WGBS Analysis Package for Custom

Contents

1 WGBS installation and usage

1. Install the required software and configure the environment. (see the prerequisites document)

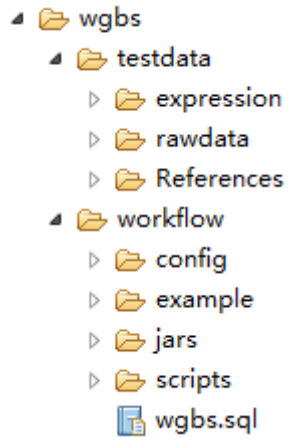
2. Download the wgbs_pipeline_custom.tar.gz from WBSA website (http://wbsa.big.ac.cn/download/download.jsp) put it under a given directory such as /home/test

Note: In the following paragraph we will use "/home/test" to make examples.

```
cd /home/test

tar xzvf wgbs_pipeline_custom.tar.gz
```

After uncompressed, you will see this directory structures like the following figure:



Note: The files in the "wgbs/workflow/example" use "/home/test" as relative directory.

Folder name	Description
testdata/expression	an expression file to test the "EXPRESSION" module in wgbs_config.xml
testdata/rawdata	**required** ,a source fastq file to test the package
testdata/References	**required** ,a test reference for test the package which need to be configured in the reference.xml
workflow/config	contains configuration template for the package
workflow/example	contains all the configuration files to test the package
workflow/jars	**required**, contains all the jars to run the package
workflow/scripts	**required**, contains all the scripts to run the package
wgbs.sql	**required** ,a database schema need to load into MySQL database

3. Import the MySQL script.

You need to login into Mysql database as root, by default the password is empty, so just click enter.

mysql -uroot -p

```
mysql> create database wgbs;
```

```
mysql>source /home/test/wgbs/workflow/wgbs.sql;

mysql> grant all on wgbs.* to ‘pipeline’@’localhost’ identified by ‘pipeline’
```

4. Run the WGBS module

Note: Before you run the following command, please ensure you have configured the files which are referred in section 2.

```
java –jar /home/test/wgbs/workflow/jars/workflow.jar    \

/home/test/wgbs/workflow/example/wgbs.properties    \

/home/test/wgbs/workflow/example/config
```

2 Configuring WGBS

The package uses four configuration files to store several parameters which needed by the WGBS pipeline; the configuration files list here from the “example” directory.

- workflow.properties
- mysql.properties
- wgbs.properties
- reference.xml

2.1 workflow.properties

Note: the parameters are identified by bold font you need to change to your real directory.

```
path.result=/home/test/wgbs/workflow/example/result

path.perl=/home/test/wgbs/workflow/scripts

path.reference=/home/test/wgbs/workflow/example/config/reference.xml

path.config=/home/test/wgbs/workflow/example/config



database.use=1

package.type=CUSTOM
```

This table lists all the parameters.

Table 1 parameters of workflow.properties

Parameter	Description	Example
path.result	Absolute path of the result directory	result
path.perl	Absolute path of the perl script directory	scripts
path.reference	Absolute path of the reference.xml file	reference.xml
path.config	Absolute path of directory of the configuration files	config
database.use	Whether or not to use database. 1: yes, 0: no. Necessary in this workflow.	1
package.type	Type of the cluster, we use CUSTOM in this workflow.	CUSTOM

2.2 mysql.properties

Note: If the Mysql database which you have installed is not located in your current machine, you need change "localhost" to the IP of that machine.

```
jdbc.driverClassName=com.mysql.jdbc.Driver

jdbc.url=jdbc\:mysql\://localhost/wgbs?user\=pipeline&password\=pipeline&useUnicode\=true&characterEncoding\=UTF-8&autoReconnect\=true&failOverReadOnly\=false
```

2.3 wgbs.properties

```
pipeline.type=wgbs

process=5

refId=1

rawDataType=single end
```

```
regElement=t-rich

tRichFile=/home/test/wgbs/testdata/rawdata/test_trich.fastq

aRichFile=

lambdaFile=

expressionFile=/home/test/wgbs/testdata/rawdata/expression/test.exp

teFile=

isTrimQV=1

isTrimAdaptor=1

minQualityValue=20

adaptorSeq= AGATCGGAAGAGC

minLength=35

minQualityValueForC=20

isFilterStartPoint=1

seedsLength=32

seedsMismatch=2

totalMismatch=4

pValue=0.005

jar.package=/home/test/wgbs/workflow/jars/wgbs.jar

jar.picture=/home/test/wgbs/workflow/jars/pipelinepicture.jar

jar.html=/home/test/wgbs/workflow/jars/pipelinehtml.jar
```

