Increased Read Length on the SOLiD™ Sequencing Platform

Eileen T Dimalanta[1], Lei Zhang[1], Cynthia L Hendrickson[1], Tanya D Sokolsky[1], Adam E Sannicandro[1], Jonathan M Manning[1], Stephen F McLaughlin[1], Haoning Fu[2], Clarence C Lee[1], Alan P Blanchard[1], Gina L Costa[1], Kevin J McKernan[1]

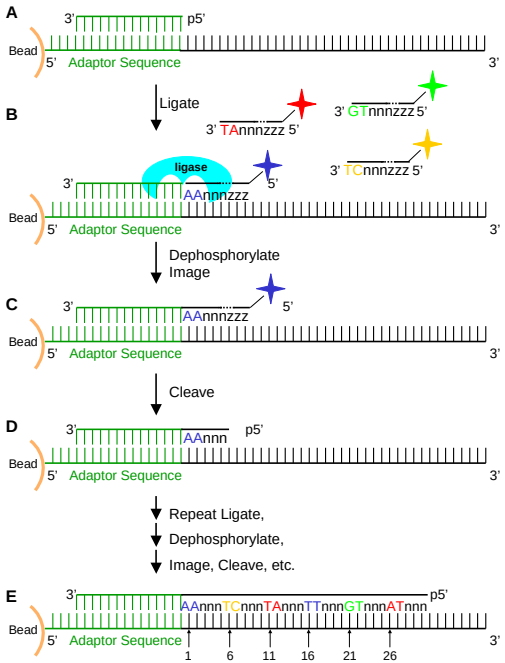
Applied Biosystems, High Throughput Discovery, [1]500 Cummings Center, Suite 2400, Beverly MA USA 01915; [2]850 Lincoln Centre Drive, Foster City, CA

ABSTRACT

The SOLiD™ Platform is an ultra-high throughput DNA sequencing system that utilizes sequential ligation of fluorescently labeled oligonucleotide probes. Improvements in the ligation biochemistry, sample preparation, and incorporation of an imaging buffer providing a higher signal to noise ratio have made it possible to routinely sequence 50 base reads from long fragment libraries, or 2x50 base reads from mate pair libraries with over 99.9% accuracy, generating over 20 gigabases of mappable sequence data per run. Currently, novel ligation protocols have been developed to support increased read lengths of 75 bases to 100 bases, as well as reverse ligations to facilitate paired end reads. These longer read lengths increase throughput per run, facilitate re-sequencing efforts of large genomes, and aid in the identification of SNPs, indels, and other structural variations. Longer reads also lend themselves to novel applications of SOLiD™ such as RNA expression analysis and *de novo* sequencing.

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. The steps involved include: **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 dephosphorylated to prevent dephasing; beads are then imaged to record fluorophore reporter. **D.**) A phosphorothiolate bond in the ligated probes is cleaved with $AgNO_3$, reducing the probe to 5 nucleotides and generating a free phosphate for the next round of ligation. **E.**) Additional cycles of ligation, dephosphorylation, imaging, and cleavage are performed until the desired read length is obtained.



75 BASE SEQUENCING

Figure 1. "Satay" Plots of cycles 1 to 15 of a single primer in a 75 base 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.

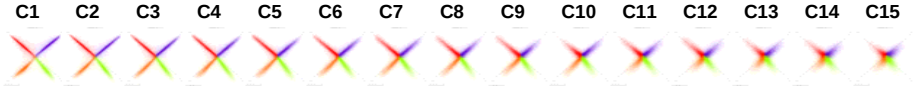


Figure 2. A) Within a few short months, improvements in chemistry for generating 75mers has led to increased accuracy and throughput. The SOLiD™ platform has the capacity to sequence 2 full arrays, resulting in over 30 Gigabases of usable data for 1x75mer run. B) Base position raw errors and colorspace error

