Automated connection of diagnostics with instant outbreak triage and live global surveillance for TB.

Zamin Iqbal, Phelim Bradley1, Mark Thomsit1, Simon Heys1, Penelope Wintringer1, Mariana López4, Martin Hunt1, Michael Hall1, Simon Grandjean2, David Moore3, Iñaki Comas4

1. EMBL-EBI
2. University of Montreal
3. London School of Hygiene and Tropical Medicine
4. Institute of Biomedicine of Valencia (IBV-CSIC)

M. tuberculosis, causal agent of the disease tuberculosis (TB) causes over a million deaths annually, and is responsible for as much as 1/3 of drug resistant infections globally. Genomics offers a route to rapid and comprehensive diagnostics, and if we could enable default (or at least common) data-sharing, there would be huge global benefits for public health epidemiology and drug discovery. However, the transition from research to routine public health usage is challenging: high and low burden countries have different needs, there is no directly useful way to share, and the act of sharing currently offers no direct benefit to the person doing it.

We set out to build a publicly available service to address this. Mykrobe Atlas provides rapid offline diagnostics for illumina or nanopore data, that run on a laptop. When a network connection is available, the raw sequence data is synched to the EBI, decontaminated of human and HIV data, and deposited in the ENA under embargo. Rapid turnaround (minutes) gives comparison with a live updated database of all TB ever deposited in NCBI/EBI, including identification of potential outbreak clusters. The user is able to filter and visualise the global set by genetic, geographical, or metadata criteria (e.g. phenotype). The global library of TB uses the BIGSI search index (Bradley et al, 2019), which scales to millions of samples, distributed across multiple disks (low RAM usage). On a slower timescale (within an hour) the data is run through the full variation analysis pipeline that is being evaluated at Public Health England. Of particular note is the system that allows countries to participate while maintaining control of their data, by supporting centralised data with access control *and* distributed databases controlled by the data owners.

I will discuss this concretely by looking at a real data from an ongoing (1993-2018) outbreak from the Canary Islands in the context of our current database of over 38,250 M. tuberculosis isolates. I will show how we find both the known outbreak samples from Valencia and Canary Islands, and also related samples from across Europe and Africa.

I will finish by detailing future plans, including public availability, incorporation of 100k genomes including thousands with phenotypes, upload of lossy compressed data when internet bandwidth is low, and integration with Nextstrain.