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

OXFORD

Subject Section

Phylotyper: *In silico* predictor of 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. In public health, subtyping has been used to infer phenotypes and distinguish bacterial strain groups. *In silico* tools that predict subtypes from sequences data are needed to transition historical data to WGS-based protocols. Phylotyper is a novel solution for *in silico* subtype prediction from gene sequences. Designed for incorporation into WGS pipelines, it is a general prediction tool that can be applied to different 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 over approaches based solely on sequence similarity and provides useful contextual feedback.
Availability and Implementation: Phylotyper is a python and R 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

WWhole-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 subtyping. Subtyping methods can broadly be categorized as phenotype-based or DNA-based. Phenotype-based subtypes are, for example, interrogated by biochemical tests (biotyping), detection of surface antigens (serotyping) or susceptibility to bacteriophage (phagetyping) (Wiedmann, 2002). Alternatively, DNA-based subtyping examines and classifies bacteria based on genetic content. DNA-based subtypes use a variety of methods from PCR to Pulse Field Gel Electrophoresis to DNA sequencing to assign groups to bacteria (Wiedmann, 2002). As a surveillance tool, subtypes provide a clear cut designation that is typically used to distinguish taxonomic groups and infer phenotypes. 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). To predict from WGS data, these tools have targeted subtypes (either Phenotype-based or DNA-based systems) that can be predicted by sequence variation in a specific region or gene in the genome. *In silico* subtype prediction is not applied to subtyping methods that already directly use the DNA sequence to assign subtype such as Multi-Locus Sequence Typing (MLST). No inference or extrapolation is needed to integrate a direct DNA sequence-based method, like MLST, into a WGS system. An example of a subtyping system that has been adapted for WGS is serotyping in *Escherichia coli*. The sequence of O-antigen processing genes in *E. coli* is used to predict O-antigen group serotype in (Joensen *et al.*, 2015; Ingle *et al.*, 2016). Another example is the Shiga toxin subtype (Stx); a DNA-based subtyping scheme generated using PCR. The tool in (Carrillo *et al.*, 2016) predicts Stx subtype by simulating PCR *in silico*.

Phylotyper is a novel *in silico* predictor of subtypes from sequence data. Similar to (Joensen *et al.*, 2015; Ingle *et al.*, 2016; Carrillo *et al.*, 2016), it also works on subtypes that can be predicted from specific, pre-selected gene or genomic region sequences. 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 then 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 (Camacho *et al.*, 2009). Inputs are in fasta format. Hits that do not align over 95% or have under 90% sequence identity with a reference sequence are discarded. Users are notified if no loci are found in genome. (2) Align input genes against a pre-aligned set of reference genes using MAFFT's `--add` feature (Katoh and Standley, 2013). (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 in text output file. Users are also provided with an image of the phylogenetic tree overlaid with the likelihood values (e.g. Figure 1).

Detailed descriptions on how to run Phylotyper are provided here: `https://github.com/superphy/insilico-subtyping`.

PPhylotyper was designed to be incorporated into a WGS workflow. The main input into Phylotyper is assembled genome sequences (in fasta format). 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.

Currently, the Phylotyper package includes the following subtype schemes for *Escherichia coli*: Stx, intimin and serotype O- and H-types (Supplementary Table 1). However, the Phylotyper software also has the capability to add new subtype schemes (instructions are provided here: `https://github.com/superphy/insilico-subtyping`). Creating a new subtype scheme will save the required reference files, allowing newly added schemes to be easily re-run from Phylotyper. To add a new subtype scheme for use in Phylotyper, users require a training set of homologous sequences with assigned subtype whose phylogenetic grouping is predictive of the subtype. Phylotyper assumes that the provided training sequences are 1) homologous, specifically, they are suitable for alignment and phylogenetic reconstruction 2) the sequence phylogeny is correlated with the subtype distribution and 3) the set is representative of the range of sequences that make up a subtype. Checks are built-in to the `new` pipeline to validate the submitted reference set. Each new subtype is subject to two tests and results are reported to the user. The first test checks that the distribution of inter-patristic phylogenetic distances between instances of the same subtype is both smaller and distinct from subtypes that are different. This test can identify isolated cases of potentially mislabelled subtype genes that are tightly clustered with other subtypes. A second check computes the accuracy measure, F-score, through a leave-one-out cross-validation; a procedure which uses each sequence in the training set as a test input to estimate positive and negative prediction rates (see Supplementary Methods for details). The performance metrics: recall and F-score rapidly decrease as the correlation between the training set phylogeny and subtype distribution decreases (Supplementary Figure S2). These checks verify that the training set sequences' phylogenetic grouping is predictive of the subtype. There is no check that can confirm the training set covers all subtypes in a particular scheme; Phylotyper can only predict subtypes that are represented in the training set. It is important that the user monitor schemes and update it as gaps are identified. Phylotyper is designed to return a `non-significant/undetermined` result when encountering an unknown subtype that has no representative in the training set.

