STEC Whole Genome Sequencing Project

Eija Trees, Ph.D., D.V.M.

Chief, PulseNet Next Generation Subtyping Methods Unit

16th Annual PulseNet Update Meeting
August 29th, 2012

Objectives of the Presentation

- ❑ Describe the STEC genome sequencing project
- ❑ Describe the sequence diversity seen in the STEC population
- ❑ Way forward

Outline

- ❑ Background and Objectives of the Project
- ❑ Materials and Methods
- ❑ Results
- ❑ Conclusions
- ❑ Future Plans
- ❑ 100K Pathogen Genome Sequencing Project

Background

- **CDC Bioinformatics Blue Ribbon Panel in June 2011**
- **2011 end-of-year funding from CDC OD, OID and NC**
  - Implemented a proof-of-concept bioinformatics core group through an outside contract
    - In 2012 focus on two pilot projects:
      - MicrobeNet: development of shared bioinformatic services and web platforms
      - Next Generation PulseNet: Whole genome sequencing, comparative genomics, molecular epi

Next Generation PulseNet Project Objectives

- Perform whole genome sequencing, assembly and comparative analysis of 250 STEC strains
  - Reference database
- Understand diversity in the STEC population and determine correlation between WGS data, PFGE and epi
- Mine the resulting data for targets to predict strain type, virulence and antimicrobial resistance quickly and at high resolution

Next Generation PulseNet Project Objectives (cont’d)

- **Public health benefits:**
  - Standardized workflows/SOPs for next generation sequencing across multiple platforms
  - Reusable tools, techniques and software “pipelines” for a wide range of comparative genomics and molecular epidemiologic applications
  - Builds capacity for rapid genome-level molecular epidemiology

MATERIALS AND METHODS

Strain Selection

- Selection criteria should help us to understand variation:
  1. within an outbreak or during a carrier state
  2. among epidemiologically unrelated isolates within a serotype and between serotypes

Strain selection (cont'd)

- **Variation within an STEC outbreak or a carrier state**
  - 10 isolates each from 5 outbreaks
    - O157:H7 (3), O111:NM (1), O145:NM (1)
    - Different types of outbreaks: clonal vs. polyclonal, short vs. long lasting, point source vs. vehicle never identified
  - 11 isolates recovered from the same patient in the course of 2½ months

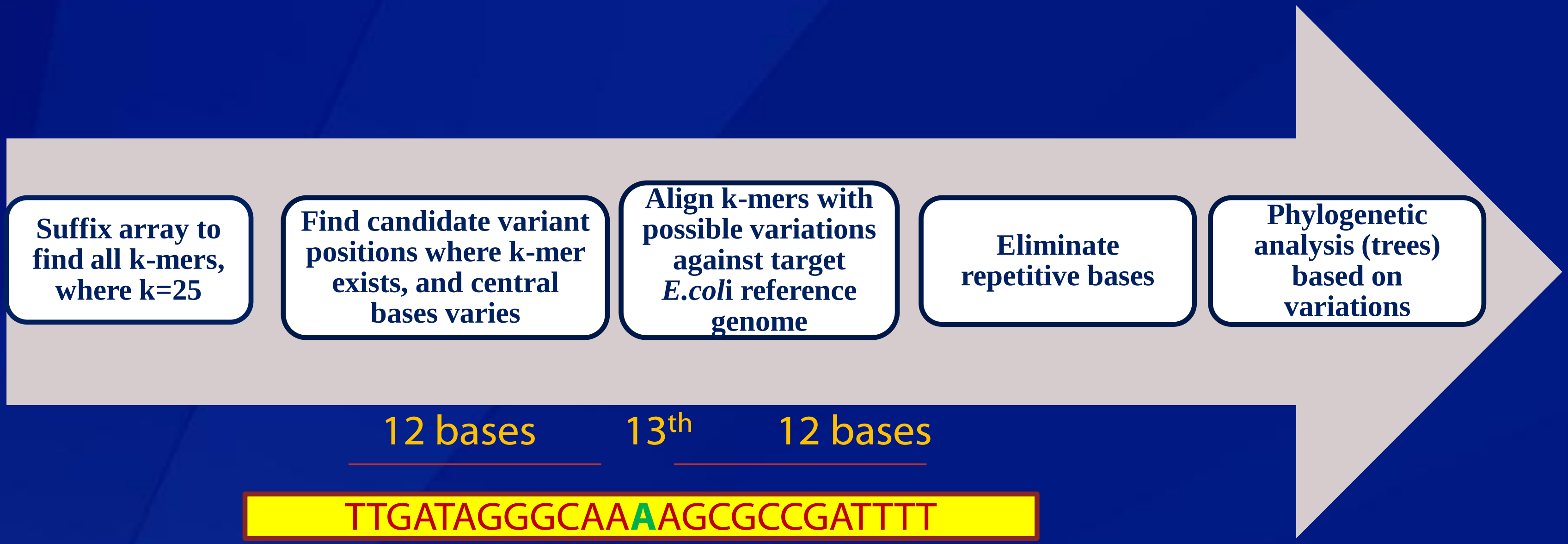
Strain selection (cont'd)

- **Variation among epidemiologically unrelated STEC strains**
  - 25 strains each from serogroups O26, O111, and O121
  - 5 strains each from serogroups O45, O69, O91, O103, O118, O145
  - 1 strain each from serogroups ranked in prevalence #11-22
  - 1 representative each from past historical O157 (and non-O157) outbreaks (36 in total)
  - Sporadic O157 isolates representing common PFGE patterns (23 in total)
- **Other *E. coli* pathotypes: ETEC (6), EPEC (3), EAggEC (3), EIEC (1)**

Whole genome sequencing and assembly

- **Illumina HiSeq 2x100 bp; average coverage 400x**
  - Pros: highly accurate reads with massive coverage
  - Cons: reads short
  - Assembled with CLC Bio; average of 200 contigs
- **Subset of 99 isolates sequenced with Pacific Biosciences SMRT (on-going)**
  - Pros: longer reads (up to 10 kb)
  - Cons: error rate high, multiple SMRT cells required per strain for adequate coverage
  - Hybrid assembly with error correction using Illumina reads being optimized

K-mer -based Approach to Identify SNPs



Suffix array to find all k-mers, where k=25

Find candidate variant positions where k-mer exists, and central bases varies

Align k-mers with possible variations against target *E.coli* reference genome

Eliminate repetitive bases

Phylogenetic analysis (trees) based on variations

12 bases 13th 12 bases

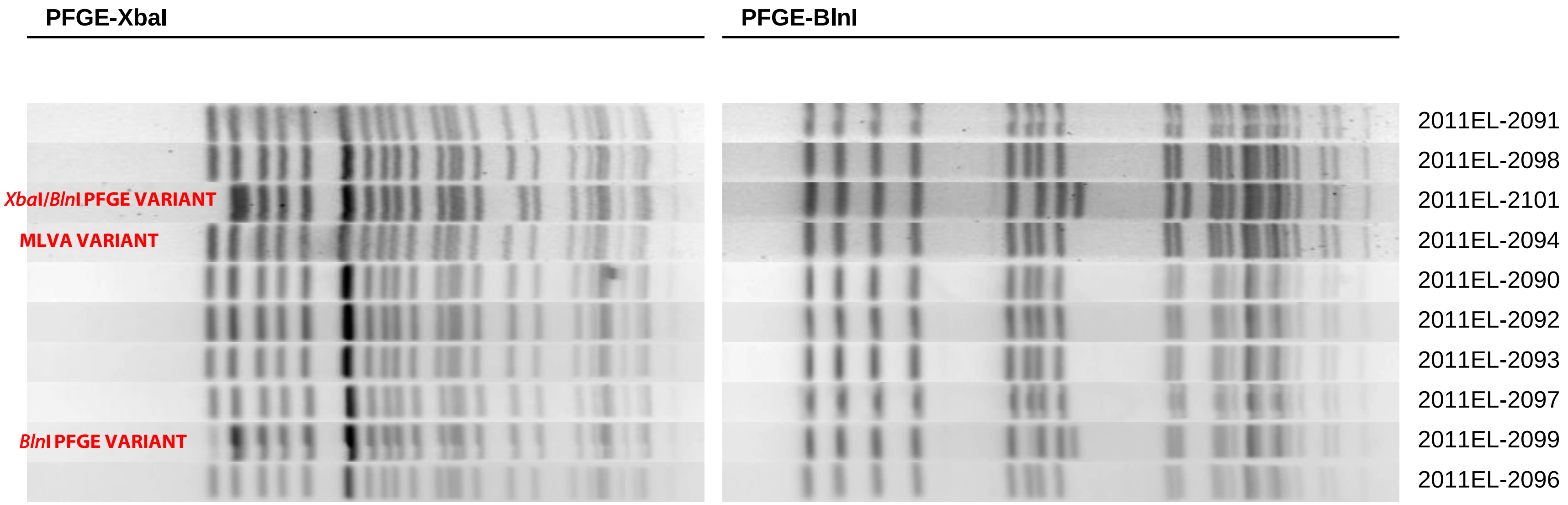
TTGATAGGGCAAAAGCGCCGATTTT

- When k-mer 25 is used, the 13th base position is used for determining if there is a variation

