Application Specific Accuracy with SOLiD™ System High Throughput Sequencing

Asim Siddiqui, Heather Peckham, Fiona Hyland, Eric Tsung, Charles Scafe, Yutao Fu, Rajesh Gottimukkala, Caleb Kennedy, Aaron Kitzmiller, Stephen McLaughlin, Onur Sakarya, Paolo Vatta, Zheng Zhang, Jeff Ichikawa, Gina Costa and Ellen Beasley, Life Techologies, 850 Lincoln Centre Drive, Foster City, CA, USA, 94404

ABSTRACT

The advent of high throughput next generation sequencing enables experiments to study genome variations, transcription of coding and non-coding RNAs and epigenetic profiles. The ability to design effective experiments that make efficient use of samples and highly parallel sequence generation requires a clear understanding of the impact of system accuracy on the power to detect a variety of meaningful biological differences between or within samples. Standard expressions of sequence quality are important, but insufficient, to support rational experimental design.

In developing the SOLiD™ high throughput sequencing system we include system validation and the release of system data sets. The SOLiD™ System, continues to improve data quality and the percentage of reads mapped in the data analysis. With mapped data accuracy at greater than 99.9%, we continue to analyze reference samples so that we can continue to report concordance to with the reference.

Using the SOLiD™ 3+ System, we generated 128GBytes of alignable sequence for the HuRef genome. Validating against the 7x Sanger and SNP chips for this individual (Levy et al.), we found 99.7% SNP concordance. Small indel concordance was 80-85% depending on mapping parameters.

We will present data that demonstrates the power of SOLiD ™ System to detect SNPs, indels and CNV in standard resequencing experiments.

We will also present data on the detection of fusion transcripts using paired end sequencing of the UHR and MCF7 RNA samples.

INTRODUCTION

Measures of sequence accuracy ultimately translate into the ability to correctly identify genomics features. Here we present work that demonstrates the performance of the SOLiD^{TM} System against orthogonal data measures.

MATERIALS AND METHODS

DNA Sequencing

The data presented are derived from 3 experiments.

1.Whole genome sequencing of the HuRef genome on the SOLiD™ 3+ System using long mate pair libraries sequenced to 50 bases on both ends.

2.Whole genome sequencing of the E. *coli* genome on the SOLiD™ 4 System using ToP chemistry on fragment libraries sequenced to 50 bases

3.Whole genome sequencing of the HuRef genome on the SOLiD™ 4 System using ToP on long mate pair libraries sequenced to 50 bases on both ends. *This data is from a single run*.

Data were mapped using Bioscope™ Software under standard settings and parameters. For E. *coli*, data were compared to the reference sequence of the sequenced strain DH10B. For ,compared to the reference sequence published by Levy et al. A gold set of SNPs was created using the concordant set of SNPs measured by the microarray and Sanger sequencing experiments performed in that paper. We also used the indels presented in that paper as a reference set.

To further investigate SNPs called in regions not covered by the microarray, we used the SNP qualities and Sanger sequencing depth of coverage to identify sets of high quality reference SNPs for validation.

Whole Transcriptome Sequencing

The data presented are derived from transcriptome paired end sequencing of UHR and MCF7 samples sequencing 50 bases forward and 25 reverse. The data were mapped Bioscope™ Software under standard settings and parameters. The software includes methods to identify fusion transcripts using paired end and single end reads. We employed the paired end method.

Fusions were validated against TaqMan® assays.

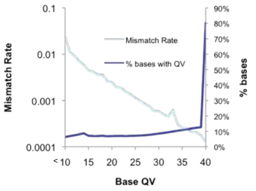
RESULTS

Table 1. Summary of experiments

Expt	Experiment
E. coli	1 slide paired end
3+ HuRef	2 slides long mate pair, 50bp; 1 slide fragment, 50 bp
4 HuRef	1 slide long mate pair, 50 bp

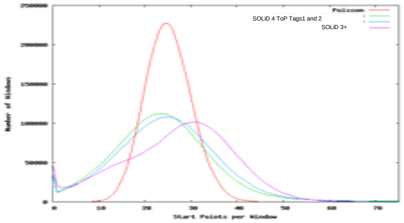
This table provides a summary of the statistics of the sequencing run for each experiment.

