EBioMedicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Harmonized Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature System

Thomas A. Kohl [a], Dag Harmsen [c], Jörg Rothgänger [d], Timothy Walker [e], Roland Diel [f], Stefan Niemann [a,b,*]

[a] Molecular and Experimental Mycobacteriology, Forschungszentrum Borstel, 23845 Borstel, Germany
[b] German Center for Infection Research, Borstel Site, 23845 Borstel, Germany
[c] Department of Periodontology and Restorative Dentistry, University Hospital Münster, 48149 Münster, Germany
[d] Ridom GmbH, 48149 Münster, Germany
[e] Nuffield Department of Medicine, University of Oxford, Oxford, UK
[f] Institute for Epidemiology, Schleswig-Holstein University Hospital, 24105 Kiel, Germany

ARTICLE INFO

Article history:
Received 27 June 2018
Received in revised form 19 July 2018
Accepted 20 July 2018
Available online xxxx

Keywords:
Whole genome sequencing
Mycobacterium tuberculosis
Core genome MLST
Molecular epidemiology
Genotyping

ABSTRACT

Background: Global tuberculosis (TB) control is challenged by uncontrolled transmission of *Mycobacterium tuberculosis* complex (Mtbc) strains, esp. of multidrug (MDR) or extensively resistant (XDR) variants. Precise analysis of transmission networks is the basis to trace outbreak M/XDR clones and improve TB control. However, classical genotyping tools lack discriminatory power due to the high similarity of strains of particular successful lineages, e.g. Beijing or outbreak strains. This can be overcome by whole genome sequencing (WGS) approaches, but these are not yet standardized to facilitate larger investigations encompassing different laboratories or outbreak tracing across borders.

Methods: We established and improved a whole genome gene-by-gene multi locus sequence typing approach encompassing a stable set of core genome genes (cgMLST) and linked it to a web-based nomenclature server (cgMLST.org) facilitating assignment and storage of allele numbers.

Findings: We evaluated and refined a previously suggested cgMLST schema by using a reference strain set ($n = 251$) reflecting the global diversity of the Mtbc. A set of 2891 genes showed excellent performance with at least 97% of the genes reliably identified in strains of all Mtbc lineages and in discriminating outbreak strains. cgMLST allele numbers were automatically retrieved from and stored at cgMLST.org.

Interpretation: The refined cgMLST schema provides high resolution genome-based typing of clinical strains of all Mtbc lineages. Combined with a web-based nomenclature server, it facilitates rapid, high-resolution, and harmonized tracing of clinical Mtbc strains needed for prospective local and global surveillance.

* Corresponding author at: Molecular and Experimental Mycobacteriology, Forschungszentrum Borstel, Parkallee 1, 23845 Borstel, Germany.
E-mail address: sniemann@fz-borstel.de (S. Niemann).

https://doi.org/10.1016/j.ebiom.2018.07.030

Please cite this article as: Kohl, T.A., et al., Harmonized Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature System, EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.07.030

Research in context

Evidence Before This Study

Whole genome sequencing (WGS) is currently widely developed for genotyping of bacterial pathogens. While an efficient tool for high resolution strain typing employing genomic data for comparison of *Mycobacterium tuberculosis* complex (Mtbc) strains and tuberculosis (TB) surveillance is urgently needed, workflow standardization and a unified nomenclature for genome-based strain typing have not yet been established. At present, this makes "genomic epidemiology" an interesting academic endeavour, but with limited practical use for cross border transmission surveillance. Here, the extension of multi locus sequence typing (MLST) which has previously been used for harmonization of classical sequencing data of few conserved genes, to the genome level has been recently suggested as an efficient approach for a unified and easily standardizable genomic typing method. Core genome MLST (cgMLST) schemes have already been suggested for some bacterial pathogens, e.g. *Staphylococus aureus*, *Legionella pneumophila*, or *Campylobacter jejuni*. Previous to the present work, there was only a provisional cgMLST scheme of 3257 genes suggested by our group for clinical Mtbc strains. This scheme had been established from a limited number of genomes, i.e. four strains from *M. tuberculosis* Lineage 4, one *M. africanum* Lineage 6 strain, and two *M. bovis* strains. While the scheme could successfully be used for the analysis of a TB outbreak caused by a *M. tuberculosis* Lineage 4 strain, the strains used to construct the scheme do not represent the global phylogenetic diversity of the Mtbc and it remained unclear whether the scheme could be used for efficient analysis of the full Mtbc species diversity. Furthermore, it has not been linked to an open web-based nomenclature system ensuring harmonized assignment of cgMLST allele types.

