Contents

1 Supplementary Methods

1.1 BLAST-based method

For comparison purposes, we developed a generalized sequence-similarity predictor of subtypes using BLAST. A training set of sequences with annotated subtypes is used to build a BLAST database (the training sets are from the the Phylotyper repository). To predict a subtype, query sequences are searched against the training BLAST database and the top hit is recorded. If the top hit's alignment with the query has at least 60% coverage with the reference sequence and is above the percent identity cutoff, then the subtype of the top hit is returned as the predicted subtype of the query sequence. The cutoff impacts performance in one of two ways: too low and non-equivalent subtype alleles may be matched with the query (i.e. Type-I error), too high and valid subtype allele BLAST hits are filtered out (i.e. Type-II error). To select the ideal cutoff that balances these competing sources of error, a cross-validation simulation was run to estimate false discovery and true positive rates across varying percent identity cutoffs. A 10-fold cross-validation was performed; 10% of the training set was randomly selected. The remaining 90% was used to build the reference BLAST database. The 10% test set was then searched against this database using BLAST and positive and negative results were recorded. This step was repeated 100 times and the F1-score, which is a weighted average of the precision and recall, was calculated. Precision is defined as:

$Precision = TP/(TP + FP)$

and recall as:

$Recall = TP/(TP + FN)$

These values are then combined to produce an F-score as follows:

$2 * (precision * recall)/(precision + recall)$

$TP = True\ Positive,\ \ FP = False\ Positive,\ \ FN = False\ Negative$

The F-score; a binary classifier metric, was recorded individually for each subtype and then averaged to create a single performance metric for this multi-class predictor. F-score was computed for increasing percent identities, and the lowest percent identity corresponding to the maximum F-score was selected as the cutoff. The percent identity cutoff was selected as 0.96.

1.2 Performance Assessment

Two separate cross-validation tests were performed to compare the performance of Phylotyper with the BLAST method. The comparison to V-typer is a separate analysis, as V-typer does not use a training set and a cross-validation test would not work in this situation. A leave-one-out cross-validation analysis was performed. For testing Phylotyper, the subtype label of each sequence in the training set was hidden, one at a time and the remaining training set sequences were used to predict the subtype state of the test gene. For the BLAST-based subtype predictor, each gene was removed from the BLAST database and used as a query input. Positive and negative results are recorded for each cross-validation iteration. The second validation iterated over each subtype and removed all genes assigned that subtype from the training set. The removed genes or test set was then used as input to identify the level of false positive predictions that occur when the tools are presented with novel subtypes not in the training set. Precision, recall and the F-score as defined earlier for a multi-classifier test, were computed for the validations.

1.2.1 V-typer Assessment

The V-typer tool predicts Stx1 or Stx2 subtypes by simulating PCR (i.e. mapping primers to the genome). Due to the unique method in V-typer, which does not use a training dataset, the cross-validation assessment was not used. Instead, a special test set of Stx1 and Stx2 genes was collected and used to evaluate the performance of V-typer relative to Phylotyper. Each experimentally-verified Stx gene sequence that makes

up the Phylotyper training dataset was searched against the NCBI nt database to retrieve surrounding sequence that may be involved in PCR primer mapping by V-typer. Specifically, sequences that have at least 5 kB up- and downstream from the Stx gene were selected for the assessment. V-typer was run on each sequence and the results recorded in Supplementary Table S3.

1.3 Randomizing Subtype Labels

The Phylotyper method assumes the reference sequence phylogeny aligns with the subtype distribution. To validate this assumption, an empirically computed F-score is generated for each new subtype scheme (the F-score is computed through a leave-one-out cross-validation test as described in the Performance Assessment section). The F-score indicates the ability of the reference sequence phylogeny to predict the subtype. To test how the F-score is impacted by the level of correlation between the subtype distribution and the subtype sequence phylogeny, we ran a simulation where we randomly assigned subtypes to a select proportion of the reference set sequences. In the simulation, we randomly selected proportions ranging from 10 to 50 percent of the reference set and then randomly assigned the selected sequence a new subtype (so selected subtypes did not match original subtype label). With the modified dataset, F-score was computed through Phylotyper's leave-one-out cross-validation test. For each proportion, 100 iterations were performed and metrics averaged.

