Reproducible Analysis of ChIP-Seq Data – GSE95632

Mengyuan Kan

October 15, 2018

Overview

Use ChIP-Seq data from SRA study SRP101282 (GEO study ID: GSE95632) as an example.

The goal of brocade pipeline [1] is to preform reproducible analysis of ChIP-Seq data:

- Download SRA .fastq data and prepare phenotype file
- Align reads to a reference genome
- Call peaks enriched in protein-DNA binding regsions
- Perform QC on aligned files
- Create a report that can be used to verify that sequencing was successful and/or identify sample outliers
- Perform differential binding analysis of reads aligned to genome according to a given reference genome
- Annotate genomic features to peaks
- Create a report that summarizes the differential binding resultss

Automatically generate LSF scripts in each step for HPC use.

Bioinformatics Tools

ChIP-Seq data analysis is performed on HPC. Directly use softwares that are already installed.

Check and load pre-installed softwares. For example:

```
module avail
module load Trimmomatic-0.32
```

Unintalled softwares or those that need re-configuration will be installed locally.

Raw reads process

trimmomatic

- usage: trim raw reads
- version: trimmomatic-0.32
- location: /opt/software/Trimmomatic/0.32/trimmomatic-0.32.jar

FastQC

Needs re-configuration. Use a customized contaminant_list.txt with updated adapter and primer sequences.

- usage: report reads quality
- version: FastQC v0.11.7
- location: /home/mengykan/.local/bin/
- local installation:

```
cd ~/softwares
wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.7.zip
unzip fastqc_v0.11.7.zip
cd FastQC/
chmod 775 fastqc
ln -s /home/mengykan/softwares/FastQC/fastqc /home/mengykan/.local/bin/
```

Important

> Prepare for the **contaminant_list.txt** in FastQC configuration with most updated adapter and primer sequences. We provide a list of sequences in **template_files/adapter_primer_sequences.txt**. Add the last two columns (i.e. Description and Sequence without header line) to the FastQC configuration file ∼/softwares/FastQC/Configu _list.txt. Users can also add the sequences from their own protocol.

Alignment

BWA

- usage: align and mapping sequencing reads
- version: 0.7.10-r789
- location: /opt/software/bwa/0.7.10/bin/bwa

```
ln -s /project/bhimeslab/STAR-2.5.2b/bin/Linux_x86_64/STAR /home/mengykan/.local/bin/
```

samtools

- usage: sort, index, statistics
- version: 1.8 (the current version has multi-thread option for .bam file index and sort)
- location: /home/mengykan/.local/bin
- local installation:

```
cd ~/softwares
wget https://github.com/samtools/samtools/releases/download/1.8/samtools-1.8.tar.bz2
tar xvjf samtools-1.8.tar.bz2
cd samtools-1.8
./configure --prefix=/home/mengykan/.local # local installation
make
make install
```

Bam file QC metrics

bamtools

- usage: manipulate bam files
- version: 2.3.0
- location: /opt/software/bamtools/2.3.0/bin/bamtools. Add bamtools library path /opt/software/bamtools/2.3.0/lib to LD_LIBRARY_PATH in .bashrc if it is not there, otherwise will get the error *error while loading shared libraries: libbamtools.so.2.3.0: cannot open shared object file: No such file or directory*

picard

- usage: stats of insertsize metrics in sequencing data
- version: picard-tools-1.96
- location: /opt/software/picard/picard-tools-1.96

Peak calling

MACS2

- usage: call peaks
- version: macs2 2.1.1.20160309
- location: /home/mengykan/.local/bin/macs2
- local installation:

```
cd /home/mengykan/softwares
wget https://files.pythonhosted.org/packages/9f/99/a8ac96b357f6b0a6f559fe0f5a81bcae12b98579551620ce07c5183aee2c/MAC\
S2-2.1.1.20160309.tar.gz
tar -zxvf MACS2-2.1.1.20160309.tar.gz
cd MACS2-2.1.1.20160309
python setup.py install --prefix /home/mengykan/.local # local installation
```