- No prior sequence assembly or multiple sequence alignment required
- Fast completion of analysis
- Phylogenies less affected by regions that have undergone strong selection, deletions or horizontal gene transfer
- Dependent on high quality data
- Reference genome required

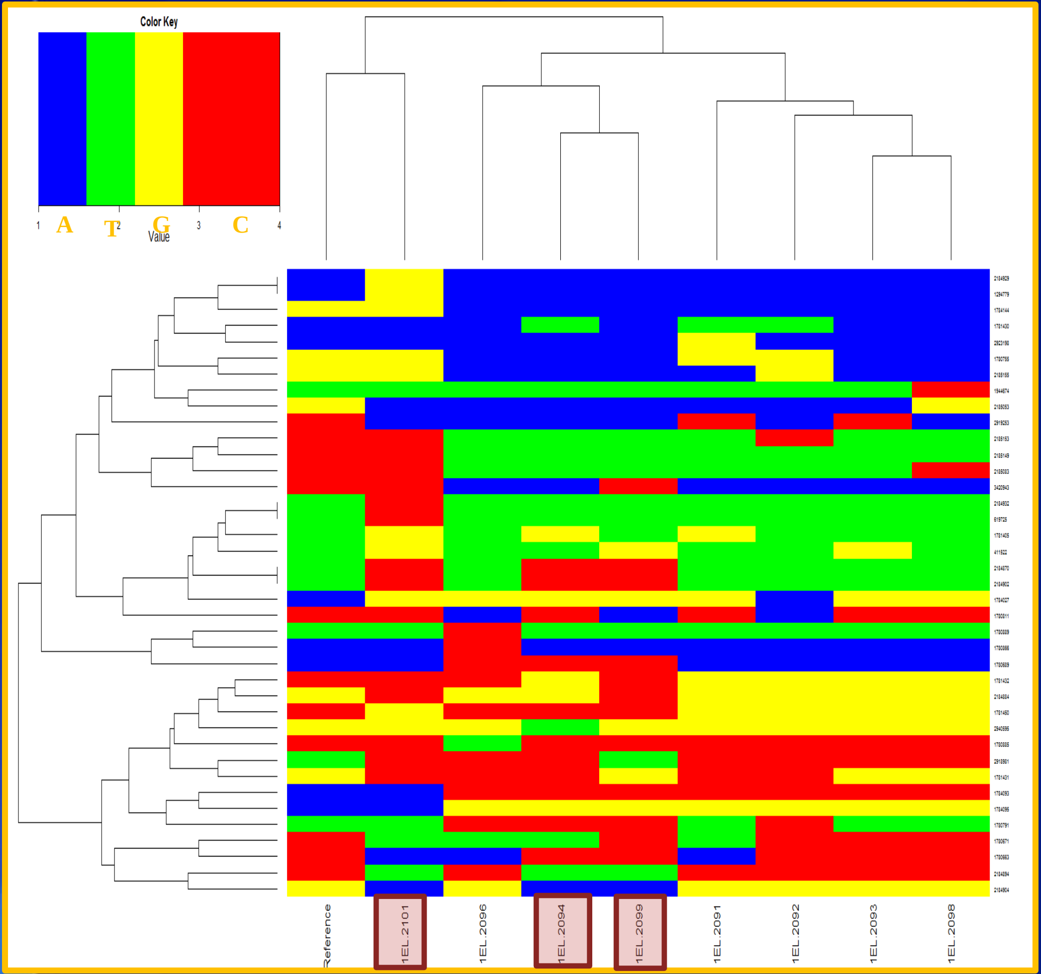
RESULTS

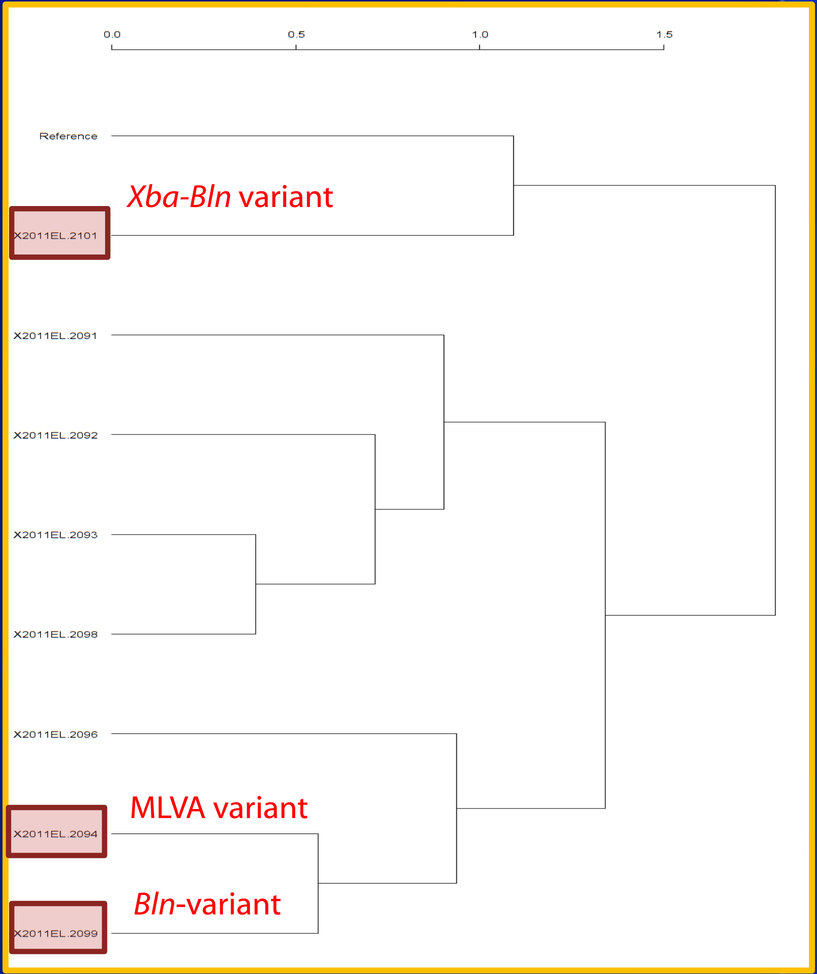
Probing Diversity within an Outbreak – O157:H7 in MN Daycare (2008)