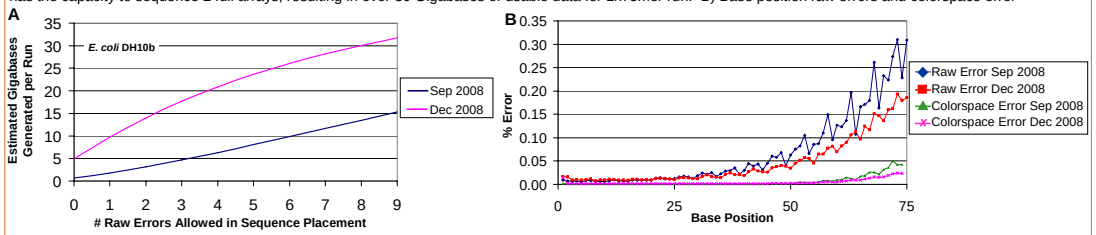
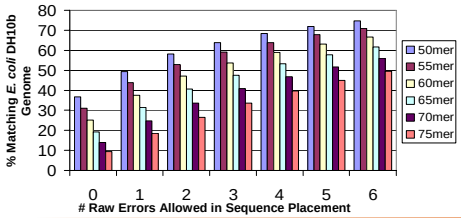


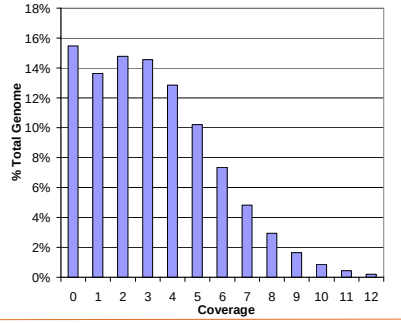
Figure 3. Matching statistics for 50 to 75 bases. Increased accuracy results in 75% raw data matching the genome with 50 base reads; 50% raw data matching with 75 base reads.



75 BASE SEQUENCING OF HUMAN

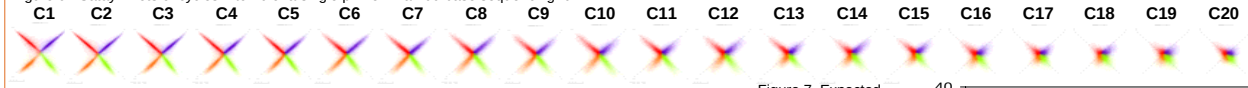
Figure 4. Over 20 Gigabases of usable data from 75 base reads from 2 full arrays can be generated for one sequenging run.

Figure 5. Coverage of the Human Genome generated from 75 base reads from 1 full array. The average coverage was 3.2 x



100 BASE SEQUENCING

Figure 6. "Satay" Plots of cycles 1 to 20 of a single primer in a 100 base sequencing run.



REVERSE LIGATIONS

Figure 8. Reverse ligations allows for the generation of paired end reads. A) Schematic of reverse ligations. Here, the ligating strand is growing from the 5' to 3' direction. B) "Satay" plots from cycles 1 to 5 of a single primer for reverse ligations.

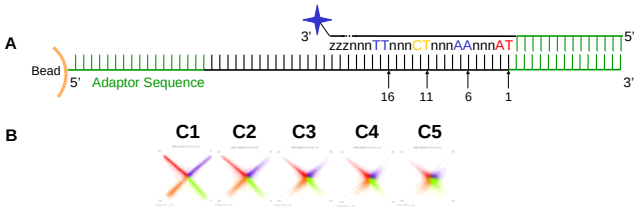
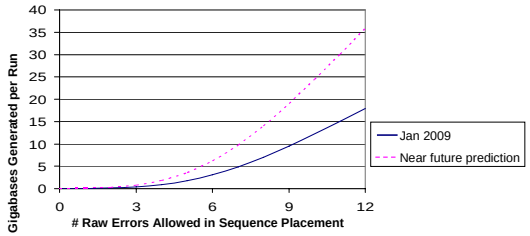


Figure 7. Expected improvements in chemistry will undoubtedly increase the accuracy and throughput of sequencing 100 bases to generate over 30 gigabases of usable data per 1x100 sequencing run.



ACKNOWLEDGEMENTS

Jer-Kang Chen	Brittney Coleman	Shaheer Khan	Steve Menchen	Ken Otteson	Krishna Upadhya
Dan Cote	Cisilya Duncan	Michel Laptewicz	Bashar Mullah	Vaishnavi Panchapakesa	Allen Wong
Christopher Clouser	Paul Kenney	Liz Levandowsky	Richard Noble	Andrew Sheridan	Ruiming Zou