A peer-reviewed version of this preprint was published in PeerJ on 15 June 2018.

View the peer-reviewed version (peerj.com/articles/5098), which is the preferred citable publication unless you specifically need to cite this preprint.

Ferrés I, Iraola G. 2018. MLSTar: automatic multilocus sequence typing of bacterial genomes in R. PeerJ 6:e5098
https://doi.org/10.7717/peerj.5098

MLSTar: automatic multilocus and core genome sequence typing in R

Ignacio Ferrés[1] **and Gregorio Iraola**[1]

[1]**Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay**

Corresponding author:
Gregorio Iraola[1]

Email address: giraola@pasteur.edu.uy

ABSTRACT

Multilocus sequence typing (MLST) is a standard tool in population genetics and bacterial epidemiology that assesses the genetic variation present in a reduced number of housekeeping genes (typically seven) along the genome. This methodology assigns arbitrary integer identifiers to genetic variations at these loci allowing to efficiently compare bacterial isolates using allele-based methods. Now, the increasing availability of whole-genome sequences for hundreds to thousands of strains from the same bacterial species has motivated to upgrade the resolution of traditional MLST schemes using larger gene sets or even the core genome (cgMLST). The PubMLST database is the most comprehensive resource of described MLST and cgMLST schemes available for a wide variety of species. Here we present MLSTar as the first R package that allows to i) connect with the PubMLST database to select a target scheme, ii) screen a desired set of genomes to assign alleles and sequence types and iii) interact with other widely used R packages to analyze and produce graphical representations of the data. We applied MLSTar to analyze a set of 400 *Campylobacter coli* genomes, showing great accuracy and comparable performance with previously published command-line tools. `MLSTar` can be freely downloaded from `http://github.org/iferres/MLSTar`.

BACKGROUND

Multilocus sequence typing (MLST) was introduced in 1998 as a portable tool for studying epidemiological dynamics and population structure of bacterial pathogens based on PCR amplification and capillary sequencing of housekeeping gene fragments (Maiden et al., 1998). In most MLST schemes, seven loci are indexed with arbitrary and unique allele numbers that are combined into an allelic profile or sequence type (ST) to efficiently summarize genetic variability along the genome. Rapidly, MLST demonstrated enhanced reproducibility and convenience in comparison with previous methods such as multilocus enzyme electrophoresis (MLEE) or pulsed-field gel electrophoresis (PFGE), allowing to perform global epidemiology and surveillance studies (Urwin and Maiden, 2003). However, as MLST started to be massively applied two main drawbacks were uncovered: i) the impossibility of establishing a single universal MLST scheme applicable to all bacteria; and ii) the lack of high resolution of seven-locus MLST schemes required for some purposes.

These problems pushed the development of improved alternatives to the original methodology. The extended MLST (eMLST) approach which is based on the analysis of longer gene fragments (Chen et al., 2011) or increased number of loci (Dingle et al., 2008; Crisafulli et al., 2013) proved to improve resolution, and the scheme based on 53 ribosomal protein genes (rMLST) was proposed as an universal approach since these loci are conserved in all bacteria (Jolley et al., 2012). Beyond these improvements, the advent of high-throughput sequencing and the increasing availability of hundreds to thousands whole-genome sequences (WGS) for many bacterial pathogens caused a paradigmatic change in clinical microbiology, making possible to use nearly complete genomic sequences to enhance typing resolution. This revolution allowed the transition from standard MLST schemes testing a handful of genes to core genome (cgMLST) approaches that scaled to hundreds of loci common to a set of bacterial genomes (Maiden et al., 2013).

The generation of this massive amount of genetic information required the accompanying development of database resources to effectively organize and store typing schemes and allele definitions.

Rapidly, the PubMLST database (http://pubmlst.org) turned into the most comprehensive and standard resource storing today schemes and allelic definitions for more than 100 microorganisms. Subsequently, the shift to WGS motivated the development of the Bacterial Isolate Genome Sequence Database (BIGSdb) (Jolley and Maiden, 2010), which now encompasses all the software functionalities used for the PubMLST. Also, many tools for automatic MLST analysis from whole-genome sequences have been developed using web servers like MLST-OGE (Larsen et al., 2012) or EnteroBase (http://enterobase.warwick.ac.uk), paid tools like BioNumerics (http://www.applied-maths.com/bionumerics) or SeqSphere+ (http://www.ridom.de/seqsphere/), and open source tools like mlst (http://github.org/tseemann/mlst) or MLSTcheck (Page et al., 2016). Here, we present MLSTar as the first tool for automatic multilocus sequence typing of bacterial genomes written in R (R Development Core Team, 2008), allowing to expand the application of MLST tools within this very popular and useful environment for data analysis and visualization.

