Getting Personal: Improved Throughput and Accuracy Toward Enhanced Understanding of Human Genome Biology

Jeffrey K. Ichikawa[1], Clarence Lee[2], Vasisht Tadigotla[2], Rachel Kasinskas[2], Eileen Dimalanta[2], Stephen Hendricks[1], Warren Tom[1], Nick Fantin[1], Maryam Shenasa[1], Andrew Hutchison[1], Jason La[1], Stephen McLaughlin[2], Heather Peckham[2], Chengyong Yang[1], Yong Chu[1], Haoning Fu[1], Christina Chung[1], Gina Costa[1], Kevin McKernan[2], and Timothy Harkins[2].

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

ABSTRACT

Next generation sequencing technologies have made it possible to readily sequence a whole human genome in a single instrument run. Generating large amounts of data is now routine and new technical advancements in accuracy and data management are becoming more critical. This sequencing data was generated using both fragment and long mate-paired library methodologies and sequenced at various read lengths. Indeed, human data supportive of paired-end sequencing using fragment libraries showed enhanced resolution across the genome with increased resolution when coupled with mate-paired annotation. In total, higher throughputs per instrument run along with improved probe and ligation chemistry have resulted in enhanced coverage with fewer sequencing gaps in both GC- and AT-rich regions, as well as a decrease in the overall G+C bias across the human genome. Sequence data of this depth allows the community to better resolve many annotation issues of the human genome. Furthermore, this has also allowed the assessment of the underlying sequencing technology including coverage and the sequencing chemistry accuracy essential to the exploration of personalized genomes. The technical advancements include new sequencing chemistries that are orthogonal to the standard probe chemistry of the SOLiD™ sequencing platform which are based upon error correcting codes developed by the telecommunications industry, higher density sequencing runs using sub-micron beads, longer sequencing reads and a new sequencing instrument, the 5500XL.

INTRODUCTION

Improvements to the SOLiD™ sequencing system will provide higher throughput and greater accuracy over the previous SOLiD™ 4 system. Higher throughput has been enabled with smaller beads, allowing greater packing and increased number of reads during sequencing. Optimized SOLiD™ ligation chemistry has been demonstrated to allow read lengths of up to 75 bp during forward ligation sequencing. Furthermore, a new proprietary ligase has shown improved reverse read chemistry for PE sequencing. The addition of the Exact Call Chemistry sequencing strategy allows for error correction and provides highly accurate sequencing data. These improvements are currently in development and will be enabled on the 5500xl series SOLiD™ sequencer.

MATERIALS AND METHODS

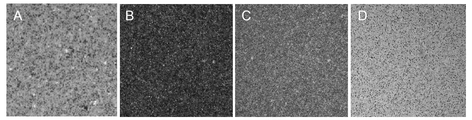
SOLiD™ Sequencing and Primary Analysis
SOLiD™ sequencing was completed on standard SOLiD 4 systems using commercially available SOLiD™ ToP reagents. SOLiD™ 5500XL sequencing was completed using development versions of Instrument Control Software (ICS) and primary analysis algorithms.

SOLiD™ Bead Production
All SOLiD™ beads were produced using standard protocols and reagents.

SOLiD™ Secondary Analysis
All secondary analysis was completed using Bioscope™ v1.3

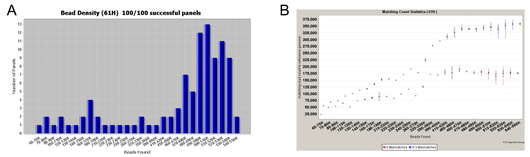
RESULTS

Figure 1. Smaller Beads Afford Higher Densities



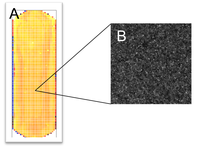
Smaller diameter beads afford higher densities. Representative panels are shown with increasing bead densities deposited on a standard SOLiD™ 4 XD slide. *Panel A* shows a comparative image with 1 µm templated beads deposited at 300,000 beads/panel. *Panels B-D* are representative images from 0.75 µm bead depositions targeting 300,000 to 500,000 beads/panel.

Figure 2. Smaller Beads Afford Higher Throughput



Smaller diameter beads afford higher densities with corresponding higher throughput. Sequencing of 0.75 µm beads with a DH10B Fragment template was sequenced. Sequencing was completed on a SOLiD™ 4 instrument with a development versions of small bead identification algorithms. *Figure A* shows a histogram plot of the distribution of bead densities across 100 randomly sampled panels. The average bead density across the slide was 490,000 beads/panel. Figure B shows the number of mapped reads to the reference genome at varying densities with both 0 mismatches and 0-3 mismatches plotted. A linear increase in mapping is observed over 60,000 to 450,000 beads per panel. Image panels with densities above 450,000 beads per panel showed decreased mapping percentages.

