

*Databases and ontologies***spolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex**Chaka Tang¹, Josephine F. Reyes^{1,2,*}, Fabio Luciani¹, Andrew R. Francis³ and Mark M. Tanaka^{1,2,*}¹School of Biotechnology and Biomolecular Sciences, ²Evolution & Ecology Research Centre, University of New South Wales, NSW 2052 and ³School of Computing and Mathematics, University of Western Sydney, NSW 1797, Australia

Received on May 27, 2008; revised and accepted on August 14, 2008

Associate Editor: Martin Bishop

ABSTRACT

spolTools is a collection of online programs designed to manipulate and analyze spoligotype datasets of the *Mycobacterium tuberculosis* complex. These tools are integrated into a repository currently containing 1179 spoligotypes and 6278 isolates across 30 datasets. Users can search this database to export for external use or to pass on to the integrated tools. These tools include the computation of basic population genetic quantities, the visualization of clusters of spoligotype patterns based on an estimated evolutionary history and a procedure to predict emerging strains – genotypes associated with elevated transmission.

Availability: Database, programs and documentation may be accessed online at <http://www.emi.unsw.edu.au/spolTools>.

Contact: j.reyes@student.unsw.edu.au; m.tanaka@unsw.edu.au

1 INTRODUCTION

Spacer oligonucleotide typing, or spoligotyping, is a genotyping method used to study the epidemiology of the *Mycobacterium tuberculosis* complex (Kamerbeek *et al.*, 1997). The resulting fingerprint, called a spoligotype, is a binary sequence of length 43, representing the presence or absence of spacers. Spoligotyping exploits polymorphism in the direct repeat region, a structure that belongs to a family of repeats called clustered repetitive interspersed palindromic repeats (CRISPRs) found in the genomes of bacteria and archaea (Pourcel *et al.*, 2005). As with other molecular markers, spoligotyping enables classification of isolates into distinct strains, and thus allows characterization of genetic diversity of *M.tuberculosis*. Spoligotyping has gained widespread use for studying tuberculosis transmission: 434 papers that mention spoligotyping have appeared in the last decade (Luciani *et al.*, 2008). Currently, the most comprehensive compilation of spoligotype datasets is SpolDB4 (Brudey *et al.*, 2006), recording 1939 distinct spoligotype patterns from 35925 isolates collected from a large number of countries. The development of an online repository of spoligotype datasets linked to bioinformatic and statistical analysis methods would further aid the tuberculosis research community. In this note, we describe spolTools, a website containing a repository of spoligotype datasets of *M.tuberculosis* complex that

are based on 43 spacers and a collection of programs for their analysis. spolTools provides an integrated framework to assist in the analysis of spoligotype datasets and automated methods to extract epidemiological information in an easily parsed format. Unpublished datasets can also be analyzed using these tools. The URL for spolTools is <http://www.emi.unsw.edu.au/spolTools>.

2 DATASETS AND FORMATTING**2.1 Repository of spoligotype data**

At present, spolTools contains a repository of 30 spoligotyping datasets taken from published studies, constituting a total of 1179 spoligotypes and 6278 isolates. The spoligotypes in these datasets can be displayed along with their corresponding shared international type (SIT) as assigned in SpolDB4 (Brudey *et al.*, 2006). New datasets can be submitted for inclusion by emailing spolTools@unsw.edu.au.

2.2 Rich spoligotype format

Each dataset is stored in a proposed working format called the rich spoligotype format (RSF). The RSF, using only ASCII characters, stores spoligotypes and associated information. RSF facilitates the handling of datasets by the programs, while keeping the information human-readable and ensuring consistent parsing and editing of spoligotype related data between research laboratories and computer platforms.

2.3 Formatting of spoligotype patterns

Several methods to represent the spoligotype patterns are in current use. These include octal, hex and binary formats (Dale *et al.*, 2001). We propose a new format called the *gap format*. This format consists of a list of missing spacers, which allows easy entering of data into electronic form (see the website for description and examples).

spolTools includes a page (Convert formats) that allows conversion between any of these representations.

