Scientific article

High-resolution outbreak tracing and resistance detection using whole genome sequencing in the case of a *Mycobacterium tuberculosis* outbreak

W. Ridderberg, F. Strino, P. Ettenhuber and A. Materna

QIAGEN Bioinformatics, Microbial Genomics, Aarhus, Denmark

Abstract: In this white paper, we demonstrate the successful utilization of CLC Genomics Workbench and CLC Microbial Genomics Module for improved outbreak tracing via whole genome SNP analysis and antimicrobial resistance detection in the case of a *Mycobacterium tuberculosis* (TB) outbreak.

Introduction

Pathogen surveillance and outbreak analysis are critical to preventing and controlling disease outbreaks and re-emergence. A major challenge in source tracing outbreaks is the ability to genotype pathogens at sufficient resolution to unravel the chain of transmissions. Accurate detection of antimicrobial resistance mechanisms is essential for administering correct and efficient treatment regimens and for monitoring multi-drug resistance.

Next-generation sequencing (NGS) of whole pathogen genomes permits transmission analysis with high discriminatory power. An additional benefit of whole genome sequencing (WGS) is its ability to simultaneously provide information on antimicrobial resistance mechanisms present in the microbe.

Using the original data from Fiebig et al. (3), we demonstrate accurate outbreak tracing via whole genome SNP analysis and improved antimicrobial resistance detection in the case of a TB outbreak using tools within CLC Genomics Workbench and CLC Microbial Genomics Module. CLC Microbial Genomics Module is an extension to CLC Genomics Workbench, and both are included in the QIAGEN® Microbial Genomics Pro Suite software solution.

Study outline/design

In March 2014, the molecular clustering of five multidrug resistant *M. tuberculosis* (MDR-TB) isolates was detected by the Austrian National Reference Laboratory (NRL). All five TB isolates shared genotypes revealed by the TB-specific typing methods, spoligotyping and the analysis of Mycobacterial Interspersed Repetitive Units – Variable Number of Tandem Repeat (MIRU-VNTR). An investigation was initiated to determine if transmission had occurred within Austria.

Patient I, II and III were diagnosed with TB between 2010–2012 and originated from the same city in Romania. Patient I and II had relocated to the same city in Austria (City 1). Patient III had moved to a second city in Austria (City 2). Patients IV and V, who were diagnosed in June 2013, were both born in Austria and both were residents of the same city as patient I and II (City 1). Patient IV and V had no history of migration or international travel. Contact investigation further revealed that patient III had a sister, also diagnosed with MDR-TB, living in Germany. ▷

Contents

This prompted the Austrian NRL to contact the NRL in Germany. In April 2014, the German NRL detected three MDR-TB isolates sharing the same MIRU-VNTR pattern ("A") as the Austrian cluster of isolates. Patient VI was the sister of patient III. Patient VII was a woman born in Romania and patient VIII was a man born in West Africa, carrying non-MDR-TB. As five patients (I, II, III, VI, VII) originated from the same city in Romania, a connection was made to the Romanian National TB Contact Point.

Due to the absence of consistent MIRU-VNTR typing of TB isolates in Romania, all five MDR-TB ever detected in that district were sent to Austria for analysis. Isolates from patients X and XI shared MIRU-VNTR pattern "A". The isolate from patient IX had a unique pattern "B", and isolates from patients XII and XIII shared pattern "C".

Materials and methods

Whole genome sequencing of one TB isolate from each of the 13 included patients was performed on the Illumina® MiSeq® platform, generating 2 x 301 bp paired-end reads. Sequence data were deposited by authors in the EMBL-EBI ENA sequence read archive (accession no. ERP013444).

We downloaded raw FASTQ files and metadata directly from the NCBI Sequence Read Archive through CLC Genomics Workbench 10.0 using the tool "Search for Reads in SRA". Figure 1 illustrates the tools and preconfigured workflows that were used for data analysis. All tools used here are available in CLC Genomics Workbench 10.0 and CLC Microbial Genomics Module 2.0 or later versions.