2 Supplementary Figures

2.1 Supplementary Figure S1

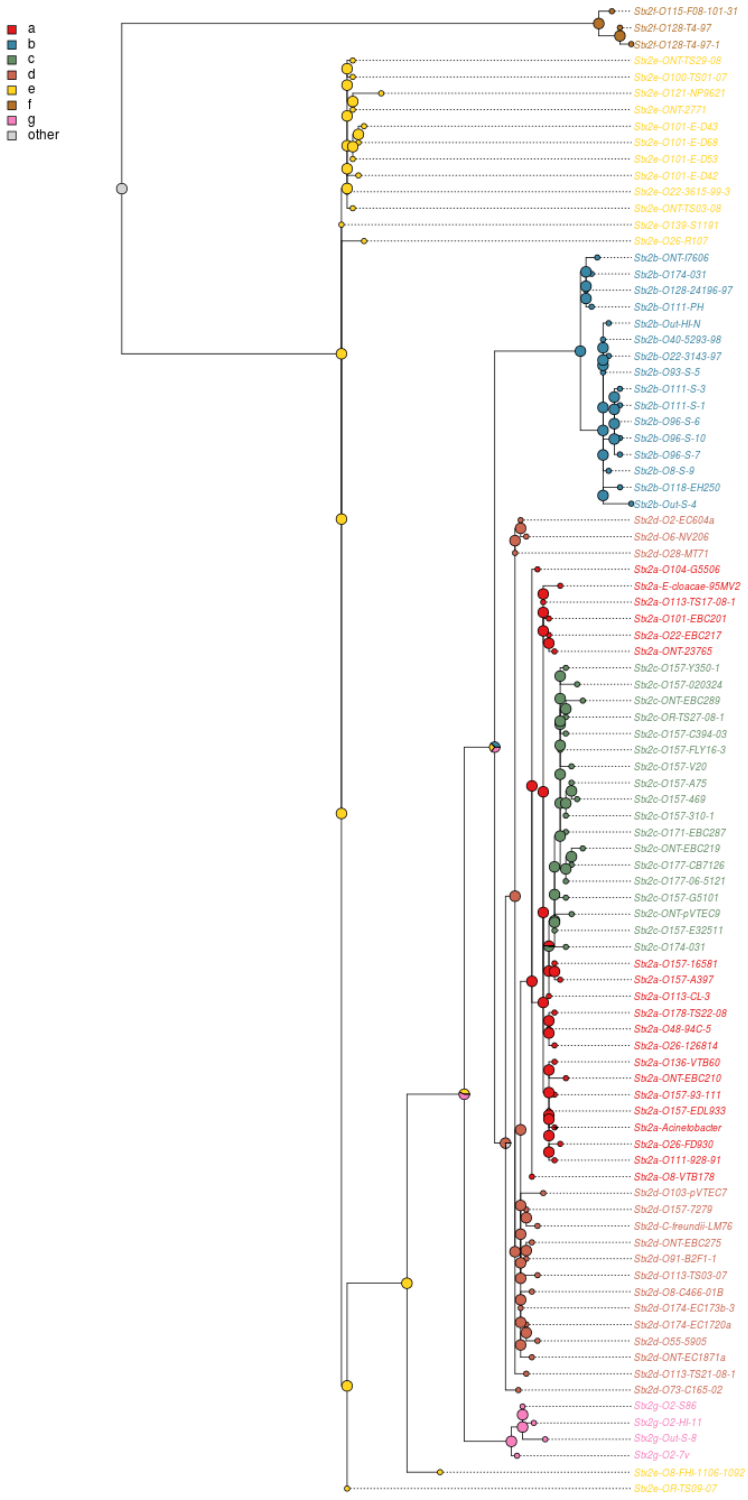

Figure S1: Stx2 Phylogenetic tree showing the Phylotyper marginal likelihoods as pie charts. This output is provided to the user.

