Comparison of eight M. tuberculosis strains in the presence of drug resistance data

Michał Woźniak*‡, Limsoon Wong†, Jerzy Tiuryn*
*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland.
†School of Computing, National University of Singapore, Singapore.
‡Corresponding author: Michał Woźniak, m.wozniak@mimuw.edu.pl.

Abstract

The first fully sequenced M. tuberculosis strain was H37Rv and since then there has been published several new MTB genomes [1, 2, 3, 4]. Progress in sequencing enables new possibilities for analysing mechanisms of drug resistance. For our experiments, we used eight fully sequenced MTB strains, of which five are susceptible to drugs: H37Rv, H37Ra, CDC1551, F11, KZN 4207; two are Multi-Drug-Resistant: KZN 1435, KZN V2475; and one is Extensively-Drug-Resistant: KZN R506.

We achieve two goals. First, we compare annotations of the eight genomes, and conclude that presumably there are some flaws. Second, we identify a set of mutations that are potentially responsible for drug resistance mechanisms. Then, we show that the level of essential point mutations among genes which are drug targets is two orders of magnitude higher on average than for all the remaining genes of the genomes.

A flawed annotations

- Annotations for genomes KZN 4207, KZN V2475 and KZN R506 contain 33, 912, and 35, annotated genes respectively whose lengths are not multiples of 3.

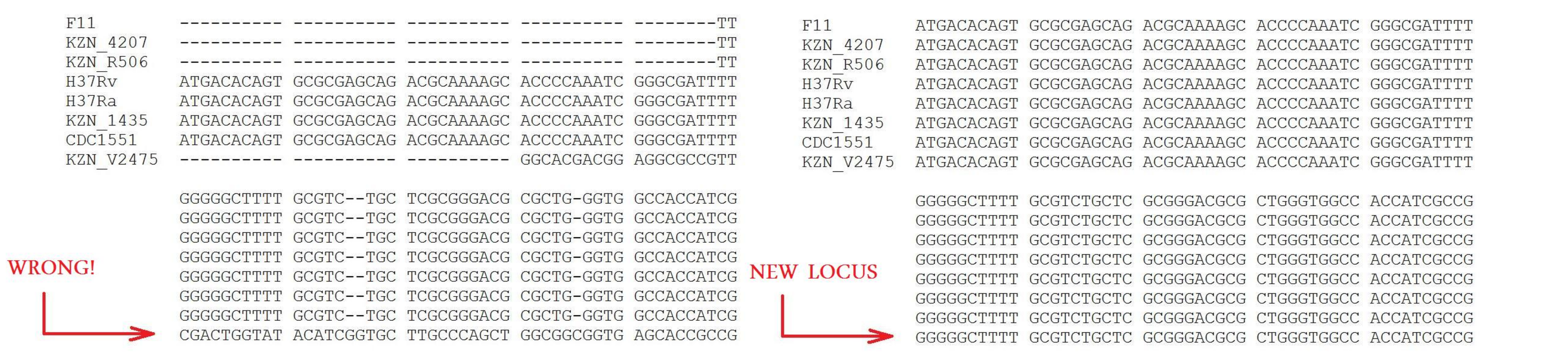

Figure 1: Left: Alignments for embB gene (drug target) sequences obtained by annotations. Right: alignments for embB gene sequences obtained by BLAST.