Differential binding analysis

- DiffBind R package usage: identify differential binding sites and visualize results
- version: 2.6.6
- location: /home/mengykan/.local/R-3.4/libs

local installation

1. set R local enviroment

add R enviromental path in .bashrc if it does not exist

```
export R_LIBS=/home/mengykan/.local/R-3.4/libs
```

2. install DiffBind

R

```r
source("https://bioconductor.org/biocLite.R")
biocLite("DiffBind")
```

Genomic feature annotation and visualization

- ChIPseeker R package usage: annotate genomic features to peaks
- version: 1.14.2
- location: /home/mengykan/.local/R-3.4/libs

local installation

1. set R local enviroment

add R enviromental path in .bashrc if it does not exist

```
export R_LIBS=/home/mengykan/.local/R-3.4/libs
```

2. install DiffBind

R

```r
source("https://bioconductor.org/biocLite.R")
biocLite("ChIPseeker")
```

Reference Genome

Human genome reference download

Download and create human reference genome hg38 if it does not exist.

```
cd /project/bhimeslab/Reference/hg38/
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
gunzip hg38.fa.gz
```

BWA index creation

Create reference genome index files for BWA if they do not exist.

```
mkdir /project/bhimeslab/Reference/hg38/bwa_index
bwa index -a bwtsw /project/bhimeslab/Reference/hg38/bwa_index/hg38 -p hg38
```

Five index files (hg38.amb, hg38.ann, hg38.bwt, hg38.pac, and hg38.sa) are created and will be used by bwa to map data

ChIP-Seq Pipeline

GitHub structure

- **pipeline_scripts:** Python scripts should be added in an executable search path.
- **template_files:** Rmd template and other files used in the pipeline. Put into a template directory specified as `template_dir`.
- **example_files:** output phenotype file, RMD scripts and output HTML report for this example.

> **Note**
> Edit **pipeline_scripts/chipseq_userdefine_variables.py** with user-defined variables before add it to an executable search path.

SRA download and fastqc

run command line

pipeline_scripts/chipseq_sra_download.py: download .fastq files from SRA.

Read in **template_files/chipseq_sra_download_Rmd_template.txt** from specified directory `template_dir` to create a RMD script.

Ftp addresses for corresponding samples are obtained from SRA SQLite database using R package SRAdb.

If .fastq files with the same names exist in the directory, skip downloading.

```
mkdir -p /home/mengykan/Projects/GSE95632/scripts/SRAdownload
cd /home/mengykan/Projects/GSE95632/scripts/SRAdownload
chipseq_sra_download.py --geo_id GSE95632 --path_start /home/mengykan/Projects/GSE95632 --project_name GSE95632 --t\
emplate_dir /home/mengykan/Projects/shared_files/ChIPSeq --fastqc
```

script options

The option `--pheno_info` refers to using user provided SRA ID for download which is included in the SRA_ID column in the provided phenotype file. If the phenotype file is not provided, use phenotype information from GEO. SRA_ID is retrieved from the field relation.1.

The option `--fastqc` refers to running FastQC for downloaded .fastq files.

submit LSF script

Generate LSF scripts for each download .fastq file. Submit LSF jobs on HPC that enables to run in parallel.

```
for i in *_download.lsf; do bsub < $i; done
cat SRR5309351_1_download.lsf
```

OUTPUT

```
#!/bin/bash
#BSUB -L /bin/bash
#BSUB -J SRR5309351_1_download
#BSUB -q normal
#BSUB -o SRR5309351_1_download_%J.out
#BSUB -e SRR5309351_1_download_%J.screen
#BSUB -M 36000
#BSUB -n 1
cd /home/mengykan/Projects/GSE95632/GSE95632_SRAdownload/
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR530/001/SRR5309351/SRR5309351.fastq.gz
fastqc /home/mengykan/Projects/GSE95632/GSE95632_SRAdownload/SRR5309351.fastq.gz -o /home/mengykan/Projects/GSE95632/
SRR5309351
```

output files

Output files are saved in `/home/mengykan/Projects/GSE95632/GSE95632_SRAdownload`:

- GEO phenotype `GSE95632_GEO_phenotype.txt`
- SRA download information `GSE95632_sraFile.info`
- The RMD and corresponding HTML report files:
  - `GSE95632_SRAdownload_ChIPSeqReport.Rmd`
  - `GSE95632_SRAdownload_ChIPSeqReport.html`
- Raw .fastq files downloaded from SRA
- FastQC results are saved under each sample folder

