Whole Genome Sequencing of E. coli Bacteria

Introduction

Whole genome analysis of bacteria is fundamental in human diagnostics, food testing, biodefense and antimicrobial research programs. Advances in automated capillary electrophoresis systems have facilitated whole genome sequencing by reducing the time and cost associated with such experiments. The SOLiD™ System, a next generation genomics platform, holds great promise to further reduce the time and cost of whole genome sequencing, yet the issue of *de novo* assembly remains. A recent approach called "nearest neighbor sequencing" or "assisted assembly"[2], whereby sequence from a closely related organism is used as a scaffold, demonstrates utility in enabling *de novo* analysis of organisms using the SOLiD™ System. The experiment below describes implementation of this method in the analysis of two strains of Escherichia coli (E. coli) and demonstrates the significant throughput advantages the SOLiD System provides for whole genome bacterial sequencing.

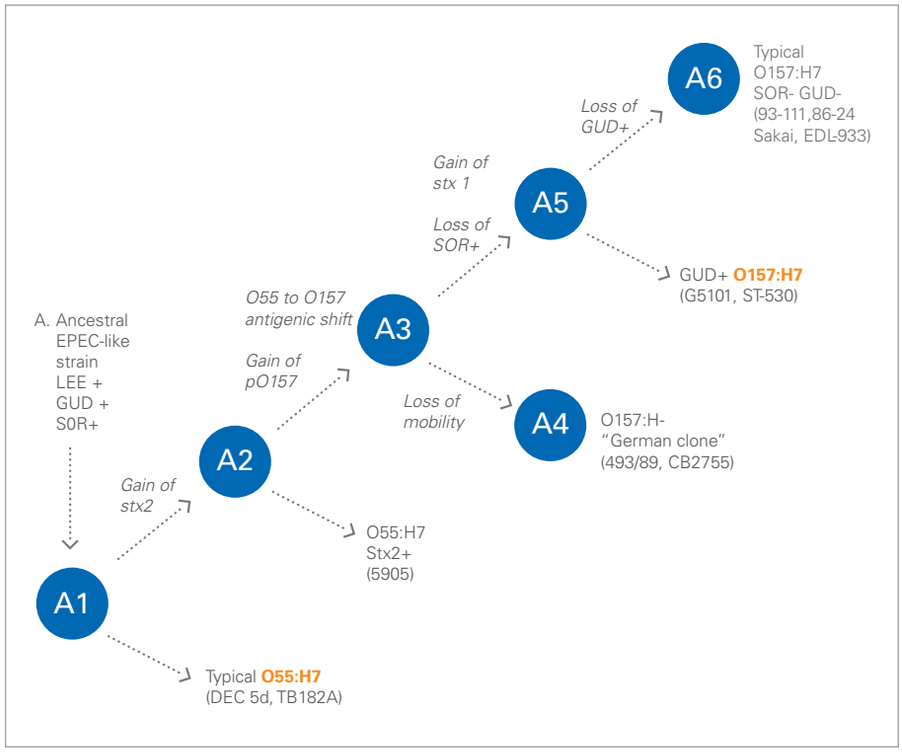

Figure 1. Evolutionary relationship between E. coli strains 055:H7 and O157:H7. This diagram illustrates the evolution of these strains from the ancestral strain and was generated based on biochemical and physical characterizations of the various subtypes of the E. coli bacteria, independent from any DNA sequencing technology[1].

Background

E. coli O157:H7 is a particularly pathogenic strain associated with eating contaminated uncooked food, drinking un-pasteurized milk, and swimming in or drinking contaminated water. In 2006, an E. coli O157:H7 outbreak in the US, linked to contaminated spinach, resulted in 205 confirmed illnesses and three deaths. With potentially deadly outbreaks associated with this strain of E. coli, assays are needed to quickly and selectively screen for its presence in food and water samples. Assay development for E. coli O157:H7 has previously been confounded by the close evolutionary relationship to another E. coli strain, O55:H7 (Figure 1). Furthermore, although different subtypes of E. coli O157:H7 were previously sequenced in the United States and Japan, E. coli O55:H7 was not fully sequenced and lacked a published reference sequence. For the development of a sensitive real time PCR assay to detect O157:H7, scientists in Applied Biosystems' Applied Markets group used the SOLiD System to sequence the genomes of E. coli O157:H7 and O55:H7, using the published O157:H7 sequence as the reference sequence for alignment of O55:H7. The two completed genomes could then be compared to identify unique O157:H7 sequences that could serve as the basis for a real-time PCR assay.