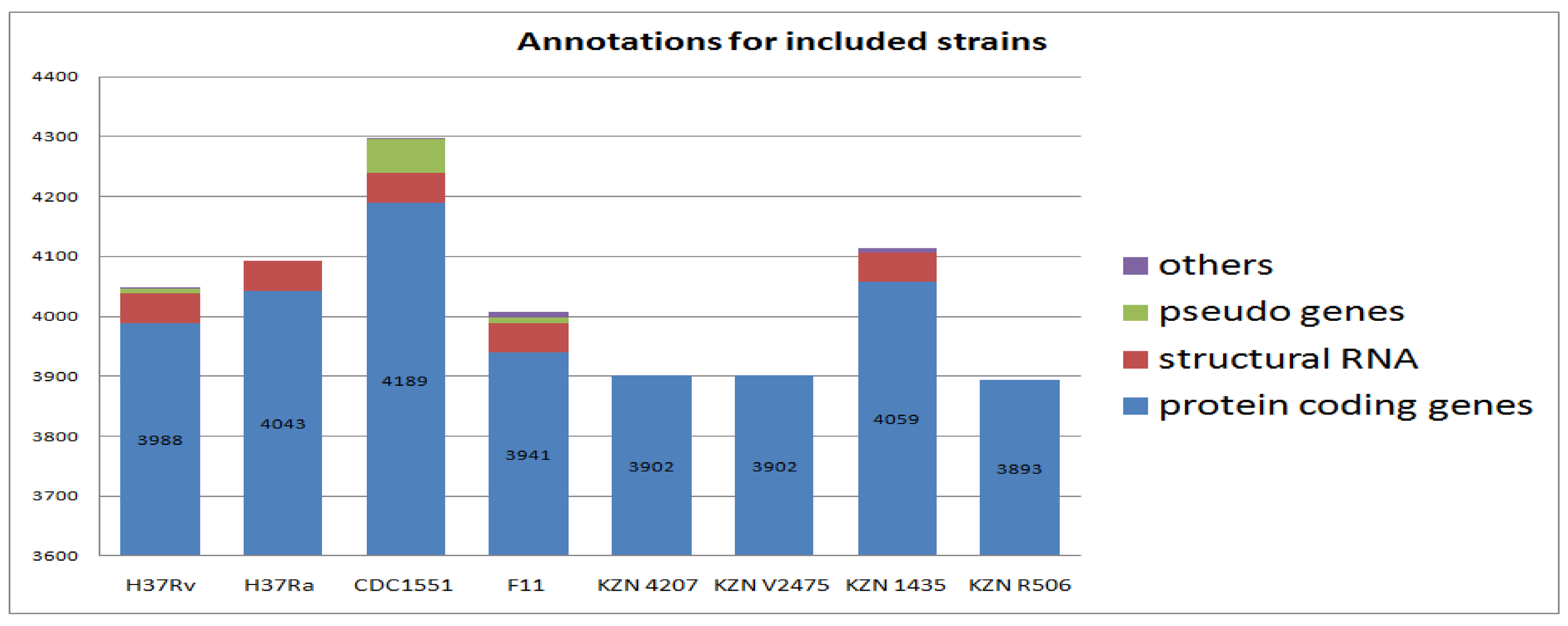

Figure 2: High variation in number of annotated genes among MTB strains.

Schema of the experiment

We work with annotations for H37Rv as a reference strain.

