Tuberculosis 94 (2014) 346–354

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

PolyTB: A genomic variation map for *Mycobacterium tuberculosis*

Francesc Coll [a,*], Mark Preston [a], José Afonso Guerra-Assunção [b], Grant Hill-Cawthorn [c,d], David Harris [e], João Perdigão [f], Miguel Viveiros [g], Isabel Portugal [f], Francis Drobniewski [h], Sebastien Gagneux [i], Judith R. Glynn [b], Arnab Pain [c], Julian Parkhill [e], Ruth McNerney [a], Nigel Martin [j], Taane G. Clark [a,b]

[a] *Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK*
[b] *Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK*
[c] *Pathogen Genomics Laboratory, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*
[d] *Sydney Emerging Infections and Biosecurity Institute and School of Public Health, Sydney, NSW 2006, Australia*
[e] *Pathogen Genomics Faculty, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA Cambridge, UK*
[f] *Centro de Patogénese Molecular, Faculdade de Farmácia da Universidade de Lisboa, 1649-003 Lisboa, Portugal*
[g] *Grupo de Micobactérias, Unidade de Microbiologia Médica, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, 1349-008 Lisboa, Portugal*
[h] *Centre for Immunology and Infectious Disease, Queen Mary University of London, E1 2AT London, UK*
[i] *Swiss Tropical and Public Health Institute, 4002 Basel, Switzerland*
[j] *School of Computer Science and Information Systems, Birkbeck College, WC1E 7HX London, UK*

ARTICLE INFO

Article history:
Received 27 October 2013
Accepted 8 February 2014

Keywords:
Mycobacterium tuberculosis
Database
Genomics
Software
Molecular epidemiology
Whole-genome sequencing

SUMMARY

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (Mtb) is the second major cause of death from an infectious disease worldwide. Recent advances in DNA sequencing are leading to the ability to generate whole genome information in clinical isolates of *M. tuberculosis* complex (MTBC). The identification of informative genetic variants such as phylogenetic markers and those associated with drug resistance or virulence will help barcode Mtb in the context of epidemiological, diagnostic and clinical studies. Mtb genomic datasets are increasingly available as raw sequences, which are potentially difficult and computer intensive to process, and compare across studies. Here we have processed the raw sequence data (>1500 isolates, eight studies) to compile a catalogue of SNPs ($n = 74{,}039$, 63% non-synonymous, 51.1% in more than one isolate, i.e. non-private), small indels ($n = 4810$) and larger structural variants ($n = 800$). We have developed the PolyTB web-based tool (http://pathogenseq.lshtm.ac.uk/polytb) to visualise the resulting variation and important meta-data (e.g. *in silico* inferred strain-types, location) within geographical map and phylogenetic views. This resource will allow researchers to identify polymorphisms within candidate genes of interest, as well as examine the genomic diversity and distribution of strains. PolyTB source code is freely available to researchers wishing to develop similar tools for their pathogen of interest.

1. Introduction

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (Mtb) is an important global health issue, being the second leading cause of death from an infectious disease worldwide. The World Health Organisation (WHO) has set the ambitious target of "elimination" of TB by 2050. Widespread (multi- and extensive) drug resistance and high HIV prevalence (13% of new infections) are a serious challenge to effective control [1]. There is an urgent need for better treatments and vaccines, which in turn require a deeper understanding of the biology of Mtb and epidemiology of TB disease. Knowledge of the genomic variability among Mtb isolates could result in such insights, as well as mechanisms of virulence and transmission. Human TB is caused by bacteria belonging to the *M. tuberculosis* complex (MTBC), predominantly *M. tuberculosis*, *Mycobacterium bovis* and *Mycobacterium africanum* with occasional cases of infection with *Mycobacterium caprae*, *Mycobacterium microti*, *Mycobacterium pinnipedii*, *Mycobacterium orygis* and *Mycobacterium canettii*

* Corresponding author. Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. Tel.: +44 (0) 20 7636 8636; fax: +44 (0) 20 7436 5389.
E-mail address: francesc.coll@lshtm.ac.uk (F. Coll).

http://dx.doi.org/10.1016/j.tube.2014.02.005

reported. They are slow growing, lipid rich actinomycetales with characteristic cell walls conferring natural resistance to many antibiotics. Members of the MTBC are indistinguishable in their 16SrRNA and *rpoB* genes, recombination does not occur between strains and have approximately the same genome length; it is their host specificity what allows their differentiation [2]. It has been speculated that MTBC emerged from a common ancestor in the Horn of Africa and spread with human migrations [3–5]. Mtb is the prevailing cause of human pulmonary disease and six major global lineages have now been identified. First sequenced in 1998 [6], Mtb has a circular genome (size 4.4 Mb, GC content 65%) containing 4111 genes. No plasmids and horizontal gene transfer have been observed after the last common ancestor of MTBC [7]. The genome is characterised by limited sequence diversity resulting from a low mutation rate [8]. Insertion sequences are also responsible for genomic variation among MTBC isolates [9]. It may be said that drug treatments are driving changes in the Mtb genomes at a faster rate than any other evolutionary force [10]. In fact, polymorphisms are elevated in genes relating to antibiotic action as resistance to anti-TB drugs is caused predominantly by point mutations that arise spontaneously and are selected during unsatisfactory drug treatment. Sequential rounds of mutation and selection result in the emergence of strains resistant to multiple drugs turning TB in an even more difficult to treat disease.

Over the last two decades, molecular typing methods such as IS6110-RFLP [11], spoligotyping [12] and MIRU-VNTR [13] have been applied and revolutionised epidemiology of TB, by providing insights into the genetic diversity and population structure of MTBC [14]. Genotyping has been used extensively with epidemiological data to further understanding of TB [15]. For example, at the individual level, cases of recurrence or treatment failure can be explained in terms of reactivation with the same strain, exogenous re-infection or due to polyclonal infection [16]. At a population level, the origins and transmission dynamics of outbreaks can be determined [17–19]; whilst at a global level, TB genotypic lineages have been defined and used to monitor their geographical distribution [15]. Nevertheless, standard genotyping methods have several limitations. First, the repetitive nature of genetic polymorphisms used by molecular techniques makes them highly prone to convergent evolution [20], reducing their usefulness as phylogenetic markers. Second, the discriminative power differs between methods, meaning that results from different techniques are not always comparable [20]. Furthermore, isolates with identical DNA fingerprints have been reported to harbour significant genomic diversity [21]. Therefore standard genotyping tools, which are based on less than 1% of the genome, may not be able to accurately resolve transmission chains and distinguish disease relapse from exogenous re-infection conclusively. However, SNPs and other genetic polymorphisms derived from whole-genome sequencing (WGS) provide enough discriminatory power to assess population natural variation and predict its host–pathogen relation including virulence factors, drug susceptibility determinants and immune modulator factors with importance on the clinical manifestations [16]. Furthermore, due to its low mutation rate [19] and limited genomic diversity, the application of WGS in clinical settings is particularly effective for Mtb [22]. With the rapid decrease in DNA sequencing costs, it is foreseen that WGS will eventually be accessible and affordable enough to be an alternative to current lab-based genotyping techniques in the context of phylogenetic and epidemiological studies [18,19,23–25].

Given the large amount of data being generated on a routine basis from Mtb WGS projects, efforts must be focused on data analysis, accessibility, visualisation and utilisation. The TB community has a number of available web-based databases and tools to exploit the existing molecular epidemiological data [26], SNP repositories [27] and manually-annotated genomes [28]. Nevertheless, there is no tool harbouring genetic polymorphisms derived from WGS projects integrated with geographic distribution, strain type information and population structure visualisation. To fill this gap, we have developed PolyTB, a web-based tool to display Mtb genetic polymorphisms derived from publicly available WGS datasets. We compile a catalogue of SNPs, small indels and large deletions by employing the state-of-the-art variation discovery software [29]. Variants can be investigated through a genome browser reporting their chromosome coordinates, and a world map showing their global allele distribution. Additionally, the construction of phylogenetic trees based on SNPs provides an additional tool to investigate the population structure. Strain genotype information is incorporated, allowing the visualisation of associations of strain types with particular polymorphisms and/or geographical locations as well as helping correlate easily with public health epidemiological data. The integration of such data into tools like PolyTB is required to fully exploit genomic variation, and potentially boost TB control research through the discovery of new drug targets, vaccine antigens and diagnostics.

2. Materials and methods

Eight publicly available Mtb WGS datasets (Table 1) were downloaded from the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/). All isolates ($n = 1627$) had been sequenced using Illumina paired-end technology (Illumina-GAII or HiSeq 2000), and were aligned to the H37Rv reference genome (Genbank accession number: NC_000962.3) using BWA [30]. SAMtools/BCFtools (SAMTOOLS) [31] and GATK [32] were used to call both SNPs and small indels. Variants were then selected as the

Table 1
Publicly available Mtb WGS datasets included in PolyTB.

Population (reference)	ENA accession number	Sample size/ post-QC	Read length	Median read depth	Post-QC No. SNPs
Samara, Russia [10]	ERP000192	329/264	49	61	18,936
Midlands, UK [18]	ERP000276	390/390	75	112	19,406
Kampala, Uganda [52]	ERP000520	51/51	75	257	8021
Global key strains [4]	ERP001731	171/166	75/100	97	29,181
Bilthoven, Netherlands [19]	ERP000111	213/153	75/100	39	10,016
Vancouver, Canada [17]	SRP002589*	36/24	50	37.5	1026
Lisbon, Portugal (J. Perdigão et al., submitted for publication)	ERP002611**	84/81	100	104	6627
Karonga, Malawi (J. Guerra-Assunção et al., in preparation)	ERP000436	353/341	75	183	19,285
Overall		1, 627/1470			74,039

A set of 8 whole-genome sequencing (WGS) studies available in the public domain were downloaded from the European Nucleotide Archive (ENA). All samples were sequenced at the Wellcome Trust Sanger Institute, except * at the Simon Fraser University and ** at the King Abdullah University of Science and Technology (KAUST); all data generated using Illumina Genome Analyzer II technology, except Malawi (Karonga Study), Portugal (Lisbon) and Uganda (Kampalan Study) obtained using Illumina HiSeq 2000.