

Stephen F. McLaughlin<sup>1</sup>, Heather E. Peckham<sup>1</sup>, Zheng H. Zhang<sup>2</sup>, Swati Ranade<sup>1</sup>, Cisilya Kosnopo<sup>1</sup>, Gina Costa<sup>1</sup>, Joel Malek<sup>1</sup>, Jon M. Sorenson<sup>2</sup>  
 1. Applied Biosystems, 500 Cummings Center, Beverly, MA 01915  
 2. Applied Biosystems, 850 Lincoln Centre Dr, Foster City, CA 94404

### ABSTRACT

The next generation of DNA sequencing platforms produces sequencing reads with different qualities from the familiar data characteristics of Sanger-based automated DNA sequencing. Reduced read lengths and lower per-base accuracy have been compensated by significant increases in the available depth of coverage. The use of such reads for whole-genome resequencing requires a re-examination of previously solved algorithmic issues such as optimal alignment, consensus calling and the incorporation of quality metrics into raw and finished results. The Applied Biosystems SOLiD™ system (a massively parallel sequencing technology based on ligation of oligonucleotides) is the only next-generation system capable of utilizing 2-base encoding to significantly reduce the raw error rate. We have developed algorithms for the SOLiD™ system that utilize 2-base encoding as well as quality values and systematic error to improve upon the raw resequencing ability of short unpaired (<50bp) reads. By incorporating per-base quality values into the consensus calling we are able to successfully discriminate between false positives and true polymorphisms. These algorithms have been tested on several bacterial genomes using a variety of data sets from the SOLiD™ system. In addition, we show that the availability of highly parallel mate-paired reads allows increased mappability resulting in more accurate characterization of single-base changes as well as the detection of large-scale rearrangements and indels. By sequencing 57 long-range PCR products of a genotyped ENCODE<sub>1</sub> region for a Yoruban individual (Coriell ID, NA18507) with SOLiD™, we were able to validate our algorithm for heterozygous and homozygous human SNP detection.

### INTRODUCTION

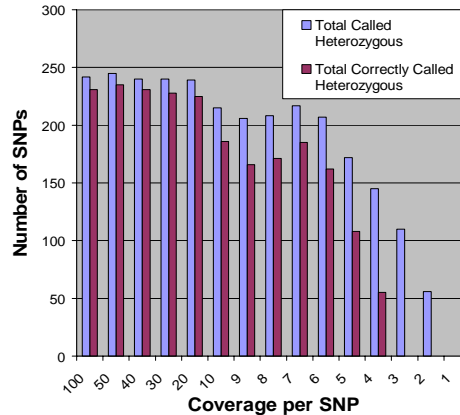
Two base encoding is uniquely enabled by the ligation-based sequencing protocol used in the SOLiD™ sequencing system. Because sequencing is carried out *via* sequential rounds of ligation, the decoding system is no longer limited to a single base being represented by a single color. In this system there are 16 dinucleotide combinations with 4 fluorescent dyes, each dye corresponding to a probe pool of 4 dinucleotides per pool. Using this dinucleotide, 4-dye encoding scheme (termed 2-base encoding or 2BE) in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions. In many resequencing projects one of the most important objectives is to measure Single Nucleotide Polymorphisms (SNPs) that may be responsible for differences in phenotype. Due to the fact that each base is measured twice, a single base change in base space leads to 2 changes in color space. Any color space change which contradicts this rule is considered invalid and is likely a measurement error. This feature of color space is very powerful in aiding SNP detection as it vastly reduces the error rate and improves consensus accuracy.

### MATERIALS AND METHODS

An algorithm was developed to detect single nucleotide polymorphisms (SNPs) from both fragment and mate-pair SOLiD™ reads. First, all of the tags are aligned to a reference sequence with enough allowed mismatches to account for at least 1 SNP per tag. Post matching, each genomic position covered by at least one tag is assessed for the number of different color-space changes which occur at that position. Positions which contain only tags which agree with the reference as well as invalid color-space changes are immediately filtered out. The remaining positions are considered SNP candidates and are rigorously evaluated employing a multi-dimensional thresholding scheme. Candidates which pass all of the threshold cutoffs are called either homozygous or heterozygous SNPs. Parameters with set thresholds include: an error-weighted and QV-weighted score, the number of tags with unique start points which cover each position, the observed allele ratio, and the ratio of valid to invalid color-space mismatches. Genotyped SNPs were downloaded from dbSNP<sub>2</sub> at NCBI as well as the International Hapmap Consortium<sub>3</sub>.