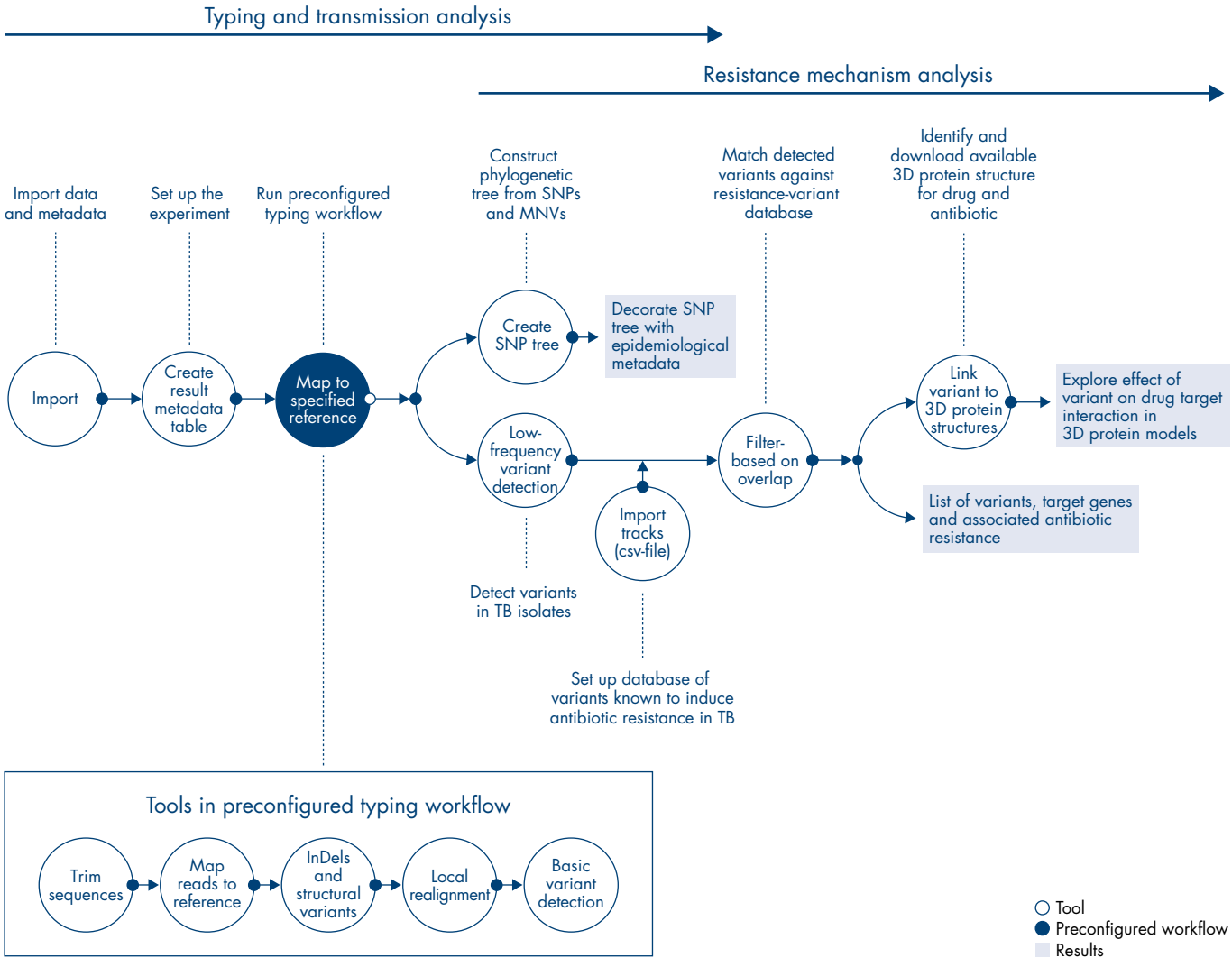


Figure 1. Steps, tools and workflows used for outbreak analysis and resistance detection.

Results

Outbreak investigation

Using whole genome data from the TB isolates under investigation, we studied the outbreak progression by analyzing accumulated mutations. Because bacteria accumulate spontaneous mutations over time, whole genome sequencing is a powerful tool for transmission analysis. Additionally, we searched for variants known to cause antimicrobial resistance and show how variants can be visualized on 3D protein models as a method for qualifying previously undescribed variants.

For greater ease of use and faster turn-around time, CLC Microbial Genomics Module 2.0 features a preconfigured workflow (Figure 1), that performs sequence trimming, mapping to a user-specified reference genome and variant detection. The workflow "Map to Specified Reference" is preconfigured with default settings, but parameters can be customized at all levels, to meet specific needs. We used the *M. tuberculosis* H37Rv genome as reference genome for variant detection.

The Workflow outputs a variant file containing all detected variant positions within the analyzed sample. With the "Batch" option, multiple samples can be processed simultaneously. For the analysis of transmission, we chose to exclude variants detected in genes known to be involved in antimicrobial resistance in TB (5). This was done using the tool "Filter Based on Overlap", which is part of our collection of "Track Tools".