IMPLEMENTATION

MLSTar is written in R and contains all data processing steps and command line parameters to call external dependencies wrapped in the package. MLSTar depends on BLAST+ (Camacho et al., 2009) that is used as sequence search engine, and must be installed locally. MLSTar is designed to work on Unix-based operating systems and is distributed as an open source software (MIT license) stored in GitHub (http://github.com/iferres/MLSTar). MLSTar contains four main functions that i) takes genome assemblies or predicted genes in FASTA format from any number of strains, ii) performs sequence typing using a previously selected scheme from PubMLST and iii) applies standard phylogenetic approaches to analyze the data. A graphical overview of the overall workflow has been outlined in Figure 1.

Interaction with PubMLST

First step in MLSTar workflow involves to interact with the PubMLST database to select a target MLST or cgMLST scheme. This interaction requires Internet connection because is performed using the RESTful web application programming interface provided by PubMLST. The listPubmlst_orgs() function allows to list the names of all microorganisms that have any scheme stored in PubMLST. Then, as some microorganisms have more than one scheme (i.e. one classical seven-loci and one core genome scheme), the listPubmlst_schemes() function lists the available schemes for any selected species.

Calling and storing alleles and sequence types

MLSTar make allele and ST calls from FASTA files containing closed genomes or contigs using BLAST+ blastn comparisons implemented by the doMLST() function. Parallelization is available as internally implemented in R by the parallel package. Also, the doMLST() function can be run at the same time for different schemes using internal R functions like lapply(). Results are stored in a S3 class object named mlst that contains two data.frame objects: one containing allele and ST assignments for the analyzed genomes (unknown alleles or STs are labeled as "u"), and the other storing known allele profiles for the selected scheme. If required, nucleotide sequences for known or novel alleles can be written as multi FASTA files.

Minimum spanning tree analysis

Allele profiles are frequently used to reconstruct phylogenetic relationships among strains. Function plot.mlst() directly takes the mlst class object to compute distances assuming no relationships between allele numbers, so each locus difference is treated equally. Then, identical isolates have a distance of 0, those with no alleles in common have a distance of 1 and, for example, in a seven-loci scheme two strains with 5 differences would have a distance of 0.71 (5/7). The resulting distance matrix is used to build a minimum spanning tree as implemented in APE package (Paradis et al., 2004). The user can choose to plot the tree only displaying the analyzed strains or them incorporated into the whole diversity of profiles of the selected scheme. The plot.mlst function also returns an igraph class object (Csardi and Nepusz, 2006) that can be used to customize graphical aspects like color, node size, etc.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26630v1 | CC BY 4.0 Open Access | rec: 7 Mar 2018, publ: 7 Mar 2018

