SomaticSniper User Manual

David E. Larson, Travis E. Abbott and Christopher C. Harris

October 26, 2011

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 $1 - 10^{(255/-10)}$ that the genotypes are different between tumor and normal. This is consistent with how the SAM format reports such probabilities.

There are two modes, the joint genotyping mode (-J) takes into account the fact that the tumor and normal samples are not entirely independent and also takes into account the prior probability of a somatic mutation. This probability can be scaled to control the sensitivity of the algorithm. An accurate value for this prior would be 0.000001, but this may result in a severe lack of sensitivity at lower depths. A less realistic prior probability will generate more sensitive results at the expense of an increase in the number of false positives. To get a similar sensitivity to the default mode, we recommend using a prior of 0.01. The default mode treats the two samples as if they came from two different individuals. This mode uses a less accurate mathematical model, but yields good results, especially if the normal may contain some tumor cells or the tumor is quite impure.

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]
-L FLAG do not report LOH variants as determined by genotypes
-G FLAG do not report Gain of Reference variants as determined by genotypes
-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) [0.01]
-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]

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). To obtain high confidence sites, we recommend also thresholding the minimum average mapping quality for the variant base to 40 for reads aligned with BWA or 70 for reads aligned with MAQ. We have not tested other aligners at this time. Disabling priors is not recommended, but may increase sensitivity at the cost of a decrease in specificity.

Basic filtering with provided Perl scripts

A small number of basic Perl scripts are included in the SomaticSniper package (located in src/scripts of the source code release) to aid in filtering out likely false positives. In order to get the recommended filtering you should do the following. Defaults are set assuming that BWA short is the aligner used. Other aligners have not been tested and recommendations are not available. Before proceeding you will need to obtain and compile bam-readcount (https://github.com/genome/bam-readcount). You will also need to generate a samtools pileup (not mpileup) indel file. Handling of indel containing VCFs is not implemented.

1. Filter on standard filters using the indel file. This will also remove LOH calls e.g. perl snpfilter.pl –snp-file your_sniper_file –indel-file your_indel_pileup

2. Adapt the remainder for use with bam-readcount e.g. perl prepare_for_readcount.pl –snp-file your_sniper_file.SNPfilter

3. Run bam-readcount (I'd recommend using the same mapping quality -q setting as you ran SomaticSniper with) e.g. bam-readcount -b 15 -f

your_ref.fasta -l your_sniper_file.SNPfilter.pos your_tumor.bam > your_readcounts.rc

4. Run the false positive filter e.g. perl fpfilter.pl –snp-file your_sniper_file.SNPfilter –readcount-file your_readcounts.rc
5. Lastly, run the "high confidence" filter which filters based on the Somatic Score and mapping quality e.g. perl highconfidence.pl –snp-file your_sniper_file.SNPfilter.fp_pass

Your final set of high confidence and highly filtered indels is now in the file your_sniper_file.SNPfilter.fp_pass.hc

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 joint prior mode)
VAQ	1	Integer	Variant quality
BQ	.	Integer	Average base quality of each base in the call, reported in alphabetical order (A,C,G,T)
MQ	1	Integer	Average mapping quality across all reads.
AMQ	.	Integer	Average mapping quality of each base in the call, reported in alphabetical order (A,C,G,T)
SS	1	Integer	Variant status relative to non-adjacent normal: 0=wildtype, 1=germline, 2=somatic, 3=LOH, 4=unknown
SSC	1	Integer	Somatic Score

User Support

Please first search Biostar (http://www.biostars.org) and then ask a question there if needed. We automatically monitor Biostar for questions related to our tools.