Methods

Two mate-paired libraries, one for E. coli O157:H7 and one for E. coli O55:H7, were constructed with ~2.5kb insert sizes and emulsion PCR was carried out using standard SOLiD protocols. Both the O157:H7 and O55:H7 libraries were deposited in duplicate onto a single slide and sequenced in a single run using the SOLiD Analyzer. Unique reads resulting from the O157:H7 library were aligned to the reference sequence (NCBI: NC_002655 and GI16445223) and a consensus assembly was generated using the SOLiD Analysis Tools. Next, the unique reads from the O55:H7 library were aligned to the reference sequence for O157:H7 and a second consensus generated. The two consensus sequences were compared and any regions of the O157:H7 that were not seen in O55:H7 were identified as unique sequences and potential targets for real-time PCR assays.

Results

Unique reads from the two libraries aligned with relatively uniform coverage across the genome (Figure 3). The accuracy of the SOLiD System has previously been demonstrated to be 99.999% at 15X coverage. In this experiment, an average coverage of 20X was achieved for both libraries indicating that the library constructions were very successful. The specific E. coli O157:H7 isolate sequenced in this experiment matched 98% of the reference sequence. The closely related E. coli O55:H7 matched ~92% of the reference sequence.

A comparison of the genomes sequenced by the SOLiD System is shown in Figure 4. As expected, E. coli O55:H7 reads did not cover parts of the O157:H7 EDL-933 reference sequence. These regions were well covered by the O157:H7 reads and provide evidence that these gaps represent true differences in sequence content rather than lack of coverage of those regions.

