Contents

1 Supplementary Figures

1.1 Supplementary Figure S1

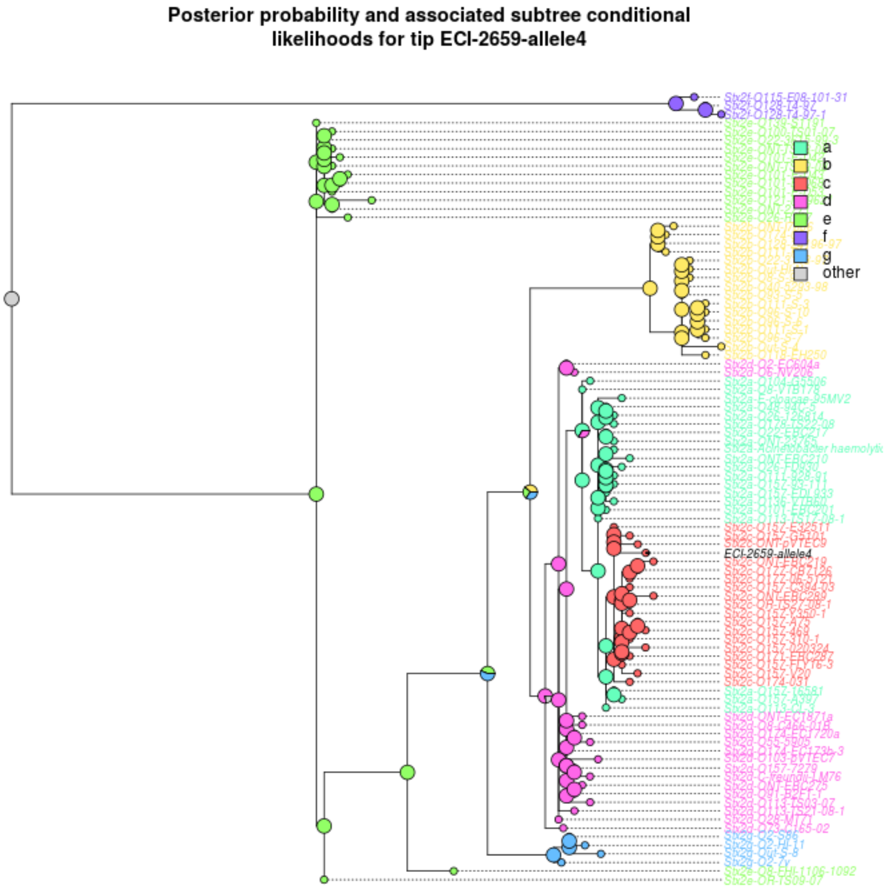

Figure S1: Stx2 Phylogenetic tree showing the Phylotyper marginal likelihoods as pie charts. This output is provided to the user.

2 Supplementary Tables

2.1 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

2.2 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 non-identical subtype gene which triggers the full phylotyper pipeline.