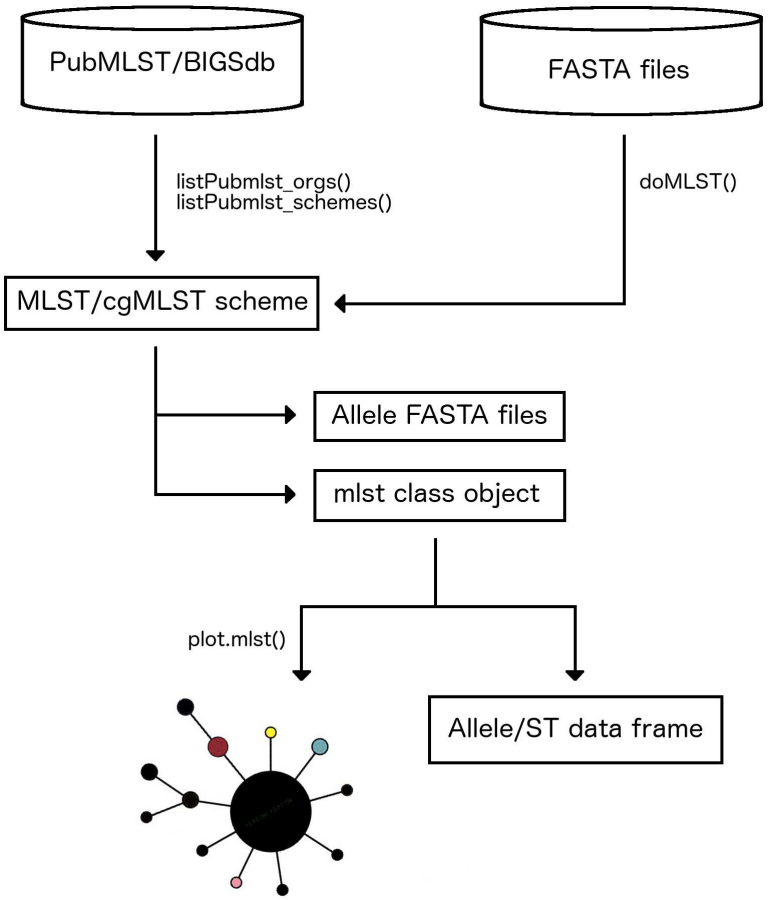

Figure 1. Main steps in `MLSTar` workflow.

VALIDATION

Comparison with reference dataset

We used a random set of 400 *Campylobacter coli* genomes downloaded from strains deposited in the `BIGSdb` (Supplemental Table S1). For this dataset, reference allele and ST assignments based on the standard *C. jejuni*/*C. coli* seven-loci MLST scheme were extracted from the `BIGSdb` and compared with results obtained running `MLSTar`. The concordance of each allele and ST is shown in Table 1, measured as the percentage of identical assignments between `BIGSdb` and `MLSTar`. In average, assignments were 99.65% and 99.5% coincident for alleles and STs, respectively.

	aspA	*glnA*	*glyA*	*gltA*	*tkt*	*pgm*	*uncA*	ST
BIGSdb	99.5	99.75	99.75	99.5	99.5	99.5	99.5	99.75

Table 1. Accuracy of `MLSTar` against reference alleles and STs from `BIGSdb`, measured as the percentage of correct calls in a seven-locus MLST scheme over 400 *C. coli* genomes.

Comparison with similar tools

Two similar command line software tools designed to screen assembled genomes based on `blastn` were selected to further test and compare `MLSTar` performance: `MLSTcheck` (Page et al., 2016) and `mlst` (`http://github.org/tseemann/mlst`). First, we compared the accuracy of `MLSTar` with respect to these tools by measuring the percentage of correctly assigned alleles and STs. Supplemental Table S2 shows that `MLSTar` presented comparable accuracy with both `MLSTcheck` (99.8%) and `mlst`

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26630v1 | CC BY 4.0 Open Access | rec: 7 Mar 2018, publ: 7 Mar 2018

(99.7%). Second, we used the same *C. coli* dataset to compare the running time between tools in a single AMD Opteron 2.1 GHz processor, by gradually increasing the number of analyzed genomes from 2 to 400 (Fig. 2. These results showed that `MLSTar` is 26-fold faster than `MLSTcheck` but is 3-fold slower than `mlst` (Supplemental Table S3).

