White Paper Application

Project Title: Comparative sequence analysis of *M.tuberculosis* isolates from Patients with multi- and extensively drug resistant tuberculosis in Belarus

Authors: **Alena Skrahina, Oksana Zalutskaya**

Primary Investigator Contact:

Name	Alena Skrahina
Position	Scientific Director
Institution	Republican Research and Practical Centre for Pulmonology and TB
Address	Dolginovski trakt 157
State	Belarus
ZIP Code	220053
Telephone	+375 17 2898356 mob. +375 29 6799871
e-mail	Alena_skrahina@tut.by

1. Executive Summary

The rise of multidrug resistant tuberculosis (MDR-TB, defined as resistant to at least rifampicin and isoniazid) and extremely drug resistant tuberculosis (XDR, defined as resistant to rifampicin, isoniazid, any fluoroquinolone and to at least 1 of the following 3 drugs: capreomycin, kanamycin, and amikacin) is a severe threat to effective TB control as well as to successful treatment of individual patients. The concern that these strains could spread around the world further stresses the need for additional control measures, such as new diagnostic methods, better drugs for treatment, and a more effective vaccine. Patients harboring MDR strains of *Mycobacterium tuberculosis* (Mtb) need to be entered into alternative treatment regimens involving second-line drugs that are more costly, more toxic, and less effective. XDR-TB now constitutes an emerging threat for the control of the disease and the further spread of drug resistance, especially in HIV-infected patients [6, 7]. The World Health Organization (WHO) estimated that globally more than 500,000 TB patients are infected with MDR strains of Mtb [1-5].

The highest incidence of MDR-TB (about 20% of new and 60% of re-treatment cases) is reported in Eastern Europe for some of the countries of the former Soviet Union. For example, in Belarus, the level of MDR-TB is found to be highest in the world. **To date there are no studies that have examined the genomic composition of *M. tuberculosis* isolates from the former Soviet Union.**

We propose sequencing 60 genomes of XDR-TB and MDR-TB from samples collected in Belarus in order to:

1. Further investigate differences between genomes of TB strains with varied virulence, clinical manifestation of disease, resistance to drugs;
2. Reveal phylogenetic/phylogeographic peculiarities of Mtb strains in Belarus;
3. Perform comparative analysis of MDR- and XDR-TB strains from Belarus with strains of *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551 and strains from other countries;
4. Add to the existing body of information and knowledge to promote research resulting in creating new generation of TB drugs, vaccines, and diagnostics.

The results of this work will be unique because:

a. XDR and MBR Mtb genomes from Belarus samples have never been sequenced before.

b. The availability of collected clinical metadata that describes patients' history will allow for the selection of strains from hundreds of samples, providing a unique opportunity to study the variability and dynamics of TB genome mutations.

2. Justification

New hot spots of MDR-TB are documented every year [8]. Countries of the former Soviet Union have been among the most severely affected by this epidemic. A survey conducted in 2010 in Minsk, the capital of Belarus, by the National TB Control Program in collaboration with the WHO showed alarming levels of drug-resistance with nearly one out of two patients with TB affected by MDR-TB. This proportion of MDR in TB patients was the highest ever recorded worldwide [9]. Sequencing of genomes for MDR-TB and XDR-TB is essential, as the presence of expected sequence diversity in *M. tuberculosis* would provide a basis for understanding pathogenesis, immune mechanisms, and bacterial evolution. The bacterial factors that contribute to disease severity and type, in addition to host genetics, and the environment, remain still largely ill-defined. Understanding the mechanisms of drug resistance, virulence, spreading of Mtb, manifestation and clinical course of TB disease, based on a full genomic analysis is extremely important both for Belarus health programs and to support worldwide efforts to combat the disease.

Huge progress in TB research was made with the availability of the genomic sequence of the Mtb H37Rv type strain. Mtb H37Rv [10] was revealed to possess a sequence of 4,411,529 bp, the second largest microbial genome sequenced at that time. The characteristically high guanine plus cytosine (G+C content; 65.5%) was found to be uniform throughout most of the genome, confirming the hypothesis that horizontal gene transfer events are virtually absent in modern Mtb [11]. One of the most thoroughly studied characteristic of Mtb is the presence and distribution of insertion sequences (IS). Of particular interest is IS6110, which has been widely used for strain typing and molecular epidemiology due to its variation in insertion site and copy number [12]. It was determined that Mtb H37Rv codes for 3,924 ORFs, accounting for 91% of the coding capacity of the genome [11]. A bias in the overall orientation of genes with respect to the direction of replication was also found. On average, bacteria such as *B. subtilis* have 75% of their genes in the same orientation as that of the replication fork, while *M. tuberculosis* only has 59%. This finding led to the hypothesis that such a bias could be part of the slow-growing phenotype of the tubercle bacillus [13].

