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

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 of 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 an 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. 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(Jenkins, 2015). Accordingly, new *in silico* tools have been developed to predict subtypes from WGS data (Joensen *et al.*, 2015; Ingle *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 state reconstruction to assign the likelihood of each subtype to the tree branch points, 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 subtype evolution model and phylogeny. In the context of Phylotyper, the marginal likelihood provides a confidence value associated with a predicted subtype.

Phylotyper is developed in python and R. The steps in the Phylotyper pipeline are: (1) Identify subtype gene loci in input genomes using BLAST