- **Relatively clonal short lasting point source outbreak**

Probing Diversity within a Clonal Point Source O157 Outbreak – MN Daycare

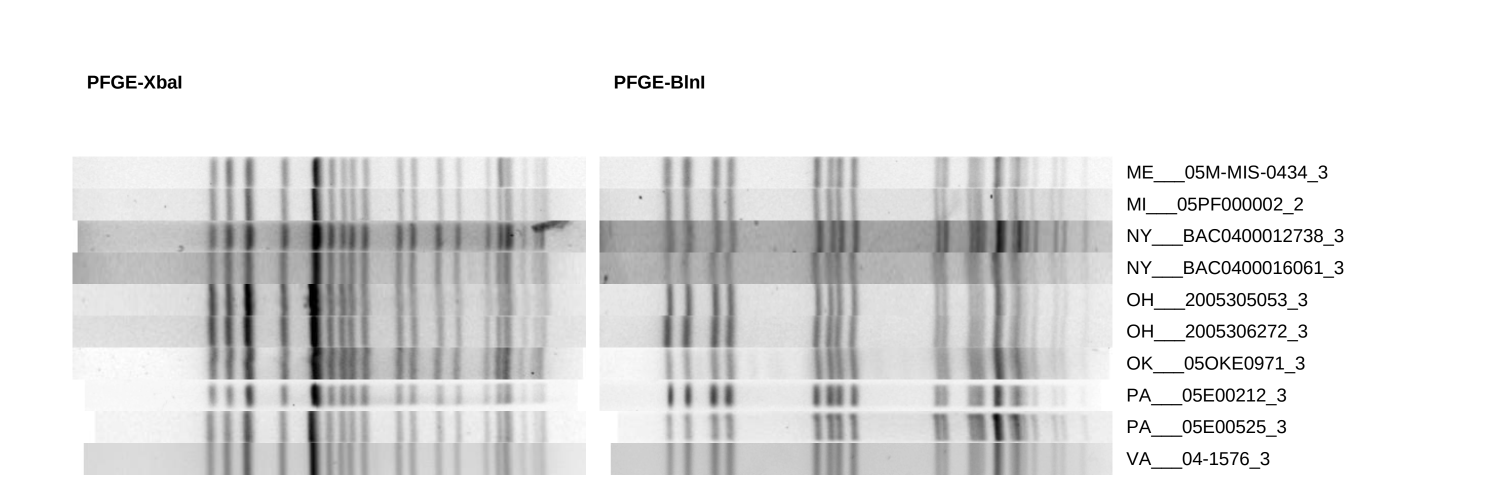




Reference Sample: 2011EL-2090_O157:H7_EXHX01.0589_EXHA26.1376_main MLVA

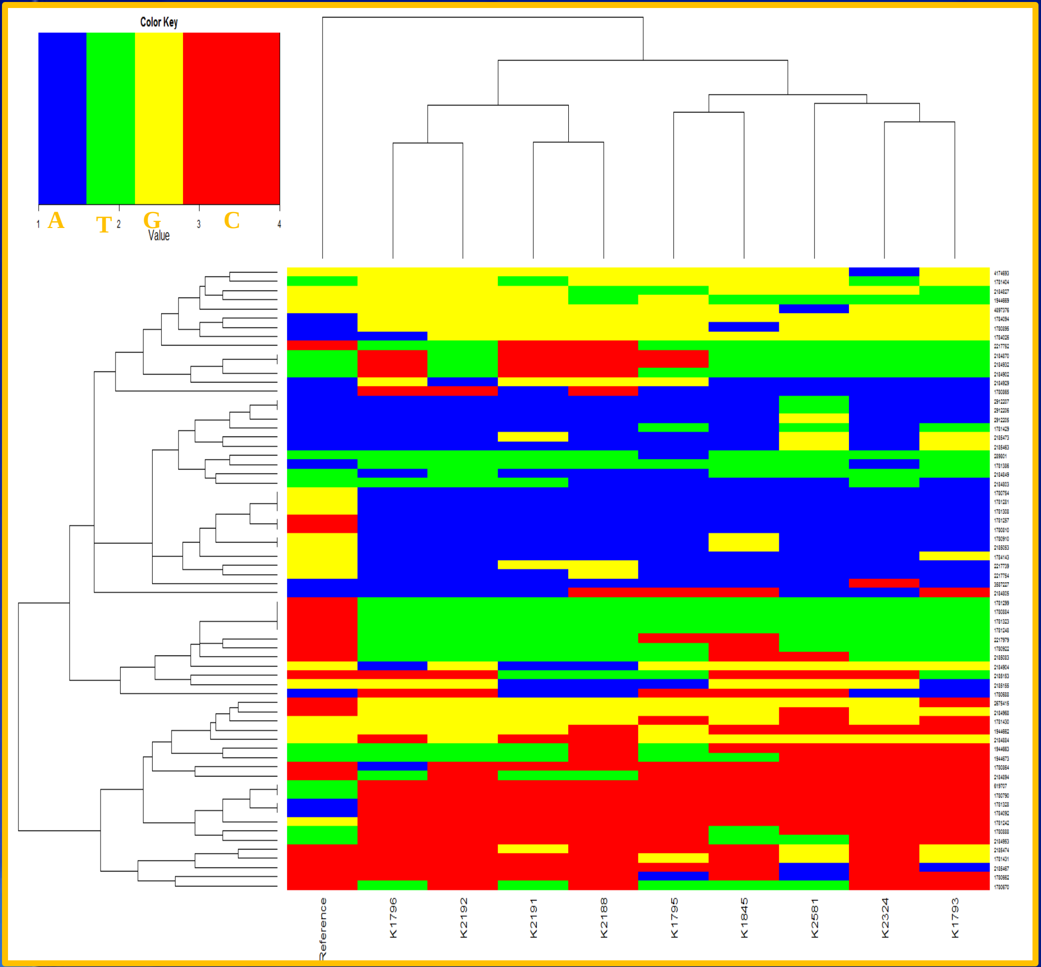
- The total number of SNPs found across all the samples from this outbreak at 100% frequency is 39.

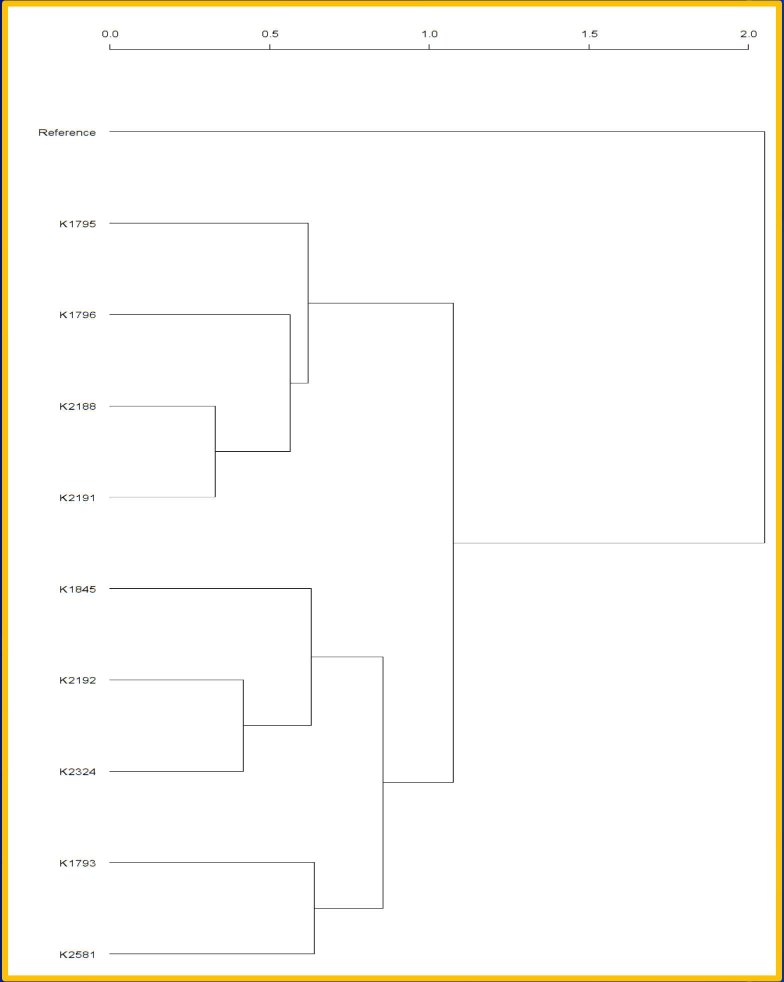
Probing Diversity within Clonal Long Lasting O157:H7 Outbreak (2004-2005)



- 110 2-enzyme matches in 20 states within 15 months
- Rare PFGE type
- An indistinguishable new MLVA pattern
- Vehicle never identified

Probing Diversity within a Clonal Long Lasting O157 Outbreak – Not resolved

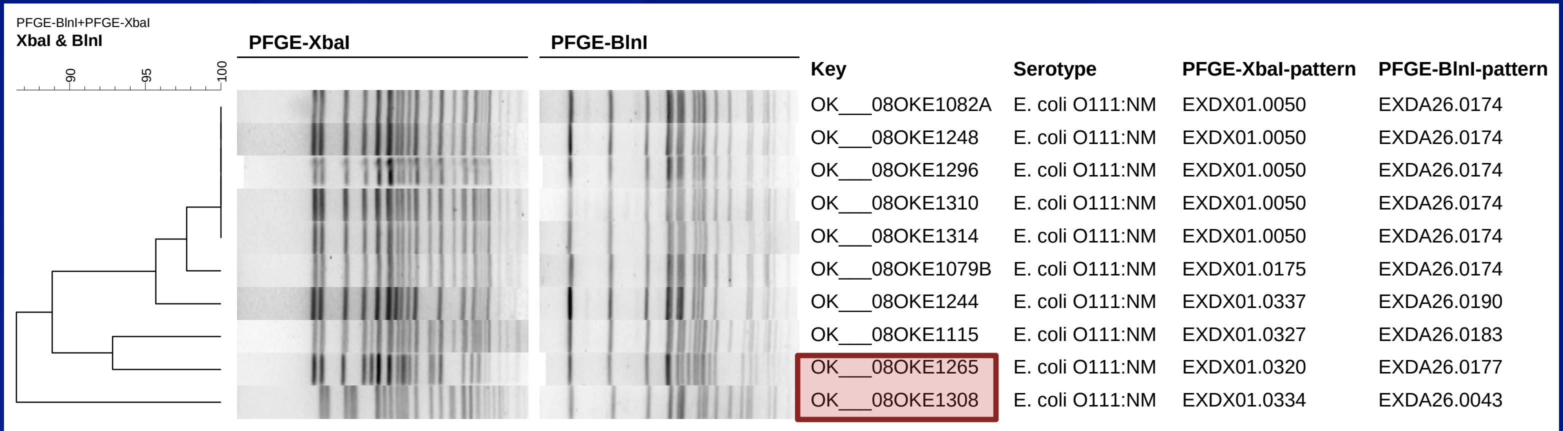




Reference Sample: **K1792_O157:H7_EXHX01.0086_EXHA26.0576**

- The total number of SNPs found across all the samples from this outbreak at 100% frequency is **68**