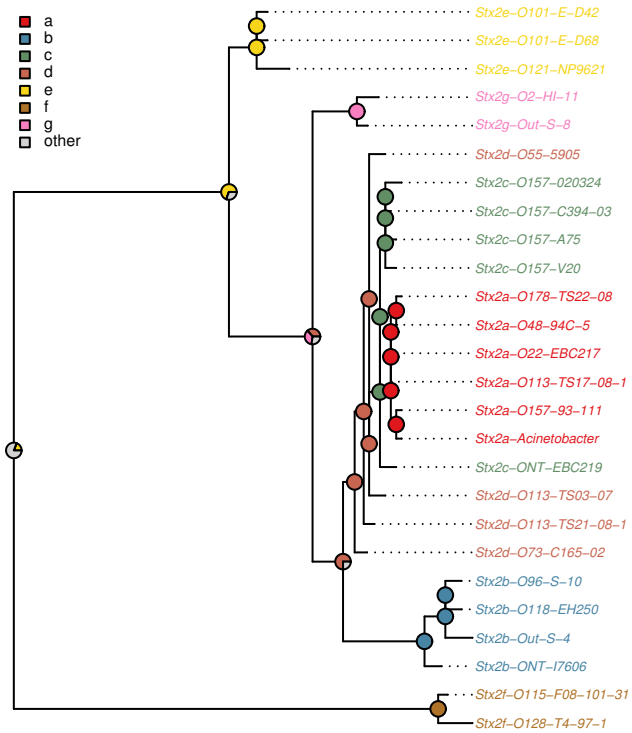

Fig. 1. Phylogenetic tree for select Stx2 genes. The subtype marginal likelihood is displayed at each node as a pie chart. Subtype is indicated by color as shown in the legend. The full Stx2 tree is displayed in Supplementary Figure S1.

3 Results

Phylotyper is a progression from the sequence-similarity approach that is the basis of current *in silico* subtype prediction strategies. To compare Phylotyper to a sequence similarity-based approach, we ran two validations that looked at how both methods perform when confronted with 1) gene sequence or 2) subtype class not present in the training set. The first validation was a leave-one-out cross-validation test that iterated through each gene in the training set, retraining the prediction tools on a reduced training set that excludes the selected test gene, and then confirming if the retrained predictor could recover the subtype of the test gene. This validation tests how the predictors perform when run on a distinct sequence that is not in the training set. The second validation examined

how the predictors perform when tested with a gene that has a subtype, not in the reference set. In this validation, each subtype was iterated over and all genes that are assigned the subtype were removed from the training set. In each iteration, we recorded the number of false-positive subtype assignments when the test sequences were used as input. The correct response for the predictors was to return a negative result since the subtype does not exist in the training set. For these assessments, we developed a sequence-similarity based tool that assigns putative subtypes using BLAST. This generalized BLAST tool, based on the approach used in (Joensen *et al.*, 2015), assigns a query sequence a subtype when the top BLAST match from an annotated reference database is above a pre-selected percent identity and alignment coverage cutoff. Details how the assessment was conducted are available in Supplementary methods. The assessment examined the five subtype schemes available in Phylotyper: Stx1, Stx2, Eae, H-type (FliC), O-type (Wzy & Wzx). When tasked with assigning a novel gene sequence not in the training set in the leave-one-out validation, Phylotyper consistently had higher precision than a top-BLAST-hit approach. The average precision in Phylotyper was 0.99 versus 0.96 in the top-BLAST-hit approach (Supplementary Table S2). The BLAST approach also often had lower recall rates; it had an average recall of 0.81 compared to 0.90 with Phylotyper. Similarly, when entire subtype classes were withheld from the training set, Phylotyper had consistently lower false positive rates for all subtypes schemes tested; the average false positive rate in this test case was 0.11, while in the BLAST approach, the average false positive rate was 0.30. A separate assessment for the V-typer tool; a Stx subtype predictor, was run using selected Stx gene sequences from the experimentally-verified Phylotyper training set (Carrillo *et al.*, 2016). The test Stx genes had sufficient surrounding DNA sequence to support *in silico* PCR. In total, 24 Stx gene sequences were tested with the V-typer tool and V-typer returned results for 7, all correct. Phylotyper correctly predicted the subtype for all these genes. Based on this level of recall, it appears conditions in the Stx subtype environment are challenging for simulated PCR.

All new or updated subtype schemes added in Phylotyper are subject to a leave-one-out cross-validation test. The test is part of the `add` pipeline and is used to estimate the F-score of the subtype scheme. The F-score reflects the predictive capability of the subtype scheme. If the associated phylogeny for the training set gene sequences is not correlated with the subtype distribution, this will be reflected in the F-score. To demonstrate this property, we randomly assigned subtypes for increasing proportions of the genes in the training set and computed the F-score with the leave-one-out validation for each proportion level. The F-score and recall rapidly decrease as the proportion of randomly altered subtypes increases (Supplementary Figure S2).