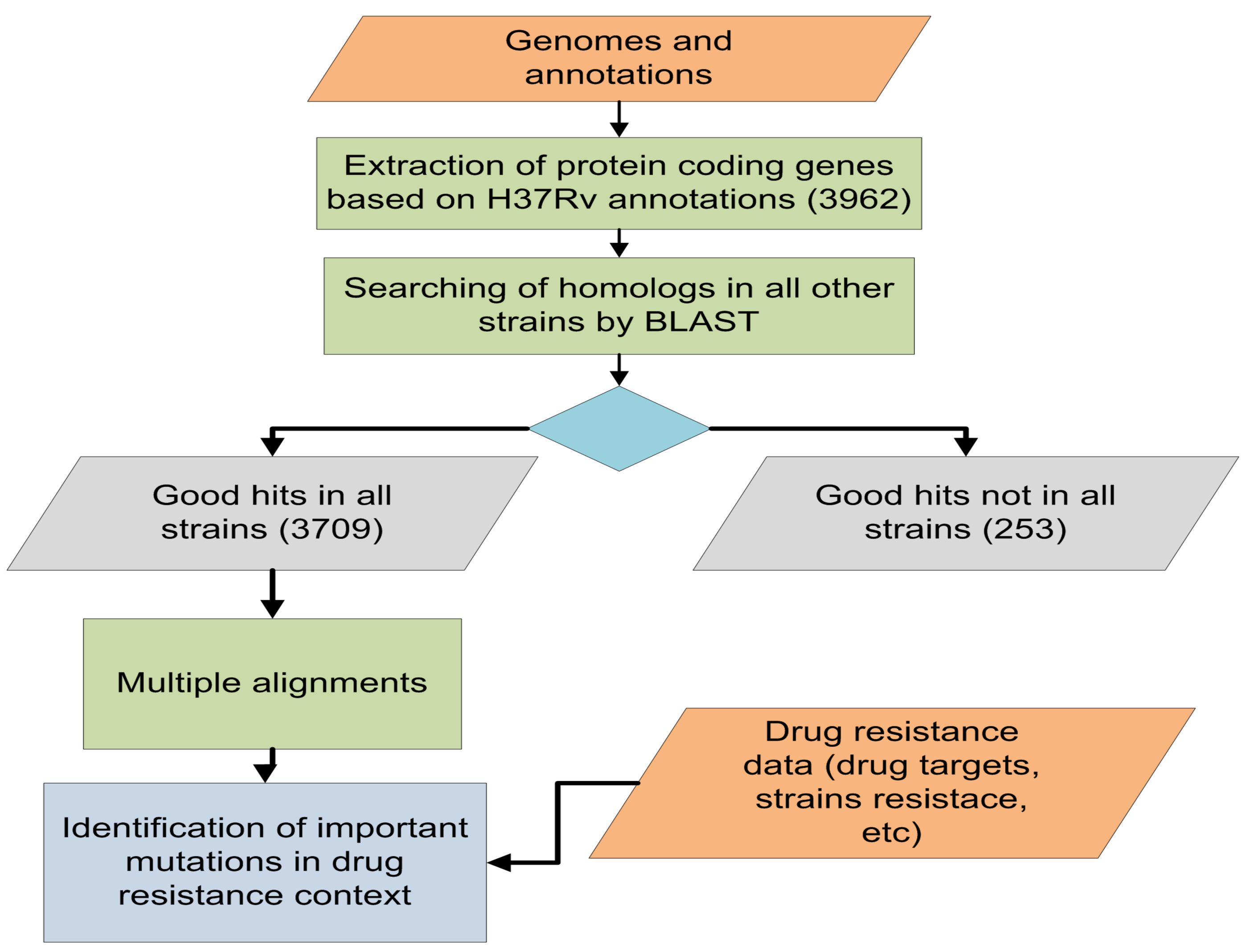

Figure 3: Schema of the experiment

In order to exclude frame-shifted genes from analysis of essential point mutations we put strong conditions on BLAST hits. We call a BLAST hit *good hit* when it meets all of the following conditions:

1. the hit has been extended to the whole query with length (excluding gaps) that is a multiple of 3
2. the hit has a typical start and stop codons with exactly one stop codon in the reading frame
3. the hit has e-value less than 10^{-10} (in our experiments all hits that meet the above conditions has e-value less than 10^{-30})

Essential mutations

Drug targets are genes or proteins perturbed by a specific drug during the therapy. For the experiment we used the list of 11 drug targets obtained from *drugbank.ca* database for 6 MTB drugs: Isoniazid, Rifampicin, Streptomycin, Ofloxacin, Kanamycin, Ethambutol.

