BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1762–1764
doi:10.1093/bioinformatics/btu113

Genome analysis

Advance Access publication February 25, 2014

CanSNPer: a hierarchical genotype classifier of clonal pathogens

Adrian Lärkeryd[1,*], Kerstin Myrtennäs[2], Edvin Karlsson[2], Chinmay Kumar Dwibedi[1,2], Mats Forsman[2], Pär Larsson[2], Anders Johansson[3] and Andreas Sjödin[2,4]

[1]Department of Clinical Microbiology, Umeå University, [2]Division of CBRN Security and Defence, FOI, Swedish Defence Research Agency, [3]Department of Clinical Microbiology, The Laboratory for Molecular Infection Medicine Sweden (MIMS) and [4]Department of Chemistry, Computational Life Science Cluster (CLiC), Umeå University, Umeå, Sweden

Associate Editor: Janet Kelso

ABSTRACT

Summary: Advances in typing methodologies have recently reformed the field of molecular epidemiology of pathogens. The falling cost of sequencing technologies is creating a deluge of whole genome sequencing data that burdens bioinformatics resources and tool development. In particular, single nucleotide polymorphisms in core genomes of pathogens are recognized as the most important markers for inferring genetic relationships because they are evolutionarily stable and amenable to high-throughput detection methods. Sequence data will provide an excellent opportunity to extend our understanding of infectious disease when the challenge of extracting knowledge from available sequence resources is met. Here, we present an efficient and user-friendly genotype classification pipeline, CanSNPer, based on an easily expandable database of predefined canonical single nucleotide polymorphisms.

Availability and implementation: All documentation and Python-based source code for the CanSNPer are freely available at http://github.com/adrlar/CanSNPer.

Contact: adrian.larkeryd@foi.se

Received on October 27, 2013; revised on February 17, 2014; accepted on February 19, 2014

1 INTRODUCTION

Whole genome sequencing of pathogen isolates will continue to improve epidemiological investigations and microbial forensics through the ability to reconstruct genetic relationships among isolates. The vast amount of information generated by whole genome sequencing can overcome the hurdles of other less-informative rich epidemiological typing methods in establishing genetic relationships among pathogens. The overall goal is to identify similarities among strains that are a result of vertically inherited variation indicative of recent common descent. Confounding sources of polymorphisms must, however, be addressed, including sequencing errors and sequence polymorphisms introduced through horizontal transfer of sequences by both homologous and non-homologous recombination. No single typing method handles all levels of bacterial diversity. Multilocus sequence typing has been used for a decade as a portable sequence-based method for identifying clonal relationships among bacteria with various amount of recombination. In highly monomorphic and slowly evolving bacterial species such as *Mycobacterium tuberculosis* or *Bacillus anthracis*, identification of single nucleotide polymorphisms (SNPs), in comparison with a defined archetypal strain, could provide more robust phylogeny and higher resolution. In such pathogens, the inheritance pattern of specific sets of SNPs is strictly clonal (Achtman, 2012). Several SNP detection technologies developed for this purpose over the past decades, e.g. based on real-time PCR or microarrays are predicted to be replaced by next-generation sequencing. For example, relationships among isolates of *Francisella tularensis*, *Coxiella burnetii*, *Yersinia pestis* and *B.anthracis* are amenable for characterization using specific sets of SNPs, often denoted canonical SNPs (canSNPs). It has been demonstrated that a small number of SNPs can be used to define genetic groups efficiently and to infer evolutionary relationships within *F.tularensis* (Chanturia *et al*., 2011; Gyuranecz *et al*., 2012; Karlsson *et al*., 2013; Svensson *et al*., 2009; Vogler *et al*., 2009, 2011a), *C.burnetii* (Hornstra *et al*., 2011; Karlsson *et al*., 2014), *Y.pestis* (Morelli *et al*., 2010; Vogler *et al*., 2011b) and *B.anthracis* (Okinaka *et al*., 2008; Simonson *et al*., 2009; Van Ert *et al*., 2007).