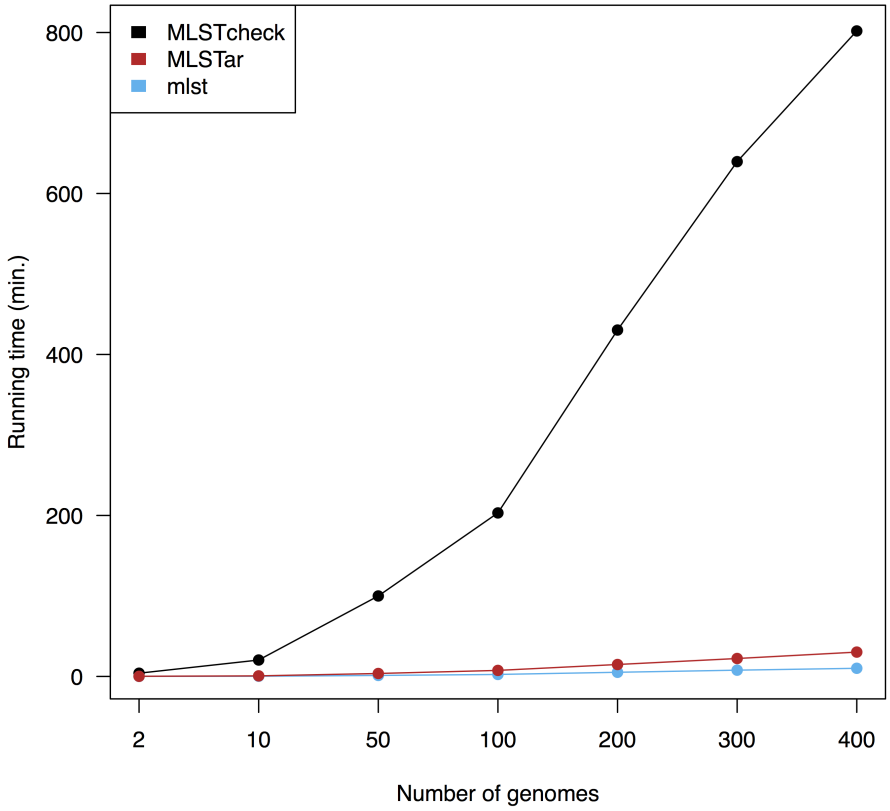

Figure 2. Comparison of running times in a single CPU between `MLSTar`, `mlst` and `MLSTcheck`.

CONCLUSIONS

The advent of WGS has now allowed to type bacterial strains directly from their whole genomes avoiding to repeat tedious PCR amplifications and fragment capillary sequencing for multiple loci. Today MLST is a valid tool which is frequently used as first-glimpse approach to explore genetic diversity and structure within huge bacterial population sequencing projects. This incessant availability of genomic information has motivated a constant effort to develop efficient analytical tools from multilocus typing data (Page et al., 2017). Here, we developed a new software package called `MLSTar` that expands the possibilities of performing allele-based genetic characterization within the `R` environment. We demonstrate that `MLSTar` has comparable performance with previously validated software tools and can be applied to analyze hundreds of genomes in a reasonable time.

ACKNOWLEDGMENTS

I.F. was supported by the Agencia Nacional de Investigación e Innovación (ANII, Uruguay) postgraduation program grant POS_NAC_2016_1_131079. We thank Daniela Costa and Cecilia Nieves for testing MLSTar.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26630v1 | CC BY 4.0 Open Access | rec: 7 Mar 2018, publ: 7 Mar 2018

REFERENCES

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421.

Chen, Y., Zhen, Q., Wang, Y., Xu, J., Sun, Y., Li, T., Gao, L., Guo, F., Wang, D., Yuan, X., et al. (2011). Development of an extended multilocus sequence typing for genotyping of brucella isolates. *Journal of microbiological methods*, 86(2):252–254.

Crisafulli, G., Guidotti, S., Muzzi, A., Torricelli, G., Moschioni, M., Masignani, V., Censini, S., and Donati, C. (2013). An extended multi-locus molecular typing schema for streptococcus pneumoniae demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. *Infection, Genetics and Evolution*, 13:151–161.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Dingle, K. E., McCarthy, N. D., Cody, A. J., Peto, T. E., and Maiden, M. C. (2008). Extended sequence typing of campylobacter spp., united kingdom. *Emerging infectious diseases*, 14(10):1620.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., Cody, A. J., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, 158(4):1005–1015.

Jolley, K. A. and Maiden, M. C. (2010). Bigsdb: scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11(1):595.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D. W., Aarestrup, F. M., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of clinical microbiology*, 50(4):1355–1361.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145.

Maiden, M. C., Van Rensburg, M. J. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., and McCarthy, N. D. (2013). Mlst revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology*, 11(10):728.

Page, A. J., Alikhan, N.-F., Carleton, H. A., Seemann, T., Keane, J. A., and Katz, L. S. (2017). Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microbial genomics*, 3(8).

Page, A. J., Taylor, B., and Keane, J. A. (2016). Multilocus sequence typing by blast from de novo assemblies against pubmlst. *The Journal of Open Source Software*, 1(8).

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Urwin, R. and Maiden, M. C. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in microbiology*, 11(10):479–487.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26630v1 | CC BY 4.0 Open Access | rec: 7 Mar 2018, publ: 7 Mar 2018