Mutation analysis of *M. tuberculosis* nucleotide sequences from patients in Belarus

R.S.Sergeev[1], I.S.Kavaliou[1,3], A. Gabrielian[2], A. Rosenthal[2], A.V.Tuzikov[1],

[1] *United Institute of Informatics Problems NASB, 6 Surganova str., Minsk, 220012, Belarus,*
roma.sergeev@gmail.com

[2] *Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD, USA*

[3] *EPAM Systems, 1/1 Academician Kuprevich str., Minsk, 220141, Belarus*

Motivation and aim: Tuberculosis (TB) is an important public health problem in Belarus and worldwide. Despite substantial progress in TB control, situation has been complicated with emergence and development of MDR (multi drug-resistant) /XDR (extensively drug-resistant) TB which require long-term treatment. According to a Belarus nationwide survey in 2010–2011 MDR-TB was found in 32.3% of new patients and 75.6% those previously treated. XDR-TB was reported in 11.9% of MDR-TB patients. Belarus is among countries having the highest multi-resistant tuberculosis incidence. The ability of TB agent to resist treatment is strongly connected with variations and mutations in specific parts of the bacteria genome. Analysis of mutations in TB sequences for genotypic predictors of drug resistance may become very valuable for choosing adequate treatment regimen and preventing therapy failure.

Experiment design: Strain selection from patients in Belarus was performed to have 79.5% of MDR and 47.5% of XDR strains. All *M. tuberculosis* strains were isolated and sequenced on Illumina HiSeq2000 instruments to generate 101bp paired-end reads at 140x coverage of the genome. Data were aligned against H37Rv reference genome to identify variants. Pilon tool was used for variant calling. Finally, we performed a genome-wide association study (GWAS) for 137 annotated nucleotide sequences.

Data analysis: Based on laboratory findings, we grouped sequences into several datasets to check for statistically significant differences between drug-resistant (cases) and drug-susceptible (control) samples under certain conditions. We analyzed known mutation lists presented by TBDreamDB database [1] and GenoType MTBDRplus/MTBDRsl assays as high-confidence mutations and discovered that resistance to most second-line drugs (ethionamide, kanamycin, amikacin, capreomycin, cycloserine) can’t be reliably explained by those lists.

Therefore, we designed our experiments to hone investigation of unexplained resistance. Lists of known associations were used as test sets to validate predictions.

Data analysis procedure comprises several steps organized into a pipeline. The initial steps are aimed to perform comparative sequence analysis and investigate population structure of TB strains. Next steps uncover associations of genome variations with results of phenotype resistance tests to known drugs. We tried a number of methods based on different principles to analyze SNP data, including single- and multi-marker association tests.

Most single-marker tests rely on contingency tables approach that compare allele frequencies, or genotypes, in sets of cases and controls. We applied the permuted versions of classical association tests, Cochran-Mantel-Haenszel test [2] and Armitage trend chi-squared statistics with Eigenstrat correction [3] to account for population structure.

A major drawback of single-marker methods is that they do not consider pairwise and higher-order interactions between genetic variants. Conversely, multi-marker association tests can realistically model the multiplicity of genotypic factors bringing a number of other challenges associated with analyses of high-dimensional data in GWAS experiments: number of SNPs (parameters) is significantly greater than the number of sequences (observations). To overcome these difficulties we experimented with feature selection methods and settings for regularized logistic regression, linear mixed model (GEMMA [4]) and mode-oriented stochastic search (MOSS [5]).

Lasso regularization showed most relevant results in logistic regression approach considering a grouping effect, where strongly correlated predictors tend to be in or out of the model together. GEMMA allowed correction for population structure due to the random effect of the linear mixed model that is calculated based on kinship/relatedness matrix. The MOSS algorithm, which is a Bayesian variable selection procedure that identifies combinations of the best predictive SNPs associated with the response and performs a hierarchical log-linear model search to identify the most relevant associations among the resulting subsets of SNPs, provided most conservative results.

Results: Principal component analysis (PCA) has shown that all strains from Belarusian patients can be segregated into five groups. Phylogeny and spoligotyping established the most prevalent sublineages: Beijing (63.6%), T1 (18.9%), H3 (5.6%) and T5 (2.8%).

