P12.045 Detection of low frequency variation in heterogeneous samples using the accuracy and sensitivity of a SOLiD™ System

Martin Storm[2], Rizza Padilla[2], Quynh Doan[2], Jeffrey Ichikawa[2], Michael Rhodes[2], Yongming Sun[2], Larry Joe[2], David W. Craig[1], Jessica Aldrich[1], Alexis Christoforides[1], Darrin Taverna[1], Tracy Moses[1], John D. Carpten[1], David Duggan[1], and Fiona Hyland[2]

1. The Translational Genomics Research Institute (TGen) 445 N. Fifth Street, Phoenix, AZ 85004;
2. Life Technologies, 850 Lincoln Center Drive. Foster City, CA. 94404

ABSTRACT

Cancers often display biological and molecular heterogeneity with a variety of genetic abnormalities. Specific genetic mutations can alter the course of disease progression, metastasis, and drug resistance. Some mutations in cancer are prevalent at a low frequency in clinical samples, often due to contamination of tumor samples with normal alleles from adjacent non-malignant cells. The identification of genetic variants resulting from sub-populations is important to better understand the causal biology of cancer. Second-generation sequencing can identify low frequency alleles resulting from cellular sub-populations. Here we describe the reliable detection of alleles occurring at very low frequencies in heterogeneous samples by utilizing the accuracy of the SOLiD™ System. We demonstrate that the SOLiD™ System with Exact Call Chemistry (ECC) provides a highly accurate approach to detecting low frequency variants in cancer research and other applications.

INTRODUCTION

Numerous studies have proven that early detection is critical to treating cancer effectively. For cancer diagnostics this is a challenge since early detection can require identification of a very small population of cancer cells from a large number of normal cells. For example, a tumor resection will often contain a complex mixture of both normal stromal cells and invasive highly mutated cells. This complexity and lack of consistent biological markers that can reliably predict disease phenotypes make the use of traditional diagnostic methods less practical. Next-generation sequencing platforms with genomic enrichment technologies allow the ability to examine the entire genetic variability of a specific genomic area of interest. The SOLiD™ System with Exact Call Chemistry (ECC) (figure 1) provides a highly accurate approach to detecting low frequency variants in heterogeneous populations.

We investigated the ability to detect lower frequencies variants using targeted amplification using PCR-based enrichment and sequencing on the SOLiD™ 4 System. For experiment 1 in collaboration with Tgen 10 HapMap individuals were pooled at equimolar concentrations (20 ploidies) and amplified 2,545 regions across a series 4 major sites. We then assessed the ability to identify low frequency variants down to 5%, as is the case when only 1 individual is heterozygote a for a known SNP. In experiment 2, we assessed the ability of the system to detect extremely rare variations, down to 0.1%, by pooling a known HapMap sample in to a background of Huref DNA.

MATERIALS AND METHODS

Enrichment, library preparation, and sequencing

Enrichment of target sequences was performed according to the RainDance RDT 1000 Sequence Enrichment Assay Manual using 2 µg of gDNA. Droplets were collected in a 0.2 mL PCR tube and amplified using 55 cycles of PCR. Amplification products were recovered by breaking the emulsions, followed by amplicon purification. Samples were then concatenated as described in the Applied Biosystems® SOLiD™ System Amplicon Concatenation Protocol. Standard fragment libraries were generated in accordance with the Applied Biosystems® SOLiD™ 4 System Library Preparation Guide.

Enriched libraries were prepared according to the Applied Biosystems® SOLiD™ 4 System Templated Bead Preparation Guide. Sequencing by ligation was carried out on the SOLiD™ 4 Analyzer in accordance with the Applied Biosystems® SOLiD™ 4 System Instrument Operation Guide. Analysis was performed using BioScope ™ 1.3, VarScan (1) and DiBayes .

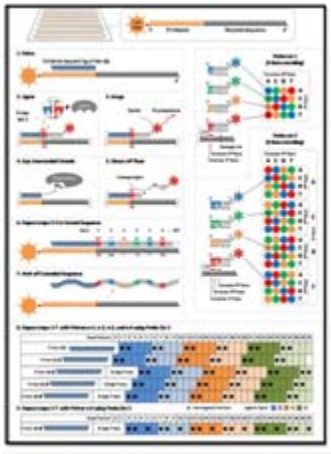


Figure 1. Ligation Based Sequencing and ECC module

Experiment 1.
Detection of SNV in pooled heterogonous samples:
A total of 1.3 Mb across several continuous regions of the genome were amplified from pooled genomic DNA of 10 HapMap individuals and sequenced using SOLiD™ v4

Table 1. Mapping characteristics of pooled heterogeneous samples

One objective of our study was to determine whether low frequency variants are readily detectable given enough coverage. Given that we have equimolar-pooled 10 individuals, we expect that SNPs will range in frequency between 5% and 100%, where a SNP only found as a heterozygote in one individual would be at 5% and a SNP homozygous in all 10 individuals relative to the reference will be 100%. As all these samples have been genotyped previously it is possible to predict for known SNPs what the frequency in the pool is expected to be. The relationship between the expected frequencies and observed frequencies is shown in figures 2a and b.

