macs2_meme

生物信息学
助教-刘柯
助教-方明昊

目录

- ❖Peak Calling (Macs2)
- ❖Motif Enrich + GO (MEME, CISTROME)

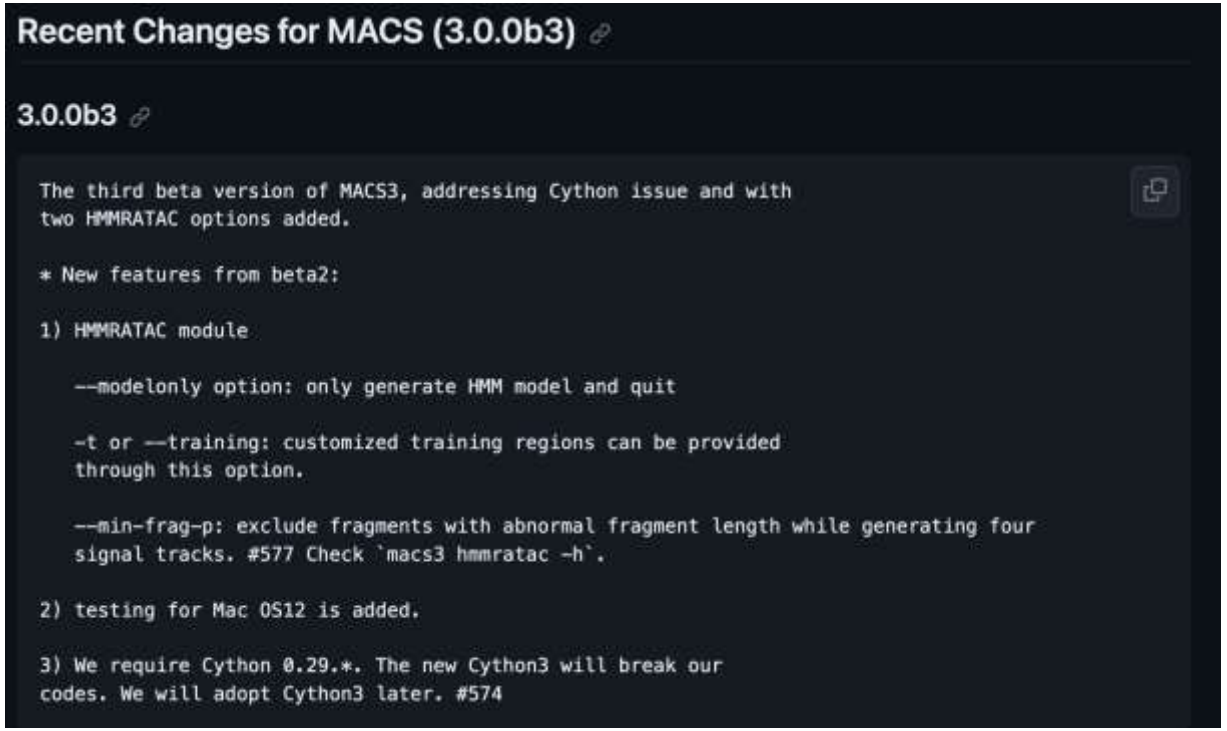
❖ Peak Calling (Macs2)

[HTML] Model-based analysis of ChIP-Seq (MACS)

Y Zhang, T Liu, CA Meyer… - Genome …, 2008 - genomebiology.biomedcentral.com

We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. MACS empirically models …

☆ 保存 99 引用 被引用次数：14081 相关文章 所有 34 个版本 »



Recent Changes for MACS (3.0.0b3)

3.0.0b3

```
The third beta version of MACS3, addressing Cython issue and with
two HMMRATAC options added.

* New features from beta2:

1) HMMRATAC module

   --modelonly option: only generate HMM model and quit

   -t or --training: customized training regions can be provided
   through this option.

   --min-frag-p: exclude fragments with abnormal fragment length while generating four
   signal tracks. #577 Check `macs3 hmmratac -h`.

2) testing for Mac OS12 is added.

3) We require Cython 0.29.*. The new Cython3 will break our
codes. We will adopt Cython3 later. #574
```

https://github.com/macs3-project/MACS

❖ Peak Calling (Macs2)

Install

The common way to install MACS is through PYPI) or conda. Please check the INSTALL document for detail.