Genomic comparison has shown that gene content can vary between strains of *M. tuberculosis*. The analysis of complete genome sequences from clinical isolates identified that single nucleotide polymorphisms (SNPs), large sequence polymorphisms (LSPs), and regions of difference (RDs) originate from small deletions, deletions in homologous repetitive elements, point mutations, genome rearrangements, frame-shift mutations, and multi-copy genes [14, 15]. Fleischeman et al. suggested that genetic variation among Mtb strains might denote selective pressure and therefore might play an important role in bacterial pathogenesis and immunity [15]. Although associations between host and pathogen populations seem to be highly stable, the evolutionary, epidemiological, and clinical relevance of genomic deletions and regions of genetic variation remain ill-defined, as do the molecular basis of virulence and transmissibility [16].

Up to six Mtb lineages adapted to specific human populations were described by Gagneux et al. using comparative genomics and molecular genotyping tools: the Indo-Oceanic lineage, the East-Asian lineage, the East-African-Indian lineage, the Euro-American lineage, and two West-African lineages [17]. One family of Mtb strains, Beijing, has attracted special attention. This hypervirulent family is reported to be common in several Asian studies and may possess selective advantages compared to other genotypes [18, 19]. This family is also more often associated with multi-drug resistance[20]. Specific deletions associated with the hypervirulent Beijing/W strains of M. tuberculosis were identified [21].

The genome of the *M. tuberculosis* laboratory strain H37Rv was completely sequenced (GenBank accession no. NC_000962) and compared to the complete genome sequence of *M. tuberculosis* strain CDC1551. (1,2). The circular representation of the *M. tuberculosis* chromosome illustrated in Fig.1 depicts the location of each predicted protein coding region as well as selected features differing between the CDC1551 and H37Rv strains, including large sequencing polymorphisms (LSPs) and single nucleotide polymorphisms (SNPs).

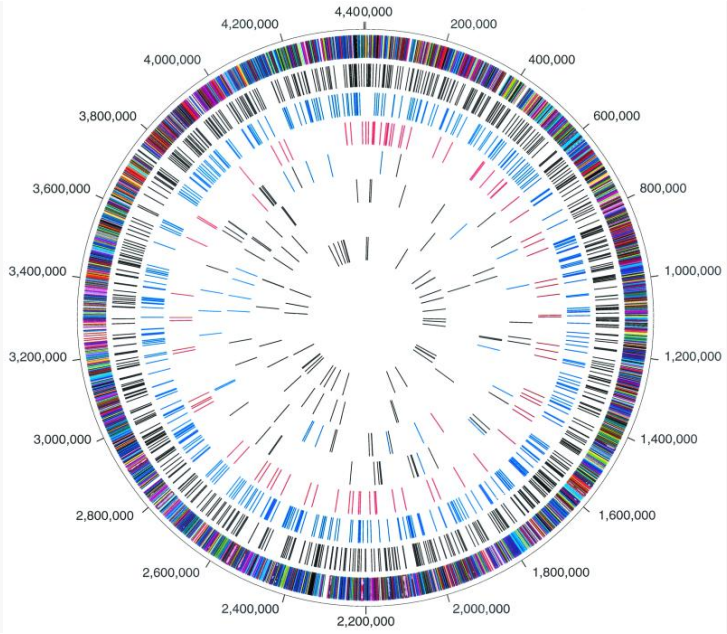

FIG.1.
Circular representation of the *M. tuberculosis* chromosome illustrating the location of each predicted protein-coding region as well as selected features differing between the CDC1551 and H37Rv strains. The outer concentric circle shows predicted protein-coding regions on both strands, color coded according to role category. The second concentric circle shows the location of nonsynonymous substitutions (black). The third concentric circle shows the location of synonymous substitutions (blue). The fourth concentric circle shows the location of substitutions in noncoding regions (red). The fifth concentric circle shows the location of insertions in strain CDC1551, including coding (black) and noncoding (blue) regions, and the location of phage phiRv1 (red). The sixth concentric circle shows the location of insertions in strain H37Rv, including coding (black) and noncoding (blue) regions, and the location of phage phiRv1 (red). The seventh concentric circle shows the location of IS*6110* insertion elements in strains CDC1551 (blue) and H37Rv (red). The eighth (innermost) concentric circle shows the location of tRNAs (blue) and rRNA (red).