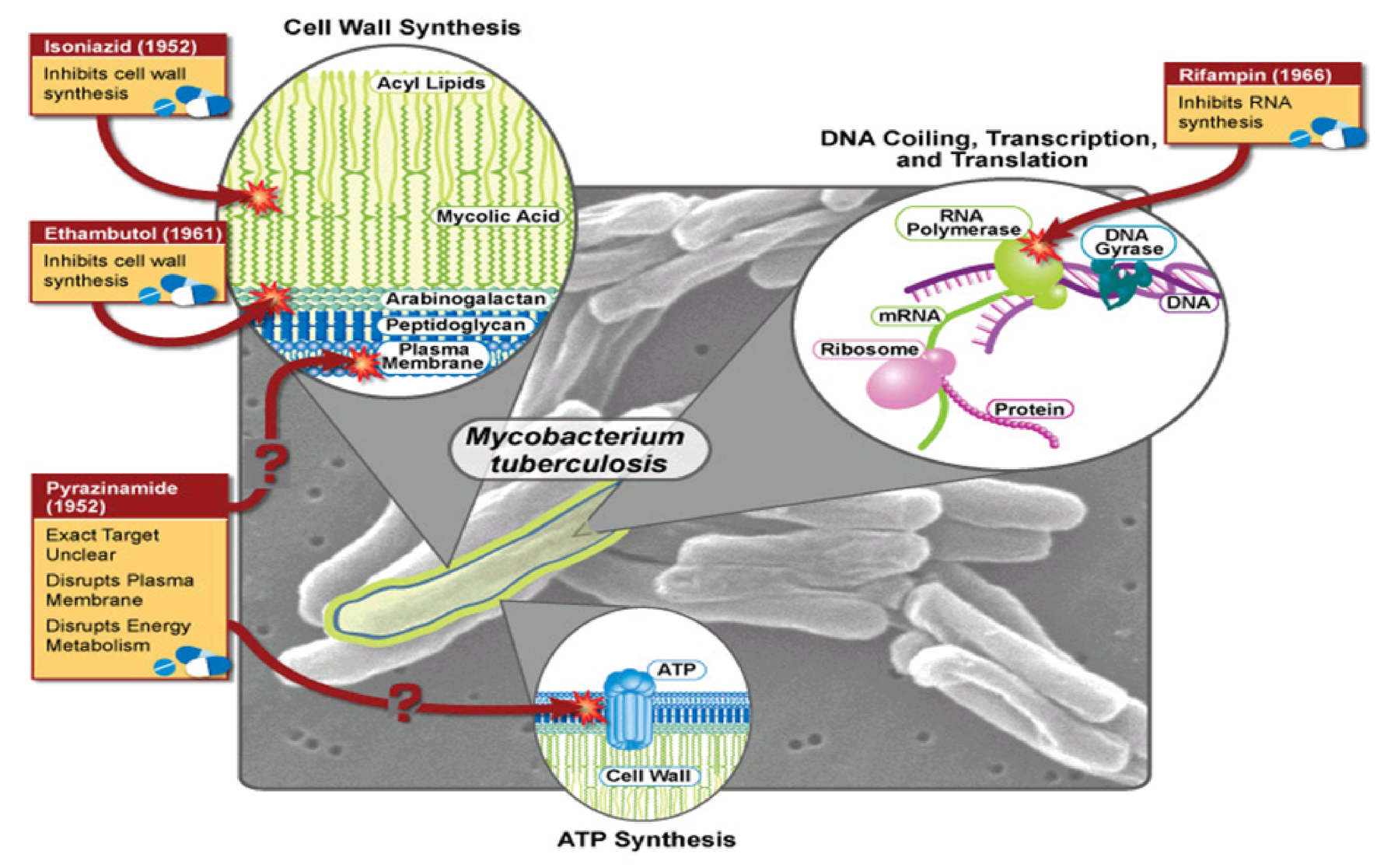

Figure 4: Action of some MTB drugs. Source: http://www3.niaid.nih.gov/

For an integer k, we call a mutation *essential* (from the standpoint of drug resistance) with a support k when in the corresponding position of the multiple alignment at least k sequences of susceptible strains have the same character, while at least one drug resistant strain has an other character.

Based on gene families with good hits for all strains we verified the hypothesis that significant mutations occur more frequently in drug target genes than in other genes.