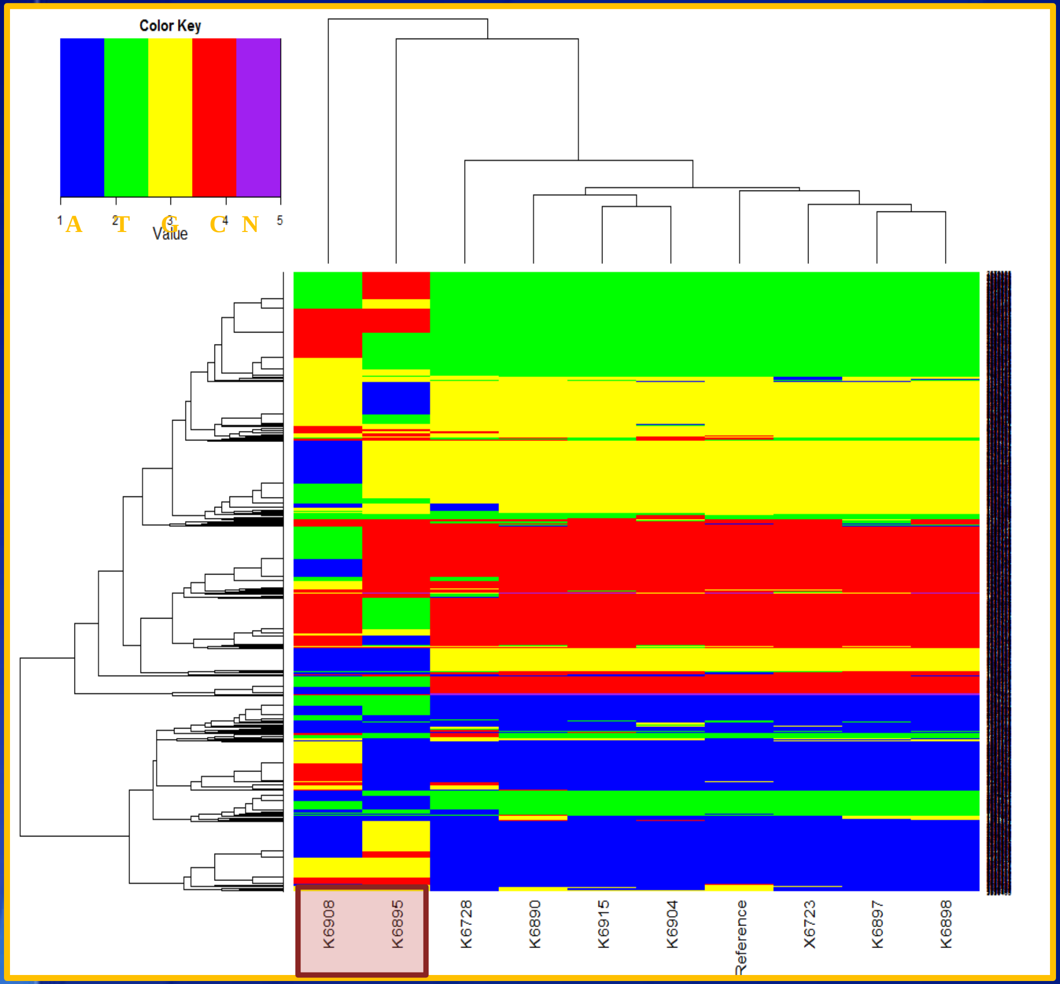
Probing Diversity within Polyclonal Outbreak – O111:NM in OK Restaurant (2008)

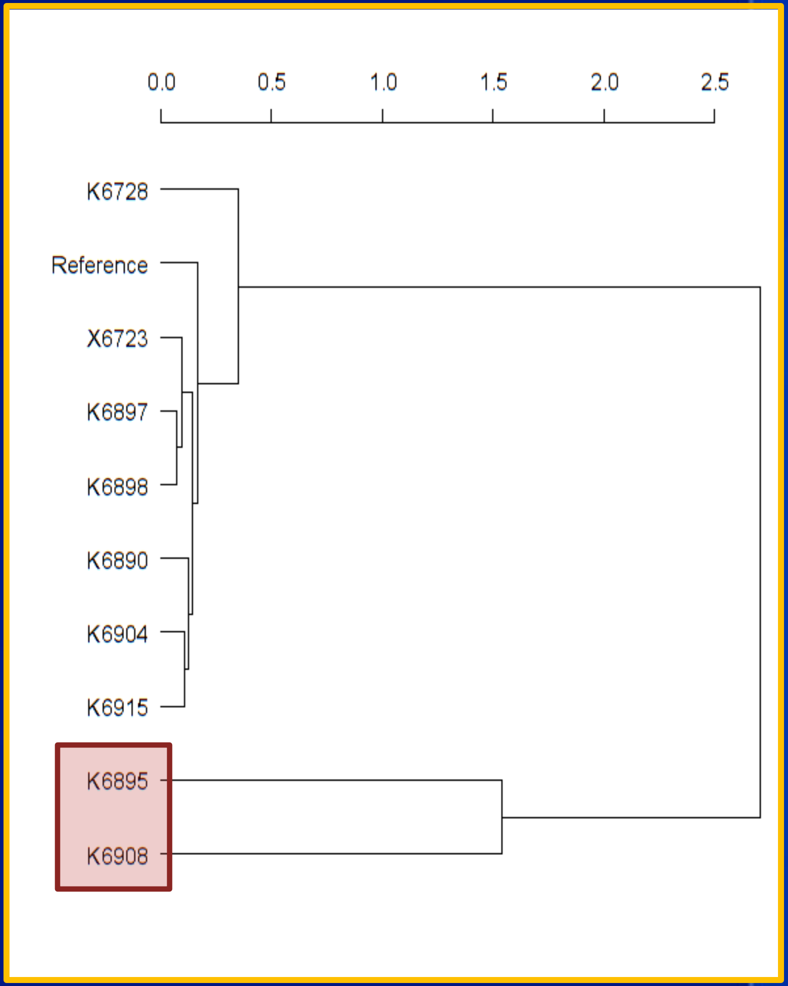


Key	Serotype	PFGE-XbaI-pattern	PFGE-BlnI-pattern
OK___08OKE1082A	E. coli O111:NM	EXDX01.0050	EXDA26.0174
OK___08OKE1248	E. coli O111:NM	EXDX01.0050	EXDA26.0174
OK___08OKE1296	E. coli O111:NM	EXDX01.0050	EXDA26.0174
OK___08OKE1310	E. coli O111:NM	EXDX01.0050	EXDA26.0174
OK___08OKE1314	E. coli O111:NM	EXDX01.0050	EXDA26.0174
OK___08OKE1079B	E. coli O111:NM	EXDX01.0175	EXDA26.0174
OK___08OKE1244	E. coli O111:NM	EXDX01.0337	EXDA26.0190
OK___08OKE1115	E. coli O111:NM	EXDX01.0327	EXDA26.0183
OK___08OKE1265	E. coli O111:NM	EXDX01.0320	EXDA26.0177
OK___08OKE1308	E. coli O111:NM	EXDX01.0334	EXDA26.0043

- 1 *Xba* variant, 4 *Xba-Bln* variants
- Vehicle not identified

Probing Diversity within Polyclonal O111 Outbreak – OK Restaurant

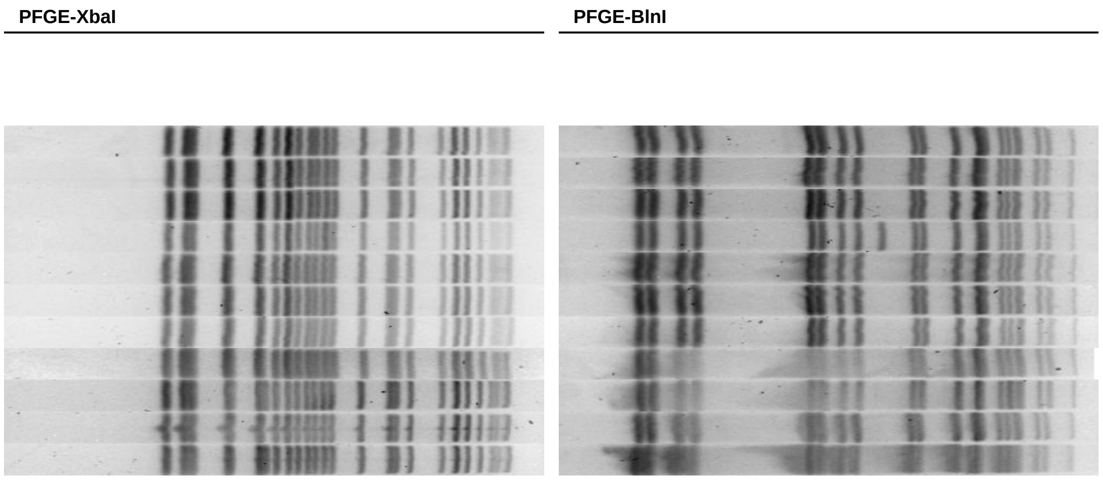




<u>Reference Sample</u>: K6722_O111:IsolatetoCDC_EXDX01.0050_EXDA26.0174

- The total number of SNPs found across all the samples from this outbreak at 100% frequency is 947

Probing Diversity During a Carrier State



LabIDcdc_id

CD. 2011EL-2103

CD. 2011EL-2104, 2010C-4979C1

CD. 2011EL-2105

CD. 2011EL-2106 *Bln* variant

CD. 2011EL-2107

CD. 2011EL-2108

CD. 2011EL-2109

CD. 2011EL-2111

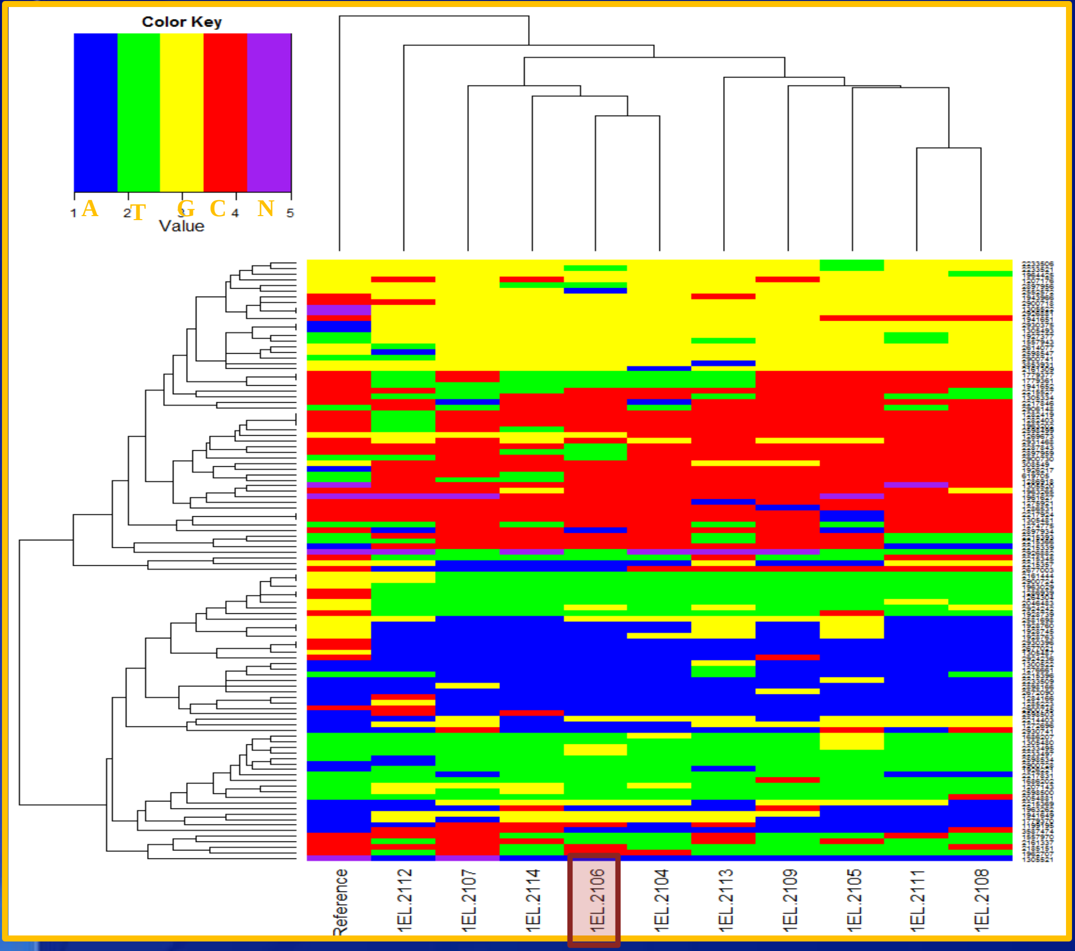
CD. 2011EL-2112

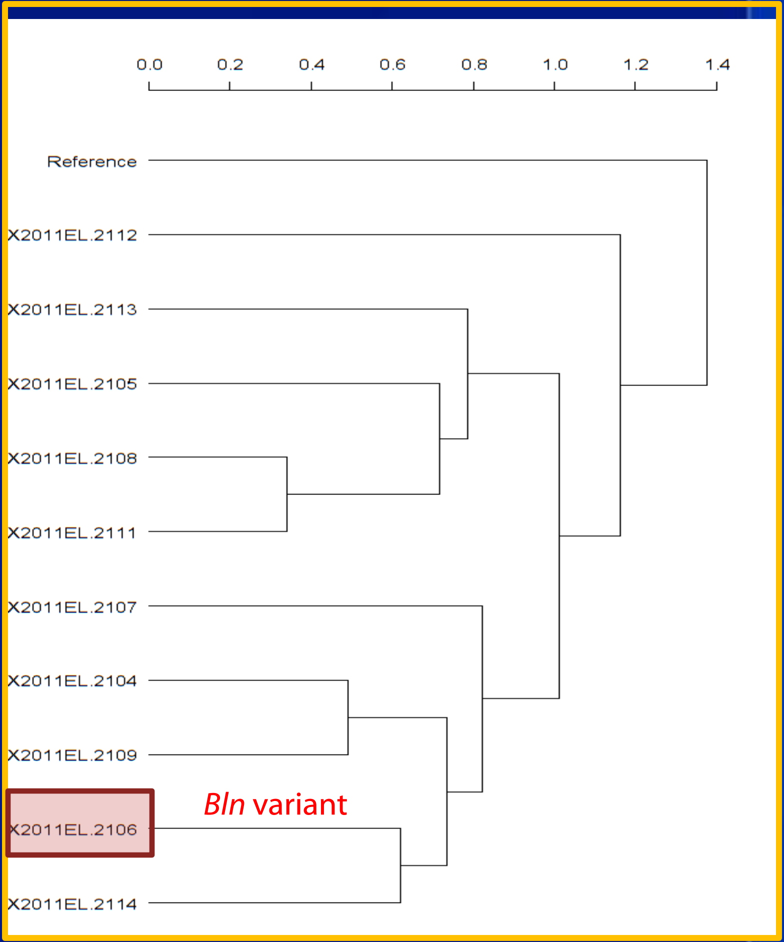
CD. 2011EL-2113

CD. 2011EL-2114

- **Collected from the same person within a period of 2½ months**
- **An indistinguishable MLVA pattern**

Probing Diversity during a Carrier State –Serial Isolates from the Same Person



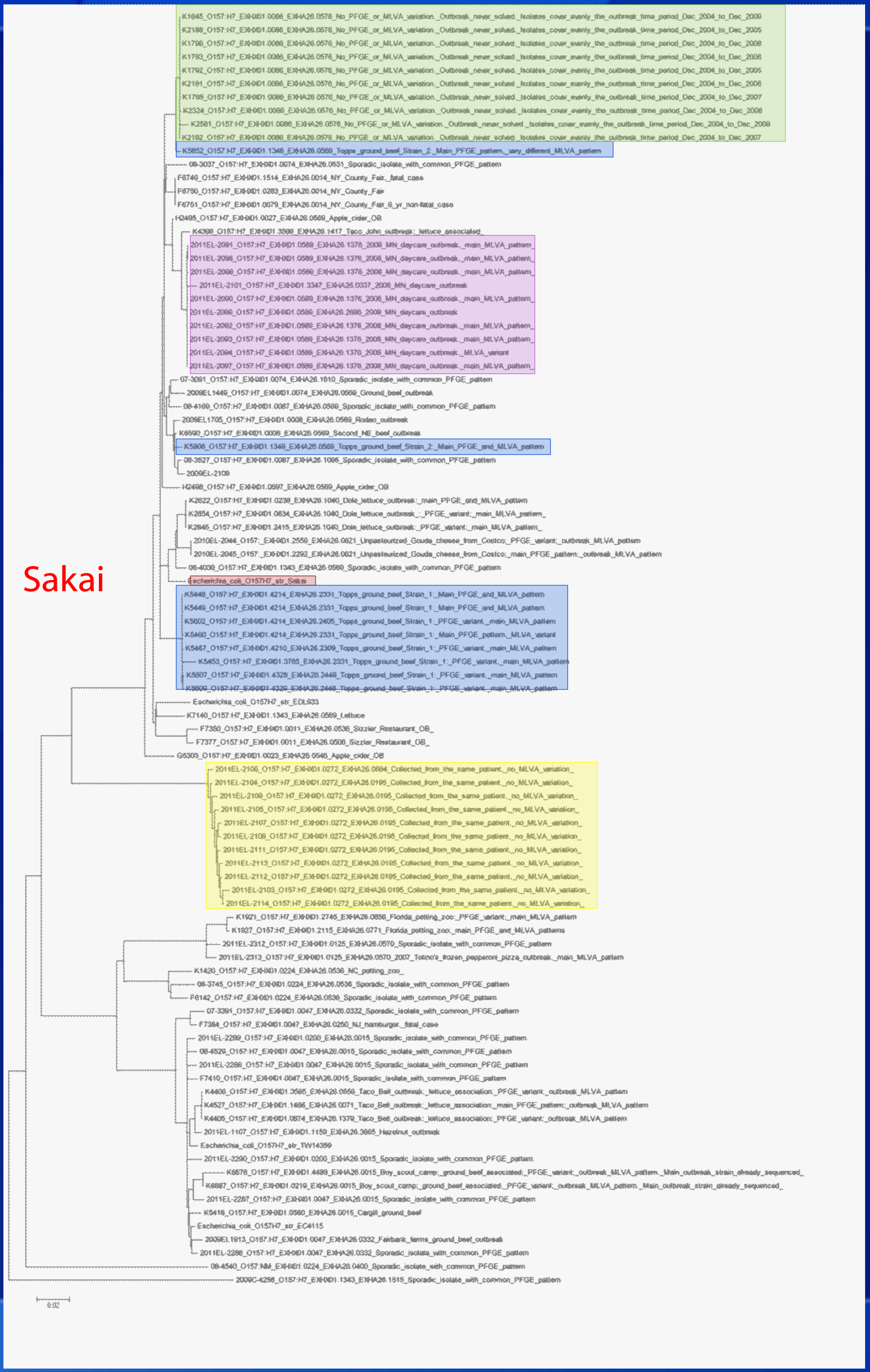


Reference Sample: 2011EL-2103_O157:H7_EXHX01.0272_EXHA26.0195

- The total number of SNPs found across all the samples from this “outbreak “at 100% frequency is 108

K-mer –based Phylogenetic Tree for STEC O157

93 to 2101 SNPs compared to the reference


K1645_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2009
K2188_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2005
K1798_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2008
K1793_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2008
K1792_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2005
K2181_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2008
K1795_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2007
K2334_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2008
K2581_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2009
K2182_O157:H7_EXHX01.0086_EXHA26.0578_No_PFGE_or_MLVA_variation._Outbreak_never_solved._Isolates_cover_evenly_the_outbreak_time_period_Dec_2004_to_Dec_2007
K5852_O157:H7_EXHX01.1348_EXHA26.0569_Topps_ground_beef_Strain_2:_Main_PFGE_pattern,_very_different_MLVA_pattern
08-3037_O157:H7_EXHX01.0074_EXHA26.0531_Sporadic_isolate_with_common_PFGE_pattern
F6749_O157:H7_EXHX01.1514_EXHA26.0014_NY_County_Fair,_fatal_case
F6750_O157:H7_EXHX01.0283_EXHA26.0014_NY_County_Fair
F6751_O157:H7_EXHX01.0079_EXHA26.0014_NY_County_Fair_6_yr_non-fatal_case
H2495_O157:H7_EXHX01.0027_EXHA26.0569_Apple_cider_OB
K4396_O157:H7_EXHX01.3596_EXHA26.1417_Taco_John_outbreak:_lettuce_associated_
2011EL-2091_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
2011EL-2096_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
2011EL-2098_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
2011EL-2101_O157:H7_EXHX01.3347_EXHA26.0337_2008_MN_daycare_outbreak
2011EL-2090_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
2011EL-2099_O157:H7_EXHX01.0569_EXHA26.2698_2008_MN_daycare_outbreak
2011EL-2092_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
2011EL-2093_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
2011EL-2094_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_MLVA_variant
2011EL-2097_O157:H7_EXHX01.0569_EXHA26.1378_2008_MN_daycare_outbreak,_main_MLVA_pattern_
07-3091_O157:H7_EXHX01.0074_EXHA26.1610_Sporadic_isolate_with_common_PFGE_pattern
2009EL1449_O157:H7_EXHX01.0074_EXHA26.0569_Ground_beef_outbreak
08-4169_O157:H7_EXHX01.0087_EXHA26.0569_Sporadic_isolate_with_common_PFGE_pattern
2009EL1705_O157:H7_EXHX01.0008_EXHA26.0569_Rodeo_outbreak
K6590_O157:H7_EXHX01.0008_EXHA26.0569_Second_NE_beef_outbreak
K5806_O157:H7_EXHX01.1348_EXHA26.0569_Topps_ground_beef_Strain_2:_Main_PFGE_and_MLVA_pattern
08-3527_O157:H7_EXHX01.0087_EXHA26.1095_Sporadic_isolate_with_common_PFGE_pattern
2009EL-2109
H2498_O157:H7_EXHX01.0697_EXHA26.0569_Apple_cider_OB
K2622_O157:H7_EXHX01.0238_EXHA26.1040_Dole_lettuce_outbreak:_main_PFGE_and_MLVA_pattern
K2854_O157:H7_EXHX01.0634_EXHA26.1040_Dole_lettuce_outbreak_:_PFGE_variant:_main_MLVA_pattern_
K2845_O157:H7_EXHX01.2415_EXHA26.1040_Dole_lettuce_outbreak:_PFGE_variant:_main_MLVA_pattern_
2010EL-2044_O157:_EXHX01.2559_EXHA26.0821_Unpasteurized_Gouda_cheese_from_Costco:_PFGE_variant:_outbreak_MLVA_pattern
2010EL-2045_O157:_EXHX01.2292_EXHA26.0821_Unpasteurized_Gouda_cheese_from_Costco:_main_PFGE_pattern:_outbreak_MLVA_pattern
08-4039_O157:H7_EXHX01.1343_EXHA26.0569_Sporadic_isolate_with_common_PFGE_pattern
Escherichia_coli_O157H7_str_Sakai
Sakai
K5448_O157:H7_EXHX01.4214_EXHA26.2331_Topps_ground_beef_Strain_1:_Main_PFGE_and_MLVA_pattern
K5449_O157:H7_EXHX01.4214_EXHA26.2331_Topps_ground_beef_Strain_1:_Main_PFGE_and_MLVA_pattern
K5602_O157:H7_EXHX01.4214_EXHA26.2405_Topps_ground_beef_Strain_1:_PFGE_variant,_main_MLVA_pattern
K5460_O157:H7_EXHX01.4214_EXHA26.2331_Topps_ground_beef_Strain_1:_Main_PFGE_pattern,_MLVA_variant
K5467_O157:H7_EXHX01.4210_EXHA26.2309_Topps_ground_beef_Strain_1:_PFGE_variant,_main_MLVA_pattern
K5453_O157:H7_EXHX01.3765_EXHA26.2331_Topps_ground_beef_Strain_1:_PFGE_variant,_main_MLVA_pattern
K5807_O157:H7_EXHX01.4328_EXHA26.2448_Topps_ground_beef_Strain_1:_PFGE_variant,_main_MLVA_pattern
K5808_O157:H7_EXHX01.4328_EXHA26.2448_Topps_ground_beef_Strain_1:_PFGE_variant,_main_MLVA_pattern
Escherichia_coli_O157H7_str_EDL933
K7140_O157:H7_EXHX01.1343_EXHA26.0569_Lettuce
F7350_O157:H7_EXHX01.0011_EXHA26.0536_Sizzler_Restaurant_OB_
F7377_O157:H7_EXHX01.0011_EXHA26.0506_Sizzler_Restaurant_OB_
G5303_O157:H7_EXHX01.0023_EXHA26.0546_Apple_cider_OB
2011EL-2106_O157:H7_EXHX01.0272_EXHA26.0694_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2104_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2109_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2105_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2107_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2108_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2111_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2113_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2112_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2103_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
2011EL-2114_O157:H7_EXHX01.0272_EXHA26.0195_Collected_from_the_same_patient,_no_MLVA_variation_
K1921_O157:H7_EXHX01.2745_EXHA26.0856_Florida_petting_zoo:_PFGE_variant:_main_MLVA_pattern
K1927_O157:H7_EXHX01.2115_EXHA26.0771_Florida_petting_zoo:_main_PFGE_and_MLVA_patterns
2011EL-2312_O157:H7_EXHX01.0125_EXHA26.0570_Sporadic_isolate_with_common_PFGE_pattern
2011EL-2313_O157:H7_EXHX01.0125_EXHA26.0570_2007_Totino's_frozen_pepperoni_pizza_outbreak,_main_MLVA_pattern
K1420_O157:H7_EXHX01.0224_EXHA26.0536_NC_petting_zoo_
08-3745_O157:H7_EXHX01.0224_EXHA26.0536_Sporadic_isolate_with_common_PFGE_pattern
F6142_O157:H7_EXHX01.0224_EXHA26.0536_Sporadic_isolate_with_common_PFGE_pattern
07-3391_O157:H7_EXHX01.0047_EXHA26.0332_Sporadic_isolate_with_common_PFGE_pattern
F7384_O157:H7_EXHX01.0047_EXHA26.0250_NJ_hamburger,_fatal_case
2011EL-2289_O157:H7_EXHX01.0200_EXHA26.0015_Sporadic_isolate_with_common_PFGE_pattern
08-4529_O157:H7_EXHX01.0047_EXHA26.0015_Sporadic_isolate_with_common_PFGE_pattern
2011EL-2288_O157:H7_EXHX01.0047_EXHA26.0015_Sporadic_isolate_with_common_PFGE_pattern
F7410_O157:H7_EXHX01.0047_EXHA26.0015_Sporadic_isolate_with_common_PFGE_pattern
K4408_O157:H7_EXHX01.3585_EXHA26.0858_Taco_Bell_outbreak:_lettuce_association:_PFGE_variant:_outbreak_MLVA_pattern
K4527_O157:H7_EXHX01.1486_EXHA26.0071_Taco_Bell_outbreak:_lettuce_association:_main_PFGE_pattern:_outbreak_MLVA_pattern
K4405_O157:H7_EXHX01.0874_EXHA26.1379_Taco_Bell_outbreak:_lettuce_association:_PFGE_variant:_outbreak_MLVA_pattern
2011EL-1107_O157:H7_EXHX01.1159_EXHA26.3665_Hazelnut_outbreak
Escherichia_coli_O157H7_str_TW14359
2011EL-2290_O157:H7_EXHX01.0200_EXHA26.0015_Sporadic_isolate_with_common_PFGE_pattern
K6676_O157:H7_EXHX01.4489_EXHA26.0015_Boy_scout_camp:_ground_beef_associated:_PFGE_variant:_outbreak_MLVA_pattern._Main_outbreak_strain_already_sequenced_
K6687_O157:H7_EXHX01.0219_EXHA26.0015_Boy_scout_camp:_ground_beef_associated:_PFGE_variant:_outbreak_MLVA_pattern._Main_outbreak_strain_already_sequenced_
2011EL-2287_O157:H7_EXHX01.0047_EXHA26.0015_Sporadic_isolate_with_common_PFGE_pattern
K5416_O157:H7_EXHX01.0580_EXHA26.0015_Cargill_ground_beef
Escherichia_coli_O157H7_str_EC4115
2009EL1913_O157:H7_EXHX01.0047_EXHA26.0332_Fairbank_farms_ground_beef_outbreak
2011EL-2286_O157:H7_EXHX01.0047_EXHA26.0332_Sporadic_isolate_with_common_PFGE_pattern
08-4540_O157:NM_EXHX01.0224_EXHA26.0400_Sporadic_isolate_with_common_PFGE_pattern
2009C-4256_O157:H7_EXHX01.1343_EXHA26.1515_Sporadic_isolate_with_common_PFGE_pattern
0.02

