1884T

Paired End Sequencing of Human Genomes on the SOLiD™ Platform

Lei Zhang[1], Eileen T. Dimalanta[1], Stephen F. McLaughlin[1], Damon S.Perez[1], Vasisht Tadigotla[1], Kara S. Eusko[1], Dalia M. Dhingra[1], Clarence C. Lee[1], Christopher Clouser[1], Jessica C. Spangler[1], Tristen N. Weaver[1], Michael R. Lyons[1], Heather E. Peckham[1], Archie Russell[2], Paul Billings[2], Nila Patil[2], Edward Kiruluta[2], George Miklos[2], Martin G. Reese[2], Alan P. Blanchard[1], Kevin J. McKernan[1]

[1]Life Technologies Corporation, 500 Cummings Center, Beverly, MA 01915

[2]Omicia, Inc., 2200 Powell Street, Emeryville, CA 94608

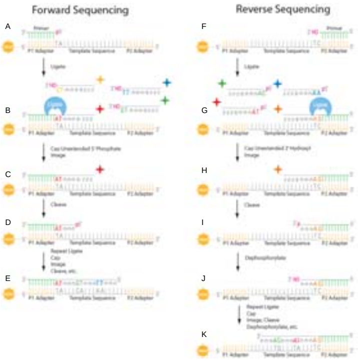
ABSTRACT

The SOLiD DNA sequencing system utilizes stepwise ligation of oligonucleotide probes and enables high fidelity, high throughput sequencing. Previous sequencing protocols for the SOLiD system have only been available in the forward direction (3' to 5'). However, fragment library paired end sequencing (in both forward and reverse directions) is highly desired to maximize sequencing capacity and to meet special research interests such as whole transcriptome and translocation studies. To achieve this using the SOLiD platform, novel ligation chemistries were developed to support 5' to 3' read lengths of up to 35 bases. Utilizing this new chemistry, we sequenced an anonymous Caucasian male to an average depth of coverage of 22.4x (2 SOLiD system sequencing runs) covering 96.55% of the genome. The purpose of any resequencing project is to measure variants against a known reference sequence, so we analyzed this data using our in-house variant detection algorithms to assess data integrity. Approximately 2.97M SNPs were discovered: 1.21M homozygotes (94.4% in dbSNP v129) and 1.76M heterozygotes (73.7% in dbSNP v129). Also detected were 103,027 small indels (73.8% in dbSNP v129). Paired end sequencing with this novel chemistry presented here is effective at variant detection in a human genome and these values are similar to variant totals from other large scale human resequencing.

INTRODUCTION

SOLiD system 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 system 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 system 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.

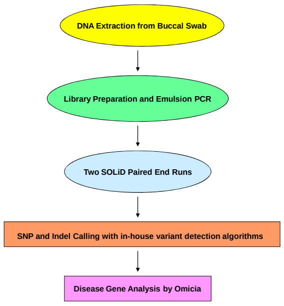


MATERIALS AND METHODS

A Buccal swab from an anonymous Caucasian male (XY) was used for DNA extraction. One µg of DNA was used to construct a fragment library according to the SOLiD™ library preparation protocol. Two SOLiD™ sequencing runs with 50 bp forward reads and 25 bp reverse reads were performed on SOLiD™ 3 plus instruments. In addition, a HuRef[1](Craig Venter**)** fragment library was sequenced on a SOLiD™ 4 instrument to obtain 75 bp forward reads and 35 bp reverse reads.

RESULTS

Figure 1. Workflow of Paired End Sequencing of Human XY Genome


DNA Extraction from Buccal Swab
Library Preparation and Emulsion PCR
Two SOLiD Paired End Runs
SNP and Indel Calling with in-house variant detection algorithms
Disease Gene Analysis by Omicia

- **Aligned Reads: 64.1 Gb (22.4x coverage)**
- **Uniquely Placed Paired Reads: 47.1 Gb (16.5x coverage)**
- **Uniquely Placed Normal Paired Reads: 46.3 Gb (16.2x coverage)**
- **Non-redundant Uniquely Placed Normal Paired Reads: 43.2 Gb (15.1x coverage)**

Figure 2. Human XY Genome Coverage with Paired End Reads

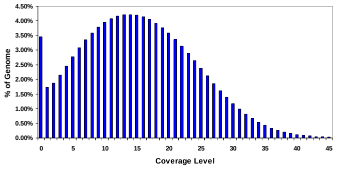


Figure 2. Distribution of coverage of uniquely placed paired reads across the human genome. 96.55% of hg18 covered with uniquely placed pairs. The average coverage was 16.5x.

Figure 3. Average Coverage Per Chromosome

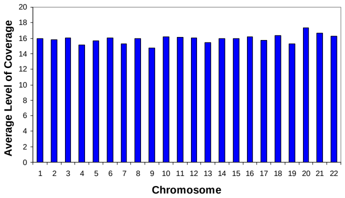


Figure3 . Average coverage for each chromosome. Coverage ranged fron 14.7x to 17.4x for the autosomal chromosomes.