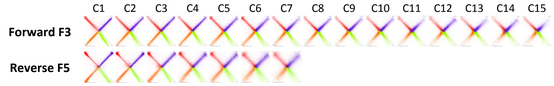
Figure 3. Small Bead Sequencing Performance



Flowcell	Density	Tag	% Map	Thru-put (Gb)	Total (Gb)
S0107_20100916_1	327K	R3	76%	26.4	50 Gb
		F3	68%	23.6	
S0107_20100916_2	323K	R3	84%	28.1	55 Gb
		F3	80%	26.7	

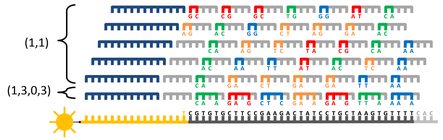
Sequencing performance of a Human Long Mate Pair template on 0.75 µm beads deposited on a SOLiD™ 4 XD slide. A 2 x 50 bp sequencing run was completed on a standard SOLiD™ 4 instrument with 2 slides. The heat map (*Panel A*) represents one full slide with an average bead density of 327,000 beads/panel. The sample image in *Panel B* shows a panel with 305,000 beads identified during the sequencing run. A portion of the image has been magnified for greater bead resolution. A total of 6.6 x 10^8 beads were identified on this slide. The bead densities and mapping rates are shown in the table. The overall throughput of this SOLiD™ sequencing run was 105 Gb, mapping to the HG18 reference genome.

Figure 4. Optimized SOLiD™ Chemistry for Longer Paired-End Read Lengths



The spectral purity or 'satay' plots are shown to highlight the SOLiD™ chemistry improvements. The first 15 cycles of forward and 7 cycles of reverse are shown from a HuRef Fragment Paired-end sequencing run. A higher proportion of beads are producing greater signal and better spectral purity across the four channels due to improved ligation chemistry for both directions of sequencing.

Figure 5. Exact Call Chemistry Strategy



The schematic in Figure 5 shows the Exact Call Chemistry (ECC) strategy. Following the normal 5 primer rounds of SOLiD 2-base encoded sequencing, an additional sixth primer round using a 3-base encoded probe set is completed, forming a redundant error correction code. This probe set (1,3,0,3) can be considered 4-base encoded (4BE) because the color does not depend on base position 3 (due to the zero at position 3) and is defined by the spacing between the first base and the last base. This strategy is based on redundant error correction codes found in digital communication systems.

Table 1. Exact Call Chemistry Mapping

	2BE Only	ECC
Total Tags Found	609,894,785	634,160,054
Read Length	75 bp	75 bp
Percent Mapped Reads	81.9%	81.2%
Coverage (Gb)	33.5 Gb	34.5 Gb
Avg no. of mapped reads per unique start points	1.12	1.12
Full length reads with 0 mismatches	135,790,270 (27.2%)	213,661,277 (41.5%)

A Human sample fragment library was sequenced to 75 bp with Exact Call Chemistry. Mapping statistics are shown with and without the addition of the sixth ECC primer. As errors are corrected, a higher number of reads are allowed through the mincall filter which resulted in a 4% increase in total tags found and a 1 Gb increase in total throughput. Although the overall increase in throughput was minimal, an increase in full length perfect reads increased by 57.3%.

Figure 6. Comparison of Mapping Distribution with ECC

The graph in Figure 6 shows the comparison of non-ECC to ECC mapping distribution from a 75 bp Human sample Long Fragment sequencing run. The mapping distribution of 75mer full length reads at 0-10 allowable mismatches are highlighted. The total number of mapped reads is increased by 57.3% at 0 mismatches to 11.2% at 10 mismatches.

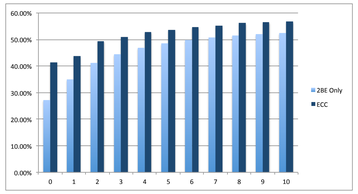


Figure 7. Improved Accuracy with Exact Call Chemistry

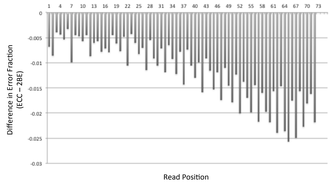
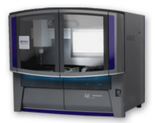


Figure 7 demonstrates the improvement in base accuracy across a 75mer Human fragment sequencing run. The difference in base accuracy between the ECC probe enabled minus the 2-base encoded probe sequencing errors are shown. The lower error rates of the ECC data results in a negative difference in error fractions. The ECC data has consistently lower error across the read, with larger improvement at the longer read lengths.

Figure 8. 5500xl Series SOLiD™ Sequencer

The new SOLiD™ system improvements are being developed for release on the new 5500 series instruments. The 5500 and 5500xl instruments will provide additional flexibility, speed and throughput enhancements in addition to read length, bead density and error correcting improvements to the SOLiD™ system.



CONCLUSIONS

Several improvements have been made to the SOLiD™ system that affords increased throughput and higher accuracy during sequencing. Higher density is achieved through smaller diameter beads along with optimized bead finding and signal identification algorithms that allow sequencing of >300,000 beads per panel. Recent improvements in ligation chemistry have yielded increased signal-to-noise ratios which allow read lengths of up to 75 bp. A new proprietary ligase for reverse read chemistry has resulted in reduced errors, more uniform coverage and increased frequency of mappable reads for paired-end sequencing. Furthermore, the addition of Exact Call Chemistry along with specialized decoding algorithms allows for error detection and correction to provide high accuracy sequencing reads. Overall, these combined improvements have resulted in enhanced sequence throughput per run and also allows progressively deeper sequencing coverage with greater accuracy of more complex model organisms.

ACKNOWLEDGEMENTS

The authors would like to thank Karun Kallakuri, Rakesh Chaparala, Yuandan Lou and Charles Scafe for their efforts to this project.

For Research Use Only. Not intended for animal or human therapeutic or diagnostic use.

© 2011 Life Technologies Corporation. All rights reserved.
The trademarks mentioned herein are the property of Life Technologies Corporation or their respective owners.