

C C A T T 0 1 0 0 0
G A G G A 0 1 1 0 1
G A A T T 0 0 1 1 0
A C A A G 0 0 1 0 0
T A C C A 0 0 1 1 0
T T A C A 0 1 0 0 0
A C C T C 0 0 0 1
A A G G A 0 0 0 0
G A T G A 0 1 1
T A G A T 0 0 1
G A T G A 1 0 1
T G T A G 1 0 0
T A G T A 0 0
G A T A T 1 0
G A G T G 1 1
A G A T T 1 1
G A G T A 1 1
T G A T G 1 1
A T T A G 1 1
T A G A T 1 1
G A G A 1 1
G T A 1 1
G A T 1 1
T A G 1 1
A G A 1 1
G A 1 1
A 1 1
T 1 1

Case Study

Detection of genomic changes in M. Tuberculosis

January 9, 2007

Detection of genomic changes in *M. Tuberculosis*

Resistance development in *Mycobacterium tuberculosis*

Tuberculosis (TB) remains the leading cause of death due to bacterial infections worldwide. About 8 million new cases of active TB arise each year resulting in 3 million annual deaths. Roughly 1 billion individuals are believed to harbour latent tuberculosis.

Primary infections with *M. tuberculosis* are generally asymptomatic but will in some cases remain as a latent infection.

Tuberculosis is a secondary disease caused by reactivation of the *M. tuberculosis* bacteria not fully eliminated after the primary infection.

Sequencing and sequence analysis of bacterial and fungal genomes and proteins/peptides have led to a better general understanding of the pathogenesis of bacterial and fungal infections. Future understanding of the regulatory events at the molecular level will increase and be accelerated by using a variety of new technologies and technology platforms within microarrays, protein chips and sequence analysis tools (CLC bio workbenches). The aim is to develop more specific and effective drugs much faster to target e.g. the expanding multi resistance *Mycobacterium tuberculosis* strains in human populations all over the world.

Multidrug resistant bacterial strains arise by sequential accumulation of resistance mutations for individual drugs. A diverse array of strategies is available to assist in rapid detection of drug resistance-associated gene mutations.

Bioinformatics (CLC bio workbenches) together with functional genomics and functional proteomics have been used to identify expression pattern (signature) changes in multi resistant *M. tuberculosis* strains.

Mycobacterium tuberculosis genome

The *M. tuberculosis* genome was sequenced in 1998 [Cole et al., 1998] (see figure 1). Today many strains of *M. tuberculosis* have been sequenced. In pest control laboratories, sequencing is routinely performed to monitor strain genomic changes.

Presently with the latest re-annotation of the *M. tuberculosis* genome, it is possible to assign a function to 2058 protein coding sequences (52% of the 3995 proteins predicted). Only 376 putative proteins share no homology with known proteins and thus could be unique to *M. tuberculosis*.

Drug resistance in *Mycobacterium tuberculosis*

Today eight agents are used in the treatment of TB. Multi drug resistant *M. tuberculosis* strains are resistant both to isoniazid (INH) and rifampicin (RFP). These agents have the most effective bactericidal activity towards *M. tuberculosis*. Nearly 95% of the RFP resistant strains possess a mutation in the *rpoB* gene encoding a DNA-dependent RNA polymerase. Approximately 90% of INH resistant strains have a mutation in the *inhA*, *katG*, and *ahpG* genes encoding enzymes related to a mycolic acid synthesis of cell wall. Pyrazinamide (PZA) resistant strains have a mutation in the *pncA* gene encoding a pyrazinamidase which degrades PZA to the bactericidal substance, pyrazinoic acid. Streptomycin resistant strains have a mutation in the *rrs* and *rpsL* gene encoding the 16S rRNA and the 12S ribosomal subunit protein, respectively. Kanamycin

