Whole Genome Sequencing of Shiga Toxin-producing *E. coli* to Identify Serogroup, Virulence Factors, and Genomic Clusters

NEW YORK STATE OF OPPORTUNITY. Department of Health | Wadsworth Center

PH: (518) 473 – 2774
Samantha.Wirth@health.ny.gov

S.E. Wirth, D.J. Baker, D. Bopp, A. Cukrovany, M.C. Dickinson, T. Halse, E. Lasek-Nesselquist, P. Lapierre, T. Quinlan, L. Thompson, D. Wroblewski, K.A. Musser, and W.J. Wolfgang

Summary of Project

Background Information

- Shiga toxin-producing *E. coli* (STEC) is responsible for ~265,000 infections each year
- Non-O157 STEC cause approximately 75% of these infections
- Our lab currently identifies genomic subtypes with pulsed-field gel electrophoresis (PFGE). Serogroup and virulence factors are identified by several real-time PCR assays.

Purpose of Study

To evaluate Whole Genome Sequencing (WGS) as a method for determining genomic subtype, serogroup, and virulence factors.

Approach

- **200** sporadic and outbreak-associated STEC were sequenced and analyzed
- Compared results from WGS analysis to the results of current gold standard assays
- For genomic subtyping:
  - Gold standard: PFGE
  - WGS analysis: Phylogenetic analysis was performed by mapping raw reads to an O26 or O121 reference genome
- For serogroup identification of **O26, O45, O103, O111, O121** and **O145**:
  - Gold standard: Two 3-plex real-time PCR assays
  - WGS analysis: Serogroup was identified by employing a kmer approach.
    - Raw sequence reads were converted into kmers of 25 nucleotides (Jellyfish).
    - **Unique** kmers were identified and extracted (kSNP3.0). Unique kmers were present in all isolates of a serogroup to the exclusion of all other serogroups. An in-house script counted the number of identical kmer matches between reads of query isolate and serogroup. Kmers from isolates shared >95% of unique kmers with one serogroup.
- For Shiga toxin genes (*stx1* and/or *stx2*):
  - Gold standard: One 2-plex real-time PCR assay
  - WGS analysis: Shiga toxin genes were identified by employing a kmer approach
    - Raw sequence reads for shiga toxin (*stx*) genes 1 and 2 were converted into kmers of 29 nucleotides (Jellyfish). Unique kmers were not shared between these genes and could not be present in *E. coli* (*stx* negative) genomes.

Results and Conclusions

Genomic subtyping

- SNP-based phylogenetic clusters, PFGE-defined clusters, and epidemiologically defined outbreaks showed a high degree of concordance.
- Isolates from known outbreaks clustered tightly with 0 to 5 SNPs difference
- Isolates belonging to the same serogroup; thousands of SNPs difference
- Reference genome (O26 or O121) did not affect the composition of genomic clusters associated with known outbreaks, but did affect deeper tree structure.

Identification of serogroup and virulence factors

- WGS analysis was able to correctly identify serogroup for **100%** of samples
- WGS analysis was able to correctly identify virulence factors for **99.5%** of samples

References and Acknowledgements

Mingle, L.A., Garcia, D.L., Root, T.P., Halse, T.A., Quinlan, T.M., Armstrong, L.A. Chiefari, A.K., Schoonmaker-Bopp, D.J. Dumas, N.B., Limberger, R.J., Musser, K.A.. Foodborne Path. and Disease. 2012. 9(11): 1028-1036. “Enhanced identification and characterization of non-O157 shiga toxin-producing *Escherichia coli*: A six-year study.”

Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) 27(6): 764-770 (first published online January 7, 2011) doi: 10.1093/bioinformatics/btr011.

Wadsworth Center Applied Genomic Technologies Core and Bioinformatics Core

Wadsworth Center Public Health Genomics Initiative

FDA Whole Genome Sequencing Program, GenomeTrakr Network

CDC Emerging Infections Program

This publication (journal article, etc.) was supported by Cooperative Agreement Number U50CK000423 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

