swDMR User Guide，version 1.0

For swDMR v1.0.7

Contents

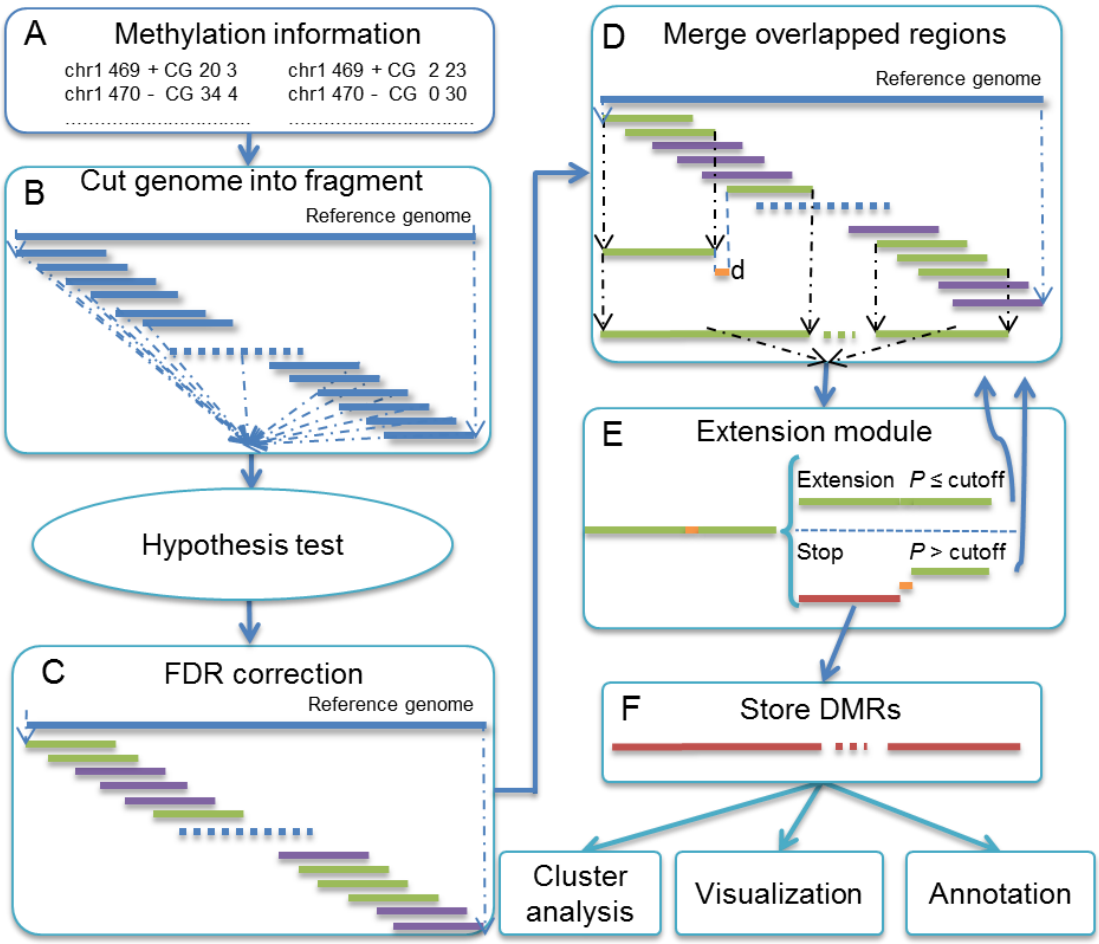
1 Introduction

swDMR is a sliding window approach, which is used to identify differentially methylated regions (DMR) from bisulfite sequencing dataset at single base resolution. The dataset, like whole genome bisulfite sequencing (WGBS) or reduced representation bisulfite sequencing (RRBS) with the same coverage region of samples, are suitable for swDMR. This software integrated several useful statistics methods, satisfying to two or multiple samples test. In addition, swDMR provides genomic features annotation with BED or GFF file. It can also produce WIG format file to upload to UCSC genome browser [1] for DMR visualization.

This software and example dataset are available to download from:

http://122.228.158.106/swDMR

2 Workflow of swDMR



3 Prerequisites

To run swDMR, you will need:

1. System: Linux or UNIX.
2. Software: Perl v5.8.8 or later， R 3.0.2 or later and BEDTools [2] are implemented.

4 Quick Start

4.1 Download swDMR

Download swDMR from: http://122.228.158.106/swDMR/latest/. Or, you can download from swDMR project : http://sourceforge.net/projects/swdmr/.