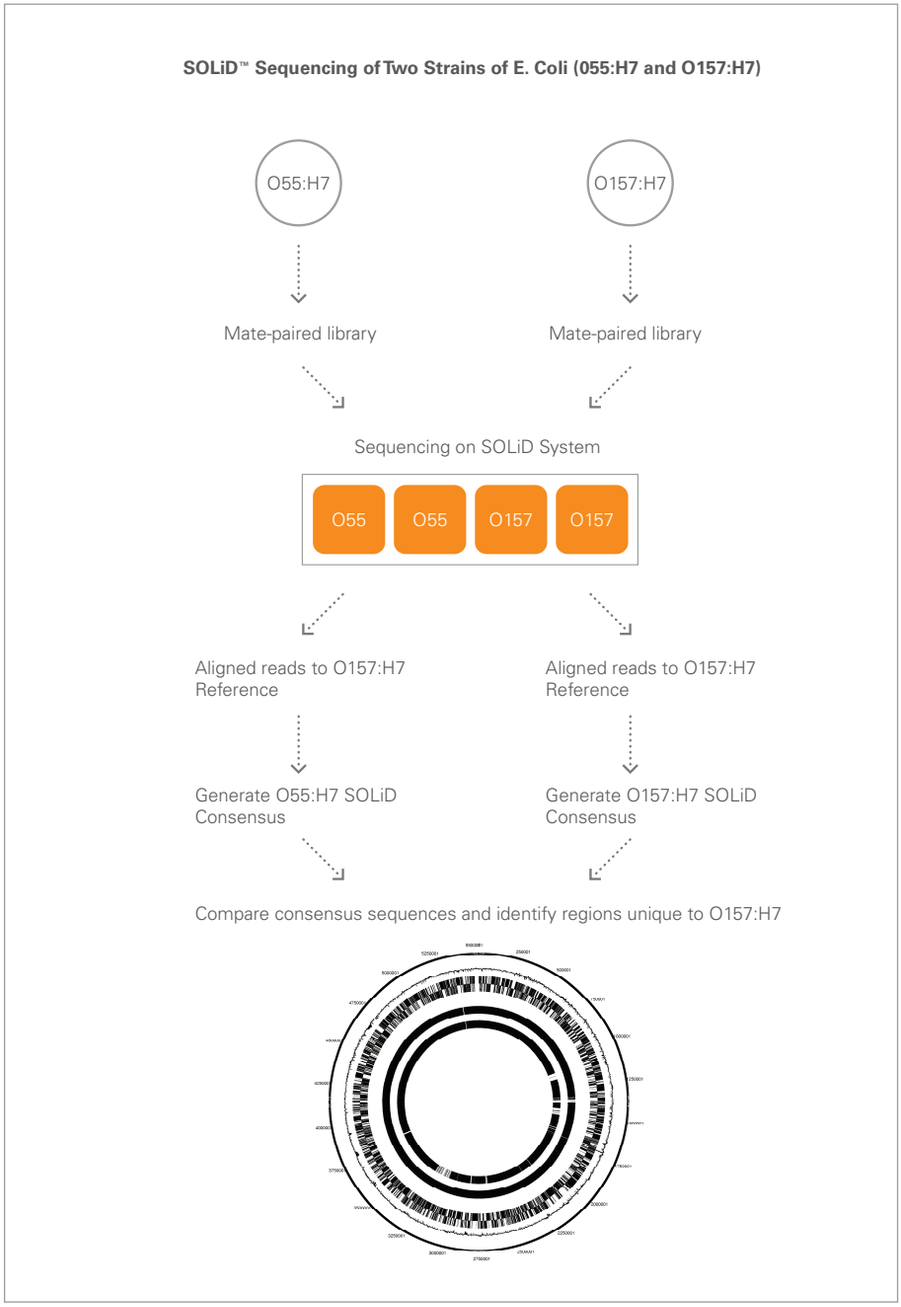

Figure 2. Workflow for 055:H7 "nearest neighbor" sequencing

The detection of these differences between the two genomes holds great promise for the development of detection assays for the dangerous E. coli O157:H7 strain.

Additionally, Figure 4 illustrates the ability of the SOLiD System to obtain uniform coverage in genomic areas of variable GC content. The GC content of the reference sequence is shown in the second circle of the figure and the data indicates there was no impact of high GC content on coverage achieved in this experiment.

Conclusion

The complete 5.5 MB E. coli O55:H7 genome was successfully sequenced on the SOLiD System using a closely related organism, E. coli O157:H7, as

