

Fiona C.L. Hyland*, Susan Tang*, Jon Sorenson*, Heather Peckhan#, Stephen F. McLaughlin#, Joel A. Malek#, Swati Ranade#, Cisilya Kosnopo#, Kevin McKernan#, Gina Costa#, and Francisco M. De La Vega*, Applied Biosystems, *Foster City, CA, and #Beverly, MA, USA.

INTRODUCTION

Important applications of ultra high-throughput, next-generation sequencing with short reads include detection of SNPs, both homozygous and heterozygous; detection of indels and other genetic rearrangements; and rare variant detection by deep re-sequencing of mixtures or heterogeneous samples. Detection of rare variant or low-frequency SNPs is important in cases including:

- Cancer tissue where only a fraction of the cells contain a somatic mutation of interest
- Pooled DNA samples for genetic epidemiology studies
- Heterogeneous colonies (bacteria, etc.)

For reliably detecting heterozygosity at low coverage, or for detecting rare variants, a low error rate is important. The two-base encoding inherent in the SOLiD™ System enables built-in error correction, producing a low post-correction error rate (< 0.06%).

MATERIALS AND METHODS

Simulation of rare variant detection

We performed a simulation study to investigate the possibilities for detecting rare variants using next-generation sequencing platforms, with specific reference to the Applied Biosystems SOLiD™ System. Our model simulates the number of reads given p, the allele frequency in the population; the number of sample pooled; the error rate; the total number of reads; the variance of coverage; and the threshold for allele detection. We investigated the relative importance of various parameters in our ability to detect rare variants. We simulated a rare variant in the presence of sequencing error, with pooled samples, allowing for a sampling of the number of molecules that are selected for sequencing.

Simulation of heterozygotes

We took mapped SOLiD reads from sequencing of a haploid organism (in this example, *S. suis*), and at every 10th genome position, we simulated a heterozygous SNP. Keeping all existing error and low-quality reads, we took only the consensus reads and replaced 50% or 30% of these with an alternate allele. We characterized the ability of the system and algorithms to detect heterozygote positions as a function of coverage and allele ratio.

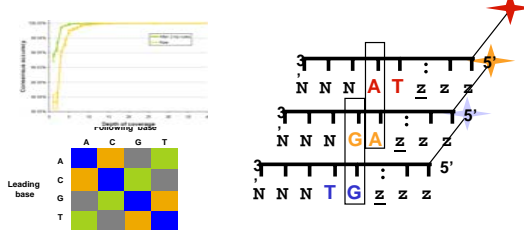
Sequencing of ENCODE region and detection of heterozygotes

We took one of the ENCODE regions (chr 4), and sequenced it on a Yoruba sample, using long-range PCR to amplify the segment. We took genotyping data from the HapMap web site and from dbSNP as independent annotation of heterozygote positions in this sample in this region to measure true positive and false negative rates of heterozygote detection. We serially randomly reduced the coverage of the SOLiD sequence data per genome position, to estimate the power to detect heterozygotes at various coverage levels. We measure the proportion of known heterozygous positions we can detect as a function of coverage.

CONCLUSIONS

Next generation sequencing has the potential to enable important applications in human genetics, including the detection of SNPs (both homozygous and heterozygotes), and detection of rare variants. We demonstrate that the SOLiD™ system has a very low rate of false positive heterozygote detection, < 10⁻⁵. We demonstrate the theoretical importance of low error rate in rare variant detection, and the observed ability of the SOLiD™ system to detect heterozygotes even with low coverage. In a sample with known genotypes, we detect 89% of known heterozygote SNPs.