MACS3 has been tested in CI for every push and PR in the following architectures:

- x86_64
- aarch64
- armv7
- ppc64le
- s390x
- Apple chips

In general, you can install through PyPI as `pip install macs3`. To use virtual environment is highly recommended. Or you can install after unzipping the released package downloaded from Github, then use `pip install .` command.

https://github.com/macs3-project/MACS

❖ Peak Calling (Macs2)

Usage

Example for regular peak calling on TF ChIP-seq:

```
macs3 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01
```

Example for broad peak calling on Histone Mark ChIP-seq:

```
macs3 callpeak -t ChIP.bam -c Control.bam --broad -g hs --broad-cutoff 0.1
```

Example for peak calling on ATAC-seq (paired-end mode):

```
macs3 callpeak -f BAMPE -t ATAC.bam -g hs -n test -B -q 0.01
```

There are currently 14 functions available in MACS3 serving as sub-commands. Please click on the link to see the detail description of the subcommands.

Subcommand	Description
`callpeak`	Main MACS3 Function to call peaks from alignment results.
`bdgpeakcall`	Call peaks from bedGraph output.
`bdgbroadcall`	Call broad peaks from bedGraph output.

`-t/--treatment FILENAME`

This is the only REQUIRED parameter for MACS. The file can be in any supported format -- see detail in the `--format` option. If you have more than one alignment file, you can specify them as `-t A B C`. MACS will pool up all these files together.

`-c/--control`

The control, genomic input or mock IP data file. Please follow the same direction as for `-t/--treatment`.

`-n/--name`

The name string of the experiment. MACS will use this string NAME to create output files like `NAME_peaks.xls`, `NAME_negative_peaks.xls` `NAME_peaks.bed`, `NAME_summits.bed`, `NAME_model.r` and so on. So please avoid any confliction between these filenames and your existing files.

`--outdir`

MACS3 will save all output files into the specified folder for this option. A new folder will be created if necessary.

`-f/--format FORMAT`

Format of tag file can be `ELAND`, `BED`, `ELANDMULTI`, `ELANDEXPORT`, `SAM`, `BAM`, `BOWTIE`, `BAMPE`, or `BEDPE`. Default is `AUTO` which will allow MACS to decide the format automatically. `AUTO` is also useful when you combine different formats of files. Note that MACS can't detect `BAMPE` or `BEDPE` format with `AUTO`, and you have to implicitly specify the format for `BAMPE` and `BEDPE`.

https://github.com/macs3-project/MACS

❖ Peak Calling (Macs2)

```
(base) [minghaofang@mgt 01.data]$ pwd
/home/qukun/minghaofang/bioinfomatics_class/CHIP-seq/01.data
(base) [minghaofang@mgt 01.data]$ ll -h
total 1.6G
-rw-r--r-- 1 minghaofang qukun 842M Oct 30 08:44 SRR520342.markdup.bam
-rw-r--r-- 1 minghaofang qukun 751M Oct 30 08:45 SRR520348.markdup.bam
(base) [minghaofang@mgt 01.data]$
```

```
# Building a small index
bowtie2-build example/reference/lambda_virus.fa example/index/lambda_virus

# Building a large index
bowtie2-build --large-index example/reference/lambda_virus.fa example/index/lamb
```