Using the new low-cost sequencing technology for SNP detection, there are vast amounts of data that can be used to characterize an isolate and its relationship with others. Whole-genome sequence typing is a powerful tool for genotyping and molecular epidemiological analysis but not completely accessible for scientists with limited computer skills (Engelthaler *et al*., 2011; Gillece *et al*., 2011; O'Farrell *et al*., 2012). The current bottleneck is time and computationally intensive generation of high-quality genome sequences. To take advantage of the new low-cost sequencing technology, sequence information must be presented and analyzed in a way that, for example, public health experts in epidemiology can easily interpret. Therefore, alternative methods are needed for Whole-genome sequence typing data to be used in real time. Here, we present an analysis pipeline for easy genotype classification of clonal pathogens. This program should have vast applications in clinical genomic research and will aid the expanding efforts to identify and trace pathogens during outbreaks.

2 IMPLEMENTATION

We developed a pipeline for the analysis of canonical SNPs in draft genomes of clonal pathogens. CanSNPer is a fast and lightweight tool written in Python for genotype classification of pathogens based on canonical SNPs. A user-customizable

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

configuration file specifies all parameter settings. Installation instructions are provided with the software distribution together with sample data and expected outputs. CanSNPer has dependencies on external software progressiveMauve (Darling *et al.*, 2010) and The Environment for Tree Exploration toolkit (Huerta-Cepas *et al.*, 2010). CanSNPer analysis could also be performed directly in a web browser as a Galaxy tool (Blankenberg *et al.*, 2010; Giardine *et al.*, 2005; Goecks *et al.*, 2010).

2.1 Building reference canSNP database

CanSNPer uses a local SQLite3 database to store information on canonical SNPs, the canSNP tree structure and reference sequences for each organism. The local database is created by the user, using a series of CanSNPer arguments. Preconstructed typing schemes for *Y.pestis*, *F.tularensis*, *B.anthracis* and *C.burnetii* are available for download with the software. Users may extend the local database by creating their own typing schemes in CanSNPer according to a predefined specification. The tree specification file must only contain the root of the tree on the first line, and each node in the tree has its own line with all its ancestors listed, separated by semicolon in order.

2.2 Classification of sequences

CanSNPer is designed to easily extract known canonical SNPs from a query draft sequence for classification of the pathogen isolate. The query is aligned to reference genomes for the selected reference organism using progressiveMauve (Darling *et al.*, 2010). CanSNPer will by default run all alignments in parallel for a significant speed increase if there are many reference sequences to a single organism. The state of each canSNP is determined, and the canSNP tree is traversed to find a classification. The pipeline uses a recursive and greedy tree walking algorithm and can classify a query sequence as either a leaf or as an intermediate node, if no more canSNPs could be determined as derived. The tree walking algorithm is also designed to handle the forcing of a number of nodes. If a region is missing in the query sequence in which one of the canSNPs is located, CanSNPer can be set to allow a mismatch in the tree and still find a classification if there are derived canSNPs further down the tree.

2.3 Visual genotype presentation

A visual representation of all SNPs in the canonical SNP tree for the query sequence can easily be generated. The visualization uses the Environment for Tree Exploration toolkit for Python (Huerta-Cepas *et al.*, 2010) to draw a phylogenetic canonical SNP tree (Fig. 1). The complete canSNP tree for the analyzed genome sequence is visualized, and the derived SNP variants are indicated by larger filled circles.

3 CONCLUSION

The ability of rapid and cost-efficient generation of whole genome sequence data of pathogens will transform treatment and epidemiological studies once clinicians have access to user-friendly analysis tools. We have presented the CanSNPer software written in Python to automatically genotype clonal bacterial pathogens using genome sequencing data. The identification of strain-specific SNP markers in pathogens permits the development of rapid diagnostics to greatly assist in the investigation of biocrimes and natural outbreaks that may require the analysis of hundreds or even thousands of specimens, including environmental and clinical samples. The database can be easily modified to handle other organisms of interest or similar SNP classification problems, such as cancer samples. Of interest for future improvements of the pipeline is the optimization of the classification algorithm and adding additional organisms.

