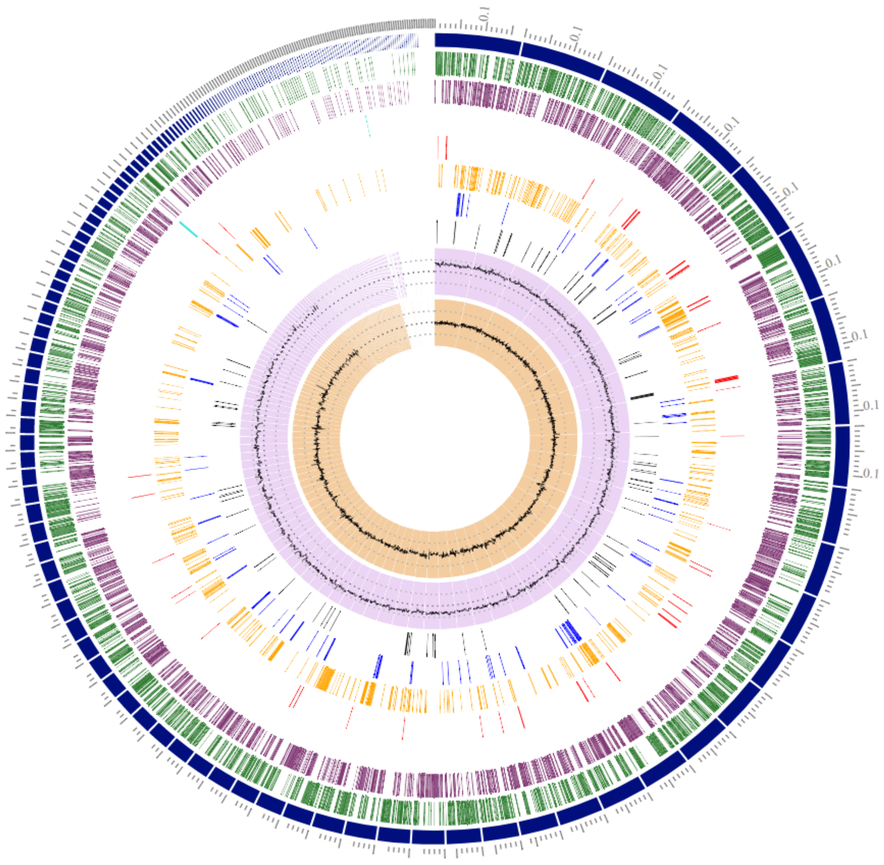
The whole-genome sequencing for 7 MDR isolates produced 667,052,381 paired reads with average read length 520 bp. The mean coverage ranged between 21-51X, and the mapping quality 94.43-97.14%. From 4385 to 4475 coding sequences CDSs and 44 tRNAs, 4 rRNA have been identified. Number of SNPs and InDels ranging from 1581 to 1952, and from 213 to 297, respectively (Table 1). Spoligotyping analysis based on NGS data revealed all MDR strains as Beijing genotype. Phylogenetic analysis (based on rpoB gene) of the five MDR strains in comparison with seven susceptible/drug-resistant strains (H37Ra, H37Rv, KZN 4207, CDC1551 and other) has been showed clustering on two main clusters (Figure 2). Among all isolates we detected several new genetic variants in drug-resistance genes which are not described/unknown in antibiotic resistance databases (ResFinder v.3.0; CARD; CASTB). Circular maps for each MDR isolate have been created (Figure 3 – representation of circular map for isolate MDR-1280).

Table 1. Genome assembling statistics of *M.tuberculosis* isolates

Isolate ID	Total raw Paired Read	Contigs	N50	Completion (%)	Full length (bp)	Coverage (×)	SNPs*	Indels*	GC (%)	Genes
1280	195,048,344	136	90951	97.14	4 287 916	51 ×	1952	225	65.46	4385
1405	126,653,863	189	50264	96.99	4 281 841	34 ×	1888	235	65.44	4414
1524	160,749,277	166	51640	96.94	4 280 264	43 ×	1908	213	65.44	4417
1525	118,335,745	213	37986	96.80	4 273 127	31 ×	1827	254	65.44	4460
1577	120,396,884	215	44245	96.91	4 278 038	33 ×	1876	239	65.44	4475
1585	85,923,603	404	18929	96.20	4 246 265	23 ×	1581	297	65.40	4623
1713	113,788,645	250	35572	96.76	4 271 719	31 ×	1763	251	65.44	4516

*The number of SNPs and Indels are assessed relative to H37Rv.