Examples

```
# Aligning unpaired reads
bowtie2 -x example/index/lambda_virus -U example/reads/longreads.fq

# Aligning paired reads
bowtie2 -x example/index/lambda_virus -1 example/reads/reads_1.fq -2 example/rea
```

https://github.com/BenLangmead/bowtie2

❖ Peak Calling (Macs2)

```
(base) [minghaofang@mgt script]$ module ava
------------------------------------------ /public/MODULES/COMPILER ------------------------------------------
cmake/3.19.0  CUDA/10.2.89  cuDNN/v7.4.2  INTEL/icc_2017_update4                openmpi/4.1.2
CUDA/8.0      CUDA/11.4.4   gcc/7.2.0     INTEL/parallel_studio_xe_2016.2.181   oracle-jdk
CUDA/9.0      cuDNN/v7.1    gcc/10.2.0    INTEL/parallel_studio_xe_2017_update4 R/4.1.2
-------------------------------------------- /public/MODULES/APPS --------------------------------------------
Gaussian/G16          MATLAB/R2017a  singularity/3.1.0           vasp/5.4.4/intel2017update4
MaterialsStudio/18.1  MATLAB/R2019a  vasp/5.4.4/intel2016withGPU
-------------------------------------------- /public/MODULES/BIO ---------------------------------------------
afnl       cryolo             Encode/npIDR                hmmer        Relion                                TRF
amber      cufflinks          Encode/PeakSeq              HOMER        Relion3                               Trinity
amber22    demuxlet           Encode/Phantompeakqualtools hotspot2     Relion3.1_beta_SinglePrecisionOnGPU   vsearch
Anaconda2  Dynamo             Encode/PIQ                  IGV          Relion3_SinglePrecisionOnGPU
Anaconda3  ea-utils           Encode/sample               IGVTools     Relion_3.0beta
bcftools   Encode/AlleleSeq   Encode/TophatBAMRepair      juicer       RepeatMasker
bedops     Encode/bismark     Encode/WASP                 kentUtils    RMBlast
blast      Encode/ChromHMM    fastqc                      lammps       samtools
bowtie     Encode/Cluster     flexbar                     macs2        scipion
bowtie2    Encode/cxrepo-bed  GATK                        meme         sratoolkit
bwa        Encode/Flux        gemtools                    modwt        STAR
cdhit      Encode/fseq        gromacs/4.5.5               nuc_dynamics subread-1.6.4
ChIA-PET2  Encode/gerp        gromacs/2016.3              picard       tantan
chilin     Encode/kentUtils   HiC-Pro                     preseq       tophat-1.4.1
chimera    Encode/mfinder     hisat2                      prinseq-lite tophat-2.1.1
------------------------------------- /public/MODULES/to_be_deleted -------------------------------------
CUDA8.0  gcc7.2.0
```

```
Traceback (most recent call last):
  File "/public/software/bio/macs2/macs2", line 31, in <module>
    from MACS2.Constants import *
ImportError: No module named MACS2.Constants
```

系统中macs2 存在调用问题，建议大家pypi/conda自行安装

❖ Peak Calling (Macs2)

核心Pbs脚本, 环境激活请改成自己路径

```
#PBS -N Macs2
#PBS -o logs/macs2.log
#PBS -e logs/macs2.err
#PBS -q batch
#PBS -l nodes=1:ppn=8
#PBS -l walltime=30:00:00

#source activate base
source /home/qukun/minghaofang/miniconda3/etc/profile.d/conda.sh
conda activate base

cd /home/qukun/minghaofang/bioinfomatics_class/CHIP-seq/01.data

##系统上macs2调用存在问题
##module load macs2

macs2 callpeak -t SRR520342.markdup.bam -c SRR520348.markdup.bam -f BAM -g hs \
        -n SRR520342 -B -q 0.01 --outdir ../02.macs2
```

```
(base) [minghaofang@mgt script]$ ll -h ../02.macs2/
total 2.4G
-rw-r--r-- 1 minghaofang qukun 1.3G Oct 30  2023 SRR520342_control_lambda.bdg
-rw-r--r-- 1 minghaofang qukun  98K Oct 30 09:35 SRR520342_model.r
-rw-r--r-- 1 minghaofang qukun 553K Oct 30  2023 SRR520342_peaks.narrowPeak
-rw-r--r-- 1 minghaofang qukun 629K Oct 30  2023 SRR520342_peaks.xls
-rw-r--r-- 1 minghaofang qukun 375K Oct 30  2023 SRR520342_summits.bed
-rw-r--r-- 1 minghaofang qukun 1.2G Oct 30  2023 SRR520342_treat_pileup.bdg
```

❖ Peak Calling (Macs2)

