Bioinformatics, 2023, **39(11)**, btad648
https://doi.org/10.1093/bioinformatics/btad648
Advance Access Publication Date: 23 October 2023
Applications Note

OXFORD

Sequence analysis

fastlin: an ultra-fast program for *Mycobacterium tuberculosis* complex lineage typing

Romain Derelle [1], John Lees [2,3], Jody Phelan [4], Ajit Lalvani[1], Nimalan Arinaminpathy[1,2], Leonid Chindelevitch [2,*]

[1]NIHR Health Protection Research Unit in Respiratory Infections, National Heart and Lung Institute, Imperial College London, London W2 1PG, United Kingdom
[2]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London , London W12 0BZ, United Kingdom
[3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus , Hinxton CB10 1SD, United Kingdom
[4]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom
*Corresponding author. MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London W12 0BZ, United Kingdom. E-mail: lchindel@ic.ac.uk (L.C.)

Associate Editor: Can Alkan

Abstract

Summary: Fastlin is a bioinformatics tool designed for rapid *Mycobacterium tuberculosis* complex (MTBC) lineage typing. It utilizes an ultra-fast alignment-free approach to detect previously identified barcode single nucleotide polymorphisms associated with specific MTBC lineages. In a comprehensive benchmarking against existing tools, fastlin demonstrated high accuracy and significantly faster running times.

Availability and implementation: fastlin is freely available at https://github.com/rderelle/fastlin and can easily be installed using Conda.

1 Introduction

Lineage typing of *Mycobacterium tuberculosis* complex (MTBC), which notably includes the aetiological agent of tuberculosis disease, has evolved from traditional phenotypic methods to more advanced molecular techniques. Early methods relied on observable characteristics in the laboratory and broad classifications, while DNA-based methods like RFLP and spoligotyping provided increased resolution but had technical limitations (Jagielski *et al.* 2014).

The advent of whole-genome sequencing (WGS) has revolutionized the comparative analysis of MTBC samples, allowing for high-resolution classification (Wyllie *et al.* 2018, Diel *et al.* 2019). While large genomic deletions were initially proposed as lineage markers, the use of single nucleotide polymorphisms (SNPs) as barcode markers has gained prominence due to their higher abundance, enabling more precise identification and classification of MTBC lineages (Coll *et al.* 2014, Cancino-Muñoz *et al.* 2019, Napier *et al.* 2020).

The identification of barcode SNPs traditionally involves the alignment of WGS reads to a reference genome, the basis of popular tools such as TB-profiler (Phelan *et al.* 2019) and TbLG (Shitikov and Bespiatykh 2023). As it is primarily designed for general variant calling, read alignment is a complex and computationally intensive approach that scales poorly to large datasets. To overcome this challenge, we propose a new bioinformatics tool, fastlin, which uses an alignment-free approach and performs MTBC lineage typing in seconds rather than minutes.

2 Materials and methods

Fastlin is written in rust and takes as input a SNP barcode file and the directory containing the fastq files to be typed. The program essentially exploits the split-k-mer approach developed in SKA (Harris 2018), which uses k-mers split over a variable middle base, but with a significant speedup for lineage typing relying on the facts that: (i) fastlin does not need to compare and store all k-mers, only those relevant to barcode SNPs, and (ii) it typically does not need to scan all sequencing reads to identify barcode SNPs, assuming that the relevant reads are randomly distributed within the input fastq files.

The input barcode file contains all barcode SNPs with corresponding lineage names, and the 50 nucleotides upstream and downstream of each SNP, extracted from the genome of *M.tuberculosis* H37Rv; NC_000962.3. Fastlin then builds the barcode k-mers at the start of each run by combining $(k-1)/2$ downstream nucleotides, the barcode SNP and $(k-1)/2$ upstream nucleotides, with the k-mer size k being an odd number between 11 and 99 defined by the user (25 by default; see the Results section). The barcode k-mers are saved in memory together with their reverse complements. The input barcode file also contains the expected genome size, allowing fastlin to

Received: 8 July 2023; **Revised:** 6 October 2023; **Editorial Decision:** 17 October 2023; **Accepted:** 20 October 2023

estimate the number of extracted k-mers needed to reach the user-defined k-mer coverage threshold (turned off by default).