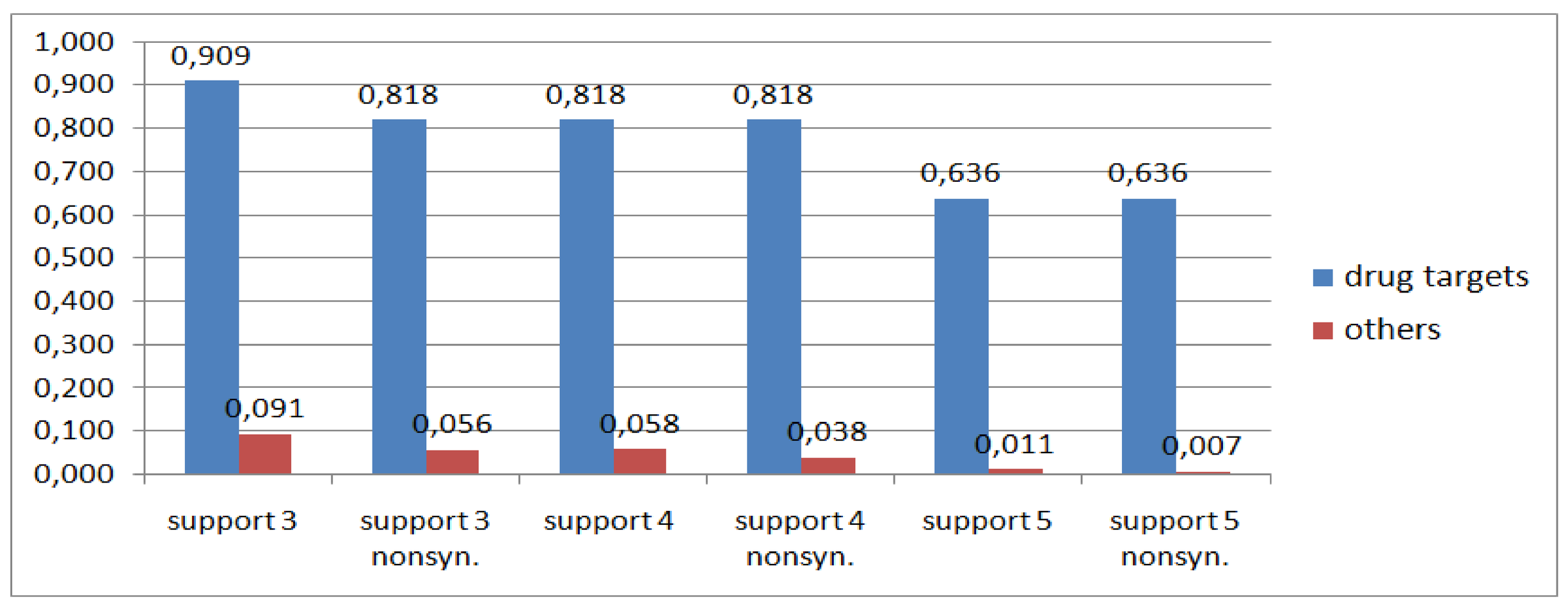

Figure 5: The average numbers of essential mutations per gene for drug targets and other genes.

Phylogenetics

The figures below show that reconstructed consensus trees based on drug targets genes (DTGs) and all genes differ. For example KZN 4207 susceptible strain is closer to other susceptible strains in the tree for DTGs. Also in the tree for DTGs the two MDR strains are closer than in the tree for all genes.