输出文件目录

```
s2/

SRR520342_control_lambda.bdg
SRR520342_model.r
SRR520342_peaks.narrowPeak
SRR520342_peaks.xls
SRR520342_summits.bed
SRR520342_treat_pileup.bdg
```

less narrowPeak

```
chr1    429980  430219  SRR520342_peak_1        305     .       15.75   35.0287 30.5704 111
chr1    665710  665949  SRR520342_peak_2        193     .       8.72527 23.5121 19.3886 112
chr1    1429805 1429957 SRR520342_peak_3        227     .       13.4147 27.0236 22.7758 59
chr1    1783066 1783228 SRR520342_peak_4        112     .       8.66847 15.098  11.2565 77
chr1    3247840 3247958 SRR520342_peak_5        103     .       8.09146 14.181  10.3788 58
chr1    3261457 3261604 SRR520342_peak_6        96      .       8.30385 13.3766 9.61339 62
chr1    3624708 3624854 SRR520342_peak_7        130     .       9.85174 16.9668 13.0541 82
chr1    6582470 6582617 SRR520342_peak_8        72      .       5.49659 10.9285 7.28692 72
chr1    8027373 8027486 SRR520342_peak_9        55      .       5.05326 9.11863 5.56791 43
chr1    8061396 8061549 SRR520342_peak_10       194     .       7.78738 23.5306 19.4067 60
chr1    8117014 8117224 SRR520342_peak_11       151     .       10.3773 19.1486 15.1581 127
chr1    8200372 8200526 SRR520342_peak_12       89      .       5.90018 12.6839 8.93433 43
chr1    8259178 8259435 SRR520342_peak_13       235     .       11.0956 27.847  23.5891 108
chr1    8263306 8263419 SRR520342_peak_14       80      .       7.50803 11.7504 8.06353 36
```

xls 输出文件格式

1. `NAME_peaks.xls` is a tabular file which contains information about called peaks. You can open it in excel and sort/filter using excel functions. Information include:
   - chromosome name
   - start position of peak
   - end position of peak
   - length of peak region
   - absolute peak summit position
   - pileup height at peak summit
   - -log10(pvalue) for the peak summit (e.g. pvalue =1e-10, then this value should be 10)
   - fold enrichment for this peak summit against random Poisson distribution with local lambda,
   - -log10(qvalue) at peak summit

   Coordinates in XLS is 1-based which is different from BED format. When `--broad` is enabled for broad peak calling, the pileup, p-value, q-value, and fold change in the XLS file will be the mean value across the entire peak region, since peak summit won't be called in broad peak calling mode.

narrowPeak 输出文件格式

2. `NAME_peaks.narrowPeak` is BED6+4 format file which contains the peak locations together with peak summit, p-value, and q-value. You can load it to the UCSC genome browser. Definition of some specific columns are:
   - 5th: integer score for display. It's calculated as `int(-10*log10pvalue)` or `int(-10*log10qvalue)` depending on whether `-p` (pvalue) or `-q` (qvalue) is used as score cutoff. Please note that currently this value might be out of the [0-1000] range defined in UCSC ENCODE narrowPeak format. You can let the value saturated at 1000 (i.e. p/q-value = 10^-100) by using the following 1-liner awk: `awk -v OFS="\t" '{$5=$5>1000?1000:$5} {print}' NAME_peaks.narrowPeak`
   - 7th: fold-change at peak summit
   - 8th: -log10pvalue at peak summit
   - 9th: -log10qvalue at peak summit
   - 10th: relative summit position to peak start

   The file can be loaded directly to the UCSC genome browser. Remove the beginning track line if you want to analyze it by other tools.

https://github.com/macs3-project/MACS/blob/master/docs/callpeak.md

目录

❖Peak Calling (Macs2)

❖Motif Enrich + GO (MEME+CISTROME)

https://meme-suite.org/meme/tools/meme-chip

http://go.cistrome.org/

https://jaspar.genereg.net/

https://meme-suite.org/meme//opal-jobs/appSTREME_5.5.4169863471743021189014 7/streme.html#inputs_sec