bam-somaticsniper User Manual

David E. Larson and Christopher C. Harris

September 29, 2010

The purpose of this program is to identify single nucleotide positions that are different between tumor and normal. It takes a tumor bam and a normal bam and compares the two to determine the differences. It outputs a file in a format very similar to Samtools consensus format. It uses the genotype likelihood model of MAQ (as implemented in Samtools) and then calculates the probability that the tumor and normal genotypes are different. This probability is reported as a somatic score. The somatic score is the Phred-scaled probability (between 0 to 255) that the Tumor and Normal genotypes are not different where 0 means there is no probability that the genotypes are different and 255 means there is a probability of $1 - 10^{\left(\frac{255}{-10}\right)}$ that the genotypes are different between tumor and normal. This is consistent with how the SAM format reports such probabilities.

Usage

bam-somaticsniper [options] -f <ref.fasta> <tumor.bam> <normal.bam> <snv_output_file>

Required Option:

-f FILE *REQUIRED* reference sequence in the FASTA format

Options:

-q INT filtering reads with mapping quality less than INT [0]

-Q INT filtering somatic snv output with somatic quality less than INT [15]

-p FLAG disable priors in the somatic calculation. Increases sensitivity for solid tumors

-T FLOAT theta in maq consensus calling model (for -c/-g) [0.850000]

-N INT number of haplotypes in the sample (for -c/-g) [2]

-r FLOAT prior of a difference between two haplotypes (for -c/-g) [0.001000]

Notes on running bam-somaticsniper

Minimally, you must provide the program a reference fasta (passed with the -f option) a tumor bam, a normal bam, and the filename of the resulting output file. We recommend filtering out reads with a mapping quality of 0 (ie use -q 1) as they are typically randomly placed in the genome. We have also found that few variants with a somatic score less than 15 validate, but you may decrease the minimum score or increase it to a higher threshold (eg -Q 40). Disabling priors is not recommended, but may increase sensitivity at the cost of a decrease in specificity.

File Format

The file output by somatic sniper consists of line for all sites whose consensus differs from the reference base. Each line contains the following tab-separated values:

1. Chromosome

2. Position

3. Reference base

4. IUB genotype of tumor

5. Somatic Score

6. Tumor Consensus quality

7. Tumor SNV quality

8. Tumor RMS mapping quality

9. Depth in tumor (# of reads crossing the position)

10. Depth in normal (# of reads crossing the position)

Note that the IUB Genotype, Consensus Quality and SNV Quality Fields were identical to what samtools pileup -c outputs on the tumor data circa Samtools version 0.1.2-20. They are not composite values based on tumor and normal data.