The remaining filtered variants were used for SNP tree calculation. The resulting SNP tree displaying the 13 TB isolates is based on 673 variant positions (Figure 2).

From the SNP tree, it is evident that isolates group in three clusters, in congruence with the findings of the original study. The first cluster contains isolates from patient XII and XIII. These isolates have a distinct genotype, MIRU-VNTR pattern "C". Both patients were born and living in Romania. The second cluster contains isolates from patients III, VI, VII and XI, all originating from the same city in Romania, but living in three different countries. Among these patients, transmission likely occurred before migration from Romania. Patient III and VI are sisters and likely had a common source of transmission. However, their isolates differ by 12 SNPs, indicating that the common source and possibly also intermediate links were absent from the investigation. The third cluster contains isolates from patients II, IV, V and X – three of the Austrian outbreak isolates. The Austrian isolates are separated by just 1–4 SNPs suggesting two transmissions of TB within Austria, from patient II to patient IV, and from patient II to patient V. These findings are consistent with the conclusions of Fiebig et al. (3).

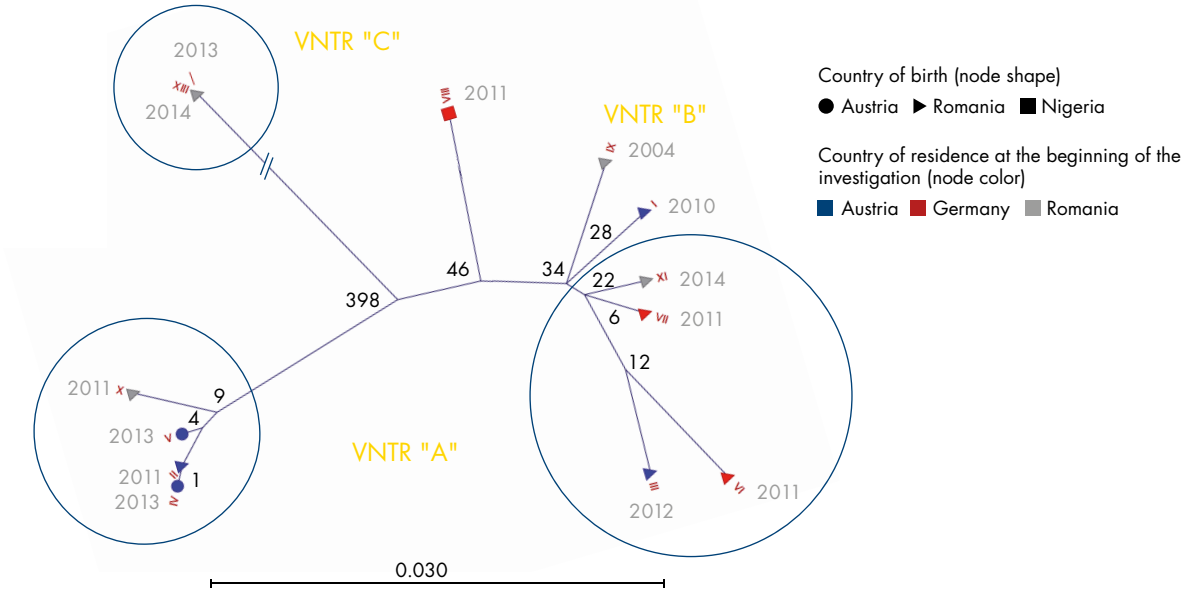


Figure 2. SNP tree. Label types show patient country of birth and label colors show patient country of residence. MIRU-VNTR genotype is marked in yellow, and date of TB diagnosis is shown in grey. The number of SNPs supporting branches are shown in black.

Detecting antimicrobial resistance

Antimicrobial resistance in *M. tuberculosis* is predominantly caused by mutations interrupting drug-target interactions rather than uptake of mobile elements carrying resistance genes. Several publicly available databases provide information on resistance-mediating mutations in *M. tuberculosis*.

We searched for resistance-causing variants in the dataset using our optimized "Low Frequency Variant Detection" tool and matched detected variants against a customized database of known resistance causing mutations (Figure 1). We created a database based on the publication by Coll et al. (2). Coll and co-workers combined variants from publically available databases TBDreaM (7) and MUBII-TB-DB (4), as well as recent literature. The database from Coll et al. (4) contains only high-confidence variants. To this, we added variants from Miotto et al. (6) and Allana et al. (1). The resulting custom database contained nearly 1500 variants in 31 loci, conferring resistance towards 15 different anti-TB drugs. **Download database**.