2.2 Supplementary Figure S2

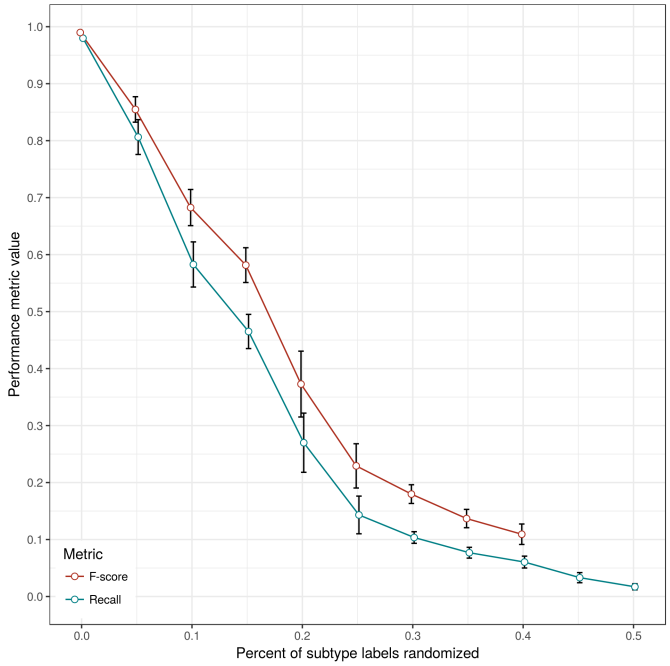

Figure S2: The effect of randomizing a specified proportion of subtype labels in the Eae phylogenetic tree on the empirically estimated performance metrics: F_1-score and recall.

3 Supplementary Tables

3.1 Supplementary Table S1

Table S1: Subtype Schemes in Phylotyper

Name	Description	Species	Loci
stx1	Shiga-toxin 1 subtype	*E. coli*	2
stx2	Shiga-toxin 1 subtype	*E. coli*	2
eae	Intimin subtype	*E. coli*	1
flic	H-serotype based on flagellin gene; fliC	*E. coli*	1
wz	O-serotype based on the wzy and wzx genes	*E. coli*	2

3.2 Supplementary Table S2

Table S2 contains performance metrics from a leave-one-out cross-validation analysis comparing Phylotyper and a top-BLAST hit approach. The analysis examines the four *E. coli* schemes available in Phylotyper. In this multi-class analysis, precision, recall and F_1 score are calculated for each individual class provided that at least one instance of the class is in the training set. The individual class positive and negatives are summed to calculate an overall precision, recall and F_1 score for the scheme.

Table S2: Leave-One-Out Cross Validation Results

Scheme	Phylotyper				Sequence-similarity		
	Precision	Recall	F_1 Score	Run-time (s) [1]	Precision	Recall	F_1 Score
E. coli Stx1	1.00	0.94	0.97	6	0.94	0.94	0.94
E. coli Stx2	1.00	0.99	0.99	32	0.93	0.93	0.93
E. coli Intimin	1.00	0.98	0.99	17	0.99	0.98	0.99
E. coli H-serotype	0.99	0.98	0.98	16	0.96	0.85	0.90
E. coli O-serotype	1.00	0.61	0.75	67	1.00	0.36	0.53

Formula:

1. $Precision = TP/(TP + FP)$
2. $Recall = TP/(TP + FN)$
3. $F_1\ score = 2 * Precision * Recall/(Precision + Recall)$

$TP = True\ Positive,\ \ FP = False\ Positive,\ \ FN = False\ Negative$

[1] Run-time is for an *Escherichia coli* genome containing a unique gene which triggers the full phylotyper pipeline.

3.3 Supplementary Table S3

Table S3: V-typer Performance

NCBI Accession	Stx Subtype	V-typer Prediciton[1]	Phylotyper Prediction
CP015020	2d	N	2d
CP015229	2g	2g	2g
CP009106	2a	N	2a
CP007133	2a	2a	2a
CP006262	2a	2a	2a
AP010960	2a	2a	2a
CP013663	2a	N	2a
CP015228	2a	N	2a
CP011331	2a	N	2a
HF572917	2a	N	2a
LN554923	2a	N	2a
LM997071	2a	N	2a
CP018250	2c	N	2c
CP018252	2c	N	2c
CP018247	2c	N	2c
CP018245	2c	N	2c
CP018243	2c	N	2c
CP017446	2c	N	2c
CP017442	2c	N	2c
CP017438	2c	N	2c
CP014314	2c	N	2c
LM997036	2a	2a	2a
LM996489	1c	1c	1c
LM997036	1c	1c	1c
Total	24	7	24

[1] N=Negative (no results returned)