K-mers are then extracted from the input fastq files and compared to the barcode k-mers until the entire fastq file is scanned, or until the number of extracted k-mers reaches the k-mer coverage threshold if one has been specified by the user. Once this process is finished, only those k-mers with a minimum number of occurrences are retained, and a lineage is validated if the number of retained k-mer barcodes reaches a user-defined threshold (3 by default). Finally, the coarser lineage names are removed (e.g. '4.1' is removed if '4.1.2' is detected), leading to the final lineage determination. If more than one lineage is detected, their abundances are calculated using the median of their barcode SNP occurrences. The relative frequency of each lineage can then be calculated as the ratio of their abundance divided by the sum of all final lineage abundances.

The barcode file used in this study and the Python scripts used to build and test it are available at https://github.com/rderelle/barcodes-fastlin. Sample analyses were performed using TB-profiler v5.0.0, QuantTB v1.01 and fastlin v0.1.0. Runtime benchmarks were performed with SKA v1.0 and TB-profiler v4.4.2 (with 6G RAM memory) and fastlin v0.1.0 on a Macbook Air laptop (M2, 8G RAM, 2022) using a single CPU. The 3139 MTBC samples were selected to represent all MTBC lineages as predicted by the TB-profiler database. The two most important contributors to this dataset were the Wellcome Trust and the UKHSA (formerly PHE), as part of the CRyPTiC consortium. All results mentioned in this manuscript, including runtimes, are available in Supplementary Data S1.

3 Results

For this case study, we used the barcodes identified in TB-profiler (Phelan *et al.* 2019). At the time of writing, this set included 1101 SNPs defining 125 MTBC major and minor lineages, with a minimum of 4 and a maximum of 10 SNPs per lineage. Since these barcode SNPs were identified using read mapping, we first determined the minimal k-mer size at which fastlin did not generate false positive SNP calls. Using k-mer sizes ranging from 11 to 99 nucleotides and three high-quality genome assemblies with known lineages, we found that only one barcode SNP was incorrectly detected with k-mer sizes above 19 (Supplementary Data S2). We discarded this problematic SNP barcode and set the default k-mer size to 25.

We then typed 3139 MTBC samples downloaded from the NCBI's SRA database. To assess the quality of fastlin's lineage predictions we compared them to the ones made by TB-profiler. Fastlin and TB-profiler predicted identical lineages for most samples, with disagreements found in only 68 samples (Supplementary Data S3). Most of these disagreements were in putative mixed samples ($n = 51$; 16 and 44 mixed samples identified by TB-profiler and fastlin, respectively) or in samples for which either TB-profiler or fastlin failed to identify any lineage (eight and one samples, respectively). Phylogenetic analyses of the eight nonmixed samples for which fastlin and TB-profiler predicted different lineages showed that fastlin accurately identified the correct lineage in all but one case (Supplementary Data S3).

We then re-analyzed the 3139 MTBC samples with a maximum k-mer coverage threshold of 80× (the median k-mer coverage of all samples was 91×). Theoretically, with a minimum number of occurrences for each barcode SNP set to four by default, fastlin should be able to type all samples at this coverage threshold. Using lineage predictions obtained by fastlin with default parameters as a reference, we only found five samples with coarser (truncated) lineage predictions (e.g. predicted lineage '2.2.1' instead of '2.2.1.2'; ERR3276004, ERR3276023, and SRR7453034) and two samples without a lineage prediction (ERR3276012 and ERR3275788), all corresponding to BAM-derived fastq files as suggested by the presence of a BAM file in the SRA database alongside the fastq files for these samples. Since BAM files are usually sorted by genomic position for further analyses, the assumption of a random read distribution in fastq files made by fastlin when a k-mer coverage threshold is set did not hold for these samples, and they consequently required more reads than expected to be correctly typed.

Next, we assessed fastlin's ability to detect mixtures of strains by analyzing a set of 47 pairs of *M.tuberculosis* samples initially produced to study cases of MTBC relapse and reinfection (Bryant *et al.* 2013) and subsequently re-analyzed to test the MTBC mixed strain detection tool QuantTB (Anyansi *et al.* 2020). We ran QuantTB, TB-profiler and fastlin on this dataset. Eleven samples were identified as mixed by at least one of these three tools, with 9, 6, and 10 mixed samples detected by QuantTB, TB-profiler, and fastlin, respectively (Fig. 1A and Supplementary Data S4). The sample identified as a mixture of strains solely by QuantTB consists of a mixture of closely related strains (Bryant *et al.* 2013), possibly belonging to the same lineage (sample 8a; Supplementary Data S4). Fastlin was the only tool to identify sample 45b as a mixed sample. We believe this sample indeed contains a strain mixture since (i) the lineage of its minor strain was also present in pure samples of this dataset, (ii) the minor strain was supported by six barcode SNPs, and (iii) this sample was also identified as mixed by Bryant *et al.* We repeated the fastlin analyses using maximum coverage thresholds ranging from 100× to 10×, in 10× increments. Fastlin successfully detected the 10 mixed samples down to a 60× maximum k-mer coverage (Supplementary Data S4). Using simulated mixtures of pairs of strains from distinct lineages, we also found that fastlin provides reliable estimates of strain frequencies down to 0.05 (RMSE = 2.47%; Supplementary Data S4).

