BS-SNPer User Guide

version 1.0, Jun 25 2015

1. Introduction

BS-SNPer is an ultrafast and memory-efficient package, a program for BS-Seq variation detection from alignments in standard BAM/SAM format using approximate Bayesian modeling.

2. System requirement

BS-SNPer works on Unix (Linux, Ubuntu, Mac OS, etc) based systems.

Hardware requirements

One computing node equipped with at least 10 GB Memory

Software requirements

GCC 4.6.0 or higher

Perl 5.16.3 or higher

zlib 1.2.8 or higher

3. Getting started

Installing

Download BS-SNPer from https://github.com/hellbelly/BS-Snper by clicking the button "Download ZIP". Run the commands below.

1. `unzip BS-Snper-master.zip`
2. `cd BS-Snper-master`
3. `sh BS-Snper.sh`

Make sure the executable files `rrbsSnp` and `chrLenExtract` are generated.

Usage

You can run BS-SNPer in Linux or MAC OS, using the command like:

```
perl BS-Snper.pl --fa <reference_file> --input <sorted_bam_file> --output <snp_result_file>
--methoutput <meth_result_file> --minhetfreq 0.1 --minhomfreq 0.85 --minquali 15 --mincover 10
--maxcover 1000 --minread2 2 --errorate 0.02 --mapvalue 20 >SNP.out 2>SNP.log
```

Attention

Both of the input and output file arguments should be passed to BS-SNPer in the form of absolute paths.

BS-SNPer requires a chromosome length file in the same folder as the reference genome file with name suffix '.len'. This file is automatically generated by BS-SNPer if it does not exist.

Options

--fa: Reference genome file in fasta format

--input: Input bam file

--output: Temporary file storing SNP candidates

--methoutput: CpG methylation information

--minhetfreq: Threshold of frequency for calling heterozygous SNP

--minhomfreq: Threshold of frequency for calling homozygous SNP

--minquali: Threshold of base quality

--mincover: Threshold of minimum depth of covered reads

--maxcover: Threshold of maximum depth of covered reads

--minread2: Minimum mutation reads number

--errorate: Minimum mutation rate

--mapvalue: Minimum read mapping value

SNP.out: Final SNP result file

SNP.log: Log file

4. Input file

Any alignments in standard sorted BAM/SAM format (see https://samtools.github.io/hts-specs/SAMv1.pdf for detailed information).

A bam file for evaluation is available at ftp://public.genomics.org.cn/BGI/BS-SNPer/example/

5. Output files

The output files include an SNP output file and a methylation output file.

The SNP output file has a tab-separated format with first 7 fields similar to VCF format (https://samtools.github.io/hts-specs/VCFv4.2.pdf):

1. **CHROM:** Chromosome.
2. **POS:** Coordinate.
3. **ID:** This field is currently not functional. When necessary, users could get the information from database like dbSNP.
4. **Ref:** Reference base(s). Each base must be one of A,C,G,T.
5. **ALT:** Alternate base(s).
6. **QUAL:** Phred-scaled quality score.
7. **FILTER:** Filter status. PASS if this position has passed all filters, i.e. a call is made at this position.
8. **GENOTYPE:** Genotype of this position.
9. **FREQUENCY:** Allele frequency.
10. **Number_of_watson:** The number of A,T,C,G in Watson strand.
11. **Number_of_crick:** The number of A,T,C,G in Crick strand.
12. **Mean_Quality_of_Watson:** Mean base quality of A,T,C,G in Watson strand.
13. **Mean_Quality_of_Crick:** Mean base quality of A,T,C,G in Crick strand.

The methylation output file has a tab-separated format same as MethylExtract (http://bioinfo2.ugr.es/MethylExtract/downloads/ManualMethylExtract.pdf):

1. **CHROM:** Chromosome.
2. **POS:** Sequence context most 5' position on the Watson strand (1-based).
3. **CONTEXT:** Sequence contexts with the SNVs annotated using the IUPAC nucleotide ambiguity code (referred to the Watson strand).
4. **Watson METH:** The number of methyl-cytosines (referred to the Watson strand).
5. **Watson COVERAGE:** The number of reads covering the cytosine in this sequence context (referred to the Watson strand).
6. **Watson QUAL:** Average PHRED score for the reads covering the cytosine (referred to the Watson strand).
7. **Crick METH:** The number of methyl-cytosines (referred to the Watson strand).
8. **Crick COVERAGE:** The number of reads covering the guanine in this context (referred to the Watson strand).

9. **Crick QUAL:** Average PHRED score for the reads covering the guanine (referred to the Watson strand).

6. Add-on script

An add-on script named "filterCG_SNP.pl" is provided to serve as a starting point for downstream applications. Using the SNP output (snp_result_file) as one of the input, the script separates the entries in the methylation output file (meth_result_file) into “CpG_meth_SNP_file” (those have been confirmed to be C>T SNPs) and “CpG_meth_filter_file” (the others).

You can run filterCG_SNP.pl in Linux or MAC OS, using the command like:

```
perl filterCG_SNP.pl <snp_result_file> <meth_result_file> <CpG_meth_filter_file> <CpG_meth_SNP_file>
```

The input files (snp_result_file and meth_result_file) are output files of BS-SNPer, which include an SNP output file and a methylation output file (see part 5).

7. Contact information

If you have any problem please do not hesitate to contact:

gaoshengjie@genomics.org.cn

zoudan_001@foxmail.com