Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Subject Section

Phylotyper: *In silico* predictor of molecular subtypes from gene sequences

Matthew D. Whiteside [1,*], Chad R. Laing [1] and Victor P.J. Gannon [1,*]

[1] National Microbiology Laboratory, Public Health Agency of Canada, Lethbridge, AB, Canada, T1J 3Z4

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Whole genome sequencing (WGS) is being adopted in public health for improved surveillance and outbreak analysis. Molecular subtyping has been used in public health to infer phenotypes and flag high-risk bacterial strain groups. *In silico* tools that predict molecular subtypes from gene sequences are needed to transition historical data to WGS-based protocols. Phylotyper is a novel solution for *in silico* molecular subtype prediction from gene sequences. Designed for incorporation into WGS pipelines, it is a general prediction tool that can be applied to most molecular subtype schemes. Phylotyper uses phylogeny to model the evolution of the subtype and infer subtypes for unannotated sequences. The phylogenic framework in Phylotyper improves accuracy, provides useful contextual feedback, and is more capable at identifying novel subtypes over approaches based solely on sequence similarity.
Availability and Implementation: Phylotyper is a python package. It is available from: `https://github.com/superphy/insilico-subtyping`.
Contact: matthew.whiteside@phac-aspc.gc.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole-genome sequencing (WGS) is transforming the public health field by providing a highly efficient method for surveying bacterial populations. The speed, discriminatory power and broad utility of WGS can improve surveillance and outbreak analysis. Adoption of WGS in public health, however, requires transitioning of historical data with the new methods (Jenkins, 2015). One of the workhorse methods in public health is molecular subtyping (such as serotyping). As a surveillance tool, subtypes provide a clearcut designation that is typically used to distinguish taxonomic groups and infer phenotypes, for example, pathogens from non-pathogens (Jenkins, 2015). A WGS-based approach to subtyping would have several benefits over current subtype systems; it would be faster, have improved discrimination and would be cheaper and easier to maintain and routinely run (Jenkins, 2015). Accordingly, new *in silico* tools have been developed to predict subtypes from WGS data (Inouye *et al.*, 2014; Joensen *et al.*, 2015; Ingle *et al.*, 2016; Lindsey *et al.*, 2016; Carrillo *et al.*, 2016).

Phylotyper is a novel *in silico* predictor of subtypes from sequence data. Phylotyper is unique in that it builds a phylogenetic tree consisting of reference sequences with known subtype and the unknown query sequences to help inform subtype prediction. Using phylogenetic ancestral reconstruction to assign the likelihood of each subtype to the branch points in the tree, Phylotyper assigns an unknown query sequence a subtype based on the extrapolated value from its ancestors in the tree.

2 Implementation

The core of Phylotyper is an ancestral state reconstruction (ASR) method that has been adapted for hidden state prediction. In phylogenetic analysis, ancestral state reconstruction involves the prediction of traits of ancestors from existent descendants. This methodology can be extended to also predict properties in a limited number of existing strains.

In Phylotyper, the `rerootingMethod` function from the phytools R package is used to perform the ASR (Revell, 2011). This function calculates the maximum marginal likelihood for unknown tip nodes in a phylogenetic tree. The likelihood reflects the most likely state for the node given the empirically estimated evolution model and phylogeny. It captures the uncertainty associated with phylogenetic branch lengths and rates of change between states. In the context of Phylotyper, the marginal likelihood provides a confidence value associated with a predicted subtype. The evolutionary model used in Phylotyper is estimated from the data,

© The Author 2017. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

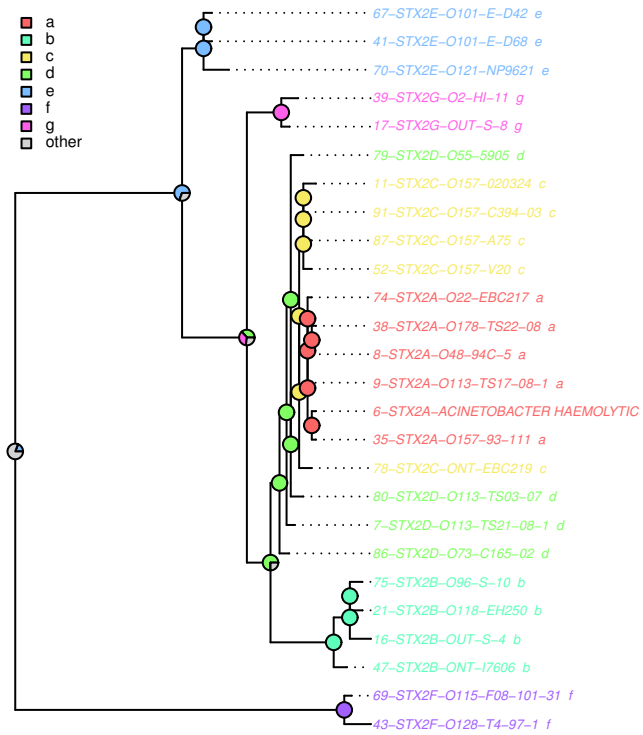

Fig. 1. Phylogenetic tree for select Stx2 genes. The subtype marginal likelihood is displayed at each node as a pie chart. The full Stx2 subtype tree is displayed in Supplementary Figure S1.

so the reference set of genes used to build a phylogenetic tree and run the ASR, is key to the accuracy. The other key assumption is that gene phylogeny is correlated with the evolution of subtype state. Because of these assumptions are central to the method, all new subtype schemes are evaluated for their predictive performance.