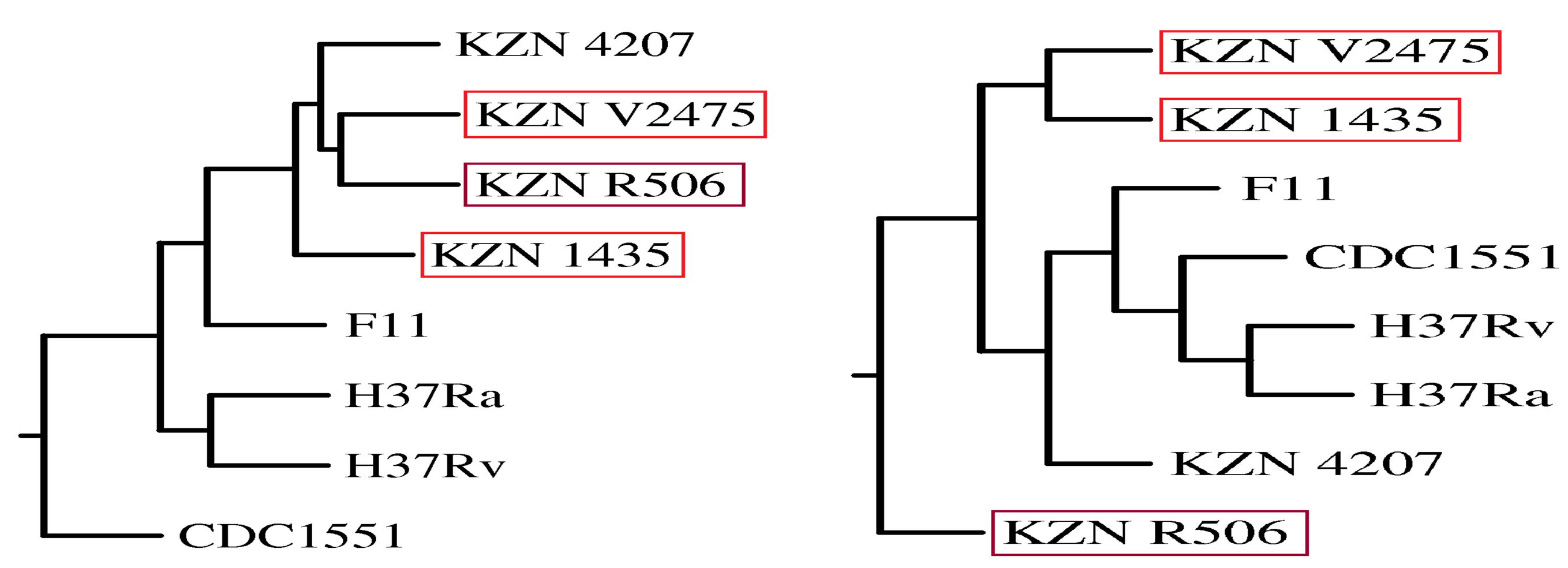

Figure 6: The first tree was constructed as consensus tree between trees constructed by maximum likelihood method with bootstraping and elimination of trees with support below 0.2 for 783 gene families. The second was constructed in the same way based on 11 drug target genes.

Frameshifts

Closer look at the set of 253 genes with good hits not in all species shows that for 205 reference genes BLAST returned hits that potentially encode a homolog proteins for all strains. Multiple alignments of those families contain 183 frameshifts of length 1. Most of those frameshifts (85%) where detected for only one strain in alignemnt. CDC1551 accumulate the largest number (61%) of single frameshifts.

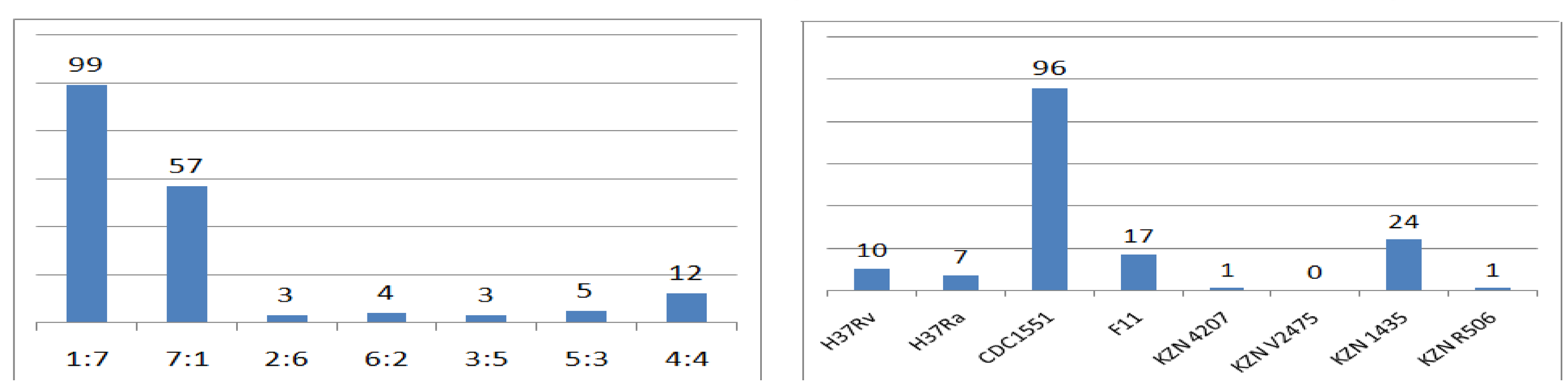

Figure 7: Left:numbers of frameshifts (length 1) cases in multiple alignments. Right: distribution of single frameshifts (1:7 or 7:1) among all strains.

The relatively high number of single frameshifts for CDC1551 suggests that perhaps most of them are sequencing errors. We repeated the experiment with excluded CDC1551, but the results did not essentially change.

References

[1] S.T. Cole et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature, June 1998.

[2] R. D. Fleischmann et al. Whole-Genome Csomparison of Mycobacterium tuberculosis Clinical and Laboratory Strains. J Bacteriol., 2002

[3] Huajun Zheng. et al. Genetic Basis of Virulence Attenuation Revealed by Comparative Genomic Analysis of Mycobacterium tuberculosis Strain H37Ra versus H37Rv. PLoS ONE, 2008

[4] Ioerger TR et al. Genome Analysis of Multi- and Extensively-Drug-Resistant Tuberculosis from KwaZulu-Natal, South Africa. PLoS ONE, 11 2009

Acknowledgements: This work is partially supported by Polish Ministry of Science and Higher Education grants no. N N301 065236 and PBZ-MNiI-2/1/2005.