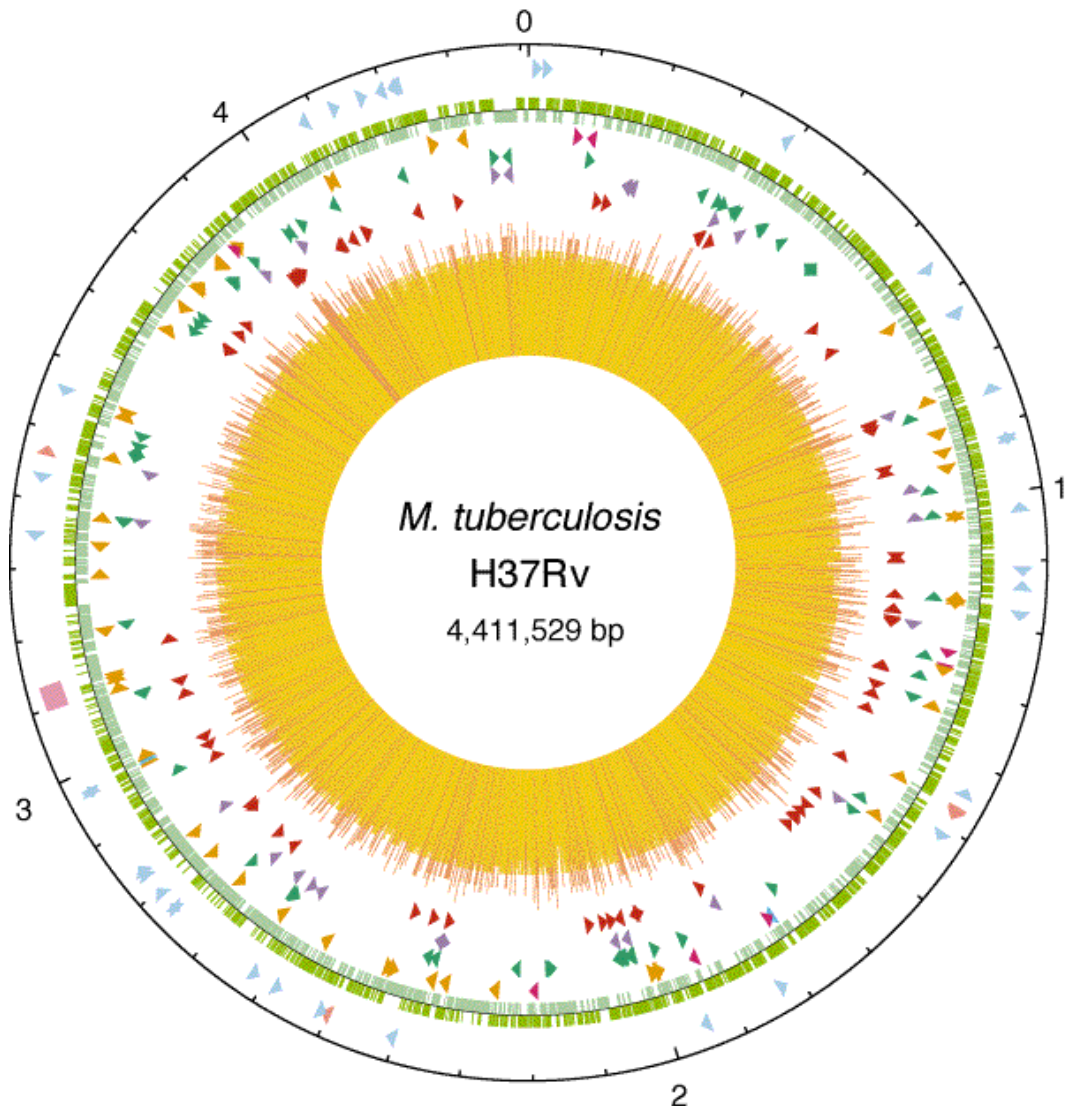


Figure 1: *The Mycobacterium tuberculosis genome [Cole et al., 1998].*

resistance is due to nucleotide substitutions in the *rrs* gene encoding 16S rRNA. Ethambutol resistant strains have a mutation in the *embB* gene encoding a arabinosyl transferase which catalyzes cell wall synthesis. Ethambutol resistance is in approximately 60% of organisms due to amino acid replacements at position 306 of an arabinosyltransferase encoded by the *embB* gene.