Figure 3. Summary of *Mycobacterium tuberculosis* sequencing and assembly data for MDR-1280 isolate. Concentric circles show the different features of isolate. From top to bottom – Position Lable (Mbp), Contigs (deep blue), CDS – forward strand (green), CDS – reverse strand (violet), Non-CDS Features (turquois), AMR Genes (red), VF Genes (orange), Transporters (blue), Drug Targets (black), GC Content (GC Skew).



By closely examining the correlation of the phenotypic drug susceptibility profiles of the strains with mutations identified in their drug resistance-associated genes, we identified a few potential new genetic determinants of drug resistance. For example, one isolate (MTB-1405) had *gyrB* D461N mutation known to confer resistance to fluoroquinolones according to ResFinder DB. But, from the other hand all 7 MDR isolates annotated by 3 DBs as susceptible strains. In our 7 MDR isolates we found unknown/novel genetic variants in gene *gyrA* (E21Q, G668D, and S95T) that can be potential determinants of susceptibility to fluoroquinolones among MDR strains. Results of drug resistance profiling among investigated 7 MDR isolates and identified known/unknown genetic variants in antimicrobial resistance genes have been shown in Table 2 and 3.

Figure 2. Phylogenetic tree showing the relationships of 7 isolates of *M. tuberculosis* with other Mycobacterium species based on aligned sequences of the *rpoB* gene.

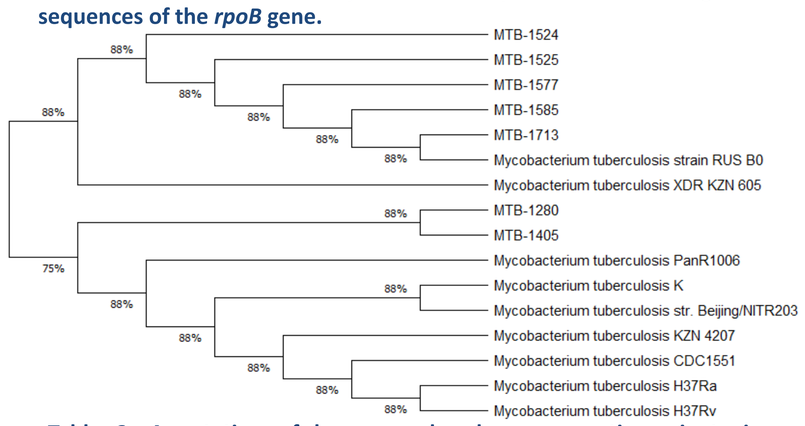


Table 3. Annotation of known and unknown genetic variants in antimicrobial resistance genes by ResFinder.

Isolate	Drug	Known target: Amino Acid Mutation	Unknown target: Amino Acid Mutation
MTB-MDR-KZ (1280)	INH [1]	katG: S315T	katG: R463L
	RIF [2]	rpoB: S450L	
	EMB [3]	embB: M306V	
			thyA: P253A
			gyrA: E21Q, S95T, G668D
			rpoC: E1092D
			ethA: T314I
			kasA: K34Q
MTB-MDR-KZ (1405)	INH [1]	katG: S315T	katG: R463L
	PZA [4]	pncA: G24D	
	SM [5]	rpsL: K43R	
	FLQ [6]	gyrB: D461N	gyrA: E21Q, S95T, G668D
	RIF [2]	rpoB: S450L	rpoB: H723D, rpoC: E1092D
	EMB [3]	embB: M306V	
MTB-MDR-KZ (1524)	INH [1]	katG: S315T	katG: R463L
	PZA [4]	pncA: L182S	pncA: R176C, T177A, A178P, S179P, V180A, E181S, V183W, C184F, S185A, S186A
	SM [5]	rpsL: K43R	
	RIF [2]	rpoB: S450L	rpoC: E1092D
	EMB [3]	embB: M306V	
			thyA: P253A
			gyrA: E21Q, S95T, G668D
			ethA: T314I
MTB-MDR-KZ (1525)	INH [1]	katG: S315T	katG: R463L
	PZA [4]	pncA: T76P	
	SM [5]	rpsL: K43R	
	RIF [2]	rpoB: S450L	rpoC: E1092D
	EMB [3]	embB: M306V	
			thyA: P253A
			gyrA: E21Q, S95T, G668D
			ethA: T314I
MTB-MDR-KZ (1577)	INH [1]	katG: S315T	katG: R463L
	PZA [4]	pncA: promoter n.-11	
	SM [5]	rpsL: K43R	
	RIF [2]	rpoB: S450L	
	EMB [3]	embB: M306V	
			thyA: P253A
			gyrA: E21Q, S95T, G668D
MTB-MDR-KZ (1585)	INH [1]	katG: S315T	katG: R463L
	SM [5]	rpsL: K43R	
	AMK [7], CM [8], KM [9]	rrs: 1401	
	RIF [2]	rpoB: S450L	rpoC: N698S, E1092D
	EMB [3]	embB: M306V	
			pncA: V155M
			rpsA: M432T
			thyA: P253A
			gyrA: E21Q, S95T, G668D
			ethA: T314I
MTB-MDR-KZ (1713)	INH [1]	katG: S315T	katG: R463L
	PZA [4]	pncA: T76P	
	SM [5]	rpsL: K43R	
	RIF [2]	rpoB: S450L	
	EMB [3]	embB: M306V	
			thyA: P253A
			fabG1: E76E – Frameshift, E77S, H78T, Q79R, G80V, P81R, V82S, E83R, V84C, L85W, V86C, S87P, N88T, A89P