Figure 1. Accuracy as a Function of QV



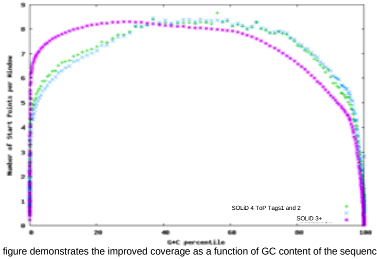
The graph shows the distribution of base mismatches as a function of quality values (QV) for the E. *coli* data for the F3 tag. The data shows a linear relationship between QV and base mismatches. Over 80% of the data has QV of 40

Figure 2. 3+ vs. 4: genome coverage



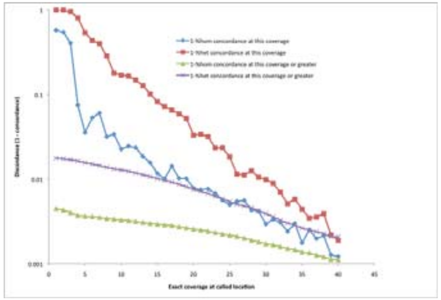
This figure demonstrates the improved genome coverage provided by the ToP sequencing chemistry employed by the SOLiD™ 4 System.

Figure 3. 3+ vs. 4: GC coverage



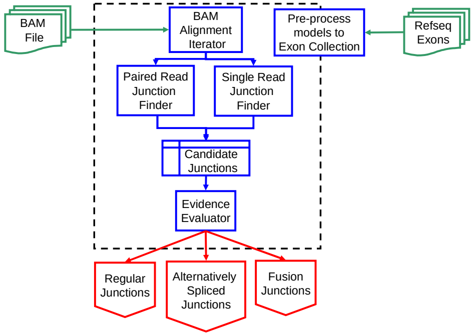
This figure demonstrates the improved coverage as a function of GC content of the sequenced region.

Figure 5. SNP concordance as a function of sequence depth for SOLiD^{TM} 3+



This graph shows how SNP concordance relative to the “gold set” of SNPs from Levey et al vary as a function of sequence depth at the SNP position. Data were generated from 2 slides of long mate pair and .

Figure 6. Transcriptome Analysis Data Flow


BAM File
BAM Alignment Iterator
Pre-process models to Exon Collection
Refseq Exons
Paired Read Junction Finder
Single Read Junction Finder
Candidate Junctions
Evidence Evaluator
Regular Junctions
Alternatively Spliced Junctions
Fusion Junctions

This figure represents the data flow for paired end reads post mapping through the junction analysis for detection of known and novel transcript detection

Table 2. Validated Fusions

Exon 1	Exon 1 Chr	Exon 2	Exon 2 Chr	Exon distance
probable read through (adjacent genes)				
SDHC-2	chr1	LOC642502-3	chr1	12097
KLHL23-3	chr2	PHOSPHO2-2	chr2	3415
ARPC4-7	chr3	TTLL3-2	chr3	6608
UBE2E2-3	chr3	UBE2E1-4	chr3	670255
ELAC1-2	chr18	SMAD4-2	chr18	72359
ZNF606-6	chr19	C19orf18-2	chr19	14004
MTG1-10	chr10	LOC619207-9	chr10	47606
PXMP2-2	chr12	PGAM5-2	chr12	24482
improbable read though (not adjacent)				
BAT3-4	chr6	SLC44A4-15	chr6	-214128
known read through				
SNHG3-RCC1-3	chr1	RCC1-2	chr1	12991
ANKHD1-35	chr5	ANKHD1-EIF4EBP3-1	chr5	3089
known read through				
BCAS4-1	chr20	BCAS3-24	chr17	N/A (different chromosome / strand)
BCR-14	chr22	ABL1-3	chr9	N/A (different chromosome / strand)
GAS6-14	chr13	RASA3-23	chr13	N/A (different chromosome / strand)
candidate in Maher set (HBR)				
C15orf38-5	chr15	AP3S2-2	chr15	14076
annotation issue (these exons are from the same gene)				
SMAGP-1	chr12	LOC57228-2	chr12	940

Of the 36 fusions testsed,16 (44%) validated against a TaqMan assay.

CONCLUSIONS

The results for human resequencing and human whole transcriptome experiments demonstrate the accuracy of the SOLiD^{TM} Sequencing System for mammalian sized genomes. The accuracy enables confident SNP calls at lower coverage.

REFERENCES

1. Levy et al (2007) The Diploid Genome Sequence of an Individual Human, PLOS Biology 5(10): e254

2. Maher et al (2009) Transcriptome sequencing to detect gene fusions in cancer, Nature 458(7234):97-101

ACKNOWLEDGEMENTS

We thank Yulei Wang, Benjamin Kong and Caifu Chen (Life Technologies) for designing and running the TaqMan validation of fusion transcripts.

TRADEMARKS/LICENSING/LEGAL

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