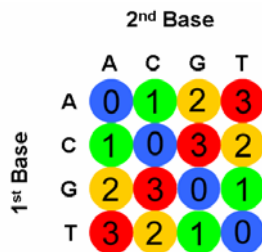
### RESULTS

**Figure 1. Heterozygous SNP Calling at Various Levels of Sequence Coverage**



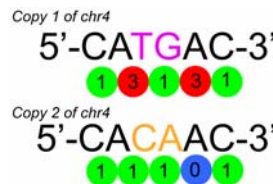
**Figure 1:** Heterozygous SNP-calling was assessed for a Yoruban sample (Coriell ID, NA18507) sequenced with SOLiD™. An ENCODE region of chromosome 4 was sequenced with high coverage and tags covering known, genotyped SNPs were randomly deprecated *in silico* from 100x down to 1x. Performance was assessed at each coverage level. Heterozygous SNPs called incorrectly were instead identified as homozygous.

**Figure 2. Color Space Changes, Coded 0-3**



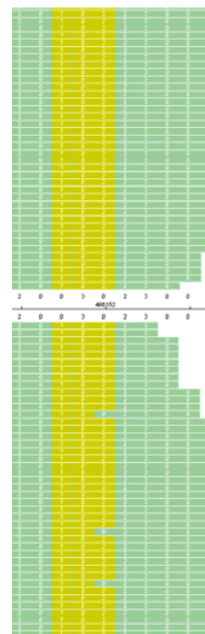
**Figure 2:** Each of the 16 dibase combinations is represented by one of 4 different-colored dyes which are in turn shared evenly between 4 dibase combinations. By convention, these are represented as numbers (shown above) which comprise the alignments of individual tags as illustrated in Figures 3 and 5.

**Figure 4. Illustration of Adjacent Heterozygous SNP Maternal/Paternal Color Space Patterns**



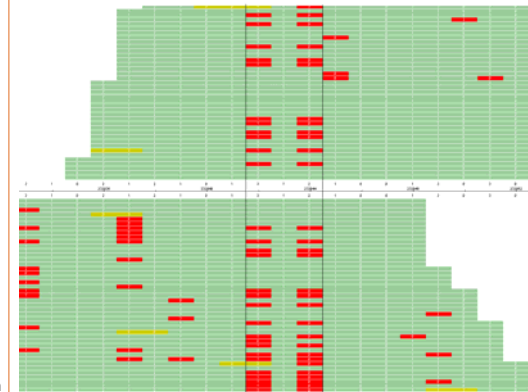
**Figure 4:** An example of the resultant color space changes for two adjacent, heterozygous SNPs detected by SOLiD™ via sequencing of long-range PCR products in NA18507 DNA. These SNPs were previously genotyped for this individual and we confirm they are correctly detected by SOLiD™. By investigating multiple alignment of SOLiD™ reads, a maternal and paternal pattern of color-space changes can be deduced (see Figure 6) and alleles for these two SNPs can be attributed to either the maternal or paternal copy of the chromosome.

**Figure 3. Adjacent Homozygous SNPs Detected by SOLiD™**



**Figure 3:** Multiple alignment of tags covering 2 adjacent known homozygous SNPs (rs4264892 and rs4446400) as displayed in the Solid Alignment Browser, a version of Apollo which has been modified to handle color space. This Yoruban individual (Coriell id, NA18507) has two adjacent homozygous differences compared to the chromosome 4, build 36 reference used to align these reads. As expected, there are 3 adjacent color space changes corresponding to the two adjacent SNPs and the uniformity of the pattern is that of homozygous SNPs.

**Figure 5. Adjacent Heterozygous Human SNPs Detected by SOLiD™ Display Maternal/Paternal Allelic Pattern**



**Figure 5:** Multiple alignment of tags covering 2 adjacent heterozygous SNPs detected by SOLiD™. The red blocks between the parallel lines represent valid color changes at positions of genotyped SNPs (rs2635348 and rs4834626) on chromosome 4. The genotypes are as expected for NA18507 (CT and AG for rs2635348 and rs4834626 respectively).

### CONCLUSIONS/FUTURE DIRECTIONS

We developed an algorithm for SNP detection using data generated from the SOLiD™ system. By sequencing long-range PCR products covering ~90% of a well sequenced 500Kb region of chromosome 4, and by choosing a region which has been extensively genotyped for the ENCODE<sub>1</sub> project specifically for our human sample (NA18507, Yoruban) we were able to validate that the algorithm successfully detects both heterozygous and homozygous SNPs. The heterozygous calls we made were always the correct alleles unless they were incorrectly identified as homozygous. Homozygous SNP calling was highly successful and performs just as well at 3x as at 100x coverage (data not shown due to space limitations). Future work will involve investigating the source of False Positives and validating any False Positives and False Negatives which we do identify via Sanger resequencing.

### REFERENCES

1. Feingold et. Al. The ENCODE (Encyclopedia Of DNA Elements) Project *Science* 22 October 2004: 636-640
2. ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1;29(1):308-11.
3. The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789-796

### ACKNOWLEDGEMENTS

A very special thanks to Georgia Giannoukos, Ph.D. at the Broad Institute, Cambridge MA for providing some of their ENCODE data as well as the primers used to amplify the long-range PCR products