Added Values of This Study

Using an extensive set of strains representing the known diversity of the Mtbc, we show the limits of the previously suggested cgMLST scheme for defined subgroups of the Mtbc. This clearly demonstrates the necessity to consider the full species/complex diversity for construction of a cgMLST scheme, even for a pathogen with a relatively stable genome such as the Mtbc. Using a set of 45 strains covering the known diversity of the Mtbc, we developed an improved cgMLST scheme consisting of 2891 genes and showed its applicability to the whole Mtbc diversity using a comprehensive set of reference strains. In order to enable its use in pathogen surveillance, we define critical thresholds for likely (at most five distinct alleles) and unlikely recent transmission events (>12 distinct alleles), the evolutionary rate of allele change, and the correlation with classical genotyping data. As a necessary prerequisite for standardized use of the cgMLST scheme for unified genotyping across different laboratories, a central web-based nomenclature server has been established.

Implications of All the Available Evidence

The improved cgMLST scheme can be directly used for standardized typing and surveillance of clinical isolates of all lineages of the Mtbc. This is especially critical for setting up a surveillance tool for multi-country, cross-border analysis of strain transmission and outbreak dynamics. WGS data are transferred into a set of 2891 numeric values that can be efficiently exchanged between laboratories and public health agencies. The thresholds for likely transmission and allele mutation rates can directly be used for the detection of recent transmission and the analysis of pathogen evolution. Importantly, we also suggest the definition of a cgMLST-based complex type, which facilitates even easier communication in outbreak situations such as the recently confirmed outbreak of a multidrug resistant Mtbc strain among Somalian refugees in the European Union. Finally, the web-based cgMLST allele server provides a universally harmonized allele nomenclature and generation of comparable cgMLST genotypes on a global scale.

1. Introduction

Tuberculosis (TB) remains a main global health challenge, with roughly a third of the human population latently infected, approx. 10.4 million new TB cases and 1.7 million deaths in 2016 [1]. The efforts in controlling TB are seriously challenged by the increasing numbers of multiple (MDR) or extensively resistant (XDR) *Mycobacterium tuberculosis* complex (Mtbc) strains [1]. Overall, approx. 490,000 new MDR-TB cases have been reported in 2016, with about 6% of them being XDR [1].

Transmission is the major driving force of the epidemic, esp. of the MDR pandemic [2–5]. As there is no environmental reservoir, controlling human-to-human transmission is key for successful local and global TB control and for achieving the targets of the "End TB strategy" proposed by the World Health Organisation [1]. Targeted interventions to stop transmission esp. of M/XDR strains requires in depth epidemiological knowledge that can only be provided by a combination of effective genotyping with classical epidemiology in molecular epidemiological studies [4, 6]. Classical genotyping techniques such as spoligotyping or MIRU-VNTR typing (Mycobacterial Interspersed Repetitive Units - Variable Number of Tandem Repeats) interrogating just a small fraction of the genome have been used for years to study transmission dynamics in low and high incidence settings, population structure of the pathogen, and global spread of particular strains/lineages [3–6]. While they provide standardized, easily computable typing results with an on-line nomenclature, recent studies indicate that their discriminatory power is too low to differentiate outbreak strains esp. in high incidence settings with the resolution needed to trace recent transmission chains [4, 7–10].

Several studies have now demonstrated that whole genome sequencing (WGS) based genotyping has an improved discriminatory power compared to classical genotyping, e.g. better discrimination of outbreak strains [7–10]. Furthermore, WGS provides a highly accurate identification of the diverse lineages of the Mtbc and potentially allows for comprehensive detection of drug resistance mediating genomic variants [6, 10–13].

While these advantages of WGS-based strain analysis are commonly accepted, there are several drawbacks limiting the routine application of WGS for transmission analysis and diagnostic questions. Most importantly, there are not yet any standardized analysis pipelines, leading to inherent problems in defining single nucleotide polymorphisms (SNPs) later included in SNP-based similarity analysis commonly used for strain comparison. Furthermore, there is not yet a common nomenclature to standardize WGS data to facilitate data exchange in an easily extendable classification scheme. The advantages and problems of WGS-based pathogen tracing have recently been demonstrated by a European wide MDR-TB outbreak, in which high resolution WGS-based investigation delineated likely transmission routes, however, analysis was done in a central laboratory to allow strain comparison [14].

These limitations can be overcome by using a whole genome gene-by-gene allele calling approach [15, 16], which extends the concept initially developed for sets limited to six or seven house-keeping genes used in multi locus sequence typing (MLST) [15], to a genome-wide

Please cite this article as: Kohl, T.A., et al., Harmonized Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature System, EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.07.030