Dibase sequencing allows accurate SNP detection at moderate and low coverage with diBayes algorithm

Hyland F.C.L., Wessel T., Scafe C.R., Yang C., Sakarya O., De La Vega F. M.

Applied Biosystems, Foster City, CA, USA

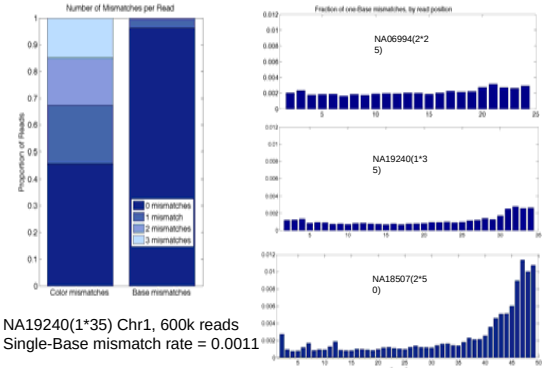
ABSTRACT

With the advent of next-generation sequencing, novel algorithms to detect heterozygous and homozygous variants from massive amounts of short-read data are needed. The dibase-coded oligonucleotides used by the Applied Biosystems SOLiD™ System allow distinguishing of SNPs and sequencing errors, and so allow sensitive and specific heterozygote detection at low coverage. We have developed diBayes, an algorithm incorporating pre-filtering followed by a Bayesian algorithm to detect SNPs in SOLiD reads; the algorithm includes a comprehensive error model including colorspace quality values and the prior probability of species heterozygosity. Colorspace quality values are predicted by training feature parameters, include image intensity, angle, and N2S, against a diverse set of annotated datasets. The QVs exhibit a linear relationship between observed and predicted phred-scale quality score, with a low variance. We also developed a tool to export SOLiD reads (including bases and base quality values) in SAM format, to facilitate analysis of SOLiD data by a growing number of SAM-compliant tools.

We evaluated the sensitivity and specificity of the diBayes algorithm using SOLiD reads from whole-genome sequencing of an African human at 20x coverage. 79% of the SNPs we detect are present in dbSNP. We compare the SOLiD genotype calls to HapMap genotypes (those that are the same in HapMap 1+2 and HapMap3); with a moderately aggressive setting of the algorithm and without using any prior information about SNPs in the reference sequence, we call 85% of heterozygotes at 9x coverage, 95% of heterozygotes at 17x coverage, and more than 99% of heterozygotes at 23x coverage. The heterozygote false discovery rate for HapMap SNPs is 8.5 * 10^{-4}. The concordance between HapMap homozygous calls and SOLiD homozygous calls is 0.9995.

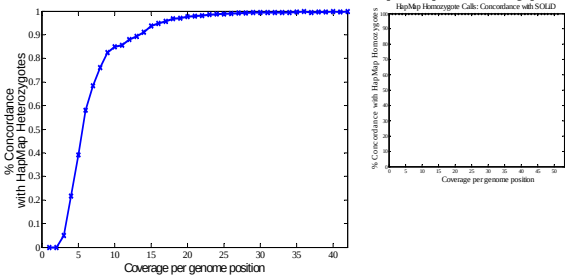
2% of SNPs are in exons. Both known and novel heterozygotes have a transition:transversion ratio of 68:32. Mitochondria have the highest SNP density, and sex chromosomes have lowest SNP density.

96% of beads have zero single-base mismatches



NA19240(1*35) Chr1, 600k reads
Single-Base mismatch rate = 0.0011

NA18507 18x: Concordance with HapMap Genotypes

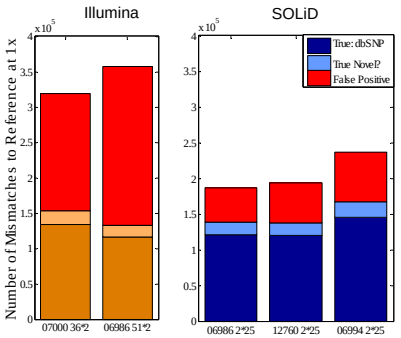


We compare the SOLiD diBayes genotype calls to HapMap (1+2 and 3) genotypes. By coverage per genome position, we call:

- 85% of heterozygotes at 11x
- 90% of heterozygotes at 14x
- 95% of heterozygotes at 17x
- >99% of heterozygotes at 26x

The concordance between HapMap and SOLiD homozygous calls is 0.9995

Most SOLiD non-reference alleles QV>30 are in dbSNP

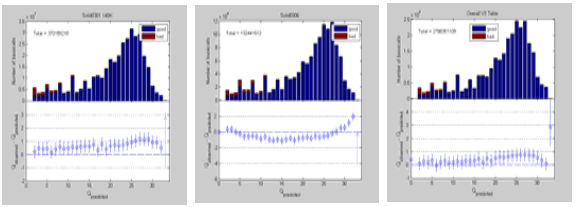


- 1000 Genome BAM files, Chr1 normalized to 1x coverage
- Mate Pair runs
- SOLiD has ~ 0.00021 non-reference alleles not in dbSNP at base QV > 30

At 5x, 48% of Hets and 95% of Homs are detected with low FP

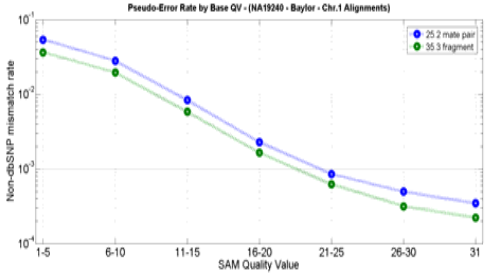
NA19240 Chr 1	5.3x	29x
# Positions with coverage	210,567,928	220,385,762
# SNPs	192,275	315,512
# Homs	111,161	110,950
# Hets	81,114	204,562
% SNPs in dbSNP	82.1%	78.6%
% Homs in dbSNP	89.7%	93.4%
% Hets in dbSNP	71.7%	70.5%
Concordance with HapMap Homozygotes (incl N)	0.95	0.9996
Concordance with HapMap Heterozygotes (incl N)	0.48	0.97
# Hets called by SOLiD	15796	32928
False Discovery Rate of Hets	0.0005	0.0007

Color Quality Values have high accuracy and discrimination



Colorcall quality value (QV) prediction is performed by querying the QV table with a set of feature parameter values, including signal, Noise2Signal, and angle. The QV table is produced by training feature parameters against a large set of annotated datasets. Training data is obtained from a variety of instruments, bead densities, read lengths, and species. This algorithm produces well calibrated QVs with high accuracy and discrimination on a range of data.

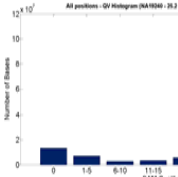
Base quality values linearly predict base accuracy



Base translations and base quality values for every uniquely-mapped read in a run are available in SOLiD SAM files. The base quality value algorithm uses color quality values and the base translation algorithm to predict the accuracy of each base call.

Base quality values allow true SNPs to be distinguished from false positives

QV for all bases
1 slide of NA19240 chr1 1000 Genome sample



Mismatch rate for bases with QV > 30:
SOLiD = 0.00021

QV for non-reference bases
1 slide of NA19240 chr1, 1000 Genome sample

Truncated at QV=31 (As released in Pilot 1)