4.2 Install

tar xzvf swDMR-x.x.x.tar.gz
cd ./swDMR-x.x.x
sh install

4.3 Download input files for test

Download test data from: http://122.228.158.106/swDMR/data-test-hg18
The following dataset are BED similar format annotation files for hg18:
gene.bed
UTR_CDS_Intron_Upstream_Downstream.bed
cpgIslandExt.txt (from: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cpgIslandExt.txt.gz)

4.4 Run swDMR program in terminal

```
cd /path/of/swDMR
perl ./swDMR --help (get help information)
perl ./swDMR --samples Fibro.chr1.cout.gz,hESC.chr1.cout.gz --name Fibro,hESC --outdir example/ChiSquare --statistics ChiSquare --cytosineType CG --window 1000 --stepSize 100 --pvalue 0.01 --coverage 4 --fold 2 --diff 0.1 --fdr 0.05 --processes 10 --Rbin /usr/bin/R --chromosome 1 --position 2 --ctype 4 --methy 7 --unmethy 8 --annotation ./UTR_CDS_Intron_Upstream_Downstream.bed --CGI cpgIslandExt.txt --gene gene.bed --left 1000 --right 1000
```

5 Usage Details

5.1 Input data format

1. The methylation input file format:

The methylation information file must contain at least six columns as the following information separated by table "\t".

Chromosome	Position	Strand	Context of C	methylated_number	unmethylated_number
chr1	4	+	CHH	0	24
chr1	5	-	CHH	0	37
chr1	6	-	CHH	0	39
chr1	10	+	CHH	0	39
chr1	11	-	CHH	0	39
chr1	12	+	CHH	0	39
chr1	16	+	CHH	0	39

chr1	17	-	CHH	0	39

2. gene.bed file format:

Chromosome	start	end	geneName	strand	NM_number
chr1	58953	59871	OR4F5	+	NM_001005484
chr1	357521	358460	OR4F29	+	NM_001005221
chr1	357521	358460	OR4F3	+	NM_001005224
chr1	357521	358460	OR4F16	+	NM_001005277
chr1	610958	611897	OR4F29	-	NM_001005221
chr1	610958	611897	OR4F3	-	NM_001005224
chr1	610958	611897	OR4F16	-	NM_001005277
chr1	850983	869824	SAMD11	+	NM_152486
chr1	869445	884542	NOC2L	-	NM_015658
chr1	885829	890962	KLHL17	+	NM_198317

3. UTR_CDS_Intron_Upstream_Downstream.bed format:

Chromosome	start	end	element	strand	geneName
chr1	58954	59871	CDS	+	OR4F5
chr1	264954	266953	upstream	+	TAS1R3
chr1	357522	358460	CDS	+	OR4F29
chr1	750927	752926	upstream	+	LOC643837
chr1	848984	850983	upstream	+	SAMD11
chr1	850984	851043	5-UTR	+	SAMD11
chr1	851044	851164	intron	+	SAMD11
chr1	851165	851184	5-UTR	+	SAMD11
chr1	851185	851256	CDS	+	SAMD11
chr1	851257	855397	intron	+	SAMD11

5.2 Statistics approaches selection

Statistics method	Statistic model	Number of samples
T test	Parametric	Two
Wilcoxon test	Non-parametric	Two
Chi-square test	Non-parametric	Two
Fisher test	Non-parametric	Two
ANOVA	Parametric	Three or more
Kruskal-Wallis test	Non-parametric	Three or more

The statistics approaches implemented in current version of swDMR

5.3 Options

Basic swDMR options:

-h | --help print this help information on screen

-sam | --samples input samples' methylation files separated bycomma "," < a,b,c,...,g,h >

-na | --name input samples' names separated by comma "," < a,b,c,...,g,h >

-chp | --chromosome column of chromosome in your file, default <1>

-pos | --position column of position in your file, default <2>

-cp | --ctypecolumn of cytosine type in your file, default <4>

-mp | --methy column of methy reads in your file, default <5>

-up | --unmethy column of unmethy reads in your file, default <6>

-o | --outdir swDMR result directory

-co | --coverage lowest coverage of cytosine reads to use, default <4>

-s | --statistics choose one method to detect DMRs

< T_test||Kolmogorov||Fisher||ChiSquare||Wilcoxon||ANOVA||Kruskal >

if input samples == 2 --statistics should choose

< T_test||Kolmogorov||Fisher||ChiSquare||Wilcoxon >

if input samples >= 3 -statistics could choose

< ANOVA||Kruskal >

-CT | --cytosineType cytosine type < C|CG||CHG||CHH >

-w | --window a sliding window to statistics, default <1000>

-N | --points lowest number of selected type of cytosine in the window, default <10>

