Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Commentary

Development of a New Genome-Wide MLST Scheme for High-Resolution Typing of Diverse *Mycobacterium tuberculosis* Complex Strains

Ronan F. O'Toole [a,b,*]

[a] School of Medicine, College of Health and Medicine, University of Tasmania, Hobart, Australia
[b] Department of Clinical Microbiology, Trinity College Dublin, Ireland

ARTICLE INFO

Article history:
Received 28 July 2018
Accepted 30 July 2018
Available online 4 August 2018

ABSTRACT

In this issue of *EBioMedicine*, Kohl and colleagues describe the development of a new core genome MLST scheme (cgMLST) for *Mycobacterium tuberculosis* complex strains based on a set of 2891 genes. Here, the application of the scheme to a number of tuberculosis surveillance studies is examined.

In their 1998 paper, Maiden and colleagues published a novel bacterial typing technique they referred to as multi-locus sequence typing (MLST) in which they amplified a core set of genes by PCR and correlated them at the level of nucleotide sequences. They based their MLST scheme on a subset of 6 housekeeping genes – a putative ABC transporter (*abcZ*), adenylate kinase (*adk*), shikimate dehydrogenase (*aroE*), glucose-6-phosphate dehydrogenase (*gdh*), pyruvate dehydrogenase subunit (*pdhC*), and phosphoglucomutase (*pgm*). This was sufficient to provide a higher level of strain differentiation than the earlier technique multi-locus enzyme electrophoresis (MLEE) [1].

Maiden et al. stated that: "The overwhelming advantage of MLST over other molecular typing methods is that sequence data are truly portable between laboratories, permitting one expanding global database per species to be placed on a World-Wide Web site, thus enabling exchange of molecular typing data for global epidemiology via the Internet." [1]. It is apparent that the authors foresaw the potential of this sequence typing technique when coupled to the rapidly growing internet. The authors also concluded that "MLST can be applied to almost all bacterial species and other haploid organisms, including those that are difficult to cultivate." [1]. It is therefore not surprising that MLST has been investigated as a possible sequence typing tool for *Mycobacterium tuberculosis*.

A study by Pitondo-Silva and co-workers directly compared the efficacy of MLST with respect to two other genotyping techniques that were in widespread use for *M. tuberculosis i.e.* spacer oligotyping or "spoligotyping" [2], and mycobacterial interspersed repetitive unit (MIRU) typing [3]. Unfortunately, using the conventional MLST approach of basing the scheme on a small set of housekeeping genes, in this case 7 *M. tuberculosis* genes (*gyrA*, *gyrB*, *katG*, *purA*, *recA*, *rpoB* and *sodA*), MLST performed least well out of the three genotyping approaches [4]. The investigators surmised that while MLST is a useful tool for many bacterial species, it had low discriminatory power for *M. tuberculosis* and may not be applicable for this species [4]. However, Pitondo-Silva et al. also noted that the high sequence conservation of housekeeping genes in *M. tuberculosis* limits the resolution of MLST but that the advent of next generation sequencing (NGS) opened up the possibility of the development of MLST schemes that were based on a large number of *M. tuberculosis* genes [4].

This set the scene for work by Kohl, Niemann and colleagues at the German Centre for Infection Research, Borstel, and partner institutes. In their paper published in this issue of *EBioMedicine*, Kohl et al. performed a number of key steps for generating a new MLST scheme from whole-genome sequence data that can be applied to the typing of *M. tuberculosis*. For the derivation of a set of core MLST loci, they included 45 genomes from isolates of *Mycobacterium tuberculosis* complex that spanned all of the Gagneux Lineages 1 to 7, as well as isolates of animal-adapted species *M. bovis*, *M. caprae*, *M. microti*, and *M. pinnipedii* [5]. This represents a significant advance on the authors' earlier study from 2014 in which a total of only 7 input genomes were used from Lineage 4, Lineage 6, and *M. bovis* [6]. A set of 2891 genes made up the new core genome MLST (cgMLST) scheme [5].