User-tailored phenotype file preparation

The sample info file used in the following steps should be provided by users.

required columns

- ’Sample’ column containing sample ID
- ’Status’ column containing variables of comparison state
- ’Antibody’ column containing antibody used for ChIP
- ’Tissue’ column containing tissue used
- ’Treatment’ column containing treatment conditions

- ’Input’ column containing sample ID used for DNA intput control that did not undergo ChIPSeq
- ’Peak’ column specify ’narrow’ or ’broad’ peak type.
- ’R1’ and/or ’R2’ columns containing full paths of .fastq files

For DNA intput controls, specify Antibody=Input, Input=NA, and Peak=NA.

other columns

’Disease’, ’Donor’ (i.e. cell line ID if the same cells were used in treated vs. untreated comparison), and ’protocol’ designating sample preparation kit information.

Index column

’Index’ column contains index sequence for each sample. If provided, trim raw .fastq files based on corresponding adapter sequences.

If use data from GEO, most GEO phenotype data do not have index information. However, FastQC is able to detect them as *Overrepresented sequences*. Users can tailor the ’Index’ column based on FastQC results. We provide a file with most updated adapter and primer sequences for FastQC detection.

In this example, we used a customized adapter file which can be found here **example_files/template_files/adapter_primer_sequences.txt**. The preparation step is shown in the section **Example of customized adapter and sequence file preparation**.

Sample file

An example phenotype file can be found here **example_files/GSE95632_Phenotype_withoutQC.txt**. Use this file in the following steps.

For demonstration, we only include wide-type cells with dex/EtOH treated.

SRA_ID	Sample	GEO_ID	Antibody	Treatment	Subject	Status	Tissue	Index	Input
SRX2609659	SRR5309351	GSM2519942	GR	EtOH	1D	GR_EtOH	ASM	AGTGAG	SRR5309352
SRX2609660	SRR5309352	GSM2519943	Input	EtOH	1D_Input	Input_EtOH	ASM	ACCTCA	NA
SRX2609661	SRR5309353	GSM2519944	GR	EtOH	1E	GR_EtOH	ASM	GCACTA	SRR5309352
SRX2609662	SRR5309354	GSM2519945	GR	Dex	2A	GR_dex	ASM	ACCTCA	SRR5309355
SRX2609663	SRR5309355	GSM2519946	Input	Dex	2A_Input	Input_dex	ASM	GTGCTT	NA
SRX2609664	SRR5309356	GSM2519947	GR	Dex	2E	GR_dex	ASM	GTGCTT	SRR5309355
SRX2609669	SRR5309361	GSM2519952	RNAP2	EtOH	5E	RNAP2_EtOH	ASM	GAGTCA	SRR5309352
SRX2609670	SRR5309362	GSM2519953	RNAP2	EtOH	5b	RNAP2_EtOH	ASM	AGCATG	SRR5309352
SRX2609671	SRR5309363	GSM2519954	RNAP2	Dex	6d	RNAP2_dex	ASM	CGTAGA	SRR5309355
SRX2609672	SRR5309364	GSM2519955	RNAP2	Dex	6e	RNAP2_dex	ASM	TCAGAG	SRR5309355

Table continued

Peak	R1
narrow	/path_to_file/GSE95632_SRAdownload/SRR5309351.fastq.gz
NA	/path_to_file/SRR5309352.fastq.gz

narrow	/path_to_file/SRR5309353.fastq.gz
narrow	/path_to_file/SRR5309354.fastq.gz
NA	/path_to_file/SRR5309355.fastq.gz
narrow	/path_to_file/SRR5309356.fastq.gz
narrow	/path_to_file/SRR5309361.fastq.gz
narrow	/path_to_file/SRR5309362.fastq.gz
narrow	/path_to_file/SRR5309363.fastq.gz
narrow	/path_to_file/SRR5309364.fastq.gz

Important

Column naming is rigid for the following columns: 'Sample', 'Status', 'Index', 'R1', 'R2', 'ERCC_Mix', 'Treatment', 'Disease', 'Donor', because pipeline scripts will recognize these name strings, but the column order can be changed.

