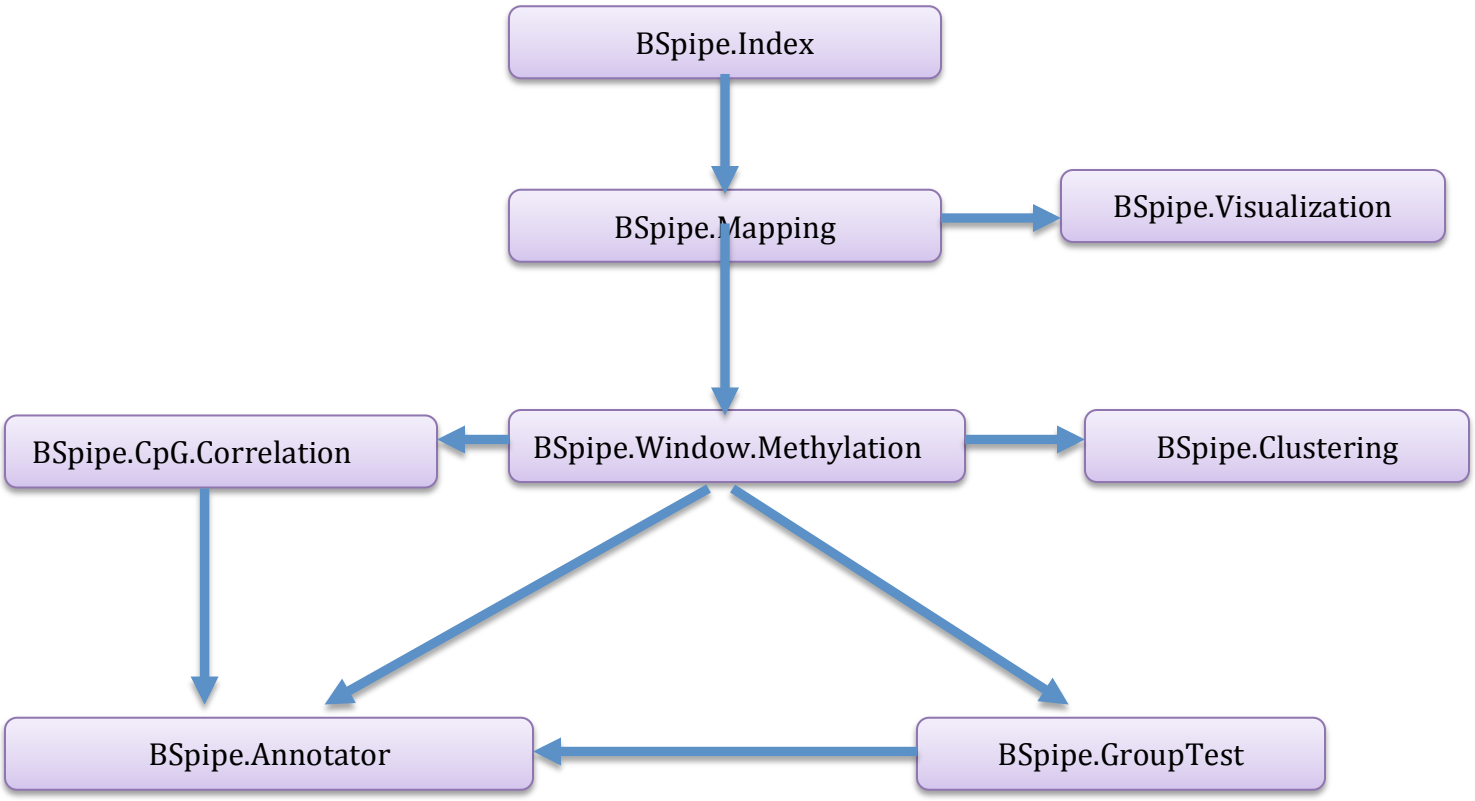
BSpipe.Mapping Documentation

Description: Performs mapping of bisulfite data
Author: GHSU COMICS
BSpipe Version: 1.0

Bisulfite sequencing is a powerful technique to study DNA cytosine methylation. Bisulfite treatment followed by PCR amplification specifically converts unmethylated cytosines to thymine. Coupled with next generation sequencing technology, it is able to detect the methylation status of every cytosine in the genome. BSpipe is an efficient bisulfite mapping software. It comes with a couple of utilities that includes mapping, annotation, calculation of CpG correlation etc.