2-base encoding produces low error rate



Low error enables detection of rare variants

Figure 1. Error rate is the main determinant of power to detect rare variants

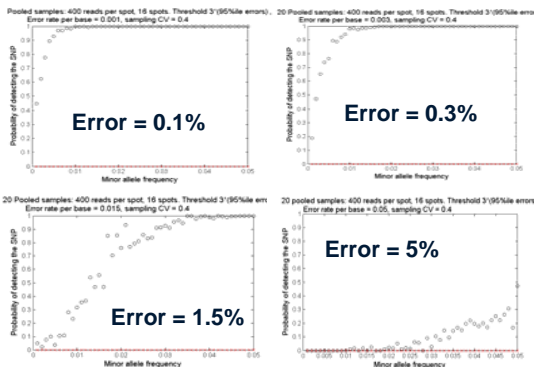
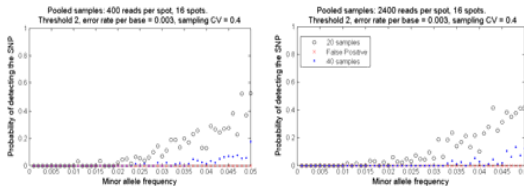


Figure 2. Further increasing coverage does not increase power to detect very rare variants, given a fixed error rate



Increasing the number of reads per sample does not change the ratio between error reads and SNP reads, and so constrains the ability to detect rare alleles even with high coverage, since the threshold for detecting an alternate allele is a function of the error rate.

TRADEMARKS/LICENSEING

Copyright © 2007 Applied Biosystems. Applied, Applied Biosystems, and AB (Design) are registered trademarks and SOLiD is a trademark of Applied Biosystems or its subsidiaries in the U.S. and/or certain other countries.

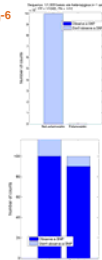
Purchase of this product alone does not imply any license under any process, instrument or other apparatus, system, composition, reagent or kit rights under patent claims owned or otherwise controlled by Applied Biosystems, either expressly or by estoppel.

Low error rate enables specific detection of heterozygous SNPs at low coverage

Figure 3. False positive rate is very low ~ 7*10⁻⁶

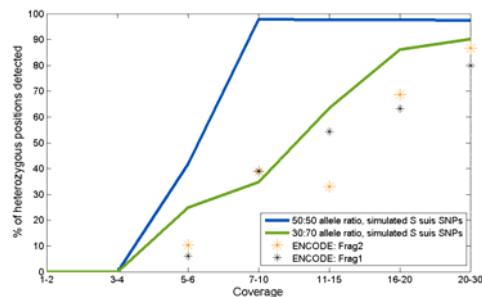
It is essential for a heterozygote detection algorithm to have a very low false positive rate, so that the false detection rate is acceptable. We sequenced a 2 Mb haploid genome, *S. Suis*, at 42x coverage, and tried to detect heterozygote SNPs in this genome.

- At p < 0.01, FP = 3 * 10⁻⁶ (6 het positions)
- At p < 0.05, FP = 7 * 10⁻⁶ (13 het positions)
- At p < 0.1, FP = 1 * 10⁻⁵ (26 het positions)
- All Hets: FP = 1.6 * 10⁻⁵ (31 het positions)



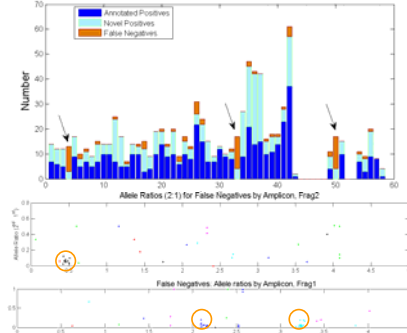
At a false positive rate of 10⁻³, half of all SNPs detected could be false positives.

Figure 4. Heterozygotes can be detected at low coverage



We simulated heterozygote SNPs against a background of real reads on a haploid genome (*S. suis* in this example), and we calculated the power of the system and algorithm to detect heterozygotes, as a function of coverage and exact allele ratios (solid lines). We took sequence data from ENCODE region (chr 4) on a Yoruba sample with known genotypes, and randomly sampled reads at various coverage levels, reporting the proportion of known heterozygotes that we detected (asterisks).

Figure 5. We detect 90% of annotated heterozygotes: Missing heterozygotes are in bad amplicons or have low coverage



We detect 352 annotated heterozygous SNPs. We miss 36 in bad amplicons, and 37 others (11%), mostly due to low coverage. We also detect 298 novel SNPs, including 49 detected multiple times in overlapping amplicons.