Results

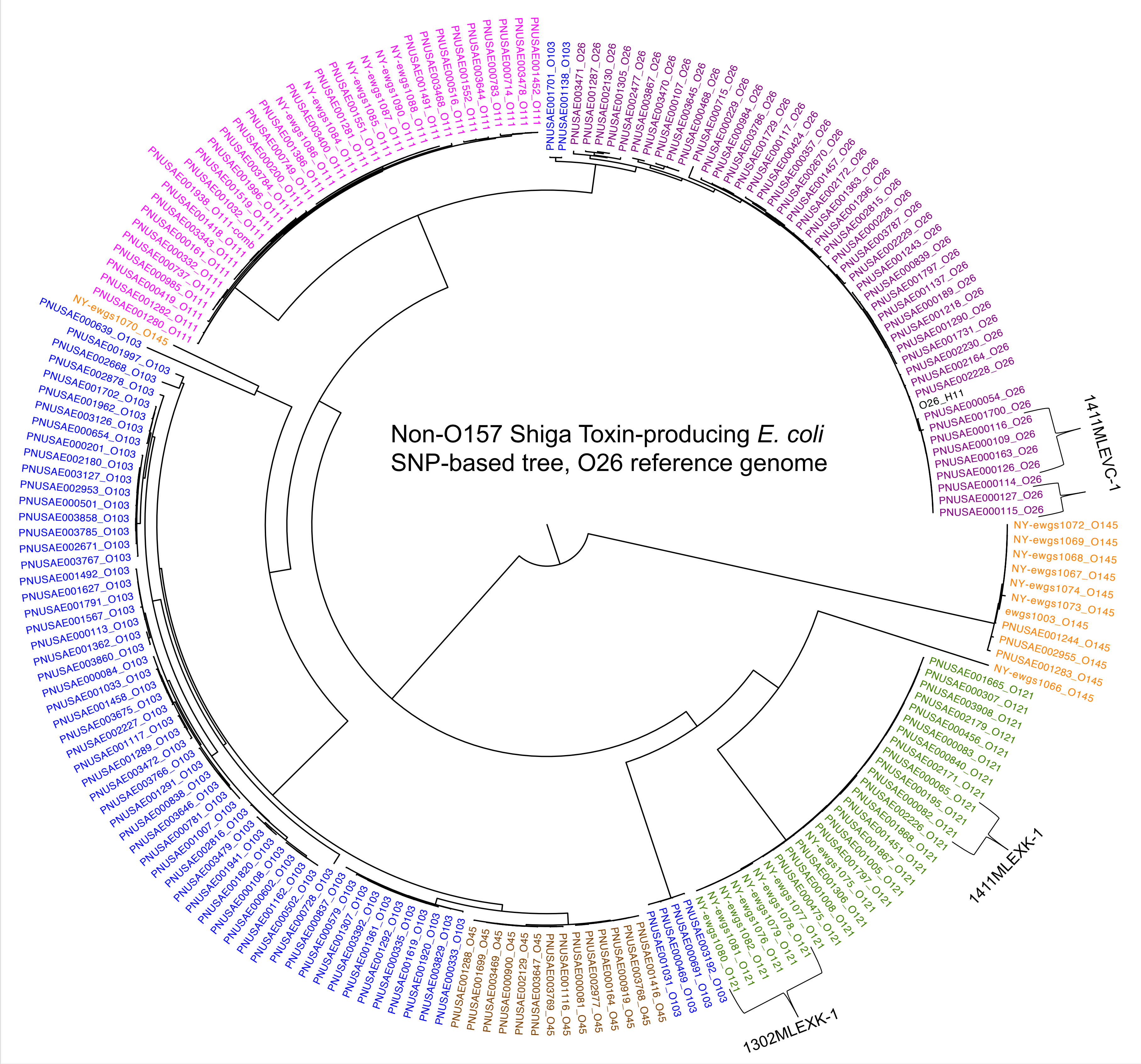
