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.

MATERIALS AND METHODS

In order to apply DiBayes to a SOLiD sequencing run, a set of colorspace reads mapped to a reference sequence. Because 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.

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.

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 template 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.

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 three allele ratios, the specificity of the algorithm is 100% and the FPR is $6.4 \times 10^{-4}$. 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 98.4% and false positive rate of $1.9 \times 10^{-3}$ 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 $2.4 \times 10^{-4}$.

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.