[1] isoniazid, [2] rifampicin, [3] ethambutol, [4] pyrazinamide, [5] streptomycin, [6] fluoroquinolones, [7] amikacin, [8] capreomycin, [9] kanamycin

REFERENCES

1. WHO. World health statistics, http://www.who.int/gho/publications/world_health_statistics/2018/en/; 2018.
2. Ministry of Health Republic of Kazakhstan. Statistic TB review (Editor Sh. Ismailov). National TB Center, Almaty, 2009, 2011 (in Russian).
3. Filliol I, Motiwala AS, Cavatore M, et al. Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol 2006;188(2):759e72.
4. 454 Sequencing System Software Manual Version 2.6. Part C: GS De Novo Assembler, GS Reference Mapper, SFF Tools. 454 Life Sciences Corp. 2011
5. World Health Organization. 2017. Global tuberculosis report 2017. World Health Organization, Geneva, Switzerland.
6. Ministry of Healthcare and Social Development of the Republic of Kazakhstan. 2014. Statistic TB review. Ministry of Healthcare and Social Development of the Republic of Kazakhstan, Almaty, Kazakhstan. (In Russian.)
7. Kozhamkulov U, Akhmetova A, Rakhimova S, Belova E, Alenova A, Bismilda V, Chingissova L, Ismailov S, Ramanculov E, Momynaliev K. 2011. Molecular characterization of rifampicin- and isoniazid-resistant Mycobacterium tuberculosis strains isolated in Kazakhstan. Jpn J Infect Dis 64:253–255.
8. Hillemann D, Kubica T, Agzamova R, Venera B, Rüsch-Gerdes S, Niemann S. 2005. Rifampicin and isoniazid resistance mutations in Mycobacterium tuberculosis strains isolated from patients in Kazakhstan. Int J Tuberc Lung Dis 9:1161–1167.
9. Kairov U. et al. Draft Genome Sequences of Two Clinical Isolates of Mycobacterium tuberculosis from Sputum of Kazakh Patients. Genome Announcements. 2015 May-Jun; 3(3): e00466-15.

Table 2. Drug resistance profiling of Mycobacterium tuberculosis isolates

Isolate fields		Drug Resistance Analysis									Microbiology	Confirmed by analysis genotype
Sample ID	Isolate ID	AMK	EMB	INH	OFX	PZA	RIF	SM	KM	CM		
MTB-1280	1280	-/S/-	R/R/R	R/R/R	-/S/S	-/-/R	R/R/R	-/R/S	-/S/-	-/S/-	MDR	MDR
MTB-1405	1405	-/S/-	R/R/R	R/R/R	-/S/S	R/-/-	R/R/R	R/R/R	-/S/-	-/S/-	MDR	MDR
MTB-1524	1524	-/S/-	R/R/R	R/R/R	-/S/S	R/-/-	R/R/R	R/R/R	-/S/-	-/S/-	MDR	MDR
MTB-1525	1525	-/S/-	R/R/R	-/R/R	-/S/S	R/-/-	R/R/R	R/R/R	-/S/-	-/S/-	MDR	MDR
MTB-1577	1577	-/S/-	R/R/R	R/R/R	-/S/S	R/-/R	R/R/R	R/R/R	-/S/-	-/S/-	MDR	MDR
MTB-1585	1585	R/R/R	R/S/R	R/R/R	-/S/S	-/-/	R/R/R	R/R/R	R/R/-	R/R/-	MDR	MDR
MTB-1713	1713	-/S/-	R/R/R	R/R/R	-/S/S	R/-/-	R/R/R	R/R/R	-/S/-	-/S/-	MDR	MDR

AMK - amikacin, EMB - ethambutol, INH - isoniazid, OFX - ofloxacin, PZA - pyrazinamide, RIF - rifampin, SM - streptomycin, KM - kanamycin, CM – capreomycin

Ulykbek Kairov, PhD, Laboratory of Bioinformatics and Systems Biology, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan
Tel: +7(7172)709312, e-mail: ulykbek.kairov@nu.edu.kz, http://www.nu.edu.kz