Using our optimized variant detector, we found 123 resistance-causing variants in the genomes of the 13 TB isolates studied (Table 1). Seventy-five of these variants were described in the original paper, while the remaining 48 were new mutations. Compared to the original study, our more sensitive approach to variant detection resulted in increased congruence between detected genotypic resistance and the independently obtained results of antimicrobial susceptibility testing (Table 2).

Table 1. Detected antimicrobial resistance causing variants

Drug	Locus	No. of variants detected in paper	No. of additional variants detected
Capreomycin	*tlyA*	0	2
Capreomycin	*rrs*	2	4
Clofazimide	*RV0678*	0	1
Ethambutol	*embB*	15	1
Ethambutol	*embC*	0	1
Ethambutol	*embR*	0	2
Ethionamide	*ethA*	5	0
Capreomycin	*gyrA*	1	1
Fluoroquinolones	*gyrB*	0	1
Fluoroquinolones	*fabG1* promoter	0	10
Isoniazid	*katG*	13	5
Isoniazid	*eis* promoter	0	6
Kanamycin	*thyA*	1	0
Fluoroquinolones	*pncA*	11	8
Fluoroquinolones	*rpoB*	14	2
Isoniazid	*rpoC*	0	4
Isoniazid	*gidB*	11	0
Kanamycin	*rpsL*	2	0

Table 2. Comparison of genotype and phenotype

	Original study		New analysis	
	No. of variants	Percentage	No. of variants	Percentage
In congruence with phenotype	61	81	30	63
Unknown phenotype (antimicrobial not tested)	8	11	14	29
In contrast with phenotype	6	8	4	8
Total	**75**		**48**	

To further explore and qualify selected variants, we searched for 3D protein models displaying detected variants, using our tool "Link Variants to 3D Protein Structure". The tool takes as input a list of variants and searches against the Protein Data Bank (PDB) for proteins of sufficient homology. A visualization of amino acid alterations is created and can be explored interactively in the Workbench.

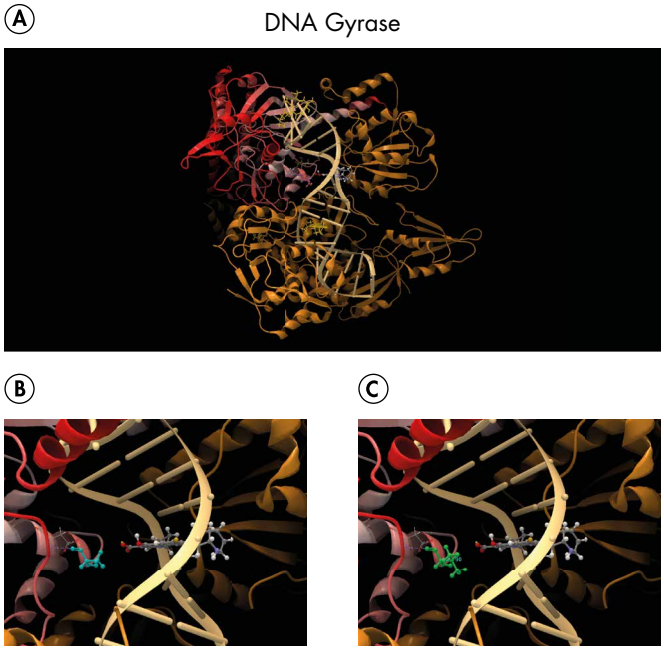
One of the additional variants we detected compared to the original study was within the DNA gyrase. DNA gyrase is the target for fluoroquinolones. DNA gyrases are involved in supercoiling of DNA, binding DNA and introducing double-stranded breaks. Fluoroquinolones bind and lock the gyrase-DNA complex. Mutation of the target region of the gyrase alters the binding affinity for the drug, resulting in resistance.

Figure 3 shows the DNA gyrase in complex with double stranded DNA and the fluoroquinolone Moxifloxacin. The wildtype at amino acid position 90 (blue) in the active site of the gyrase is shown in the lower left panel. The lower right panel shows the corresponding variant position (green). As shown in the lower right panel, the variant is positioned in the drug-binding site and likely confers resistance.

Another variant we investigated in more detail was the Ser431Asn substitution in Subunit B of the RNA Polymerase. RNA polymerase is the target for the Rifamycin antimicrobials, which bind inside the RNA/DNA channel, physically blocking elongation. Antimicrobial resistance arises from amino acid alterations in the binding channel, decreasing the affinity for the drug.