Kohl et al. then evaluated the cgMLST scheme against a reference collection of 251 strains that consisted of an extensive range of *M. tuberculosis* lineages including also *M. canettii* and *M. orygis*. As shown in Tables 1 and 2 of their *EBioMedicine* paper, they found that at least 97.4% of the genes included in their core set were present in all of the mycobacterial strains analysed [5] compared to 94.2% of

DOI of original article: https://doi.org/10.1016/j.ebiom.2018.07.030.

* Corresponding author at: School of Medicine, College of Health and Medicine, University of Tasmania, Hobart, Australia.
E-mail address: ronan.otoole@utas.edu.au.

https://doi.org/10.1016/j.ebiom.2018.07.038

genes from their earlier scheme [6]. This emphasises the importance of incorporating a large and diverse array of input genomes when generating a cgMLST scheme for application to medical and veterinary *M. tuberculosis* complex isolates from different geographical regions.

Kohl and co-workers then tested their cgMLST scheme against the genomes of 390 previously sequenced cross-sectional, longitudinal, household, and community *M. tuberculosis* isolates from an earlier study conducted in the UK. As shown in Fig. 2a, they found that a threshold of ≤ 5 different cgMLST alleles applied to epidemiologically-linked isolates, similar to a previously-described single polymorphism nucleotide (SNP) based threshold of ≤ 5 SNPs [7], with a threshold of >12 different cgMLST alleles used to exclude isolates from a cluster. To further validate the performance of the scheme, Kohl et al. applied it to 52 isolates from seven clusters in a prospective longitudinal surveillance study undertaken in Hamburg from 1997 onwards. In Fig. 3 of their *EBioMedicine* paper, the authors illustrate that the cgMLST scheme positioned isolates within clusters in a similar manner to SNP-based minimum spanning trees utilising a maximum distance of 12 distinct cgMLST alleles/12 SNPs to the nearest group member [5]. The authors report that at least 98.6% of their cgMLST scheme targets met quality thresholds for cluster detection.

In summary, SNP thresholds have proved useful in establishing recent TB transmission and in identifying epidemiologically-linked cases [7, 8], but SNP calling can be a relatively complicated process and is subject to inter-laboratory variation and the influence of genomic features such as repetitive regions. By refining the variant analysis to a common set of *M. tuberculosis* complex genes, the work by Kohl et al. has helped simplify and standardise the analysis of genome-wide data for TB surveillance purposes. Their findings should widen the current bottleneck associated with the translation of whole-genome sequencing from a research tool to a routine TB diagnostic and public health molecular epidemiological technique. The open availability of their cgMLST scheme may also lead to a more unified approach with regard to comparing genomic data from different reference laboratories and jurisdictions. The apparent cost of proprietary software used by the authors for their cgMLST data analysis may still be prohibitive to potential users in low resource settings and it is hoped that such users will have access to the necessary software through discount, subsidy or similar initiatives in the future.

Conflicts of Interest

The author declares no financial or other conflicts of interest.

References

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., et al., 1998]. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95, 3140–3145.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., et al., 1997]. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 35, 907–914.

Supply, P., Magdalena, J., Himpens, S., Locht, C., 1997]. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. Mol Microbiol 26, 991–1003.

Pitondo-Silva, A., Santos, A.C.B., Jolley, K.A., Leite, C.Q.F., ALDC, Darini, 2013]. Comparison of three molecular typing methods to assess genetic diversity for Mycobacterium tuberculosis. J Microbiol Methods 93, 42–48.

Kohl, T.A., Harmsen, D., Rothgänger, J., Walker, T., Diel, R., Niemann, S., 2018]. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. EBioMedicine (in press).

Kohl, T.A., Diel, R., Harmsen, D., Rothganger, J., Walter, K.M., Merker, M., et al., 2014]. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 52, 2479–2486.

Walker, T.M., Ip, C.L., Harrell, R.H., Evans, J.T., Kapatai, G., Dedicoat, M.J., et al., 2013]. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13, 137–146.

Nikolayevskyy, V., Kranzer, K., Niemann, S., Drobniewski, F., 2016]. Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: a systematic review. Tuberculosis (Edinb) 98, 77–85.