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 (or, in theory, any two bam files). 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.
-J FLAG Use prior probabilities accounting for the somatic mutation rate
-s FLOAT prior probability of a somatic mutation (implies -J) [0.000001]
-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]
-F STRING select output format (vcf or classic) [classic]

Notes on running SomaticSniper

Minimally, you must provide the program the reference fasta the bams were aligned against (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 (i.e. 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 Formats

The output by SomaticSniper consists of line for all sites whose consensus differs from the reference base. Each of the three available output formats is described below

Classic:

Each line contains the following tab-separated values:

1. Chromosome
2. Position
3. Reference base
4. IUB genotype of tumor
5. IUB genotype of normal
6. Somatic Score
7. Tumor Consensus quality
8. Tumor variant allele quality
9. Tumor mean mapping quality
10. Normal Consensus quality
11. Normal variant allele quality
12. Normal mean mapping quality
13. Depth in tumor (# of reads crossing the position)
14. Depth in normal (# of reads crossing the position)
15. Mean base quality of reads supporting reference in tumor

16. Mean mapping quality of reads supporting reference in tumor
17. Depth of reads supporting reference in tumor
18. Mean base quality of reads supporting variant(s) in tumor
19. Mean mapping quality of reads supporting variant(s) in tumor
20. Depth of reads supporting variant(s) in tumor
21. Mean base quality of reads supporting reference in normal
22. Mean mapping quality of reads supporting reference in normal
23. Depth of reads supporting reference in normal
24. Mean base quality of reads supporting variant(s) in normal
25. Mean mapping quality of reads supporting variant(s) in normal
26. Depth of reads supporting variant(s) in normal

VCF

VCF output from SomaticSniper conforms to version 4.1 of the VCF specification. Hence, each non-header output line contains the following fields:

1. Chromosome
2. Position
3. ID (unused)
4. Reference base
5. Alternate bases (comma separated)
6. Quality (unused)
7. Filters (unused)
8. INFO (unused)
9. FORMAT specification for each sample
10. NORMAL sample data
11. TUMOR sample data

The following FORMAT fields will be populated for each of NORMAL and TUMOR.

ID	Number	Type	Description
GT	1	String	Genotype
IGT	1	String	Genotype when called independently (only filled if called in joint prior mode)
DP	1	Integer	Total read depth
DP4	4	Integer	# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases
BCOUNT	4	Integer	Occurrence count for each base at this site (A,C,G,T)
GQ	1	Integer	Genotype quality
JGQ	1	Integer	Joint genotype quality (only filled if called in join prior mode)
VAQ	1	Integer	Variant allele quality
BQ	.	Integer	Average base quality
MQ	.	Integer	Average mapping quality
SS	1	Integer	Variant status relative to non-adjacent normal: 0=wild-type, 1=germline, 2=somatic, 3=LOH, 4=unknown
SSC	1	Integer	Somatic Score