Next Generation Sequencing of the *Escherichia coli* O55:H7 Genome and Comparison with the Closely Related Enterohemorrhagic *Escherichia coli* O157:H7

Paolo Vatta[1*], Craig Cummings[2*], Lily Wong[1], Melissa Barker[2], Janet Ziegle[2], Lee Jones[2], Jason Chin[2], Pius Brzoska[2], Michael Rhodes[2], Manohar Furtado[1], and Olga Petrauskene[1]

[1]Applied Markets R&D, and [2]Molecular & Cellular Biology Division, Applied Biosystems, Foster City, CA, USA
*These authors contributed equally

Introduction

Detection in the food supply of pathogenic *E. coli*, particularly strains that cause hemorrhagic colitis (HC), has become a public health priority. The O157:H7 serotype of *E. coli* has been responsible for most HC outbreaks to date, so detection of this type is critically important. The ideal assay must detect O157:H7, but not any other serotypes, including the vast majority of commensal *E. coli* that are not pathogenic. *E. coli* O157:H7 is very closely related to the O55:H7 serotype, which does not frequently cause HC outbreaks (Figure 1). Numerous lines of evidence indicate that O55:H7 is the nearest phylogenetic neighbor of O157:H7, making the design of O157:H7-specific assays challenging. The *E. coli* O55:H7 genome sequence would be a valuable tool for identification of assay target sequences unique to O157:H7, but no such sequence was available. To this end, the genome of one *E. coli* O55:H7 strain, and another O157:H7 strain, were sequenced by oligonucleotide ligation and detection using the next-generation AB SOLiD™ platform. Comparison of the O55:H7 and O157:H7 genomes identified 500 kb of sequence that is present on the O157:H7 chromosome and absent or divergent in O55:H7. Comparison of these putative O157:H7-specific sequences against the publicly available genome sequences of other pathogenic and non-pathogenic *E. coli* and *Shigella* strains identified regions that are conserved beyond the O157:H7 lineage, further narrowing the list of putative assay design targets. The short time requirement (two to three weeks from library construction to sequence) and deep coverage obtained (>20X), makes the SOLiD™ system ideally suited for microbial genome sequencing when a closely related reference genome sequence is available. In particular, this method can be sufficiently robust to permit genome sequencing of a reference organism's nearest phylogenetic neighbors. Importantly, short-read mapping methods can define regions of difference between the query and reference genomic sequences, which is fundamental to the definition of specific target sequences for differential assay design.

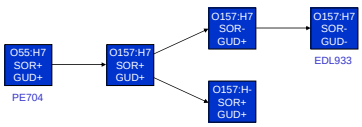


Figure 1. Model for evolution of *E. coli* O157:H7 from O55:H7 ancestor. Adapted from Wick, L. M., et al., (2005) *J Bacteriol* **187**: 1783-1791.

Scope

Design, through genomic comparisons that include "nearest neighbor" genome *E. coli* O55:H7, highly selective and specific TaqMan® assays for the rapid detection of *E. coli* O157:H7.

Experimental design

Step 1
All available *E. coli* genome sequences were compared to identify target regions specific to *E. coli* O157:H7.

Step 2
The genome of the "nearest neighbor", *E. coli* O55:H7, was sequenced with SOLiD technology.

Step 3
The *E. coli* O157:H7 (EDL933) genome seqeunce was used as a reference for assembly of the *E. coli* O55:H7 SOLiD sequence.

Step 4
Specific TaqMan® assays were designed in selected O157:H7-specific regions, and then tested for selectivity and specificity.

1 Comparison of O157:H7 to existing genome sequences

Selection criteria for O157:H7-specific sequences

• Present and at least 97% identical in 15/15 *E. coli* O157:H7 complete and draft genome sequences (by BLASTN search)

• Less than 80% identical to any other genome sequence, including 22 non-O157:H7 *E. coli* and 10 *Shigella* genomes

Figure 2. Multiple sequence alignment of a representative O157:H7-specific region and flanking sequences with *E. coli* and *Shigella* genome sequences. All BLASTN hits of at least 50 nt with more than 80% identity are depicted. O157:H7 strains are identical in this region. Some non-O157:H7 strains possess the conserved flanking sequence, but lack the O157:H7-specific region.

2 SOLiD™ sequencing of *E. coli* O55:H7

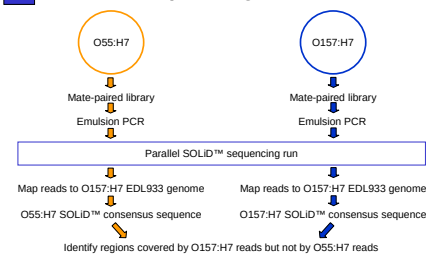


Figure 2. Schematic of SOLiD™ sequencing strategy. One O55:H7 strain and one O157:H7 strain were analyzed in parallel.

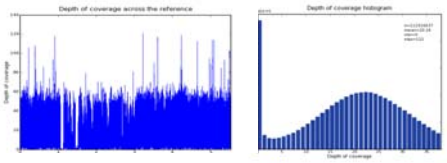


Figure 3. Mate-paired libraries (insert size 2.6 kb), were constructed for each of *E. coli* O157:H7 and *E. coli* O55:H7. Emulsion PCR was carried out using standard SOLiD™ protocols. Each library was sequenced on one eigth of a slide. The SOLiD™ System's flexible bead deposition formats allowed the sequencing of both *E. coli* genomes simultaneously on one SOLiD™ slide. Data analysis utilized the published sequence of *E. coli* O157:H7 (EDL933), as the reference sequence for alignment and assembly. Coverage of the reference sequence was 98% for the *E. coli* O157:H7 (PE30) sequence and 91% for the *E. coli* O55:H7 (PE704) genome (Table 1). Mean depth of mapped coverage for these sequences was 20.24 and 21.28, respectively.

