DiBayes: A SNP Detection Algorithm for Next-Generation Dibase Sequencing

Susan Tang*, Fiona C.L. Hyland*, Thomas C. Wessel*, Jon Sorenson*, Heather Peckham, Francisco M. De La Vega***
Applied Biosystems, Foster City, CA* , and Beverly, MA, USA**

INTRODUCTION

With the advent of next-generation sequencing by ligation, there is a need to design algorithms that can establish variations between the sequenced genome and reference sequence. Because the SOLiD™ System uses a novel 2 base color-encoding scheme to better differentiate true sequence differences from error, data is produced in the form of color calls. We have developed algorithms for SNP detection on SOLiD sequencing data.

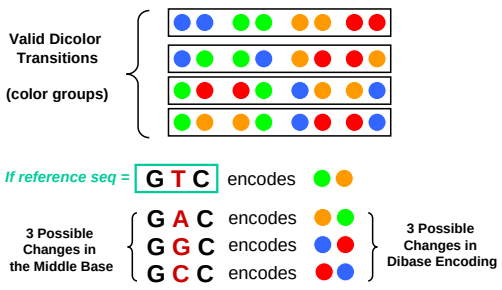


Figure 1. Effect of Base Substitutions in Colorspace

MATERIALS AND METHODS

In order to apply DiBayes to a SOLiD sequencing run, a set of colorspace reads mapped to a reference sequence is required. For each covered genome position and read that spans it, we record the observed dicolor call, the quality value of the dicolor call, and the position of the dicolor call on the read. In the case of a mate pair experiment, we also record whether the read is a forward or reverse tag. Using this set of information, in conjunction with QC reports generated from the SOLiD system, we are able to evaluate positions using preliminary evidence and distill all covered genome positions to a small subset of candidate heterozygous positions. Depending on the coverage density of the candidate heterozygous position, one of two SNP detection methods will be applied. We developed a Bayesian algorithm that formally incorporates prior probabilities of heterozygosity, error, and GC content. Its time-accuracy profile makes it ideal for low coverage reads. For higher coverage reads, we use a fast and accurate frequentist statistical method for SNP detection.

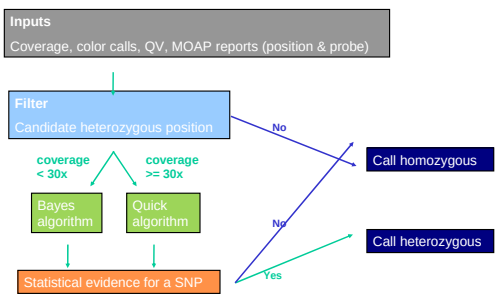


Figure 2. DiBayes SNP Calling Workflow

Filtering

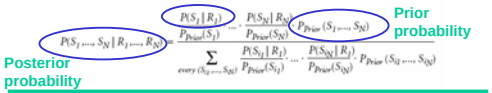
Filtering is necessary for both SNP calling accuracy as well as reducing the computation time of DiBayes. There are three different set of filtering settings that can be applied (stringent, moderate, and permissive). Relaxing the criteria for identifying potential heterozygotes provides the ability to call more rare variants and/or positions with noisier background, but at the expense of increased false positives.

Rule	Stringent	Moderate	Permissive
2nd allele must be a valid dicolor substitution	Yes	Yes	Yes
Read must be seen on both strands	Yes	No	No
Minimum coverage at this genome position	4	2	2
Coverage for the 2nd SNP allele	2	1	1
At least half of all reads should be one of the two candidate alleles	Yes	Yes	No
Both candidate SNP alleles were seen at more than one read position	Yes	Yes	No
2nd allele reads > 3rd allele reads	Yes	No	No
Number of unique start positions required	4	2	1

Figure 3. Screen for Candidate Heterozygosity

Bayesian Algorithm

Our implementation evaluates posterior probabilities for different permutations of dicolor calls that can be true for a given genome position and coverage. To compute such probabilities, we use known sources of error, GC content, expected polymorphism rate in the sample genome, as well as the observed color calls. Possible sources of error include imprecise 6mer probe annealing, reduced accuracy towards the end of a read, as well as low quality values associated with color calls. Permutations are grouped according to their reduced pair of dicolors, and a position is called heterozygous if the sum posterior probability for the best group exceeds a threshold value. Our implementation is a Bayesian approximation that evaluates probabilities for candidate permutations, as opposed to the exhausted set of possible dicolor permutations.