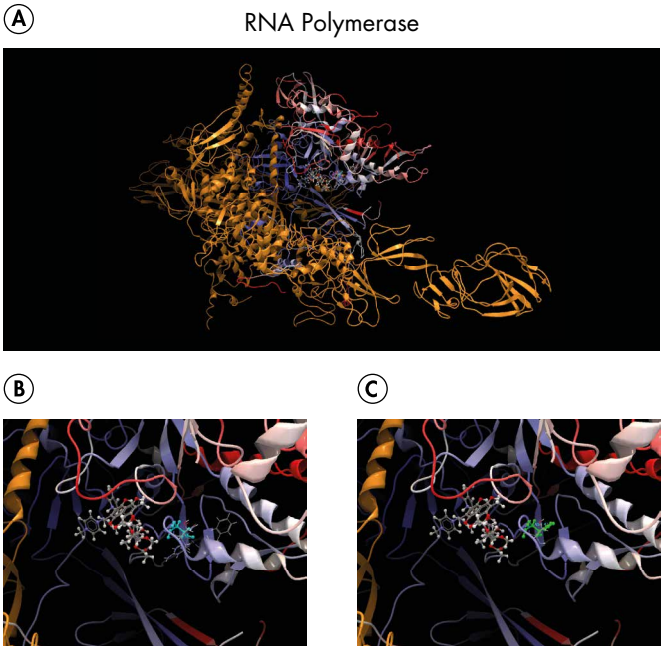
Figure 4 displays the RNA Polymerase with Rifampicin bound in the active site. The lower left panel shows the wildtype at amino acid position 431 (blue) and the lower right panel shows the corresponding variant position (green). As shown in the lower right panel, the variant is located in the drug-binding site, likely altering drug affinity.

In both cases, the detected amino acid changes are in line with antimicrobial susceptibility test results, showing resistance towards the antimicrobial. 3D visualization of variants can be an advantageous method for predicting and qualifying the effect of previously undescribed variants on antimicrobial susceptibility.



A GyrA
B Reference model
C Variant model – Ala90Val

Figure 3. 3D protein model. DNA gyrase in complex with double stranded DNA and Moxifloxacin displaying the consequences of the substitution Ala90Val in subunit A is shown.



A RpoB
B Reference model
C Variant model – Ser431Asn

Figure 4. 3D protein model. RNA Polymerase Subunit B in complex with Rifampicin displaying the consequences of the substitution Ser431Asn is shown.

Conclusion

This study highlights that whole genome sequencing provides significantly higher resolution in outbreak analyses compared to traditional genotyping methods. In addition to revealing the sequence of transmission events, whole genome sequencing additionally provides information on antimicrobial resistance mechanisms present in the bacterium. We have demonstrated how the tools and workflows of CLC Genomics Workbench and CLC Microbial Genomics Module streamline accurate analysis of pathogen transmission and can greatly improve detection of genotypic antimicrobial resistance. Preconfigured workflows ensure ease of use, and built-in data visualization allows clear and interactive graphical representations of results and associated metadata.

References

1. Allana, S. et al. (2017) pncA gene mutations associated with Pyrazinamide resistance in drug-resistant Tuberculosis, South Africa and Georgia. Emerg Infect Dis. 3, 491–495.
2. Coll, F. et al. (2015) Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 7(1), 51.
3. Fiebig, L. et al. (2017) A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt. Euro Surveill. 22(2).
4. Flandrois, J., Lina, G., and Dumitrescu, O. (2014) MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*. BMC Bioinformatics. 15,107.
5. Islam, M. M. et. al. (2017) Drug resistance mechanisms and novel drug targets for tuberculosis therapy. J Genet Genomics. 44(1), 21–37.
6. Miotto, P. et. al. (2014) *Mycobacterium tuberculosis* pyrazinamide resistance determinants: a multicenter study. MBio. 5(5).
7. Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B. K., Church, G. M., Murray, M. B., (2009) Tuberculosis Drug Resistance Mutation Database. PLoS Med. 6(2).

To learn more, have a look at these informative tools:

Product sheets
CLC Genomics Workbench **https://www.qiagenbioinformatics.com**
CLC Microbial Genomics Module **https://www.qiagenbioinformatics.com**

Tutorials
Typing and epidemiological clustering of pathogens
http://resources.qiagenbioinformatics.com/tutorials/Typing_Epidemiological_Clustering.pdf

Discover more at **www.qiagenbioinformatics.com**.