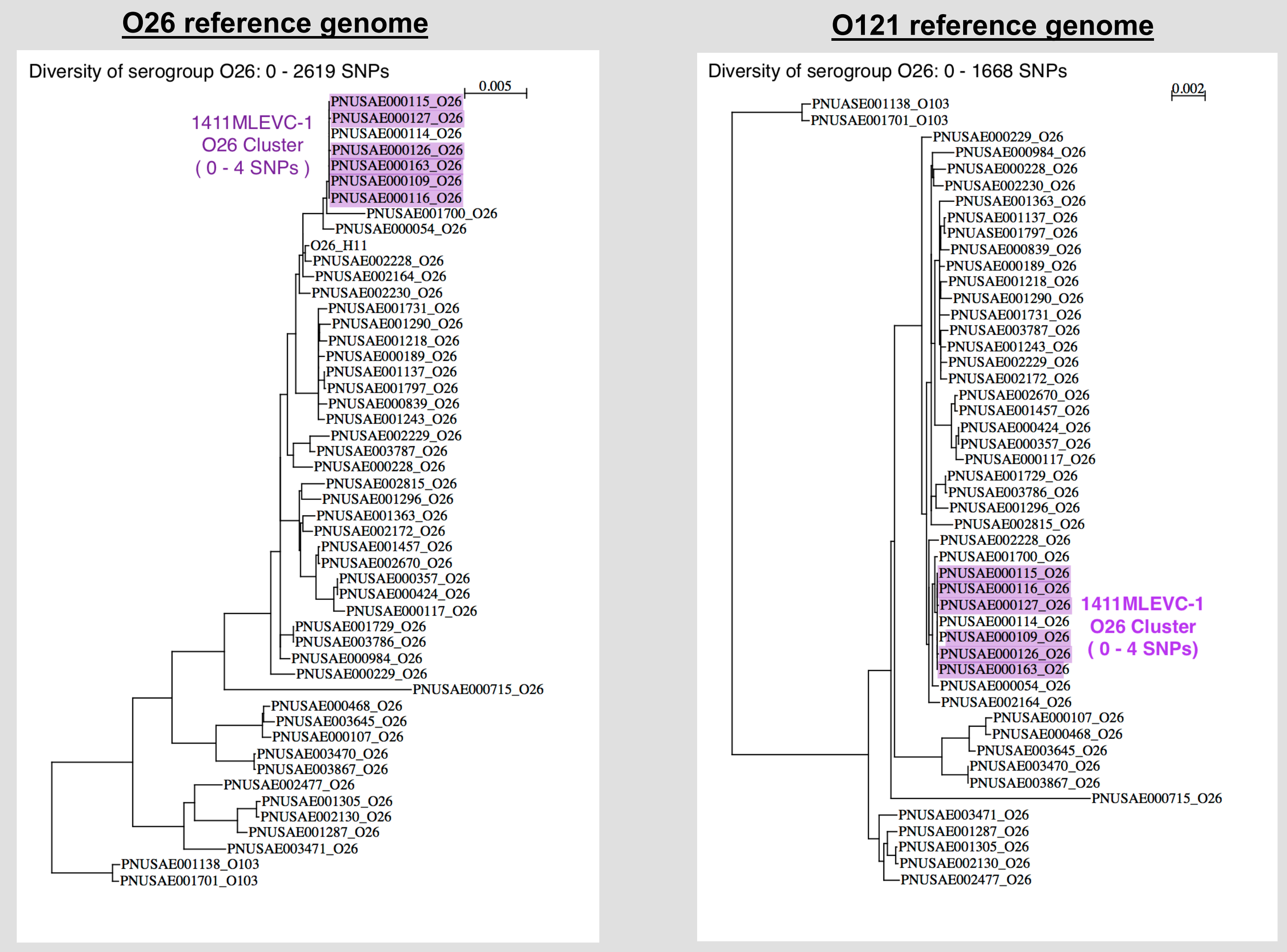