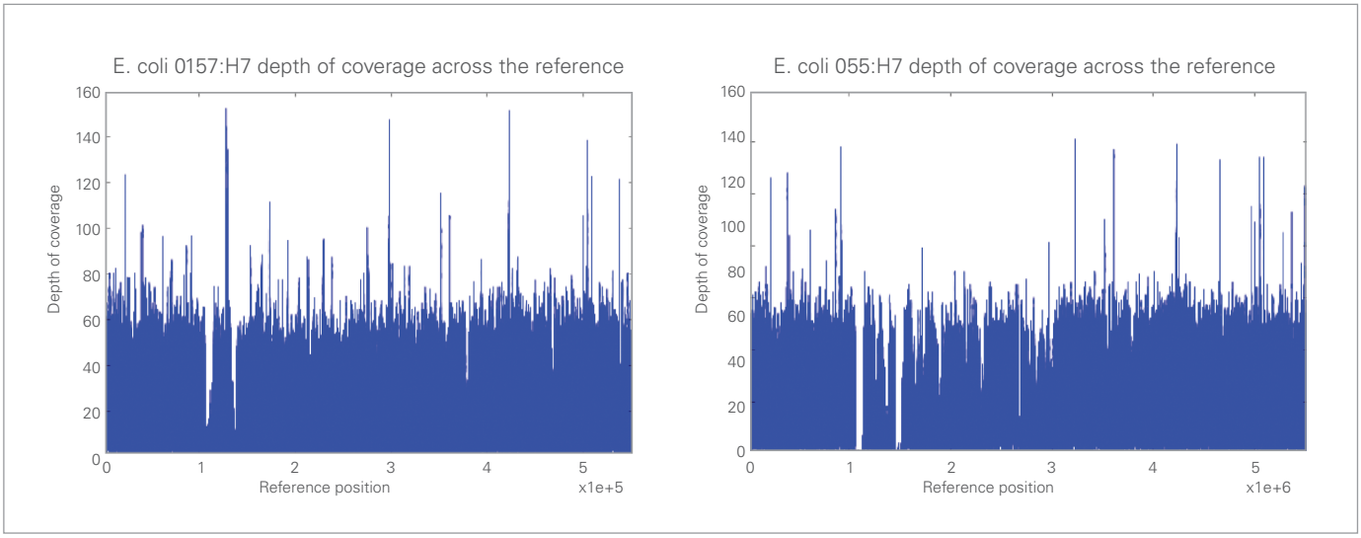

Figure 3. Coverage Map of E. Coli O157:H7 and E. Coli O55:H7 from SOLiD™ Experimental Tracking Software.

the reference. This method of "nearest neighbor" or "assisted assembly" sequencing holds great promise in enabling *de novo* analyses using the SOLiD System. Comparison of the consensus sequences for both E. coli O157:H7 and E. coli O55:H7 identified regions unique to the O157:H7 strain which are potential targets for a highly specific screening assay for the pathogenic strain. The SOLiD System's ultra high throughput and flexible bead deposition formats allowed researchers to sequence both E. coli genomes, in duplicate, on a single slide. These attributes enable the SOLID System to significantly reduce the time and the cost to sequence any bacterium. Multiple whole genome bacterial sequencing projects can now be completed in less than two weeks instead of months.

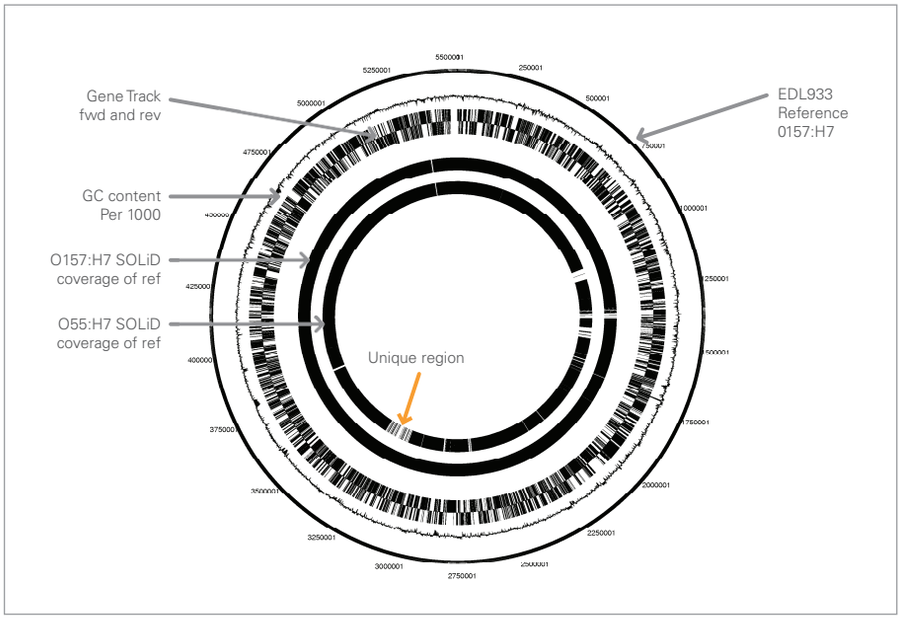

Figure 4. O157:H7 and O55:H7 Genome Comparison — Gaps seen only in O55:H7 (orange arrow) indicate that sequence is missing in this strain but present in O157:H7. This unique region in the O157:H7 genome is suitable for further assay development.

References

[1]Wick, L., *Journal of Bacteriology* 17:1547-1549 (2007).

[2]Green P., *Genome Res.* 187:1783-1791 (2005).

Printed in the USA, 02/2008, Publication 139AP09-01

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm