Paired End Sequencing on the SOLiD™ Platform

Eileen T. Dimalanta[1], Lei Zhang[1], Lele Sun[1], Kara S. Eusko[1], Dalia M. Dhingra[1], Damon S. Perez[1], Michael R. Lyons[1], Adam E. Sannicandro[2], Jonathan M. Manning[1], Heather E. Peckham[1], Eric Tsung[1], Steve M. Menchen[2], Alan P. Blanchard[1], and Kevin J. McKernan[1]

[1]Life Technologies, 500 Cummings Center, Suite 2400, Beverly, MA 01915 [2]Life Technologies, 850 Lincoln Centre Drive, Foster City, CA 94404

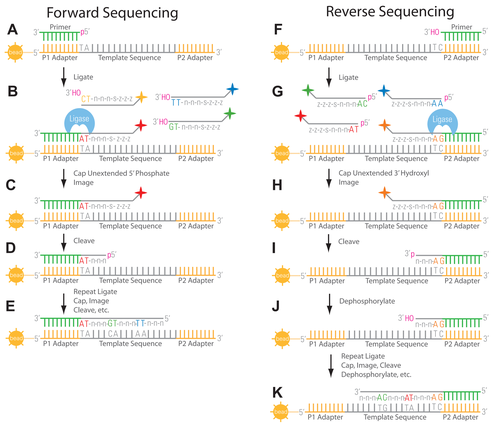
ABTSRACT

The SOLiD™ platform is a revolutionary sequencing system that utilizes sequential ligation of fluorescently labeled oligonucleotide probes, enabling high fidelity and ultra-high throughput sequencing. Previous sequencing protocols for the SOLiD™ system have only been available in the forward direction (3' to 5'). Sequencing in the reverse direction (5' to 3') is ideal as it enables fragment library paired end sequencing. To this end, novel ligation chemistries were developed to support 5' to 3' read lengths of up to 35 bases. This paired end sequencing technology will be incorporated into the SOLiD V4 platform, increasing effective read length, maximizing throughput per run, and meeting special research interests, such as whole transcriptome studies.

SOLiD™ Overview

SOLiD™ Sequencing involves the serial ligation of probes in which a dye reports the subset of four possible dibase pairs at the 1st and 2nd positions from the ligation junction. SOLiD™ Sequencing in the forward direction involves: A.) 5'-phosphorylated primer is hybridized to the adapter region of the templates to be sequenced; B.) Fluorophore labeled 8-mer complementary probes, containing 3 universal bases to decrease complexity, are ligated to the primer; a second round of ligation is performed with unlabeled probe to increase the amount of primer extended per bead; C.) Any remaining unextended primer is capped by dephosphorylation to prevent dephasing; beads are then imaged to record fluorophore reporter; D.) A phosphorothiolate bond in the ligated probes is cleaved with AgNO3, reducing the probe to 5 nucleotides and generating a free phosphate for the next round of ligation; E.) Additional cycles of ligation, capping, imaging, and cleavage are performed until the desired read length is obtained.

SOLiD™ Sequencing in the reverse direction involves: F.) 3'-hydroxylated primer is hybridized to the adapter region of the templates to be sequenced; G.) Fluorophore labeled 8-mer complementary probes (5' phosphorylated), containing 3 universal bases to decrease complexity, are ligated to the primer; a second round of ligation is performed with unlabeled probe to increase the amount of primer extended per bead; H.) Any remaining unextended primer is capped by polymerase incorporation of a ddNTP to prevent dephasing; beads are then imaged to record fluorophore reporter; I.) A phosphorothiolate bond in the ligated probes is cleaved with AgNO3, reducing the probe to 5 nucleotides and generating a free 3' phosphate; J.) The 3' phosphate is removed for the next round of ligation; K.) Additional cycles of ligation, capping, imaging, cleavage, and dephosphorylation are performed until the desired read length is obtained.



Schematic of Paired End SOLiD™ Sequencing

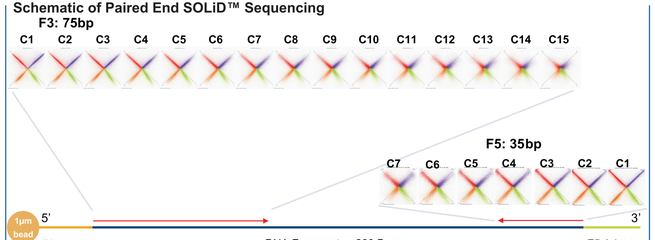


Figure 1. Paired end sequencing on the SOLiD™ platform. Priming from the 5' end of the template for forward reads (F3 Adapter) and from the 3' end for the reverse reads (F5 adapter) of a standard fragment library. Novel reverse ligation chemistry enables paired end sequencing without requiring synthesis of the complementary strand of the DNA template.

Reverse Sequencing *E coli* DH10b "Satay" Plots

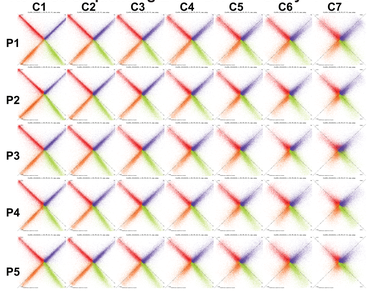


Figure 2. "Satay" Plots for an *E. coli* DH10b 35 base reverse sequencing run. These plots show the spectral quality and intensity of the sample. The axes correspond to the 4 different fluorochromes used in SOLiD™ sequencing: FAM, CY3, TXR, CY5. Each dot on the plot represents the fluorescent wavelength and intensity of multiple copies of the bead bound DNA template. Beads that fall on or near an axis are monoclonal (i.e. they contain multiple copies of a single DNA template), and beads that are far from the origin are high intensity beads.

Matching Statistics for Paired End Sequencing

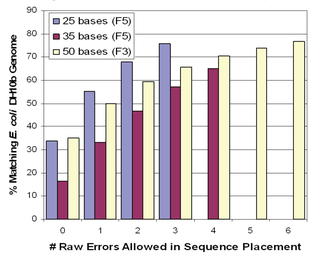


Figure 3. Matching Statistics for paired end sequencing of an *E. coli* DH10b genome. For reverse sequencing (F5), 65% of the reads matched the genome for 35 bases allowing up to 4 mismatches; 77% of the reads matched the genome for 25 bases allowing up to 3 mismatches. For forward sequencing (F3) 76% of the reads matched the genome for 50 bases allowing up to 6 mismatches.

Paired End Sequencing Throughput

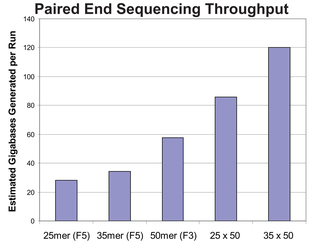


Figure 4. Estimated throughput of paired end sequencing based on SOLiD™ 4 bead densities. A full sequencing run can generate 28 Gigabases of data for 25mers (reverse reads), 35 Gigabases of data for 35mers (reverse reads), and 57 Gigabases of data for 50mers (forward reads). A 25mer (reverse) x 50mer (forward) can generate up to 86 Gigabases of data, and up to 120 Gigabases for 35mer (reverse) x 50mer (forward).

DNA Template Size Distribution

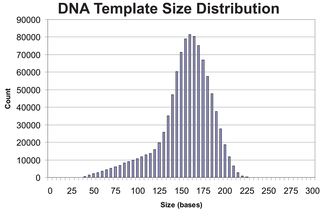


Figure 5. Size distribution of DNA template determined from paired reads.

E coli DH10b Pairing Rates

Read Length	Pairing Rate
25 x 50	98.94%
35 x 50	99.05%

Table 1. Pairing rates for 25mer (reverse read) x 50mer (forward read) and 35mer (reverse read) x 50mer (forward read).

Paired End Sequencing of the Human Genome

Reverse Sequencing Human "Satay" Plots

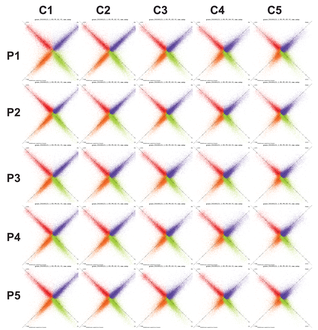


Figure 6. "Satay" Plots for a 25 base reverse sequencing run of a human genome.

- DNA extracted from buccal swab
- Constructed paired end library using 1 µg of DNA
  1 day library protocol
- 2 SOLiD Sequencing Runs
  50 bp Forward Reads
  25 bp Reverse Reads
- Generated over 88 Gigabases of aligned reads

Human Genome Distribution of Coverage

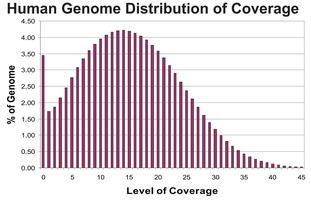


Figure 7. Distribution of coverage of uniquely placed paired reads across the human genome. The average coverage was 14.5x.

Average Coverage per Chromosome

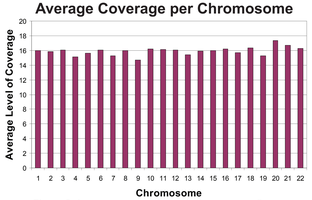


Figure 8. Average coverage for each chromosome. Coverage ranged from 14.7x to 17.4x for the autosomal chromosomes.

Summary of SNPs/Indels

	Count	% in dbSNP
Total SNPS	2972853	82.16%
Heterozygous SNPs	1963475	73.70%
Homozygous SNPs	1213219	94.42%
Small Indels *	103027	73.80%

Table 2. Summary of identified SNPs and small indels, and their concordance with dbSNP.

* stringent indel calling conditions

Distribution of Small Insertions

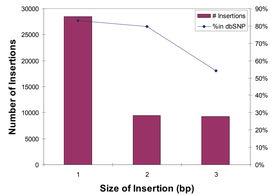


Figure 9. Distribution of small insertions (1-3 bases); 76.9% were in dbSNP.

Distribution of Small Deletions

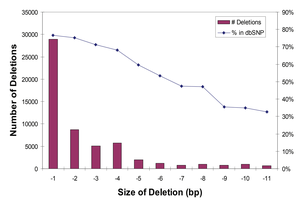


Figure 10. Distribution of small deletions (1-11 bp); 71.2% were in dbSNP.

ACKNOWLEDGEMENTS

Christopher Clouser	Nick Fantin	Rachel Kasinskas	Bashar Mullah	Jessica Spangler
Brittney Coleman	Tamara Gilbert	Mike Laptewicz	Dick Noble	Dean Tsou
Gina Costa	Sherry Hansen	Clarence Lee	Ken Otteson	Tristen Weaver
Tenzin Dawoe	Stephen Hendricks	Liz Levandowsky	Vaishnavi Panchapakesa	Allen Wong
Dave Dupont	Jeffrey Ichikawa	Stephen McLaughlin	Andrew Sheridan	Joon Yang

TRADEMARKS/LICENSING