Table 1. Summary of SOLiD run results for the *E. coli* O157:H7 and *E. coli* O55:H7 genomes.

	O55:H7 (PE704)	O157:H7 (PE30)
Mean depth mapped coverage	20.24	21.28
Percentage shared with O157:H7 EDL933 reference	91.0%	98.0%
Number of SNPs	15,259 (0.28%)	806 (0.015%)

3 Mapping and assembly of SOLiD™ reads

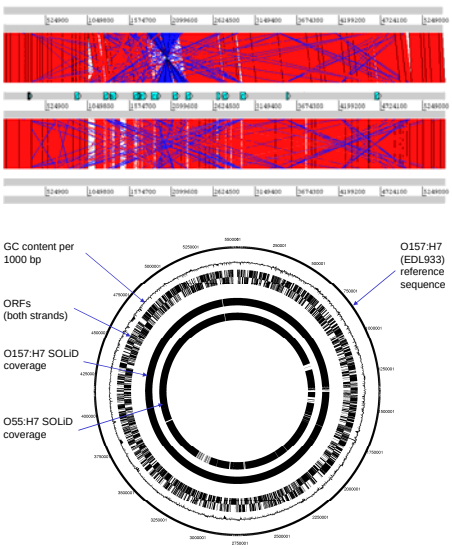
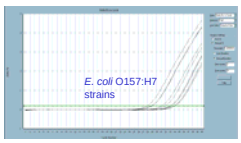


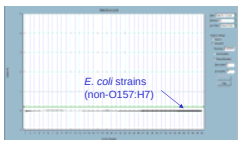
Figure 4. Top, linear map comparing SOLiD sequences of O157:H7 Sakai and O55:H7 to O157:H7 EDL933. Red and blue blocks between genomes indicate conservation on the same and opposite strands, respectively. Positions of prophage are indicated in cyan on the EDL933 track. Bottom, circular map comparing the SOLiD sequences for O157:H7 and O55:H7 to the reference sequence of *E. coli* O157:H7 (EDL933).

Analytical step	% reduction of putative O157:H7-specific regions
Alignment of 10 O157:H7 and 17 non-O157:H7 *E. coli*	-
⇩	
Stringent BLASTN validation against 23 non-O157:H7 *E. coli*, 10 *Shigella*	52
⇩	
Map SOLiD™ reads from O55:H7 and O157:H7 to Sakai genome	76

Figure 5. Schematic representation of the percent reduction in the number of putative target areas obtained by introducing the nearest neighbor sequence comparison. Alignment of various *E. coli* genomes identified putative O157:H7 specific target areas; a stringent BLASTN validation of these against other *E. coli* and *Shigella* genomes reduced the number of original targets by half. Further comparison with the newly obtained *E. coli* O55:H7 genome reduced this number by another 75%.

4 Design and validation of *E. coli* O157:H7-specific TaqMan® assays





Strain ID	*E. coli* serotype	Assay C_T	
PE30	O157:H7	33.24	+
PE677	O157:H7	28.98	+
PE678	O157:H7	28.91	+
PE679	O157:H7	31.13	+
PE700	O157:H7	33.06	+
PE701	O157:H7	31.95	+
PE702	O157:H7	34.00	+
PE703	O157:H7	34.04	+
PE40	O157:HNM	27.96	+
PE674	O55:H7	No signal	
PE704	O55:H7	No signal	
PE705	O55:H7	No signal	
PE706	O55:H7	No signal	
PE735	O55:H9	No signal	
PE684	O137:H41	No signal	
PE37	O145:NM	No signal	
PE826	O154:H25	No signal	
PE827	O156:H8	No signal	
PE731	O26:H32	No signal	
PE688	O28:H35	No signal	
PE687	O48:H21	No signal	
PE673	O5:NM	No signal	
PE38	O78:K80:H12	No signal	
Ground beef (20 samples)		No signal	

Figure 6. For each of the *E. coli* O157:H7-specific regions, multiple TaqMan® assays were designed. Each of these assays was screened by real-time PCR against an "inclusion set" of O157:H7 strains, and an "exclusion set" comprising strains of non-O157:H7 *E. coli, Shigella, Salmonella*, and a selection of other bacterial genera. Assays were also tested against ground beef samples to assess background signal in the presence of a complex food matrix. Assays were run on the AB 7500 Fast Sequence Detection System. Representative data from an acceptable assay are shown here. Amplification plots (left) show the detection of fluorescent signal in the presence of *E. coli* O157:H7 (top) or non-O157:H7 *E. coli* (bottom) genomic DNA. Each curve represents a different isolate. The assay is considered positive when the fluorescence curve crosses the threshold (green line) and the PCR cycle at which the curve crosses the threshold is C_T. These data are summarized in the table on the right. Experiments that do not yield a signal after 40 cycles of PCR are considered to be negative.

Conclusions

- A rapid and efficient methodology for the definition of highly specific and selective assays was developed.
- Nearest neighbor sequences are shown to be very important in the precise definition of highly specific and selective target regions in the genome of interest.
- The SOLiD™ next generation sequencing technology was proven to be an enabling platform for the generation of "nearest neighbor" sequences.
- The *E. coli* O55:H7 genomic sequence obtained with the SOLiD™ platform allowed the identification of several highly specific and selective target regions in the O157:H7 genome
- E. coli O157:H7 TaqMan® assays have been designed and proven to be highly specific and selective for this serotype