CLC bio provide some important bioinformatics solutions for detection and identification of mutations in selected genes of new strains of *M. tuberculosis* and other important pathogenic microbes. An example of this is described below.

Detection of mutations in *M. tuberculosis embB, rrs* and *pncA* genes

The work flow for detecting and identifying mutations using the *CLC Combined Workbench* is summarized in figure 2.

The multi resistance *M. tuberculosis* strains are identified and the genomic DNA extracted.

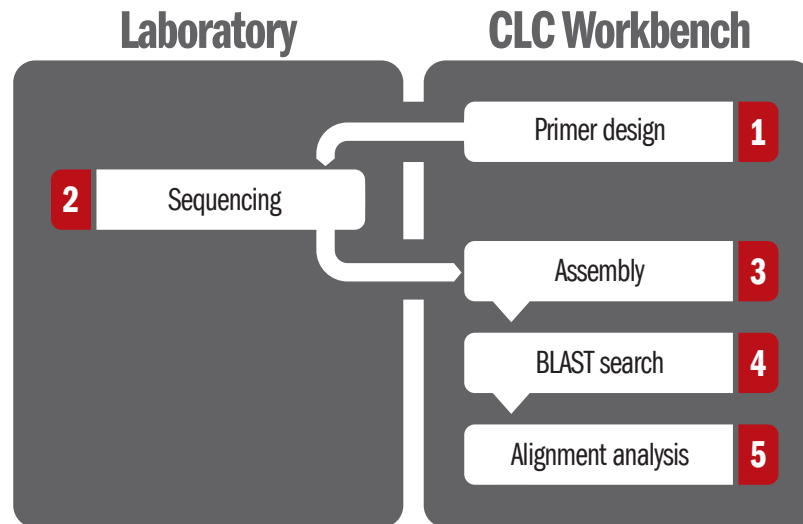


Figure 2: Illustration of work flow.

The encoding region of the *embB*, *rrs* and *pncA* genes are marked with primers using the graphical primer design. The relevant regions are sequenced automatically and assembled into contigs. BLAST searches in selected databases (both local and NCBI databases) are performed. Subsequently homologous sequences are aligned and annotations are transferred to the contig sequence.

In the next two sections, two of the steps in the work flow are described in further detail.

Zooming in on Primer design

Using the integrated GenBank search function, the annotated *M. tuberculosis* is downloaded. In a few clicks, the relevant genes (*embB*, *rrs* and *pncA*) are extracted. Guided by the CDS annotations, a set of primers are calculated for each gene (see figure 3). Using the interactive primer design functionality of the *CLC Combined Workbench*, the primers can be designed to match very specific needs.

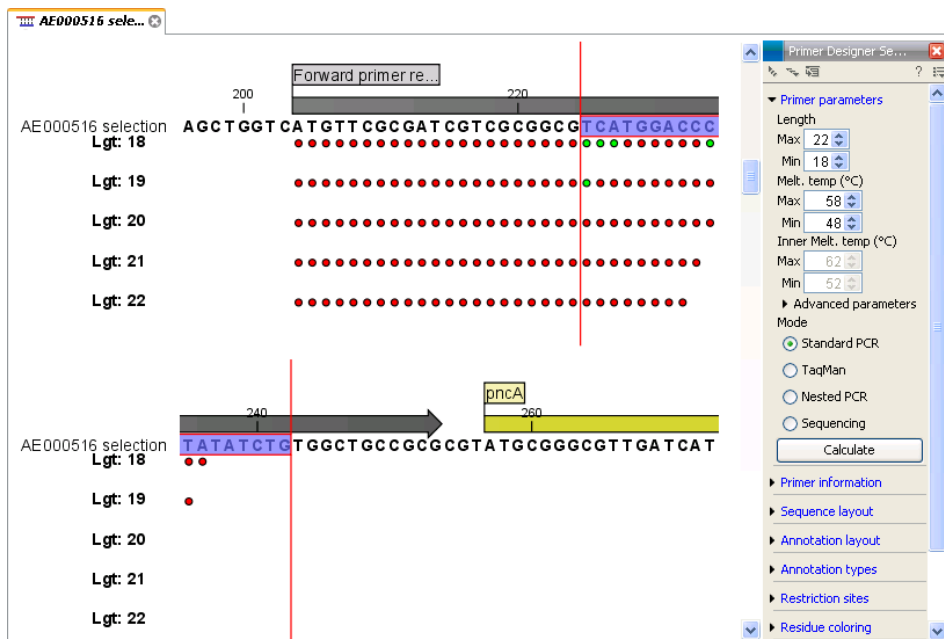


Figure 3: Primer design guided by CDS annotations (the yellow boxes).

Zooming in on Assembly

The sequencing data is imported into the CLC Workbench, and during the assembly, the genes are automatically divided into separate contigs.

The inconsistencies that exist between different reads are inspected using the variance table (see figure 4). The reads are both forward and reverse, which are automatically detected during

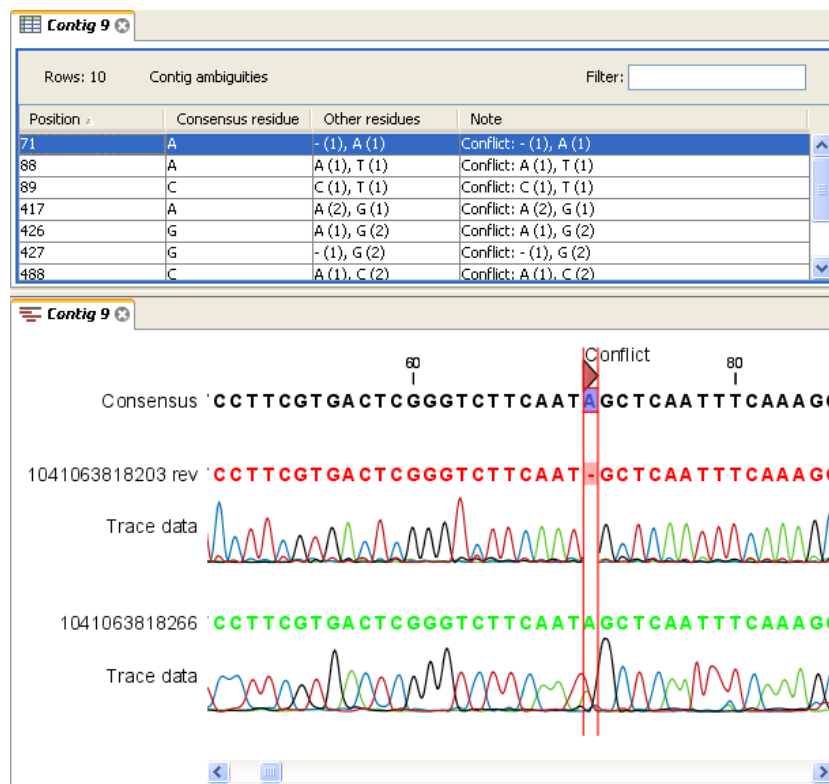


Figure 4: Assembling of raw sequence data. The upper image is the contig variance table, showing where there are conflicts/variation between the reads. When selecting a row in the table, the bottom view automatically selects this position in the contig, and it can be inspected in more detail.

the assembly. The orientation is reflected in the color of the reads (in figure 4, the read at the top is reversed which is indicated by the red color of the residues). Quality scores are assigned in order to trim low-quality trace data, so that reads align properly. The quality scores can also be shown graphically below the sequence.

The result of the assembly is a contig sequence which is used in the following steps for BLAST searches and alignments.

The description above is only a brief example of how bioinformatics software, such as the CLC Workbenches, can be used in clinical microbiology. Although remarkable advances have been made, much remains to be learned about the molecular genetic basis of drug resistance development including M. tuberculosis. New therapeutics will be developed based on improved bioinformatic data-mining of results from bacterial drug resistance studies.

References

[Cole et al., 1998] Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. (1998). Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