Finally, we compared the running times of fastlin with those of SKA and TB-profiler, chosen as representatives of k-mer-based and read-mapping tools, respectively. It is important to note that these two bioinformatics tools are not fully dedicated to lineage typing and generate more information than fastlin does (e.g. all SNPs, drug resistance predictions). For this benchmark we used 16 SRA samples representing a large range of sequencing depths, all composed of paired-end 150 bp reads. We observed that fastlin was at least an order of magnitude faster than SKA and TB-profiler, with running times ranging from 2 to 14 s per sample under default settings (Fig. 1B). Using a maximum k-mer coverage threshold of 80×, fastlin's running times were consistently under 5 s. Fastlin also exhibited minimal memory footprint, reaching a maximum use of only 4 MB in this benchmark, whereas both SKA and TB-profiler require several GB.

4 Discussion

While primarily developed to achieve ultra-fast lineage typing, fastlin also shows accurate MTBC lineage prediction,

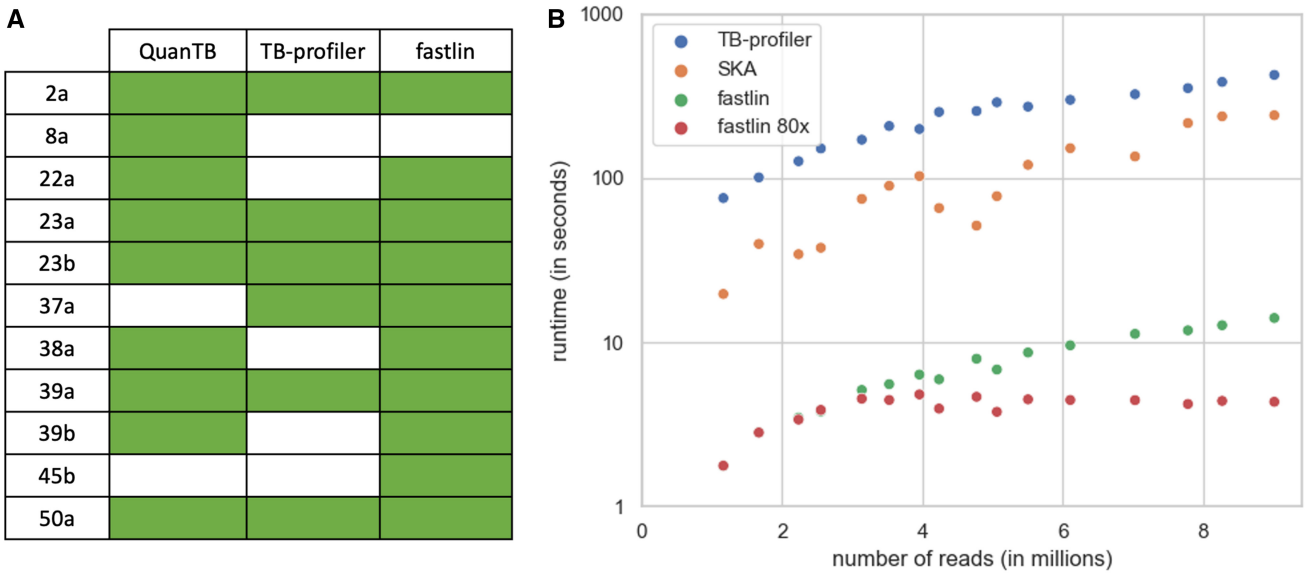

Figure 1. Fastlin benchmarking. (A) Detected mixed samples in Bryant *et al.* dataset. Samples with detected mixture are indicated with a colored box. (B) Runtimes obtained by TB-profiler, SKA (using $k = 27$), and fastlin (with default parameters in green, and with a maximum k-mer coverage of 80× in red). Note the logarithmic scale on the *y*-axis.

comparable and in some cases superior to that of standard tools. Its lineage predictions are nearly identical to those made by TB-profiler, with the few discrepancies observed in non-mixed samples being mostly attributed to TB-profiler errors. Fastlin also shows an accurate identification of mixed-lineage samples and estimation of the associated lineage proportions. We believe that the identification of low-frequency strains is facilitated in fastlin by the absence of background noise related to read-alignments, which mapping-based methods need to mitigate with stringent parameters. Fastlin's simplicity and efficiency make it a powerful tool for rapidly typing large MTBC datasets, with the potential for broader applications in lineage typing of other pathogens determined by a set of known barcodes.