$P(S_1,...,S_N \mid R_1,...,R_N)$ = Posterior probability for a particular permutation of dicolors

$P(S_i \mid R_i)$ = Likelihood, or probability of seeing a particular dicolor in read given various error parameters

$P_{Prior}(S_1,...,S_N)$ = Prior probability of observing a specific permutation of dicolors (taking into consideration the expected polymorphism rate in these sequences)

$P_{Prior}(S_1)$ = Prior probability of observing a particular dicolor in cross-section

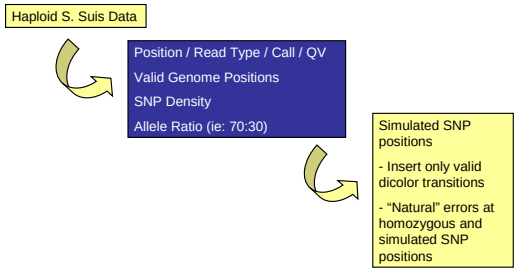
Figure 4. Bayesian Probability Inference for Dicolor Permutations

Frequentist Statistical Algorithm

This algorithm is based on the null hypothesis that a given position is homozygous, where only one allele exists and any other observations are errors. It fits a Poisson cumulative distribution function to the observed data and for each covered position, computes the probability of having the observed number of the second most popular allele under the null hypothesis. Similar to the Bayesian algorithm, known sources errors are factored into the probability calculation. A position is called as heterozygous if the probability exceeds a threshold value.

Data Simulation

We evaluated the accuracy of our algorithms with sequence data from a haploid organism (S. suis), using real reads and their respective quality values. Heterozygotes were simulated at every 10th genome position, with allele ratios of 90:10, 80:20, and 70:30.



RESULTS

Using simulated SNP data at various allele ratios (90:10, 80:20, 70:30) with permissive filtering, we evaluated the sensitivity, specificity, and false positive rate (FPR) of DiBayes. For all 3 allele ratios, the specificity of the algorithm is 100% and the FPR is 6.4 x 10^{-5}. We characterized sensitivity of the algorithm to be 47.0%, 86.0%, and 95.3% for allele ratios 90:10, 80:20, and 70:30, respectively.

At allele ratio 70:30, DiBayes can detect heterozygotes with a sensitivity of 98.8 % and false positive rate of 1.9 x 10^{-5} at >15x coverage. For positions with 6-15x coverage, we are able to detect heterozygotes with a sensitivity of 76.6 % and false positive rate of 2.4 x 10^{-4}.

Allele Ratio	Sensitivity	Specificity	FPR
70:30	0.953	1.000	6.4 x 10^{-5}

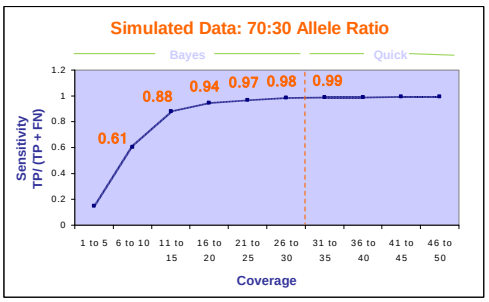


Figure 5. Effects of coverage on sensitivity using simulated data at 70:30 allele ratio

CONCLUSION

The low dibase error rate of SOLiD makes this next-generation sequencing platform particularly suitable for SNP detection. We have successfully designed and implemented a suite of algorithms that can accurately detect SNPs at low coverage, at disparate allele ratios, and with a very low false positive rate.