Outbreak 1 (Not solved, 2004-05)

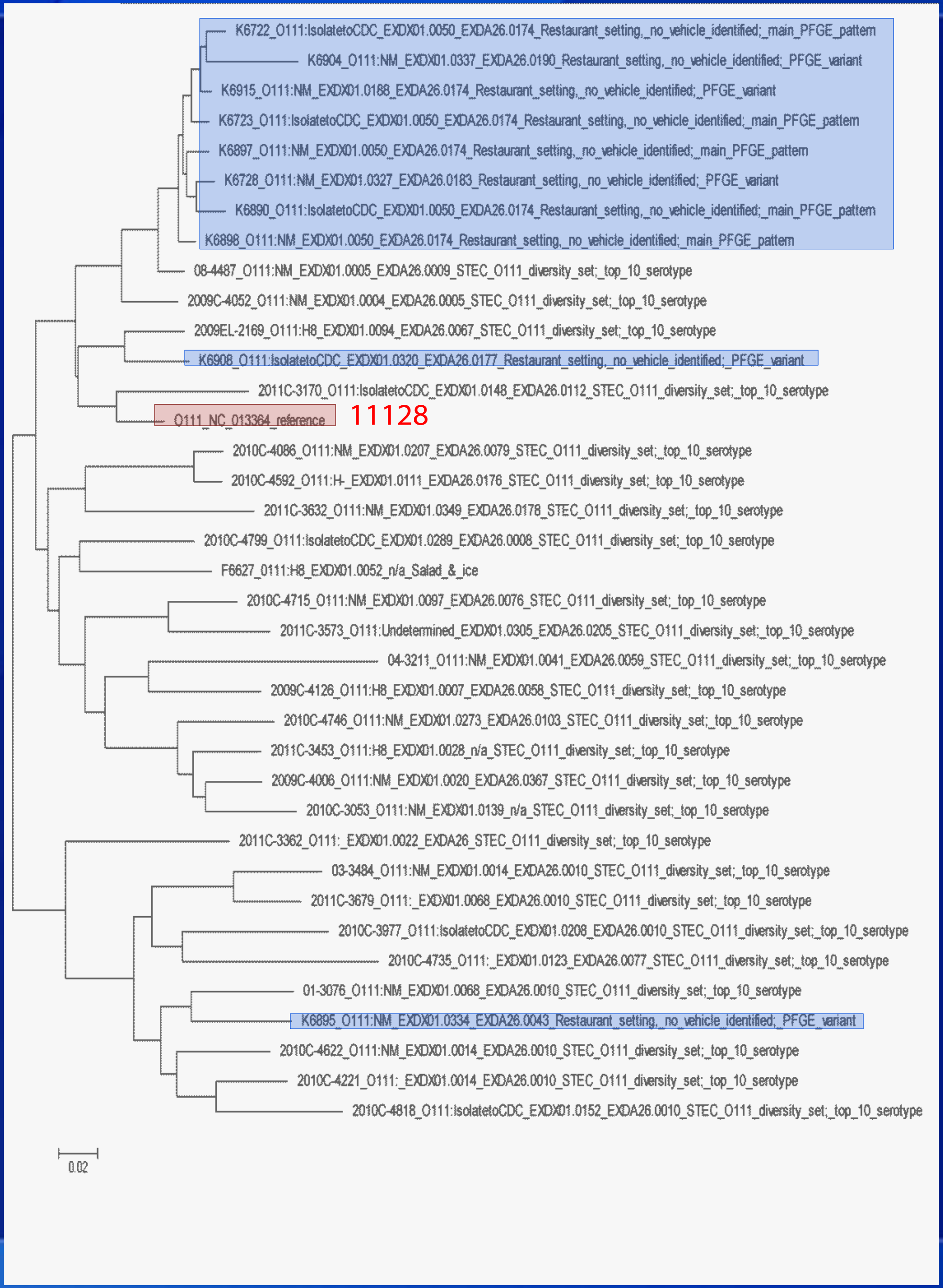
Outbreak 2 (Daycare, 2008)

Outbreak 3 (Topps ground beef, 2007)

Long term carrier patient

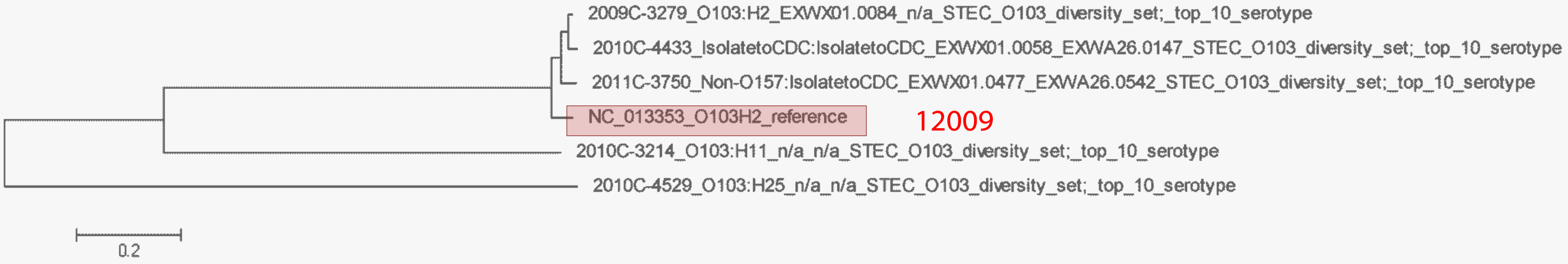
K-mer -based Phylogenetic Tree for STEC O111