The ability to detect rare variants will depend on the accuracy of the platform. The spectrum of base-space quality scores is shown in figure 3, where 82% of mapped bases had a base-space quality score above 30, as well as the sequence coverage over the regions of interest shown here in figure 4.

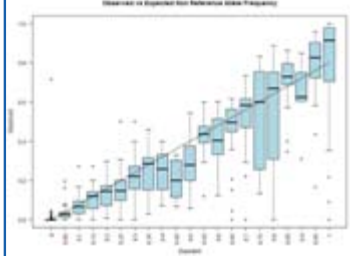


Figure 2a. The hapmap SNP genotype information was obtained from the Broad Institute HapMap Database. The expected frequency is calculated from the 10 hapmap individuals (2 individuals have some missing genotypes, resulting in some increase in variance). The observed the frequency is the number of non-reference allele divided by the coverage for that position.

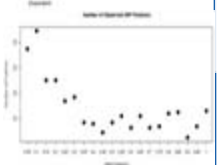


Figure 2b. Number of observations for each of the allele frequencies. The expected allele frequency is calculated from hapmap genotype information.

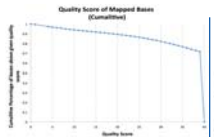


Figure 3. Percentage of mapped bases with base quality score above a threshold

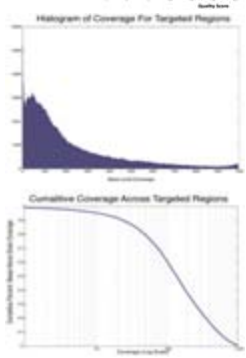
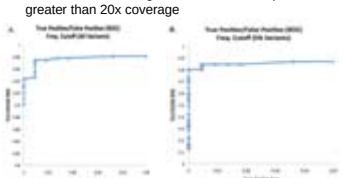


Figure 4. Per-base coverage of targeted amplicons. Overall, 91.8% of targeted bases were sequenced to greater than 20x coverage



The ROC curves shown in figures 5a and 5b , clearly show the ability to discriminate true positives from false positives. The subset of data for SNPS who were expected to be seen at 5% (heterozygotes in a single individual) show the ability of the SOLiD™ system to discriminate real variation from measurement errors

Experiment 2. Detection low frequency SNVs in normal background:
The objective of this experiment was to determine at which frequency we would be able to reliably detect the frequencies of non reference alleles in a background of genetic material. In order to investigate this we set up a dilution experiment where two libraries were pooled at known concentrations to create known allele frequencies. In this experiment we pooled down to 0.1% of the non reference allele in 99.9% of reference background

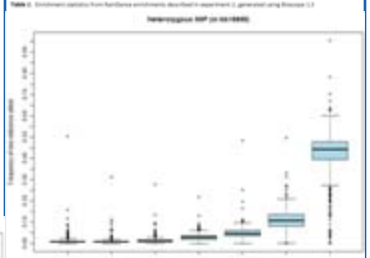


Figure 6. Boxplot of non-reference allele frequency for heterozygous SNP. A total of 613 heterozygous SNP positions are found for NA18858 within the target regions that are not shared with huref sample. The y-axis is the observed allele frequency that is calculated from number of reads showing the non-reference allele divided by the coverage for that position. The x-axis denotes the content of NA18858 in the mixed samples.

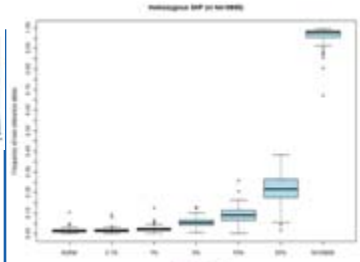


Figure 7. Boxplot of non-reference allele frequency for homozygous SNP. A total of 112 homozygous SNP positions are found for NA18858 within the target regions that are not shared with huref sample. The y-axis is the observed allele frequency that is calculated from number of reads showing the non-reference allele divided by the coverage for that position. The x-axis denotes the content of NA18858 in the mixed samples.

5500 Genetic Analyzer and LifeScope Software

ECC data is supported by the SOLiD 5500 system, and is analyzed with LifeScope Software.

CONCLUSIONS

We have described how we used the unique chemistry and resulting high accuracy of the SOLiD™ system to determine the allele frequency of rare alleles in a background of reference genetic content in targeted resequencing experiments. We have demonstrated feasibility in both a highly heterogeneous sample (experiment 1) as well as the identification of specific allele frequencies in a dilution series, containing a high amount of reference background (experiment 2). We have also demonstrated the ability to call alleles at very high coverage of the targeted areas. This work demonstrates early feasibility for low frequency variant detection in controlled samples. Ongoing improvements to the workflows, sequencing accuracy and detection algorithms extend the promise of detecting extremely rare variations in biologically relevant samples.

REFERENCES

1. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England), 25* (17), 2283-5 PMID: 19542151
URL: http://varscan.sourceforge.net

ACKNOWLEDGEMENTS

We would like to thank the High-Performance Biocomputing Center of TGen for providing the clustered computing resources used in this study, this includes the Saguaro-2 cluster supercomputer, a collaborative effort between TGen and the ASU Fulton High Performance Computing Initiative. We would also like to thank Warren Tom and Rachel Fish of Life Technologies for generating the sequencing data described in experiment 2.

TRADEMARKS/LICENSING