4 Discussion

From assembled WGS data, Phylotyper can assign unclassified strains a subtype. Currently, the Phylotyper software offers subtyping schemes for *E. coli*. It can, however, be applied to other subtype schemes and Phylotyper includes functionality to build new schemes. Phylotyper can produce predicted subtypes from any input sequence that is strongly correlated with the subtype distribution, however, input sequences with a direct biological causal link to the subtype will have fewer caveats; A gene sequence that is causal cannot become disassociated from the subtype through recombination or horizontal gene transfer. Outside of *E. coli*, the PCR-based capsular typing system for *Haemophilus influenzae*, Neurotoxin serotyping in *Clostridium botulinum* and the haemagglutinin and neuraminidase types in Influenza A virus are all examples of potential future subtype schemes that we are incorporating into Phylotyper. We plan on expanding the Phylotyper resource by adding and updating high-quality subtype schemes for other pathogens. We encourage users to contact us with their new subtype schemes or updates to schemes (`https://github.com/superphy/insilico-subtyping`).

The main strategy currently in use by other *in silico* tools for predicting subtypes is to use sequence similarity to annotated gene alleles. Query genomes or genes are matched to alleles that are attributed the subtype phenotype or are correlated with the distribution of the subtype. For example, SerotypeFinder uses BLAST to find the top matches based on sequence similarity to O-antigen processing genes for *in silico* O-typing and the flagellin genes for H-typing *E. coli* genes (Jenkins, 2015). O-type and H-type are transferred from the top matches to the queries provided they are above coverage and percent identity thresholds. This general strategy of allele matching is also applied in the EcOH tool (Ingle *et al.*, 2016), however, the EcOH tool can directly use unassembled sequence reads as input. The EcOH tool aligns reads to alleles linked to *E. coli* O-types and H-types, and identifies the top candidates that have an alignment score above pre-defined thresholds. Phylotyper is comparatively more robust as it generates fewer Type-I errors when encountering novel alleles or subtypes not present in the training set. With the allele matching strategy, the reference set make-up can have a greater impact on performance. When alleles or even subtypes are missing in the reference database, the sequence similarity approach more frequently generates false positive predictions. Alternatively, Phylotyper computes an empirical model of subtype evolution to predict subtypes for unclassified sequences. By estimating the phylogenetic distribution of each subtype, Phylotyper is less likely to make a Type-I error when encountering a novel subtype or allele. The empirical testing we performed demonstrated this behavior; the rate of false positive classifications was significantly lower than in a sequence-similarity approach in validations where we withheld an allele or an entire subtype from the training set and used it as a test input. V-typer takes a distinct approach; it directly simulates the *in vitro* wet-lab PCR procedure used to perform Stx subtyping (Carrillo *et al.*, 2016). V-typer's direct replication of a wet-lab method *in silico* means it can only be applied to subtypes schemes that use PCR. Additionally, we found in our evaluation of Stx subtypes that it failed to generate predictions for most test cases. From a methodology standpoint, Phylotyper has an additional benefit over current methods; the phylogenetic framework in Phylotyper provides a statistical likelihood for interpreting results. In comparison, there is no built-in mechanism in the sequence similarity approach to quantify the level of confidence in assigning alleles a subtype.

Subtypes are mainly used as a proxy for evolutionarily-related bacterial strain groups or to infer phenotypes. A recent analysis of O-antigen serotypes and their associated O-antigen gene sequences in *E. coli* found that the sequence data indicated several changes to the organization of the O-groups (DebRoy *et al.*, 2016). There are potentially other subtype schemes that would show discrepancies between genetic data and subtype grouping. A tool that can evaluate the ability of a genotype to predict a subtype would be better equipped for developing the new subtype schemes or updating current schemes for WGS workflows. Phylotyper's `add` pipeline tests subtype schemes for their predictive accuracy by returning an F-score based on a cross-validation assessment. The validation verifies that the phylogeny generated by the training set sequences can be used to predict the subtypes with a high level of accuracy. We showed in empirical tests that the farther a subtype is dissociated from the input gene's phylogeny, the lower the F-score computed in the Phylotyper `add` pipeline. In addition to this subtype-level verification, Phylotyper also computes confidence scores for each individual prediction that reflect the rate of subtype change occurring at the input gene's phylogenetic locale. If an input sequence falls in a region in the phylogenetic tree where the subtype is highly fluid due to evolution or poor subtype-genotype correlation, users would be notified in the confidence score and in the phylogenetic tree visualization output by Phylotyper. Phylotyper's ability to inform users about the level

of agreement between the subtype assignments and genotype makes it uniquely capable of transitioning historical subtype data to new whole genome sequence-based 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. *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, **10**(1), 421.

Carrillo, C. D. *et al.* (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. *et al.* (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. *et al.* (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).

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. *et al.* (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.

Price, M. N. *et al.* (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.

Wiedmann, M. (2002). Subtyping of bacterial foodborne pathogens. *Nutr. Rev.*, **60**(7 Pt 1), 201–208.