The two genomes contained notable differences. The genetic variability in *M. tuberculosis* arises through a complex evolutionary process that involves recombination or multiple insertion-deletion events occurring independently at the same locus. The H37Rv strain contained 37 insertions (greater than 10 bp) relative to strain CDC1551. Twenty-six insertions affected open reading frames (ORFs) and 11 were intergenic. The insertions in strain H37Rv included tandem repeats, additions to the 5′ or 3′ ends of ORFs, and the addition of complete ORFs. Complete ORFs included three encoding hypothetical proteins (Rv0793, Rv3427c, Rv3428c), two encoding PPE proteins (Rv3425, Rv3426), one encoding a PE_PGRS protein (Rv3519), and two encoding proteins with putative functions (Rv0794c, a dihydrolipoamide dehydrogenase, and Rv0792c, a putative transcriptional regulator). Forty-nine

insertions were identified in strain CDC1551 relative to strain H37Rv. Thirty-five insertions affected ORFs and 14 were intergenic.

The IS3-type insertion sequence IS*6110* is the principal epidemiological marker for *M. tuberculosis*. A number of the insertions and deletions were associated with this insertion sequence, suggesting a role for this element in genome plasticity [22, 23]. Studies have shown that homologous recombination between nearby copies of IS*6110* may result in genomic deletions and can be a mechanism for generating genomic diversity [24].

There are no studies that have examined the genomic composition of *M. tuberculosis* isolates from the high MDR- and XDR-TB burden countries (former Soviet Union republics). The bacterial factors that contribute to disease severity and type, in addition to host genetics, and the environment remain largely unknown. Understanding the mechanisms of drug resistance, virulence, spreading of Mtb, manifestation and clinical course of TB disease, based on a full genomic analysis is extremely important both for Belarus health programs and to support worldwide efforts to combat the disease.

The development of MDR- and XDR-TB is the result of a number of mutational events which lead to the formation of resistance to antituberculosis drugs. During anti-TB treatment, *Mycobacterium tuberculosis* faces selective pressure of anti-TB drugs. Resistance to TB drugs provides a significant benefit of Mtb for survival within the host. As a rule, the identification of drug resistance of *Mycobacterium tuberculosis* is carried out using phenotypic methods and an assumption regarding the presence or absence of specific mutations could be made on the basis of Mtb resistance or susceptibility to anti-TB drugs. Whole genome sequencing could detect possible mutational events (SNPs in important proteins, regulatory regions, rearrangements, duplications, etc.) in other genes of mutant drug resistant Mtb strains and reveal possible relationships between the spectrum of Mtb drug resistance and other characteristics of mycobacteria, for example, virulence. It is known that some Mtb families, in particular, Beijing (widespread in Belarus, according to our data), demonstrate significant drug resistance and high virulence. Nevertheless, the relationship of these two Mtb characteristics is not fully resolved for all known Mtb families, despite numerous experimental studies. Sequencing Mtb strains genomes could determine the possible correlation between drug resistance and features of genes involved in mycobacteria-host interaction. Genetic variation may have an important role in disease pathogenesis and immunity. Putative virulence genes identified by homology and sequence analysis will later be studied using classical bacterial pathogenesis techniques, e.g., gene knockout experiments, to determine their contribution to pathogenesis.

3. Rationale for strain selection and the number of strains proposed in the study.

During the nationwide TB drug resistance survey conducted between mid-2010 and mid-2011, 934 new and 410 previously treated cases (both *M.tb*. microscopy and culture positive) were detected. MDR-TB was found in 32.3% (95% CI: 29.7-35.0) and 75.6% (95% CI: 72.1-78.9) of new and previously treated patients, respectively. Of the 612 patients with MDR-TB, 11.9% (95% CI: 9.7-14.6) had extensively drug-resistant (XDR) TB. Past treatment and HIV infection were the strongest independent predictors of MDR-TB. Young age (less than 35 years), history of imprisonment, disability to work, alcohol abuse and smoking were also associated with MDR-TB.