Within the descriptive analysis, we discovered pairwise correlations of drug susceptibility testing (DST) results to investigate cross-resistance between drugs (table 1). As anticipated, the highest levels of correlations were detected between amikacin, capreomycin and inside groups of first-line drugs.

Table 1. Correlations between results of TB drug susceptibility testing.

	EMB	INH	RIF	PZA	STM	CYCL	ETH	PARA	AMIK	CAPR	OFLO
EMB	1.00	0.90	0.90	1.00	0.80	0.36	0.34	0.25	0.53	0.59	0.55
INH	0.90	1.00	1.00	0.91	0.89	0.36	0.31	0.23	0.48	0.52	0.53
RIF	0.90	1.00	1.00	0.91	0.89	0.37	0.31	0.22	0.48	0.53	0.53
PZA	1.00	0.91	0.91	1.00	0.72	0.38	0.38	0.18	0.38	0.46	0.49
STM	0.80	0.89	0.89	0.72	1.00	0.33	0.27	0.20	0.42	0.45	0.47
CYCL	0.36	0.36	0.37	0.38	0.33	1.00	0.33	0.32	0.27	0.35	0.23
ETH	0.34	0.31	0.31	0.38	0.27	0.33	1.00	0.06	0.33	0.46	0.14
PARA	0.25	0.23	0.22	0.18	0.20	0.32	0.06	1.00	0.07	0.13	0.23
AMIK	0.53	0.48	0.48	0.38	0.42	0.27	0.33	0.07	1.00	0.90	0.57
CAPR	0.59	0.52	0.53	0.46	0.45	0.35	0.46	0.13	0.90	1.00	0.61
OFLO	0.55	0.53	0.53	0.49	0.47	0.23	0.14	0.23	0.57	0.61	1.00

Interesting results showed analysis of the proportion of phenotypic variance explained (PVE) by SNP genotypes (table 2), which can be summarized as $k_{SNP}^2 = \sigma_G^2/(\sigma_G^2 + \sigma_E^2)$, where σ_G^2 is variance due to genotypic markers and σ_E^2 is influence of the environmental factors.

Table 2. PVE values for anti-TB drugs analyzed.

Drug	PVE, %	Standard Error, %
First-line drugs:		
ISON (isoniazid)	99.997	0.021
RIFM (rifampicin)	99.997	0.021
PYRA (pyrazinamide)	99.997	0.049
STRE (streptomycin)	99.997	0.036
ETHA (ethambutol)	97.119	1.695
Second-line drugs:		
CYCL (cycloserine)	75.716	12.386
CAPR (capreomycin)	73.903	11.048
AMIK (amikacin)	69.831	11.925
OFLO (ofloxacin)	58.682	12.922

ETHI (ethionamide/prothionamide)	45.680	24.906
PARA (para-aminosalicyclic acid)	29.998	17.010

According to these results, SNPs do not fully explain resistance to most of second-line TB drugs in our datasets due to issues of DST protocols [6] or other factors.

Genotype/phenotype association tests resulted in high-confidence mutation lists that were ordered according to mutation significance values for each drug and annotated using NCBI databases.

Availability: Elements of this approach are used in current to establish the Belarus tuberculosis portal (http://tuberculosis.by) and conduct comprehensive study of obtained MDR and XDR TB strains. In the nearest future we plan to significantly expand the functionality of the portal to share our bioinformatics data on this research.

Acknowledgements: We express our thanks to the Republican Research and Practical Center for Pulmonology and Tuberculosis (Belarus) for their collaboration.

1. A.Sandgren et al. (2009) Tuberculosis drug resistance mutation database, *PLoS Med,* **6**(2):132-136.

2. S.Purcell et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet*, **81**(3):559-575.

3. A.L.Price et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies, *Nat Genet*, **38**(8):904-909.

4. X.Zhou, M.Stephens (2012) Genome-wide efficient mixed-model analysis for association studies, *Nature Genetics*, **44**:821–824.

5. A. Dobra et al. (2010) The mode oriented stochastic search (MOSS) for log-linear models with conjugate priors, *Statistical Methodology*, **7**:240-253.

6. D.J.Horne et al. (2013) Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs, *J Clin Microbiol*, **51**(2):393-401.