Example of customized adapter and sequence file preparation

Although GEO phenotype provides barcode information for each sample, overrepresented sequences from fastqc results of raw .fastq files do not 100% match to any known Illumina adapters.

However, the pattern of overrepresented sequences are the same across different samples: i.e. GATCGGAAGAGCACACGTCAGAACTCCAGTCAC[index]ATCTCGTATGCC

The adapter sequence before index has one base different from Illumina truseq single adapter: GATCGGAAGAGCACACGTC[A]GAACTCCAGTCAC for the current study, while GATCGGAAGAGCACACGTC[T]GAACTCCAGTCAC for Illumina standard adapter sequences.

The adapter sequence after index is exactly a part of Illumina standard adapter sequences: ATCTCGTATGC for the current study, and ATCTCGTATGCCGTCTTCTGCTTG for Illumina truseq single adapter sequences.

Thus, create user customized sequences, and append these sequences to **template_files/adapter_primer_sequences.txt**. This file should be saved under **template_dir** to perform adapter trimming.

Type	Index	ID	Sequence
GSE95632	AGTGAG	GSE95632_1	GATCGGAAGAGCACACGTCAGAACTCCAGTCACAGTGAGATCTCGTATGCCGTCTTCT
GSE95632	ACCTCA	GSE95632_2	GATCGGAAGAGCACACGTCAGAACTCCAGTCACACCTCAATCTCGTATGCCGTCTTCT
GSE95632	GCACTA	GSE95632_3	GATCGGAAGAGCACACGTCAGAACTCCAGTCACGCACTAATCTCGTATGCCGTCTTCT
GSE95632	GTGCTT	GSE95632_4	GATCGGAAGAGCACACGTCAGAACTCCAGTCACGTGCTTATCTCGTATGCCGTCTTCT
GSE95632	GAGTCA	GSE95632_9	GATCGGAAGAGCACACGTCAGAACTCCAGTCACGAGTCAATCTCGTATGCCGTCTTCT
GSE95632	AGCATG	GSE95632_10	GATCGGAAGAGCACACGTCAGAACTCCAGTCACAGCATGATCTCGTATGCCGTCTTCT
GSE95632	CGTAGA	GSE95632_11	GATCGGAAGAGCACACGTCAGAACTCCAGTCACCGTAGAATCTCGTATGCCGTCTTCT
GSE95632	TCAGAG	GSE95632_12	GATCGGAAGAGCACACGTCAGAACTCCAGTCACTCAGAGATCTCGTATGCCGTCTTCT
GSE95632	Read1Primer	Multiplexing_Read_1_Sequencing_Primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
GSE95632	Read2Primer	Multiplexing_Read_2_Sequencing_Primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Align and obtain bam statistics

run command line

Run **pipeline_scripts/chipseq_align_and_qc.py** to: 1) trim adapter and primer sequences if index information is available, 2) run FastQC for (un)trimmed .fastq files, 3) align reads mapped to genes/transcripts, and 5) obtain various QC metrics from .bam files.

Edit **pipeline_scripts/chipseq_userdefine_variables.py** with a list user-defined variables (e.g. paths of genome reference file, paths of bioinformatics tools, versions of bioinformatics tools), and save the file under an executable search path.

If perform adapter trimming, read in **template_files/chipseq_adapter_primer_sequences.txt** from specified directory `template_dir` used as a reference list of index and primer sequences for various library preparation kits.

```
mkdir /home/mengykan/Projects/GSE95632/scripts/align
cd /home/mengykan/Projects/GSE95632/scripts/align
chipseq_align_and_qc.py --project_name GSE95632 --samples_in /home/mengykan/Projects/GSE95632/files/GSE95632__Pheno\
type_withoutQC.txt --ref_genome hg38 --library_type SE --index_type GSE95632 --path_start /home/mengykan/Projects/G\
SE95632 --template_dir /home/mengykan/Projects/shared_files/ChIPSeq
```

script options

The `--library_type` option refers to PE (paired-end) or SE (single-end) library.