Based on collected information, we propose sequencing 10 XDR-TB samples, 10 MDR-TB samples, 10 previously treated (PT) XDR-TB samples, 10 PT MDR-TB samples, 10 new “susceptible” TB samples and 10 PT susceptible TB samples - a total of 60 samples (Table 3). For each category, five samples are obtained from patients younger than 35 years of age, and five from patients older than 35 years.

Table 3. Sample selection for sequencing

	new XDR-TB	new MDR-TB	PT XDR-TB	PT MDR-TB	new susceptible TB	PT susceptible TB
Samples	10	10	10	10	10	10

M.tb strains were isolated from TB patients in various regions of Belarus. Strains were identified with standard biochemical tests. Drug susceptibility testing was performed using Lowenstein Jensen medium and automated system BACTEC MGIT 960.

Every DNA sample has information about the patient's age, gender, place of birth, place of living, other demographic characteristics, education, living and employment conditions, history of imprisonment, use of alcohol and illegal drugs, smoking, TB disease history, clinical presentation, radiology and laboratory (routine blood work, routine blood chemistry) findings, Mtb microscopy and culture results, drug susceptibility testing, receiving drugs, HIV status and other co-morbidity (see Appendix 1).

4. Data analysis.

Our goal is to perform comparative analysis of all existing TB genomes, find SNPs and find correlations of genome variations with a) patients' medical history, and b) resistance of corresponding bacteria to known drugs. For annotation purposes, we will map sequenced reads to the *M. tuberculosis* reference genome (H37Rv). We will use filtered, high quality variants for genome-wide association study (GWAS) using the Efficient Mixed-Model Association (EMMA) and haplotype likelihood ratio (HLR) tests. We will look for variants with high association to the MDR and XDR phenotypes, as well as resistance to specific drugs using the clinical and *in vitro* data available for these samples. We will also look for evidence of genomic changes over time that appear to have given rise to stronger resistance to various drugs.

5. **Data release policy.**

All sequences generated under this proposal will be submitted to GenBank. Assembled contigs and scaffolds will be deposited in the Whole Genome Shotgun section of GenBank within 45 calendar days of completing the shotgun or high-throughput sequencing. If it is determined that the final assembly can be significantly improved, an updated record will be deposited in the appropriate part of Genbank when complete.

Also, we intend to make all genomic data and associated de-identified clinical metadata public through the NIAID Bioinformatics Resource Center PATRIC (patricbrc.org).

Annotation data will be made available via GenBank and PATRIC BRC web sites. Annotation data will be released within 45 calendar days of being generated.

6. **Literature cited.**

1. Anti-tuberculosis drug resistance in the world: the WHO/IUATLD Global Project on Anti-Tuberculosis Drug Resistance Surveillance / Geneva, World Health Organization. – 1997. - Document WHO/TB/97.229.

2. Anti-tuberculosis drug resistance in the world: the WHO/IUATLD Global Project on Anti-Tuberculosis Drug Resistance Surveillance. Report 2: Prevalence and trends. / Geneva, World Health Organization. – 2000. - Document WHO/CDS/TB/2000.278.
3. Anti-tuberculosis drug resistance in the world: third global report. The WHO/IUATLD Global Project on Anti-Tuberculosis Drug Resistance Surveillance 1999–2002. / Geneva: WHO. - 2004. – Document WHO/HTM/TB/2004.343.
4. Cohn D, Bustreo F, Raviglione M. Drug resistance in tuberculosis: review of the worldwide situation and WHO/IUATLD global surveillance project / Clin. Infect. Dis. – 1997. – Vol. 24 (Suppl. 1). - S121–S130.
5. Espinal M et al. Global Trends in resistance to anti-tuberculosis drugs / New Engl. J. Med. – 2001. – Vol. 344. – P. 1294–1302.
6. Extensively Drug-Resistant Tuberculosis / K. Dheda [et al.] // N. Engl. J. Med. – 2008. – Vol. 359, № 27. – P. 2390.
7. Gandhi NR, Moll A, Sturm AW, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lancet 2006; 368: 1575-80.
8. Zignol M, van Gemert W, Falzon D, Sismanidis C, Glaziou P, Floyd K, Raviglione M. Surveillance of anti-tuberculosis drug resistance in the world: an updated analysis, 2007-2010. Bull World Health Organ 2012; 90:111-9.
9. Skrahina A, Zalutskaya A, Sahalchyk E, Astrauko A, van Gemert W, Hoffner S, Rusovich V, Zignol M. Alarming levels of drug-resistant tuberculosis in Belarus: results of a survey in Minsk. Eur Respir J 2011; Oct 20. [Epub ahead of print].
10. Cole ST, Saint Girons I. Bacterial genomics. FEMS Microbiol Rev 1994; 14: 139-60. 30. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 1998a; 393: 537-44.
11. Sreevatsan S, Pan X, Stockbauer KE, et al. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc NatlAcadSci U S A 1997; 94: 9869-74.
12. vanEmbden JD, Cave MD, Crawford JT, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J ClinMicrobiol 1993; 31: 406-9.
13. Cole ST. Learning from the genome sequence of Mycobacterium tuberculosis H37Rv. FEBS Lett 1999; 452: 7-10.
14. 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: 759-72.