3 TOOLS FOR ANALYZING SPOLIGOTYPES

spolTools implements several tools for analyzing *M.tuberculosis* complex data from spoligotyping, and an interface between these

*To whom correspondence should be addressed.

tools and the dataset repository. In addition to computing basic population statistics, spolTools includes two programs. First, DESTUS is an implementation of a method for predicting emerging strains in a sample of spoligotypes; second, the spoligoforest program provides visualizations of the probable relationships among spoligotypes in a given sample. To access these tools for datasets in the repository, first retrieve a dataset under *Search*, then proceed to *Get Details*. By processing data appropriately through *Run Programs* users can also analyze their own data with these programs.

3.1 Population statistics

spolTools computes a set of summary statistics that provide epidemiological information. These include the size of the sample n , number of genotypes in the sample g , number of unique spoligotype patterns (singletons) and other indices traditionally used to estimate the diversity of spoligotype patterns and to infer the extent of recent transmission in a given sample. It also computes a maximum likelihood estimate of θ , a measure of genetic diversity that is proportional to both the mutation rate of the marker and the effective population size. For further details of these quantities see Tanaka and Francis (2005) and Luciani et al. (2008). Where information is available, the proportion of strains that are drug resistant is also given.

3.2 Spoligoforests

Like other molecular techniques, spoligotyping can be used for phylogenetic analysis (Guernier et al., 2008); its usefulness for this purpose is discussed, for example, by Eldholm et al. (2006). A method to visually represent relationships among spoligotypes is described in J.F.Reyes et al.'s. *Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes* (submitted for publication, 2008). The output graph, called a *spoligoforest*, shows a plausible history of mutation events and therefore relationships among spoligotypes in a sample of isolates. This method makes use of a model that considers mutation by irreversible deletions of spacers and assigns probabilities to the lengths of these deletions. The size of each node is an increasing function of the number of isolates (the cluster size); edges between nodes reflect evolutionary relationships between spoligotypes with arrowheads pointing to descendants.

Two layout methods for spoligoforests are provided in spolTools: a hierarchical layout and a 'burst' layout based on a Fruchterman-Reingold algorithm.

3.3 Emerging strains

A method to determine if any strains of *M.tuberculosis* are spreading faster than the background rate was described by Tanaka and Francis (2006). A program implementing this method, DESTUS

(detecting emerging strains of tuberculosis using spoligotypes), is included in spolTools. Note that this method is intended to be used on self-contained datasets corresponding to specific regions, rather than composite data from different countries or collection periods.

4 CONCLUSION

In presenting spolTools, our goal is to make available a suite of tools for analyzing spoligotype patterns from *M.tuberculosis* complex datasets. Although it is not intended to be a comprehensive database like SpolDB4, we welcome submission of new data for inclusion in the spolTools repository. We will add new or refined computational tools to spolTools as they are developed. At present, spolTools is designed for spoligotype data based on the 43 spacers described by Kamerbeek et al. (1997). The methods and tools described here should be generalizable to similar direct repeat structures in other organisms or a different number of repeats in the *M.tuberculosis* complex.

ACKNOWLEDGEMENTS

We thank Z. Aandahl for technical assistance on the website; A. McLean and S. Kinathil for the feedback.

Funding: Australian Research Council through Discovery Grant (DP0556732).

Conflict of interest: none declared.

REFERENCES

- Brudey,K. et al. (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.*, **6**, 23.
- Dale,J. et al. (2001) Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standardised nomenclature. *Int. J. Tuberc. Lung Dis.*, **5**, 216–220.
- Eldholm,V. et al. (2006) *BMC microbiology*. *BMC Microbiology*, **6**, 76.
- Guernier,V. et al. (2008) Use of cluster-graphs from spoligotyping data to study genotype similarities and a comparison of three indices to quantify recent tuberculosis transmission among culture positive cases in French Guiana during a eight year period. *BMC Infect. Dis.*, **8**, 46.
- Kamerbeek,J. et al. (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.*, **35**, 907–914.
- Luciani,F. et al. (2008) Interpreting genotype cluster sizes of *Mycobacterium tuberculosis* isolates typed with IS6110 and spoligotyping. *Infect. Genet. Evol.*, **8**, 182–190.
- Pourcel,C. et al. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.
- Tanaka,M.M. and Francis,A.R. (2005) Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information. *Infect. Genet. Evol.*, **5**, 35–43.
- Tanaka,M.M. and Francis,A.R. (2006) Detecting emerging strains of tuberculosis by using spoligotypes. *Proc. Natl Acad. Sci. USA*, **103**, 15266–15271.