The `--index_type` option refers to index used in sample library preparation. The index types provided in **template_files/adapter_primer_sequences.txt** are: truseq_single_index (TruSeq Single Indexes), illumina_ud_sys1 (Illumina UD indexes for NovaSeq, MiSeq, HiSeq 2000/2500), illumina_ud_sys2 (Illumina UD indexed for MiniSeq, NextSeq, HiSeq 3000/4000).

template_files/adapter_primer_sequences.txt:

- contains four columns (i.e. Type, Index, Description, Sequence). Sequences in the Index column is used to match those in Index column in sample info file. This column naming is rigid.
- based on the following resources:
    - illumina adapter sequences [2]
- If users provide new sequences, add the new index type in the 1st column 'Type' and specify it in `index_type`. In this example, use `--index_type GSE95632` which was defined in **example_files/adapter_primer_sequences.txt**.

submit LSF script

LSF scripts are generated for each sample. Submit LSF jobs on HPC that enables to run in parallel.

BASH

```bash
for i in *_align.lsf; do bsub < $i; done
```

Check one sample LSF

```
cat SRR1039508_align.lsf
```

OUTPUT

```
#!/bin/bash
#BSUB -L /bin/bash
#BSUB -J SRR5309351_align
#BSUB -q normal
#BSUB -o SRR5309351_align_%J.out
#BSUB -e SRR5309351_align_%J.screen
#BSUB -M 36000
#BSUB -n 12
cd /home/mengykan/Projects/GSE95632/SRR5309351/
java -Xmx1024m  -classpath /opt/software/Trimmomatic/0.32/trimmomatic-0.32.jar org.usadellab.trimmomatic.TrimmomaticS
E -phred33 /home/mengykan/Projects/GSE95632/GSE95632_SRAdownload/SRR5309351.fastq.gz /home/mengykan/Projects/GSE95632
/SRR5309351/SRR5309351_R1_Trimmed.fastq ILLUMINACLIP:/home/mengykan/Projects/GSE95632/SRR5309351/SRR5309351_adapter.f
a:2:30:10 MINLEN:40
fastqc -o /home/mengykan/Projects/GSE95632/SRR5309351/ /home/mengykan/Projects/GSE95632/SRR5309351/SRR5309351_R1_Trim
med.fastq
cat /home/mengykan/Projects/GSE95632/SRR5309351/SRR5309351_R1_Trimmed.fastq | awk '((NR-2)%4==0){read=$1;total++;coun
t[read]++}END{for(read in count){if(count[read]==1){unique++}};print total,unique,unique*100/total}' > /home/mengykan
/Projects/GSE95632/SRR5309351/SRR5309351_ReadCount
mkdir /home/mengykan/Projects/GSE95632/SRR5309351/bwa_out
cd /home/mengykan/Projects/GSE95632/SRR5309351/bwa_out
bwa mem -t 12 /project/bhimeslab/Reference/hg38/bwa_index/hg38 /home/mengykan/Projects/GSE95632/SRR5309351/SRR5309351
_R1_Trimmed.fastq | samtools view -S -b - | samtools sort -@12 -T SRR5309351.tmp -o SRR5309351.bam -
samtools index -@12 SRR5309351.bam
samtools idxstats SRR5309351.bam > SRR5309351.stats
bamtools stats -in SRR5309351.bam > SRR5309351.bamstats
```

output files

Various output files will be written for each sample in directories structured such as:

- Sample-level directory `/home/mengykan/Projects/GSE95632/SRR5309351`
- Trimemd and FastQC files are:
  - `SRR5309351_R1_Trimmed.fastq`
  - `SRR5309351_R1_Trimmed_fastqc.zip`
  - `SRR5309351__ReadCount`
- Aligned .bam and QC metrics files are saved in
  - `bwa_out/`

Call peaks

run command line

```
mkdir /home/mengykan/Projects/GSE95632/scripts/macs2
cd /home/mengykan/Projects/GSE95632/scripts/macs2
chipseq_peakcaller.py --project_name GSE95632 --samples_in /home/mengykan/Projects/GSE95632/files/GSE95632_Phenotyp\
e_withoutQC.txt --ref_genome hg38 --path_start /home/mengykan/Projects/GSE95632 --template_dir /home/mengykan/Proje\
cts/shared_files/ChIPSeq
```