The following is the flowchart of use of utilities in BSpipe package.



BSpipe.Mapping is specific utility under BSpipe software for mapping bisulfite reads to large genomes.

Parameters:

Name	Type	Description
Input.fastq.file	fq / fq.gz	Input fastq file with reads. This option is used only when single file needs to be mapped.
Read.type	text	Use this option only if Input.fastq.file has been provided. If the reads are not paired , choose Single/Pair1. If sample configuration file has been provided, leave blank.
Sample.configuration	tab delimited txt (.conf)	Refer next section (for multiple I/P files)
Mapping.configuration	tab delimited txt (.conf)	Refer next section
Reference.configuration	tab delimited txt (.conf)	Refer next section

Name	Type	Description
Error	float	Error rate of mapped reads [default- 0.05]
Minimum.mapping.quality	int	Minimum mapping quality score [default -0]
Coverage	int	Minimum read coverage for CpGs [default -1]
RRBS.configuration.file	tab delimited txt (.conf)	Refer next section
RRBS.name.in.RRBS.conf.file	text	RRBS name in RRBS configuration file.
Sample.file.directory	directory	Base directory for sample files
Reference.file.directory	directory	Base directory for reference fasta files
Minimum.base.quality	int	Minimum base quality for cytosines to call methylation [default -20]

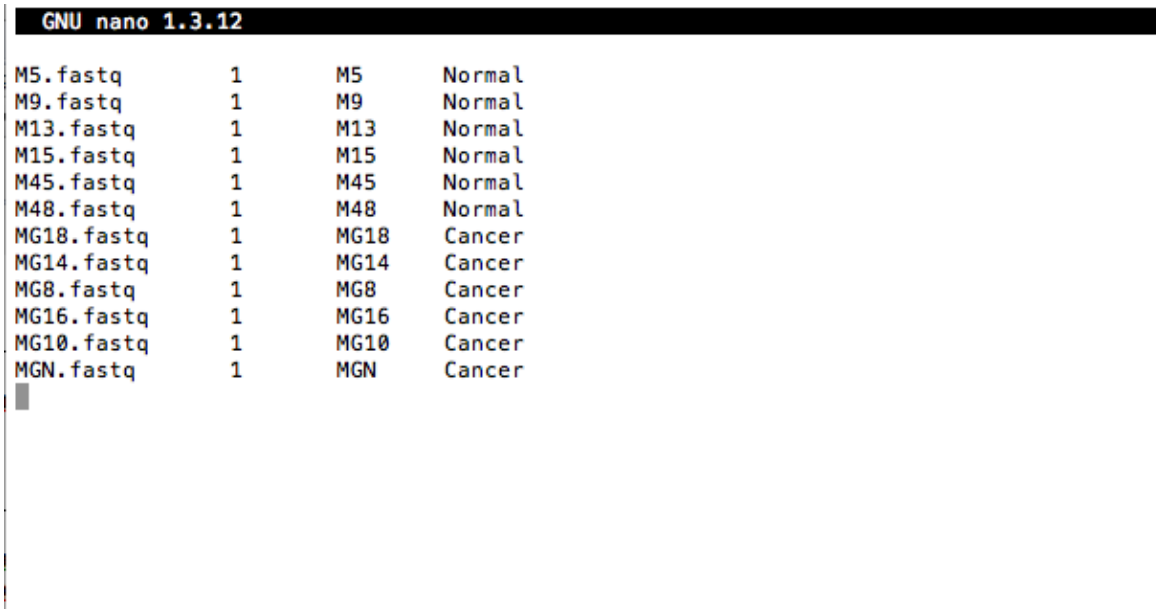
Nucleotides.to.follow.cytosine	text	Nucleotide sequence that follows cytosine. By default the value is G:H. For plants or stem cells, the sequence is G:HG:HH.
Number of sequences	int	Number of sequences to be used to calculate quality score type [default – 1000]
Threads	int	Number of worker threads [default – 1]

Configuration files:

- **Sample Configuration file**

  This option is used when multiple samples with groups have to be mapped. The sample conf file is a tab-delimited file in the following format:
  Column 1: Full path of input fastq file
  Column 2: File type (either 1 / 2 if paired file or 1 if single reads)
  Column 3: Sample Name
  Column 4: Sample Group Name

A snapshot of sample configuration file is provided below.