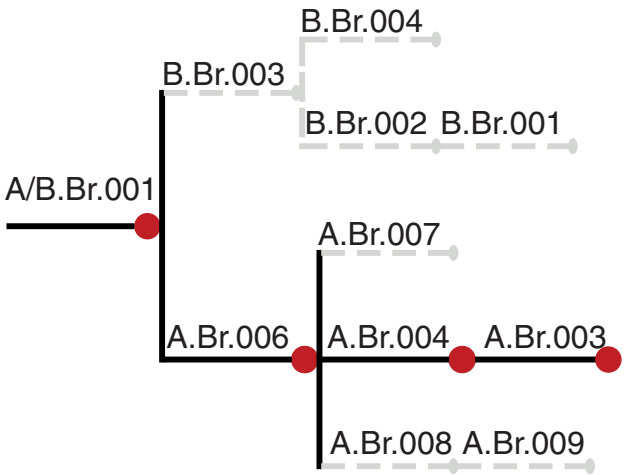

Fig. 1. Visual presentation of CanSNPer output. This example shows the genotyping of *B.anthracis*

In conclusion, our strategy permits the identification of SNPs that are diagnostic for narrowly defined pathogen genetic lineages. The use of strain-specific SNPs for high-throughput genotyping may prove to be a valuable tool to assist in real-world forensic and epidemiological studies.

Funding: This project was funded by the Swedish Ministry of Foreign Affairs, project A4952, the Swedish Civil Contingencies Agency, project B4055, and the Swedish Ministry of Defence, project A4040. Umeå University and Västerbotten County Council through a translational research grant to Anders Johansson.

Conflict of Interest: none declared.

REFERENCES

Achtman,M. (2012) Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **367**, 860–867.

Blankenberg,D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, **Chapter 19**, Unit 19.10.1–Unit 19.10.21.

Chanturia,G. *et al.* (2011) Phylogeography of *Francisella tularensis* subspecies holarctica from the country of georgia. *BMC Microbiol.*, **11**, 139.

Darling,A.E. *et al.* (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.

Engelthaler,D.M. *et al.* (2011) Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. *Emerg. Infect. Dis.*, **17**, 227–232.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Giardine,B. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
Gillece,J.D. *et al.* (2011) Whole genome sequence analysis of *Cryptococcus gattii* from the Pacific Northwest reveals unexpected diversity. *PLoS One*, **6**, e28550.
Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
Gyuranecz,M. *et al.* (2012) Phylogeography of *Francisella tularensis* subsp. holarctica, Europe. *Emerg. Infect. Dis.*, **18**, 290–293.
Hornstra,H.M. *et al.* (2011) Rapid typing of *Coxiella burnetii*. *PLoS One*, **6**, e26201.
Huerta-Cepas,J. *et al.* (2010) Ete: a python environment for tree exploration. *BMC Bioinformatics*, **11**, 24.
Karlsson,E. *et al.* (2014) Eight new genomes and synthetic controls increase the accessibility of rapid melt-mama snp typing of *Coxiella burnetii*. *PLoS One*, **9**, e85417.
Karlsson,E. *et al.* (2013) The phylogeographic pattern of *Francisella tularensis* in Sweden indicates a Scandinavian origin of Eurosiberian tularaemia. *Environ. Microbiol.*, **15**, 634–645.
Morelli,G. *et al.* (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat. Genet.*, **42**, 1140–1143.
O'Farrell,B. (2012) Transforming microbial genotyping: a robotic pipeline for genotyping bacterial strains. *PLoS One*, **7**, e48022.
Okinaka,R.T. *et al.* (2008) Single nucleotide polymorphism typing of *Bacillus anthracis* from sverdlovsk tissue. *Emerg. Infect. Dis.*, **14**, 653.
Simonson,T.S. *et al.* (2009) *Bacillus anthracis* in China and its relationship to worldwide lineages. *BMC Microbiol.*, **9**, 71.
Svensson,K. *et al.* (2009) A real-time PCR array for hierarchical identification of Francisella isolates. *PLoS One*, **4**, e8360.
Van Ert,M.N. *et al.* (2007) Global genetic population structure of *Bacillus anthracis*. *PLoS One*, **2**, e461.
Vogler,A.J. *et al.* (2009) Phylogeography of *Francisella tularensis*: global expansion of a highly fit clone. *J. Bacteriol.*, **191**, 2474–2484.
Vogler,A.J. *et al.* (2011a) Phylogeography of *Francisella tularensis* ssp. holarctica in France. *Lett. Appl. Microbiol.*, **52**, 177–180.
Vogler,A.J. *et al.* (2011b) Phylogeography and molecular epidemiology of *Yersinia pestis* in Madagascar. *PLoS Negl. Trop. Dis.*, **5**, e1319.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016