15. Fleischmann RD, Alland D, Eisen JA, et al. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J Bacteriol 2002; 184: 5479-90.
16. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc NatlAcadSci U S A 2004; 101: 4871-6.
17. Gagneux S, DeRiemer K, Van T, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc NatlAcadSci U S A 2006; 103: 2869-73.
18. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D (2002) Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review. Emerg Infect Dis 8: 843–849.
19. Phyu S, Jureen R, Ti T, Dahle UR, Grewal HM (2003) Heterogeneity of Mycobacterium tuberculosis isolates in Yangon, Myanmar. J ClinMicrobiol 41: 4907–4908.
20. Lillebaek T, Andersen AB, Dirksen A, Glynn JR, Kremer K (2003) Mycobacterium tuberculosis Beijing genotype. Emerg Infect Dis 9: 1553–1557.
21. Tsolaki AG, Gagneux S, Pym AS, et al. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. J ClinMicrobiol 2005; 43: 3185-91.
22. McHugh, T. D., and S. H. Gillespie**.** 1998. Nonrandom association of the IS*6110* and Mycobacterium tuberculosis: implications for molecular epidemiological studies. J. Clin. Microbiol.36**:**1410-1413.
23. Warren, R. M., S. L. Sampson, M. Richardson, G. D. Van Der Spuy, C. J. Lombard, T. C. Victor, and P. D. van Helden. 2000. Mapping of IS*6110* flanking regions in clinical isolates of Mycobacterium tuberculosis demonstrates genome plasticity. Mol. Microbiol. 37**:**1405-1416.
24. Fang, Z., C. Doig, D. T. Kenna, N. Smittipat, P. Palittapongarnpim, B. Watt, and K. J. Forbes.1999. IS*6110*-mediated deletions of wild-type chromosomes of *Mycobacterium tuberculosis*. J. Bacteriol. 181**:**1014-1020.

Appendix 1. Clinical sample form. All samples are going to have such forms attached.

TB-disease form Patient ID code:
Center/Cohort name:
_____
_____
Please read Instructions before completing the form
This form is completed by:
Name (in print):
Position: Physician: Nurse: Other, describe:
Date of completion of this form (dd-mm-yyyy) - -
_____
_____

Date when the patient was last seen at the clinic (dd-mm-yyyy) - -
Section A. Background demographics and basic clinical information
Date of Birth (dd-mm-yyyy): - - Gender: 1=Male, 2=Female
Risk factors for
TB acquisition
(tick all that
applied) (x)
___(1) History of injecting
drug user (IDU)
___(2) In prison within last 2 years
___(3) Alcohol abuse
/surroundings
___(4) Recent TB in the family
___(5) Travelling in TB endemic area
specify: _____
___(9) Other
specify:_____
Originating from (x)
___(1) Same country as centre
___(2) Other European country
Specify:_____
Ethnicity(x)
___(10) White
___(20) Black
___(21) Black African
___(3) Any other country
Specify: _____
___(9) Unknown
___(22) Black Caribbean
___(30) Hispanic
___(60) Indigenous
___ (#) Combination of
any of the previous,
specify numbers:
_____
___(98) Data collection
prohibited
___(40) Asian
___(50) American
___(99) Unknown
If the patient is IDU
Is the patient currently active IDU?
Is the patient currently receiving methadone?
if Yes, please indicate dose mg
if No, is the patient currently receiving any other substitution therapy?
Yes No Unknown
if Yes, please indicate drug name: _____ and dose: mg
Height (999cm = unknown) cm Most recently measured weight kg
Date of measurement (dd-mm-yyyy) - -