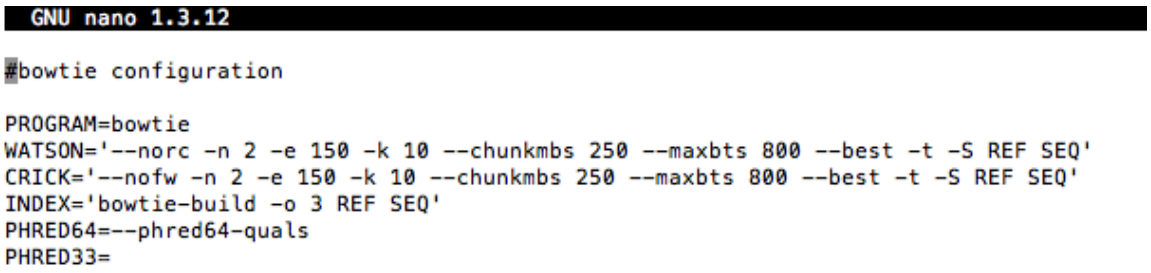
- **Mapping Configuration file**

  This file saves information and parameters for the mapping program to be used (either BWA or BOWTIE or SOAP2). The rows in the file are as follows (a sample bowtie configuration)

  # bowtie configuration

  PROGRAM=bowtie
  WATSON=*parametes in quotes*
  CRICK=*parameters in quotes*
  INDEX=*parameters in quotes*
  PHRED64= --phred64-quals
  PHRED33=

The above format makes sure different parameters be set for WATSON strand and CRICK strand. A snapshot of the mapping configuration file for bowtie is shown below.



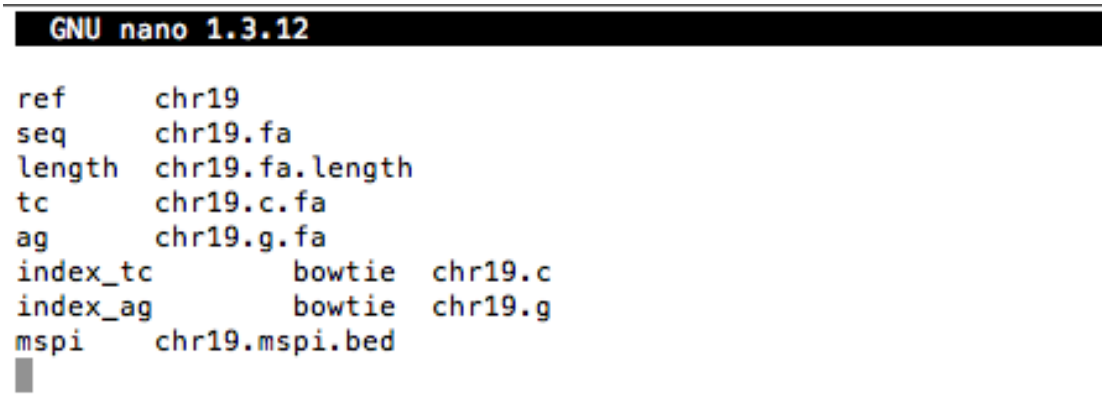
N.B: For convenience, conf files for all 3 mapping programs are available with the uploaded scripts in genepattern.

- **Reference Configuration file**

This file saves information of the reference file. For convenience, BSpipe.Index can be run with the reference fasta file, which creates all necessary files and a reference configuration file.

The file is organized as below:
ref *reference name*
seq *reference fasta (full path)*
length *reference fasta.length (full path)*
tc *reference c.fasta (full path)*
ag *reference g.fasta (full path)*
index_tc *program name* *bsbowtie reference.c prefix*
index_ag *program name* *bsbowtie reference.g prefix*
mspi *path to mspi.bed file.*

A sample reference configuration file for hg19 is shown below.



- **RRBS Configuration file**

  This file stores information about the rrbs file. The first column represents rrbs name, and the second column represents the rrbs file.

- **Running jobs on SGE**

  Users can submit mapping jobs to the Sun Grid Engine (SGE), by including the SGE configuration file (optional). This option is not set by default. Administrators can edit the wrapper shell script to include the sge.conf file using the prefix '-c' in the command line.

  N.B: For convenience, sge.conf file is available with the uploaded scripts in genepattern, which can be edited by the administrators.

Output files:

For each input file in the sample configuration file, an alignment file in bam format is created with a bai file. 2 bed files, namely *cg.bed.gz and *ch.bed.gz are also created along with a bam summary file that summarizes the alignment statistics.