SomaticSeq Manual

Li Tai Fang

August 13, 2015

SomaticSeq

SomaticSeq is a flexible workflow that uses multiple somatic mutation callers to obtain a combined call set, and then use machine learning to distinguish true mutations from false positives from the call set. The manuscript is in preparation. The source code is deposited at *https://github.com/bioinform/somaticseq/*.

SomaticSeq.Wrapper.sh is a bash script that calls a series of scripts to combine the output of the somatic mutation caller(s), after the somatic mutation callers are run. Then, depending on what files are fed to SomaticSeq.Wrapper.sh, it will either train the call set into a classifier, predict high-confidence somatic mutations from the call set, or do nothing.

SomaticSeq.Wrapper.sh Commands

To train data set into a classifier

To create a trained classifier, ground truth files are required for the data sets. There is also an option to include a list of regions to ignore, where the ground truth is not known in those regions.

```
# -M/-I/-V/-v/-J/-S/-D/-U are output VCF files from individual
    callers.
# -i is also optional.
SomaticSeq.Wrapper.sh -M MuTect/variants.snp.vcf -I Indelocator/
    variants.indel.vcf -V VarScan2/variants.snp.vcf -v VarScan2/
    variants.indel.vcf -J JointSNVMix2/variants.snp.vcf -S
    SomaticSniper/variants.snp.vcf -D VarDict/variants.vcf -U MuSE/
    variants.snp.vcf -N matched_normal.bam -T tumor.bam -R
    ada_model_builder.R -g human_b37.fasta -c cosmic.b37.v71.vcf -d
     dbSNP.b37.v141.vcf -s $PATH/TO/DIR/snpSift -G $PATH/TO/
    GenomeAnalysisTK.jar -i ignore.bed -Z truth.snp.vcf -z truth.
    indel.vcf -o $OUTPUT_DIR
```

SomaticSeq.Wrapper.sh supports any combination of the somatic mutation callers we have incorporated into the workflow, so -M/-I/-V/-v/-J/-S/-D/-U are all optional parameters. SomaticSeq will run based on the output VCFs you have provided. It will train SNV and/or INDEL if you provide the truth.snp.vcf and/or truth.indel.vcf file(s).

To predict somatic mutation based on trained classifiers

```
# The *RData files are trained classifier from the training mode.
SomaticSeq.Wrapper.sh -M MuTect/variants.snp.vcf -I Indelocator/
    variants.indel.vcf -V VarScan2/variants.snp.vcf -v VarScan2/
    variants.indel.vcf -J JointSNVMix2/variants.snp.vcf -S
    SomaticSniper/variants.snp.vcf -D VarDict/variants.vcf -U MuSE/
```

```
variants.snp.vcf –N matched_normal.bam –T tumor.bam –R
ada_model_predictor.R –C sSNV.Classifier.RData –x sINDEL.
Classifier.RData –g human_b37.fasta –c cosmic.b37.v71.vcf –d
dbSNP.b37.v141.vcf –s $PATH/TO/DIR/snpSift –G $PATH/TO/
GenomeAnalysisTK.jar –o $OUTPUT_DIR
```

The Workflow

The SomaticSeq.Wrapper.sh calls a series of programs and procedures.