The following table describes how to configure the parameters in wgbs.properties

Table 2 parameters of wgbs.properties

Parameter	Description	Example
pipeline.type	The pipeline type of WGBS, **the value must be wgbs**	wgbs
process	Number of processes used during the processing.	5
refId	Reference ID for the data, which is included in the file reference.xml.	1
rawDataType	The type of raw data, single end or pair end.	single end|pair end
regElement	Regulate element of the raw data including: t-rich, a-rich or t-rich a-rich.	t-rich|a-rich|t-rich a-rich
tRichFile	Absolute path of the t-rich file.	trich.fastq
aRichFile	Absolute path of the a-rich file.	arich.fastq
isTrimQV	Whether or not to filter the low quality bases of the raw data. 1:yes, 0: no.	1|0
isTrimAdaptor	Whether or not to filter the adaptor sequences	1|0
minQualityValue	The user could trim low quality bases from two ends if the base quality value is less than a threshold	20
adaptorSeq	Adaptor sequences	AGATCGGAAGAGC
minLength	If the read length is less than a preset value after above trimming and filtering, the read will be discarded.	35
minQualityValueForC	The minimum quality value is a threshold, less than which the C base is not considered to be a methylcytosine.	20
isFilterStartPoint	Whether or not to remove the duplicated reads.	1|0

lambdaFile	Absolute path of lambda file.	lambda.fa
seedsLength	BWA parameter: seed length.	32
seedsMismatch	BWA parameter: maximum difference in the seed.	2
totalMismatch	BWA parameter: maximum differences in total reads.	4
pValue	p-value, if not choose to use lambda sequence to calculate it, users should offer it directly.	0.005
expressionFile	A file with gene information and expression value. #gene id #express value #chromosome #gene start position (count from 0) #gene end position (count from 1) #strand	
teFile	The upload TE data file format is as below: #species #TE_id #chromosome #strand #start position (count from 1) #end position (count from 1)	
jar.package	for WGBS pipeline, **the value must be wgbs.jar**	
jar.picture	to generate pictures, **the value must be pipelinepicture.jar**	
jar.html	to generate html result page, **the value must be pipelinehtml.jar**	

2.4 reference.xml

Note: The parameter which identified by bold font you need to change to satisfy your real demands, for the left parameters and the whole XML format you should not change it in order to run the package correctly.

If you have multiple references, just add the <reference> element as the example file.

```
<?xml version="1.0" encoding="UTF-8"?>

<references>

<reference id="1" name="Rice">

   <params>

      <param name="refDir">/home/test/testdata/References/rice/ref</param>

      <param
name="refC2TG2AFile">/home/test/wgbs/testdata/References/rice/Ref_C-T_G-A/ref_all_C-T_G-A.fa</param>

      <param name="genelistChromDir">/home/test/wgbs/testdata/References/rice/genes</param>

      <param name="repeatDir">/home/test/testdata/References/rice/repeats</param>

      <param name="goNumberFile">/home/test/wgbs
/testdata/References/rice/GO/rice_gene_GO.txt</param>

    <param
name="geneOntologyFile">/home/test/wgbs/testdata/References/rice/GO/gene_ontology_ext.obo</param>

    <param name="refCGIDir">/home/test/wgbs/testdata/References/rice/cpgislands</param>

    <param name="ideogramDir">/home/test/wgbs/testdata/References/rice/ideogram</param>

</params>

</reference>

</references>
```

Table 3 attributes of element <reference>

Attribute	Description
Id	ID of the species. **The value must be unique, it will be used by the parameter "refId" in the wgbs.properties**
Name	Name of the species.

Table4 parameters of a reference < reference >

Param	Description	Example
refDir	Absolute path of directory of raw reference files that must be standard .fa format and named chr*.fa.	/home/test/wgbs/testdata/References/rice/ref
refC2TG2AFile	Absolute path of reference file that has merged all the raw reference files and converted C to T and G to A. The file must have been indexed by BWA before used.	/home/test/wgbs/testdata /References/rice/Ref_C-T_G-A/ref_all_C-T_G-A.fa
genelistChromDir	Absolute path of gene file directory. The gene file is downloaded from UCSC. The genes on forward strand of each chromosome are stored in a file named chr*_C-T.gene. The genes on reverse strand of each chromosome are stored in a file named chr*_G-A.gene. #geneid #geneid #chromosome #strand #start #end #start #end #exon number #first exon start pos #second exon start pos #first exon end pos #second exon end pos All the start positions are count from 0 and the end positions are count from 1.	/home/test/wgbs/testdata /References/rice/genes
repeatDir	Absolute path of repeat file directory. It is downloaded from UCSC and in the	/home/test/wgbs/testdata

	standard format.	/References/rice/repeats
goNumberFile	File with GO information. Each line begins with a gene id followed by the GO number.	/home/test/wgbs/testdata /References/rice/GO/rice_gene_GO.txt
geneOntologyFile	Full ontology file, including cross-products, inter-ontology links, and has part relationships. It could be downloaded at http://www.geneontology.org	/home/test/wgbs/testdata /References/rice/GO/gene_ontology_ext.obo
refCGIDir	Absolute path of CpG islands file directory. It is downloaded from UCSC and in the standard format.	/home/test/wgbs/testdata /References/rice/cpgislands
ideogramDir	Absolute path of ideogram files of the species including: chromosomes.png, chromosomes.map.	/home/test/wgbs/testdata /References/rice/ideogram

3 Contact and Support

WGBS Analysis package is developed and maintained by Beijing Institute of Genomics(BIG), Chinese Academy of Sciences. If you have feedback or questions, please feel free to contact us at wbsa@big.ac.cn.