Non-O157 Shiga Toxin-producing E. coli
SNP-based tree, O26 reference genome
1411MLEVC-1
1411MLEXK-1
1302MLEXK-1

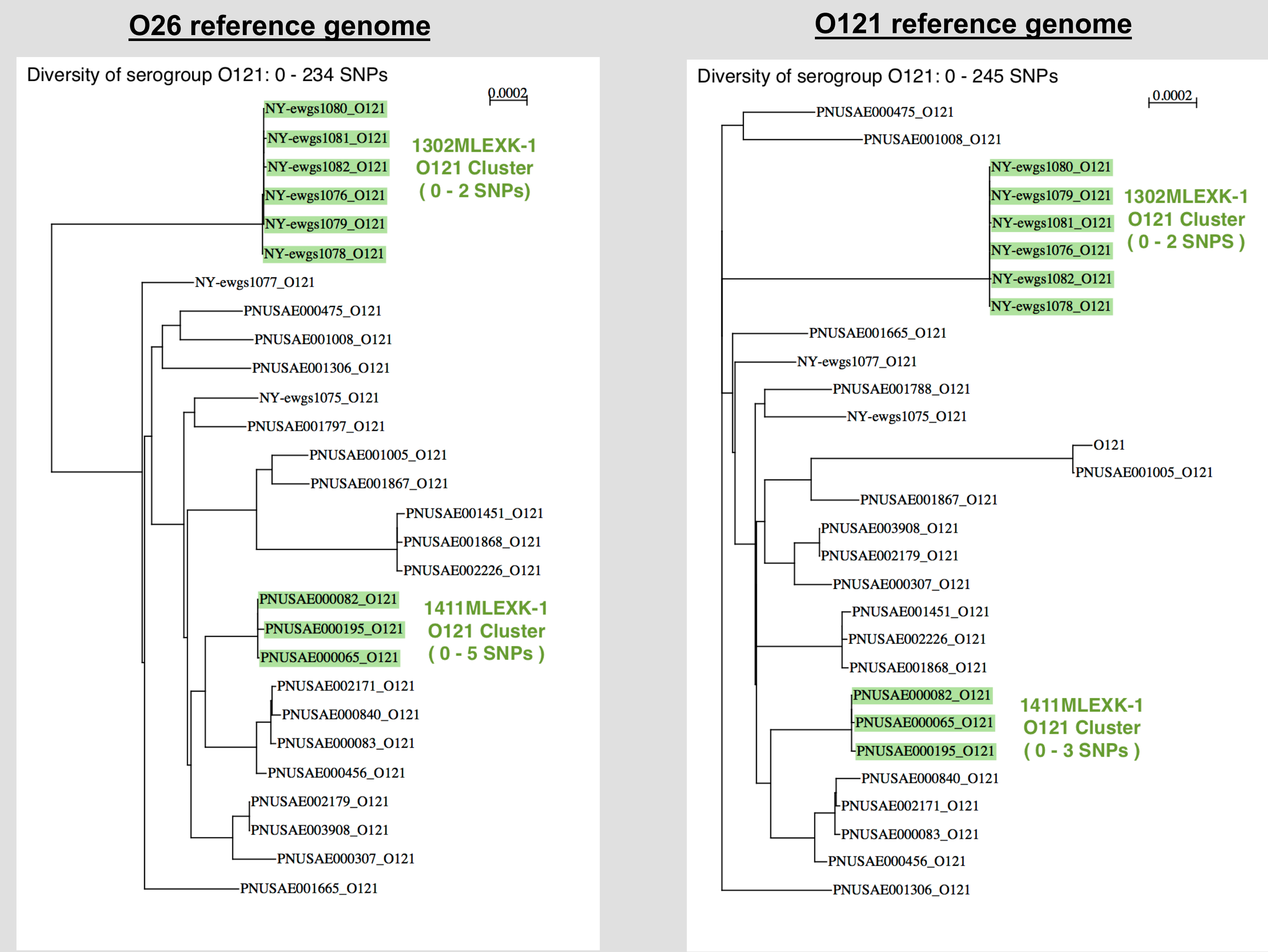
Figure 1. SNP-based phylogenetic tree of 200 Non-O157 Shiga Toxin-producing *E. coli*. Serogroups O26, O45, O103, O111, O121 and O145 were included in this analysis. Raw reads were mapped to a single O26 reference genome. Isolates generally clustered according to serogroup, with known outbreaks forming very tight clusters.

The serogroup of reference genome <u>does not</u> affect the composition of outbreak-associated genomic clusters, but does affect deeper tree structure.

O26 clade phylogenies using two different reference genomes



O121 clade phylogenies using two different reference genomes



WGS can be used to Identify serogroup

Identity by real-time PCR assay	% concordance of WGS kmer approach to real-time PCR assay (# samples)
O26	100% (48)
O45	100% (14)
O103	100% (63)
O111	100% (35)
O121	100% (28)
O145	100% (12)
Total	**100% (200)**

Table 1. Groups of unique kmers (25 nucleotides) were used to identify the serogroup of each sample. An in-house script counted the number of identical kmer matches between reads of query isolate and serogroup. Kmers from isolates shared >95% of unique kmers with one serogroup.

WGS can be used to Identify virulence factors

Identity by real-time PCR assay	% concordance of WGS kmer approach to real-time PCR assay (# samples)
Shiga Toxin 1	100% (137)
Shiga Toxin 2	97.6% (41)
Shiga Toxin 1 and 2	100% (21)
Total	**99.5% (199)**

Table 2. Unique kmers (29 nucleotides) were used to bioinformatically identify the presence of genes for Shiga Toxin 1 and/or Shiga Toxin 2. There was only one discrepancy between real time PCR and WGS analysis; one sample was *stx2* (+) by real time PCR, but *stx2* (-) by WGS analysis.