Phylotyper is developed in python and R. Outlined below are the steps and tools used in the Phylotyper pipeline:

1. Identify subtype gene loci in input genomes using BLAST (Camacho *et al.*, 2009).
2. Align input genes against a pre-aligned set of reference genes using the tool MAFFT's `--add` feature (Katoh and Standley, 2013; Capella-Gutierrez *et al.*, 2009).
3. If multiple loci are involved, concatenate individual alignments into superalignment.
4. Generate maximum likelihood phylogenetic tree of aligned genes with FastTree (Price *et al.*, 2010).
5. Run phytools `rerootingMethod` using the phylogenetic tree and assigned subtypes (Revell, 2011).
6. Identify the subtype with maximum marginal likelihood for the unknown genes and report to user. Users are provided with an image of the phylogenetic tree overlaid with the likelihood values (e.g. Figure 1).

Phylotyper was designed to be incorporated into a WGS public health workflow. The main input into Phylotyper is assembled genome sequences. Putative loci needed for the subtype scheme are identified in the input genomes using BLAST (Camacho *et al.*, 2009). The identified loci are then sent to the Phylotyper subtype prediction module. It is possible in Phylotyper to use multiple loci for subtype prediction. Individual loci alignments are concatenated to form a single superalignment that is used to build the phylogenetic tree. The Stx1 and Stx2 schemes are examples of a multi-loci subtype scheme; the Stx toxin A and B subunit genes are the inputs in these schemes.

Currently, subtype schemes for *Escherichia coli* (*E. coli*) are available in the Phylotyper package (listed in Supplementary Table 1). However, included in the Phylotyper software, is the capability to add new subtype schemes. Creating a new subtype scheme will save the required reference files in a data directory, allowing newly added schemes to be easily re-run from Phylotyper. Checks are built-in to the new subtype pipeline to ensure the assumptions of the Phylotyper approach are not violated.

3 Results

To assess the performance of the phylogenetic-based method used in Phylotyper and compare it to a sequence similarity-based approach, we ran a leave-one-out cross validation analysis. For this assessment, we developed a mock sequence-similarity based tool that assigns putative subtypes using BLAST. The leave-one-out analysis examined the four subtype schemes available in Phylotyper. The results of the analysis indicated that the precision of the Phylotyper method is consistently higher than in a top-BLAST hit approach (see Supplementary Table S2 for performance metrics).

4 Discussion

From assembled WGS data, Phylotyper can assign unclassifed strains a molecular subtype. Currently the Phylotyper software offers subtyping schemes for *E. coli*. It can, however, be applied to most molecular subtype schemes and Phylotyper includes functionality to build new schemes.

Performance testing showed that Phylotyper is more robust than an approach based solely on sequence similarity. Phylotyper computes an empirical model of subtype evolution to predict subtypes for unclassified sequences. By estimating the phylogenetic distribution of each subtype, Phylotyper can more easily identify a novel subtype or accurately classify a novel sequence allele. The phylogenetic framework in Phylotyper provides a statistical likelihood for interpreting novel alleles. In comparison, the performance of a sequence similarity approach is highly dependent on the reference database containing the exact allele. In a direct sequence search approach, there is no inherient mechanism to interpret novel alleles.

Subtypes are mainly used as a proxy for evolutionarily-related bacterial strain groups or to infer phenotypes. Recent analysis of serotype data revealed several inconsistencies between molecular subtype and genomic data (DebRoy *et al.*, 2016). Subtypes predicted using a phylogenetic framework is more consistent with these main uses of subtype information, so Phylotyper is uniquely capable to transition historical subtype data to new WGS systems.

Funding

This work is funded in part by the Public Health Agency of Canada and a grant from the Genomics Research and Development Initiative

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.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**(15), 1972–1973.

Carrillo, C. D., Koziol, A. G., Mathews, A., Goji, N., Lambert, D., Huszczynski, G., Gauthier, M., Amoako, K., and Blais, B. W. (2016). Comparative evaluation of

genomic and laboratory approaches for determination of shiga toxin subtypes in escherichia coli. *Journal of Food Protection*, **79**(12), 2078–2085.

DebRoy, C., Fratamico, P. M., Yan, X., Baranzoni, G., Liu, Y., Needleman, D. S., Tebbs, R., O'Connell, C. D., Allred, A., Swimley, M., Mwangi, M., Kapur, V., Garay, J. A. R., Roberts, E. L., and Katani, R. (2016). Comparison of o-antigen gene clusters of all o-serogroups of escherichia coli and proposal for adopting a new nomenclature for o-typing. *PLOS ONE*, **11**(1), e0147434.

Ingle, D. J., Holt, K. E., Levine, M. M., Kuzevski, A., Valcanis, M., Robins-Browne, R. M., Tauschek, M., Inouye, M., and Stinear, T. (2016). In silico serotyping of e. coli from short read data identifies limited novel o-loci but extensive diversity of o:h serotype combinations within and between pathogenic lineages. *Microbial Genomics*, **2**(7).

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., and Holt, K. E. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*, **6**(11).

Jenkins, C. (2015). Whole-genome sequencing data for serotyping escherichia coli—it's time for a change! *Journal of Clinical Microbiology*, **53**(8), 2402–2403.

Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M., and Scheutz, F. (2015). Rapid and EasyIn SilicoSerotyping of escherichia coli isolates by use of whole-genome sequencing data. *Journal of Clinical Microbiology*, **53**(8), 2410–2426.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**(4), 772–780.

Lindsey, R. L., Pouseele, H., Chen, J. C., Strockbine, N. A., and Carleton, H. A. (2016). Implementation of whole genome sequencing (WGS) for identification and characterization of shiga toxin-producing escherichia coli (STEC) in the United States. *Frontiers in Microbiology*, **7**.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**(3), e9490.

Revell, L. J. (2011). phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**(2), 217–223.