Our analyses also uncover two limitations of fastlin. First, the presence of sorted reads in BAM-derived fastq files restricts the use of the maximum coverage threshold implemented in fastlin. Although this issue only affects the lineage prediction of seven files out of 3139 at a maximum coverage of 80×, this option is disabled by default, enabling fastlin to process all types of fastq files. However, in the case of in-house fastq files or fastq files of known provenance, we recommend specifying a maximum coverage threshold to further reduce fastlin's runtime. Second, the detection of strain mixtures is intrinsically limited to the lineages defined by the SNP barcodes, and fastlin will fail to identify any mixture of strains belonging to the same lineage. If the identification of mixed samples is essential, we recommend the use of alternative tools with either larger reference datasets such as QuantTB (Anyansi *et al.* 2020) or database-free tools such as SplitStrains (Gabbassov *et al.* 2021).

While we tested fastlin with one specific set of MTBC barcode SNPs, new lineage classifications and their associated barcode SNPs have recently been proposed (Coscolla *et al.* 2021, Netikul *et al.* 2022, Shuaib *et al.* 2022, Shitikov and Bespiatykh 2023). To accommodate the evolving landscape of MTBC lineage typing, we made the scripts required to build and test custom barcode files available on GitHub. This will enable end-users to leverage fastlin with their barcode SNPs of choice. We recommend including a minimum of 4–5 SNPs per lineage, and considering all mutations rather than only synonymous SNPs, to ensure robust lineage classification.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the NIHR Health Protection Research Unit in Respiratory Infections, in partnership with the UK Health Security Agency [NIHR200927]. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. L.C. and N.A. acknowledge funding from the MRC Centre for Global Infectious Disease Analysis [MR/R015600/1], jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

Anyansi C, Keo A, Walker BJ *et al.* QuantTB—a method to classify mixed *Mycobacterium tuberculosis* infections within whole genome sequencing data. *BMC Genomics* 2020;**21**:80.

Bryant JM, Harris SR, Parkhill J *et al.* Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med* 2013;**1**:786–92.

Cancino-Muñoz I, Gil-Brusola A, Torres-Puente M *et al.* Development and application of affordable SNP typing approaches to genotype *Mycobacterium tuberculosis* complex strains in low and high burden countries. *Sci Rep* 2019;**9**:15343.

Coll F, McNerney R, Guerra-Assunção JA *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 2014;**5**:4812.

Coscolla M, Gagneux S, Menardo F *et al.* Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microb Genom* 2021;**7**:000477.

Diel R, Kohl TA, Maurer FP *et al.* Accuracy of whole-genome sequencing to determine recent tuberculosis transmission: an 11-year population-based study in Hamburg, Germany. *Eur Respir J* 2019;**54**:1901154.

Gabbassov E, Moreno-Molina M, Comas I *et al.* SplitStrains, a tool to identify and separate mixed *Mycobacterium tuberculosis* infections from WGS data. *Microb Genom* 2021;**7**:000607.

Harris, S.R., SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology, *bioRxiv*, 2018, https://doi.org/10.1101/453142.

Jagielski T, van Ingen J, Rastogi N *et al.* Current methods in the molecular typing of *Mycobacterium tuberculosis* and other mycobacteria. *Biomed Res Int* 2014;**2014**:645802.

Napier G, Campino S, Merid Y *et al.* Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med* 2020;**12**:114.

Netikul T, Thawornwattana Y, Mahasirimongkol S *et al.* Whole-genome single nucleotide variant phylogenetic analysis of *Mycobacterium tuberculosis* Lineage 1 in endemic regions of Asia and Africa. *Sci Rep* 2022;**12**:1565.

Phelan JE, O'Sullivan DM, Machado D *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* 2019;**11**:41.

Shitikov E, Bespiatykh D. A revised SNP-based barcoding scheme for typing *Mycobacterium tuberculosis* complex isolates. *mSphere* 2023;**8**:e0016923.

Shuaib YA, Utpatel C, Kohl TA *et al.* Origin and global expansion of *Mycobacterium tuberculosis* complex lineage 3. *Genes (Basel)* 2022;**13**:990.

Wyllie DH, Davidson JA, Grace Smith E *et al.* A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying *Mycobacterium tuberculosis* transmission: a prospective observational cohort study. *EBioMedicine* 2018;**34**:122–30.