Tuberculini: combining targeted sequencing and machine learning to optimize antibiotic therapy in tuberculosis patients.

Dr. Sebastian Dümcke (Clemedi AG; Universität Zürich), Dr. Prajwal (Clemedi AG; Universität Zürich), Prof. Thorsten Buch (Insitut für Labortierkunde, Universität Zürich), Dr. Peter Keller Insititut für Infektionskrankheiten, Universität Bern)

Introduction

Each year, there are more than 10 M new cases of tuberculosis (TB) and a steady increase in rates of drug-resistant TB (DR-TB) is observed. To reduce the spread of hard to treat DR-TB strains, novel methods to rapidly detect DR-TB and provide guidance on antibiotic efficacy are needed. Clemedi is developing a targeted sequencing pipeline "Tuberculini" to optimise treatment for drug-resistant tuberculosis. It comprises reagent kits and machine learning algorithms, which are described here.

Methods:

14'190 publicly available WGS data sets with annotated antibiotic susceptibility information were aligned against the tuberculosis reference genome using bwa and variants called using freebayes on genomic regions which are part of the Tuberculini enrichment panel (incl. 138 genes covering 208 kilo-bases). 95 % of panel had to be covered at least 20-fold to include the data in the machine learning set. We chose the best performing combination from 11 model/parameter combinations, based on 10-fold cross validation.

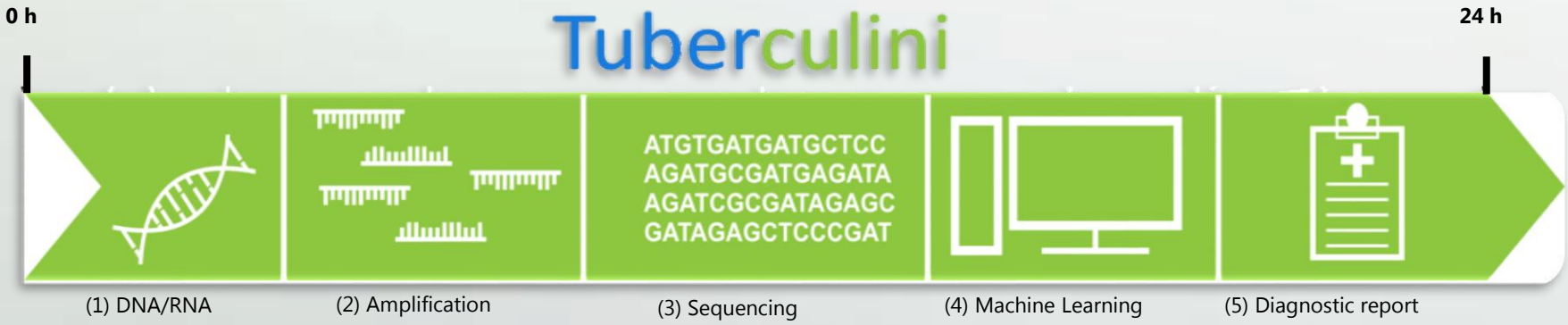
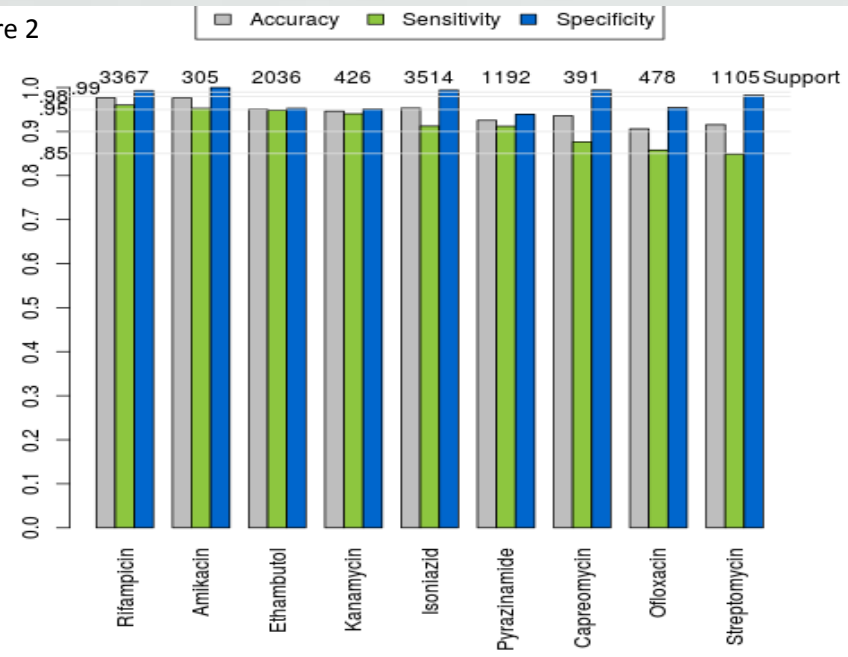


Figure 1

Figure 1: end-to-end workflow to diagnose tuberculosis. 1) DNA is extracted from the patient sample, 2) then amplified using a highly multiplex PCR reaction (930 targets) and the resulting amplicons are 3) sequenced using off-the-shelf sequencing instruments. 4) sequencing data is analyzed with machine learning algorithms to 5) generate a diagnostic report comprising the tuberculosis lineage, NTM detection and a full antibiotic susceptibility profile.

Figure 2: cross-validated performance of selected machine learning models with the highest diagnostic-odds ratio for each antibiotic. Support indicates the number of resistant cases available to each model. Horizontal lines show 85%, 90%, 95% and 99% thresholds.

Figure 2



Results

We could generate models for a total of 8 antibiotics: Isoniazid, Rifampicin, Ethambutol, Pyrazinamide, Ofloxacin, Capreomycin, Kanamycin, Amikacin. Both sensitivity and specificity in predicting antibiotic resistance from whole genome sequences of isolates were > 90% for 6 antibiotics and > 95% for 2 antibiotics (see figure 1). Models that performed best were regularized logistic regression and support vector machines.

Discussion

Implementation of Tuberculini can help improve patient management by providing comprehensive information about drug susceptibility and resistance within 24-48 hours. This allows a rapid switch from empirical to targeted therapy, shortens isolation period and time to recovery and reduces mortality and morbidity.