195 to 1229 SNPs compared to reference


K6722_O111:IsolatetoCDC_EXDX01.0050_EXDA26.0174_Restaurant_setting,_no_vehicle_identified;_main_PFGE_pattern
K6904_O111:NM_EXDX01.0337_EXDA26.0190_Restaurant_setting,_no_vehicle_identified;_PFGE_variant
K6915_O111:NM_EXDX01.0188_EXDA26.0174_Restaurant_setting,_no_vehicle_identified;_PFGE_variant
K6723_O111:IsolatetoCDC_EXDX01.0050_EXDA26.0174_Restaurant_setting,_no_vehicle_identified;_main_PFGE_pattern
K6897_O111:NM_EXDX01.0050_EXDA26.0174_Restaurant_setting,_no_vehicle_identified;_main_PFGE_pattern
K6728_O111:NM_EXDX01.0327_EXDA26.0183_Restaurant_setting,_no_vehicle_identified;_PFGE_variant
K6890_O111:IsolatetoCDC_EXDX01.0050_EXDA26.0174_Restaurant_setting,_no_vehicle_identified;_main_PFGE_pattern
K6898_O111:NM_EXDX01.0050_EXDA26.0174_Restaurant_setting,_no_vehicle_identified;_main_PFGE_pattern
08-4487_O111:NM_EXDX01.0005_EXDA26.0009_STEC_O111_diversity_set;_top_10_serotype
2009C-4052_O111:NM_EXDX01.0004_EXDA26.0005_STEC_O111_diversity_set;_top_10_serotype
2009EL-2169_O111:H8_EXDX01.0094_EXDA26.0067_STEC_O111_diversity_set;_top_10_serotype
K6908_O111:IsolatetoCDC_EXDX01.0320_EXDA26.0177_Restaurant_setting,_no_vehicle_identified;_PFGE_variant
2011C-3170_O111:IsolatetoCDC_EXDX01.0148_EXDA26.0112_STEC_O111_diversity_set;_top_10_serotype
O111_NC_013364_reference
11128
2010C-4086_O111:NM_EXDX01.0207_EXDA26.0079_STEC_O111_diversity_set;_top_10_serotype
2010C-4592_O111:H-_EXDX01.0111_EXDA26.0176_STEC_O111_diversity_set;_top_10_serotype
2011C-3632_O111:NM_EXDX01.0349_EXDA26.0178_STEC_O111_diversity_set;_top_10_serotype
2010C-4799_O111:IsolatetoCDC_EXDX01.0289_EXDA26.0008_STEC_O111_diversity_set;_top_10_serotype
F6627_O111:H8_EXDX01.0052_n/a_Salad_&_ice
2010C-4715_O111:NM_EXDX01.0097_EXDA26.0076_STEC_O111_diversity_set;_top_10_serotype
2011C-3573_O111:Undetermined_EXDX01.0305_EXDA26.0205_STEC_O111_diversity_set;_top_10_serotype
04-3211_O111:NM_EXDX01.0041_EXDA26.0059_STEC_O111_diversity_set;_top_10_serotype
2009C-4126_O111:H8_EXDX01.0007_EXDA26.0058_STEC_O111_diversity_set;_top_10_serotype
2010C-4746_O111:NM_EXDX01.0273_EXDA26.0103_STEC_O111_diversity_set;_top_10_serotype
2011C-3453_O111:H8_EXDX01.0028_n/a_STEC_O111_diversity_set;_top_10_serotype
2009C-4006_O111:NM_EXDX01.0020_EXDA26.0367_STEC_O111_diversity_set;_top_10_serotype
2010C-3053_O111:NM_EXDX01.0139_n/a_STEC_O111_diversity_set;_top_10_serotype
2011C-3362_O111:_EXDX01.0022_EXDA26_STEC_O111_diversity_set;_top_10_serotype
03-3484_O111:NM_EXDX01.0014_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
2011C-3679_O111:_EXDX01.0068_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
2010C-3977_O111:IsolatetoCDC_EXDX01.0208_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
2010C-4735_O111:_EXDX01.0123_EXDA26.0077_STEC_O111_diversity_set;_top_10_serotype
01-3076_O111:NM_EXDX01.0068_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
K6895_O111:NM_EXDX01.0334_EXDA26.0043_Restaurant_setting,_no_vehicle_identified;_PFGE_variant
2010C-4622_O111:NM_EXDX01.0014_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
2010C-4221_O111:_EXDX01.0014_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
2010C-4818_O111:IsolatetoCDC_EXDX01.0152_EXDA26.0010_STEC_O111_diversity_set;_top_10_serotype
0.02

Restaurant outbreak, 2008

K-mer -based Phylogenetic Tree for STEC O103


2009C-3279_O103:H2_EXWX01.0084_n/a_STEC_O103_diversity_set;_top_10_serotype
2010C-4433_IsolatetoCDC:IsolatetoCDC_EXWX01.0058_EXWA26.0147_STEC_O103_diversity_set;_top_10_serotype
2011C-3750_Non-O157:IsolatetoCDC_EXWX01.0477_EXWA26.0542_STEC_O103_diversity_set;_top_10_serotype
NC_013353_O103H2_reference
12009
2010C-3214_O103:H11_n/a_n/a_STEC_O103_diversity_set;_top_10_serotype
2010C-4529_O103:H25_n/a_n/a_STEC_O103_diversity_set;_top_10_serotype
0.2

1789 to 20,094 SNPs
compared to the reference

Conclusions Based on the Illumina Data

- K-mer –based clustering appears to have good correlation with epidemiological and PFGE data
- Variation within a (relatively) clonal outbreak < 100 SNPs
- The variation between unrelated strains of the same serotype hundreds to ~ 2000 SNPs
- The variation between serotypes > 10,000 SNPs

Future Plans

- Improve the assemblies using PacBio sequences
  - Make the draft genomes publicly available
- Additional sequencing to close reference genomes
- Define a cross serotype gene set to be used to determine strain subtype, virulence and antimicrobial resistance profiles (‘super’ MLST)
  - Presence / absence
  - SNPs
- Add prospective strains from outbreaks to the database

Questions to be Answered

- **Which clustering method will work for surveillance?**
  1. K-mer –based clustering as a primary tool with gene-based ‘super’ MLST as a secondary tool OR
  2. Gene-based ‘super’ MLST as a primary tool
- **How to standardize data across the platforms?**
  - Different error profiles
- **How to get an actionable report out from raw reads?**

100K Pathogen Genome Sequencing Project

- **Dilemma: the safety and security of the world food supply is hindered by lack of food-related genomes**
- **Partnership between BGI and UC Davis**
  - http://100kgenome.vetmed.ucdavis.edu
- **Goal: sequence 100,000 foodborne pathogen isolates within 5 years**
  - Sequences will be made publicly available
- **Steering committee**
  - BGI
  - UC Davis School of Veterinary Medicine
  - FDA
  - CDC
  - Agilent Technologies, Inc.
  - Mars, Inc.

100K Pathogen Genome Sequencing Project (cont'd)

Isolate priority list

Tier 1	Tier 2	Tier 3
Salmonella	*Yersinia*	Toxigenic bacilli
Campylobacter	*Shigella*	Norovirus
E. coli	*Clostridium*	Hepatitis
Vibrio *Listeria*	*Enterococcus* *Cronobacter*	Rotovirus

- Seeking isolates from diverse types of food, distinct environments, diverse regions of the world and longitudinal isolate collections

Acknowledgements

- CDC/EDLB
  - Peter Gerner-Schmidt
  - John Besser
  - Efrain Ribot
  - Nancy Strockbine
  - Cheryl Tarr
  - Lee Katz
  - Amber Schmidtke
  - Ashley Sabol
  - Haley Martin
  - Devon Stripling
- MN Dept. of Health
  - Dave Boxrud
- CDC Bioinformatics Core Group
  - Duncan MacCannell
  - Ryan Weil
  - Shankar Changayil
  - Satishkumar Ranganathan
  - Kun Zhao
- Expression Analysis
  - Stephen Sifed
  - Victor Weigman
  - Steve Mcphail

Thank You for Your Attention!
Questions?

For more information please contact Centers for Disease Control and Prevention

1600 Clifton Road NE, Atlanta, GA 30333
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
E-mail: cdcinfo@cdc.gov   Web: http://www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.