Table 1. Summary of SNPs/Indels

	# Identified	Concordance with dbSNP[1]
Total SNPs	2,972,853	82.2%
Heterozygous SNPs	1,759,634	73.7%
Homozygous SNPs	1,213,219	94.4%
Small Indels *	103,027	73.8%

* stringent indel calling conditions

Table 1. Summary of identified SNPs and small indels, and their condordance with dbSNP. There are 17,489 coding SNPs, 15.0% are novel. There are 501 coding indels.

Figure 4. Small Insertions and Deletions

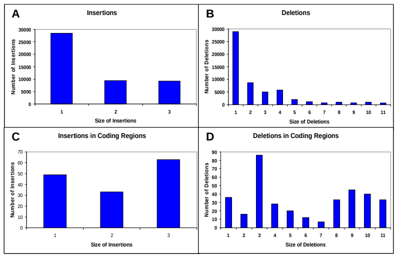


Figure4. Distribution of small insertions (1-3 bases) and small deletions (1-11 bases). A and B, Total small indels. C and D, small indels in coding regions.

Figure5. Distribution of total XY Non-Synonymous variants in disease genes.

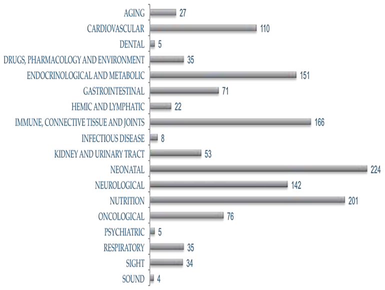


Figure5. Distribution of total 1,369 Non-Synonymous variants in disease genes. Data was analyzed by Omicia, Inc.

Figure 6A. HuRef Reverse Sequencing Spectral Purity Plots

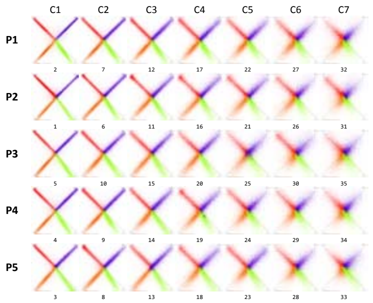


Figure 6B. HuRef Forward Sequencing Spectral Purity Plots

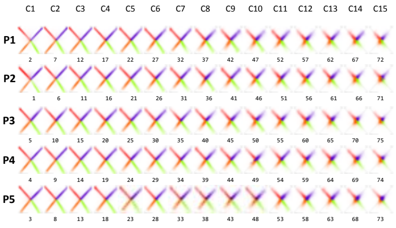


Figure 6. SOLiD system spectral purity plots for HuRef 35 base reverse (**A**) and 75 base forward (**B**) sequencing run. These plots show the spectral quality and intensity of the sample. The axes correspond to the 4 different fluorochromes used in the SOLiD sequencing system : FAM™, Cy3®, Texas Red®, Cy5® dyes. 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.

Table 2. Analysis of 75x35 HuRef Paired End Reads: SNP Identification

SNPs	# Identified	Concordance with dbSNP
Total SNPs	2,737,493	91.1%
Heterozygous SNPs	1,419,423	85.6%
Homozygous SNPs	1,318,070	97.0%

Figure 7. Human XY/HuRef SNP Overlap

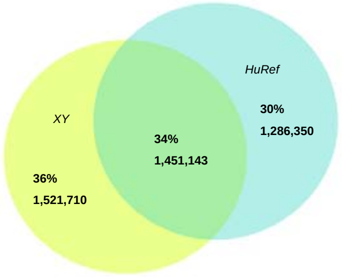


Figure 7. Venn diagram showing the overlap in reference variant SNP calls made for XY and HuRef relative to hg18 NCBI build 36.

CONCLUSIONS

Sequencing of human fragment libraries from both forward and reverse directions has been successfully performed on the SOLiD platform. Optimization of the chemistry resulted reverse sequencing reads up to 35 bases and forward sequencing reads up to 75 bases. The longer paired end read lengths increase throughput per run, facilitate re-sequencing 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 system such as whole transcriptome analysis and *de novo* sequencing.

REFERENCES

1. The diploid genome sequence of an individual human Levy S et al. *PLoS Biol.* (2007) Sep 4;5(10):e254

ACKNOWLEDGEMENTS

Timothy Harkins
Luisa Andruzzi
Liz Levandowsky
Brittney Coleman
Gina Costa
Bashar Mullah
Tenzin Dawoe
Dick Noble
Ken Otteson
Vaishnavi Panchapakesa
Matthew Dindinger
Theodore Foss
Sherry Hansen
Andrew Sheridan
Stephen Hendricks
Jeffrey Ichikawa
Dean Tsou
Rachel Kasinskas
Mike Laptewicz
Allen Wong
Tamara Gilbert
Marian Peris

TRADEMARKS/LICENSING

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