Peaks within 'blacklist regions' in hg38 and hg19 reference genome were filtered out. Blacklist regions for hg38 and hg19 are provided in **templete_files** (**hg19_blacklist.bed** and **hg38_blacklist.bed**) and should be saved in user-defind `emplate_dir`.

submit LSF script

LSF scripts are generated for each sample. Submit LSF jobs on HPC that enables to run in parallel.

```
for i in *_macs2.lsf; do bsub < $i; done
```

Check one sample LSF

Use GR antibody

```
cat SRR5309351_macs2.lsf
```

OUTPUT

```
#!/bin/bash
#BSUB -L /bin/bash
#BSUB -J SRR5309351_macs2
#BSUB -q normal
#BSUB -o SRR5309351_macs2_%J.out
#BSUB -e SRR5309351_macs2_%J.screen
#BSUB -M 36000
#BSUB -n 1
mkdir /home/mengykan/Projects/GSE95632//SRR5309351/macs2_out/
macs2 callpeak -c /home/mengykan/Projects/GSE95632/SRR5309352/bwa_out/SRR5309352.bam -t /home/mengykan/Projects/GSE95
632/SRR5309351/bwa_out/SRR5309351.bam -n SRR5309351 --outdir /home/mengykan/Projects/GSE95632//SRR5309351/macs2_out/
-f BAM -g hs -B -q 0.01
bedtools intersect -v -a /home/mengykan/Projects/GSE95632/SRR5309351/macs2_out/SRR5309351_peaks.narrowPeak -b /home/m
engykan/Projects/shared_files/ChIPSeq/hg38_blacklist.bed | grep -P 'chr[\dXY]+[ \t]' | awk 'BEGIN{OFS="\t"} {if ($5>1
000) $5=1000; print $0}' > /home/mengykan/Projects/GSE95632/SRR5309351/macs2_out/SRR5309351.blackfilt.bed
```

Whether to perform narrow or broad peak calling was specified in Peak column of sample info file. Generally apply narrow peak calling for transcription binding site. Encode provides narrow or broad marks for histone modifications [3] (the Target-specific Standards section).

output files

Peak files will be written for samples that underwent ChIPSeq in the macs2 output directory:

- Sample-level directory `/home/mengykan/Projects/GSE95632/SRR5309351`
- Peak files are saved in
  - `macs2_out/`

Generate summary report of QC metrics

run command lines

Run **pipeline_scripts/chipseq_align_and_qc_report.py** to create an HTML report of QC and alignment summary statistics for ChIP-seq samples.

Read in **template_files/chipseq_align_and_qc_report_Rmd_template.txt** from specified directory `template_dir` to create a RMD script.

This script uses the many output files created in align and quantification step, converts these sample-specific files into matrices that include data for all samples, and then creates an Rmd document.

```
mkdir /home/mengykan/Projects/GSE95632/scripts/qc_report
cd /home/mengykan/Projects/GSE95632/scripts/qc_report
chipseq_align_and_qc_report.py --project_name GSE95632 --samples_in /home/mengykan/Projects/GSE95632/files/GSE95632\
_Phenotype_withoutQC.txt --ref_genome hg38 --library_type SE --path_start /home/mengykan/Projects/GSE95632 --templa\
te_dir /home/mengykan/Projects/shared_files/ChIPSeq
```

submit LSF script

Generate a single LSF script GSE95632_qc.lsf. This is a single-node analysis, but we recommend running it on HPC as the step of count normalization for PCA plots takes a lot of memory.

BASH

```
bsub < GSE95632_qc.lsf
```

```
cat GSE95632_qc.lsf
```

OUTPUT

```
#!/bin/bash
#BSUB -L /bin/bash
#BSUB -J GSE95632_qc
#BSUB -q normal
#BSUB -o GSE95632_qc_%J.out
#BSUB -e GSE95632_qc_%J.screen
#BSUB -M 36000
#BSUB -n 1
cd /home/mengykan/Projects/GSE95632/GSE95632_Alignment_QC_Report/; echo "library(rmarkdown); rmarkdown::render('GSE95
632_QC_ChIPSeqReport.Rmd')" | R --no-save --no-restore
```

output files

Output files are saved in `/home/mengykan/Projects/GSE95632/GSE95632_Alignment_QC_Report`

The RMD and corresponding HTML report files are:

- `GSE95632_QC_ChIPSeqReport.Rmd`
- `GSE95632_QC_ChIPSeqReport.html`

Differential binding analysis and feature annotation

run command lines

Run **pipeline_scripts/chipseq_diffbind_report.py** to perform differential binding site analysis and create an HTML report of differential expression summary statistics.