-z | --stepSize step size of the sliding processes, default <100>

-f | --fold max/min methylation level difference

-d | --diff value of max-min methylation level

-len | --length lowest length to join two fragment into one, default <100>

-V | --pvalue p value to judge as a DMR, default <0.01>

-R | --Rbin R bin

-fdr fdr to adjust DMR p value, default <0.05>

-pro | --processes parallel processes for DMR test and annotation, default <1>

Annotation and DMR maps:

-a | --annotation your annotation file should be bed of GFF format < BED||GFF >

-g | --Gene BED file of gene coordinates

--CGI cpgIsland file:

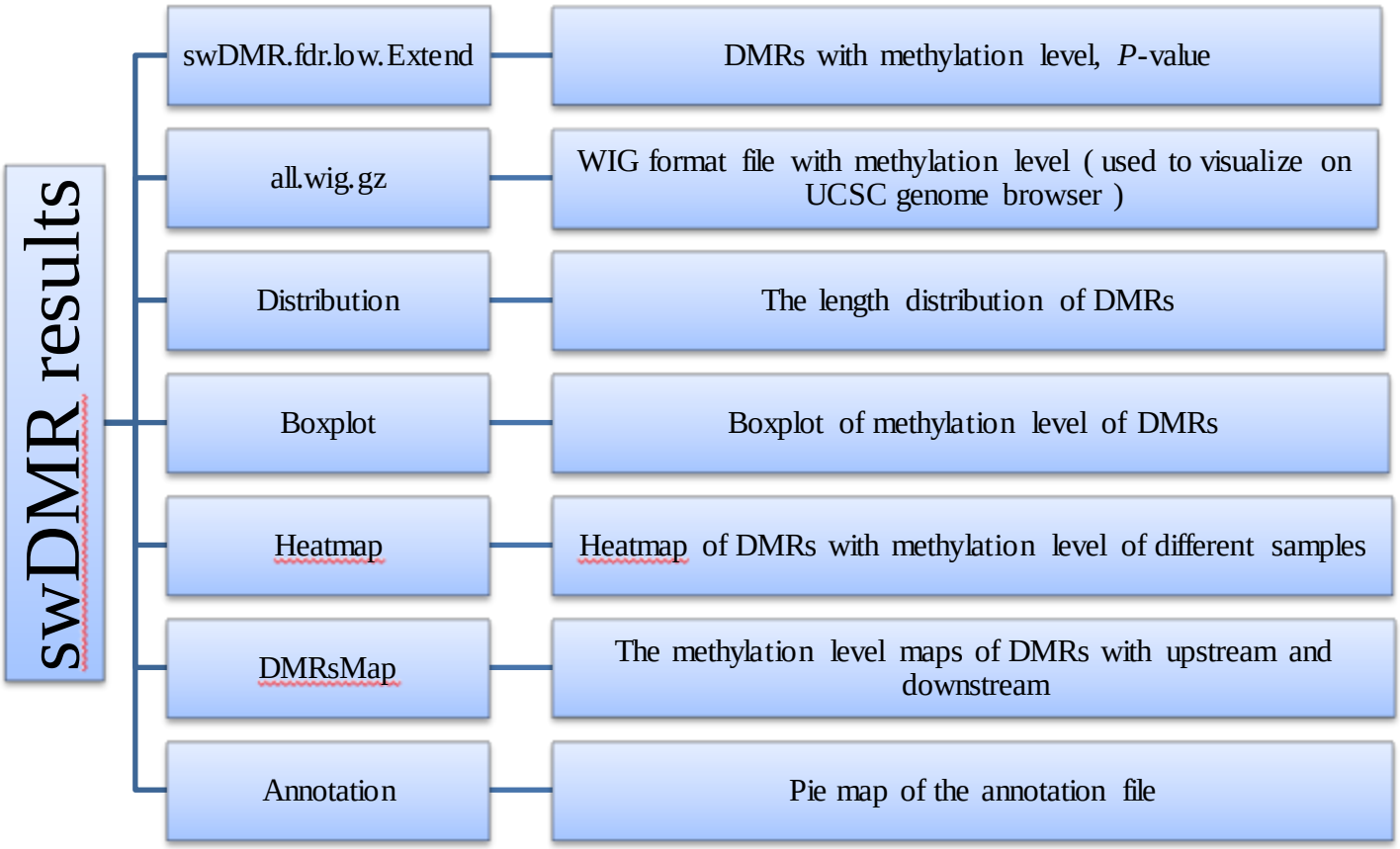
http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cpgIslandExt.txt.gz

The other release or species cpgIsland file can also be downloaded from UCSC genome browser

--left left side length of DMR in a map, default <1000>

--right right side length of DMR in a map, default <1000>

6 Output results



7 References

[1] http://genome.ucsc.edu/.

[2] Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

8 Connect for help

For any question on swDMR, please send email to xfliwz@gmail.com.