Read in **template_files/chipseq_diffbind_report_Rmd_template.txt** from specified directory `template_dir` to create a RMD script.

```
mkdir /home/mengykan/Projects/GSE95632/scripts/diffbind_report
cd /home/mengykan/Projects/GSE95632/scripts/diffbind_report
chipseq_diffbind_report.py --project_name GSE95632 --samples_in /home/mengykan/Projects/GSE95632/files/GSE95632_Phe\
notype_withQC.txt --comp /home/mengykan/Projects/GSE95632/files/GSE95632_comp_file.txt --ref_genome hg38 --path_sta\
rt /home/mengykan/Projects/GSE95632 --template_dir /home/mengykan/Projects/shared_files/ChIPSeq
```

script options

The `--sample_in` option specifies user provided phenotype file for differential binding analysis (the example file *example_files/GSE95632_Phenotype_withQC.txt*). The columns are the same as *example_files/GSE95632_Phenotype_withoutQC.txt* but with an additional column *QC_Pass* designating samples to be included (QC_Pass=1) or excluded (QC_Pass=0) after QC. This column naming is rigid which will be recoganized in pipeline scripts, but column order can be changed. In the current example, all samples pass QC.

The `--comp` option specifies comparisons of interest in a tab-delimited text file with one comparison per line with three columns (i.e. Condition1, Condition0, Design), designating Condition1 vs. Condition0. The current version of differential analysis does not support paired analysis.

Find the example comp file here **example_files/GSE95632_comp_file.txt**.

Condition1	Condition0	Design
GR_dex	GR_EtOH	unpaired
RNAP2_dex	RNAP2_EtOH	unpaired

submit LSF script

Generate a single LSF script GSE95632_diffbind.lsf. This is a single-node analysis, but we recommend running it on HPC as the steps of count normalization for pairwise comparisons take a lot of memory.

```
bsub < GSE95632_diffbind.lsf
cat GSE95632_diffbind.lsf
```

OUTPUT

```
#!/bin/bash
#BSUB -L /bin/bash
#BSUB -J GSE95632_diffbind
#BSUB -q normal
#BSUB -o GSE95632_diffbind_%J.out
#BSUB -e GSE95632_diffbind_%J.screen
#BSUB -M 36000
#BSUB -n 1
cd /home/mengykan/Projects/GSE95632/GSE95632_diffbind_out/; echo "library(rmarkdown); rmarkdown::render('GSE95632_Dif
fBind_Report.Rmd')" | R --no-save --no-restore
```

output files

Output differential binding results are saved in `/home/mengykan/Projects/GSE95632/GSE95632_diffbind_out`:

- Sample sheet input for DiffBind

    - e.g. GSE95632_GR_dex_vs_GR_EtOH.sampleinfo.csv

- Pairwise differential binding comparisons. Only save significant binding sites.
    - e.g. GSE95632_GR_dex_vs_GR_EtOH_sig_diffbind_results.csv
- Normalized counts of significant binding site in samples for each pairwise comparison
    - e.g. GSE95632_GR_dex_vs_GR_EtOH_sig_counts_normalized_by_diffbind.csv

The RMD and corresponding HTML report files are:

- `GSE95632_DiffBind_Report.Rmd`
- `GSE95632_DiffBind_Report.html`

References

1. [brocade pipeline]
   https://github.com/HimesGroup/brocade
2. [illumina adapter sequences]
   https://www.nature.com/protocolexchange/system/uploads/6661/original/SupplementaryDocument2-illumina-adapter-sequences-Feb2018.pdf?1530635414
3. [Encode provides narrow or broad marks for histone modifications]
   https://www.encodeproject.org/chip-seq/histone/