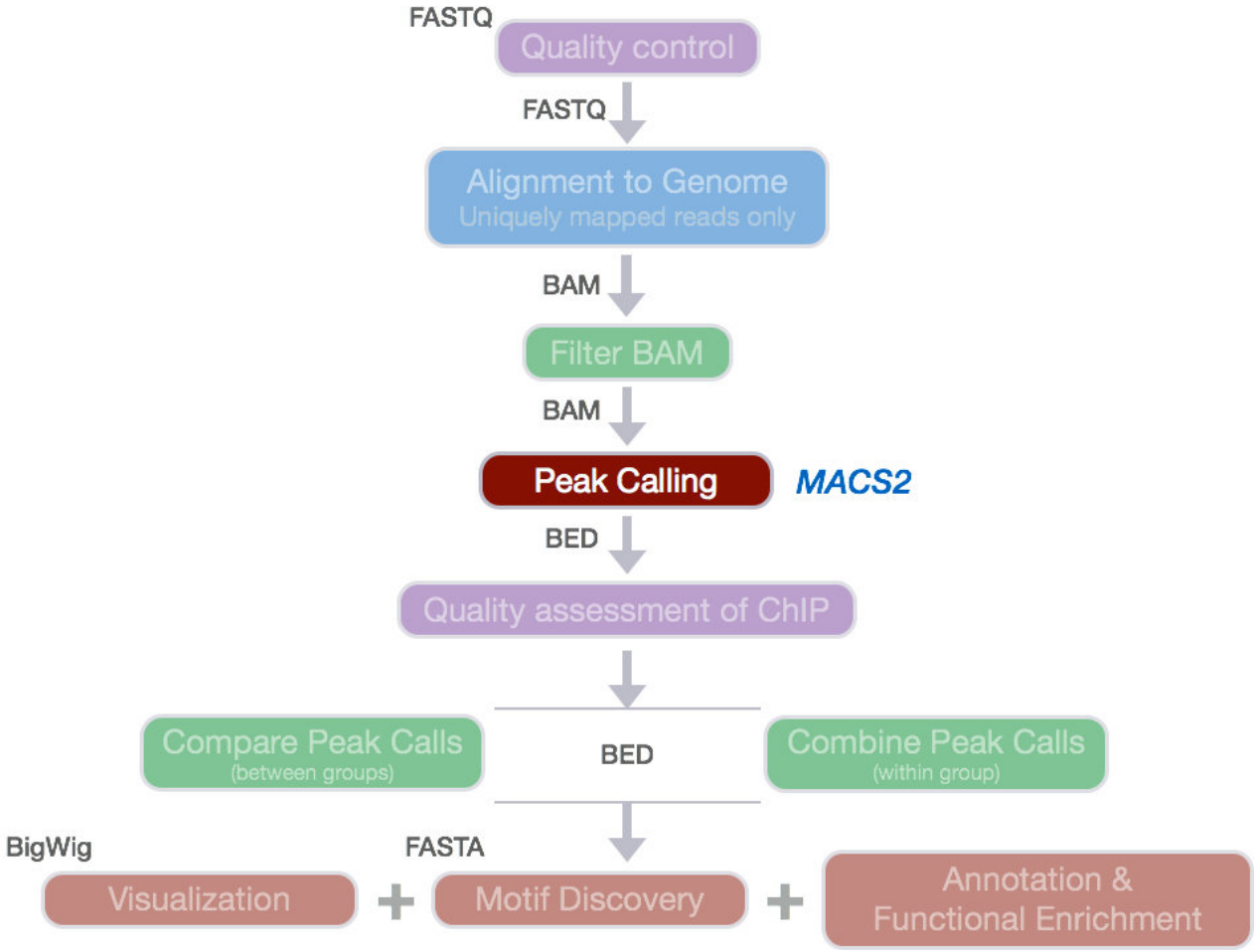
[MACS2][https://github.com/taoliu/MACS]

Model-based Analysis of ChIP-Seq(MACS), 识别转录因子结合位点. 通过集合序列标签位置和方向提高结合位置的空间分辨率. 可简单用于单个ChIP-Seq数据或通过指控样本增加检出特异性.

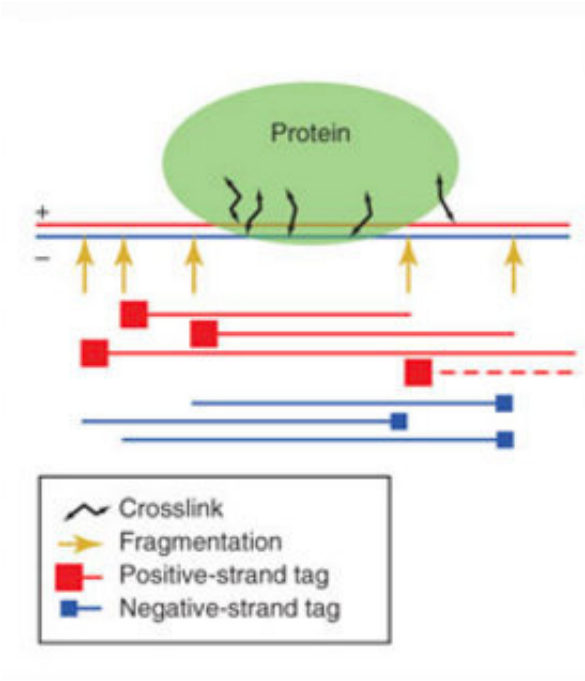


Moreove, as a general peak-caller, MACS can also be applied to any 'DNA enrichment assays' if the question to be asked is simply: where we can find significant reads coverage than the random backgound.

Usage

```
macs2 [-h] [--version]

{callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdiff,filterdup,pred
ictd,pileup,randsample,refinepeak}
```



ChIP-Seq的分析方法可以鉴定两种类型的富集模式:broad domains和narrow peaks. broad domains, 如组蛋白修饰在整个基因body区域的分布; narrow peak, 如转录因子的结合. narrow peak相对于broad或分散的marks更易被检测到. 也有一些混合的结合图谱, 如PolI包括narrow和broad信号.

regular peak calling:

```
macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01
```

broad peak calling:

```
macs2 callpeak -t ChIP.bam -c Control.bam --broad -g hs --broad-cutoff 0.1
```

MACS2包含12个功能, 略, 这里仅介绍`callpeak`, 其余`macs2 COMMAND -H`

Enssential Options

-t / --treatment FILENAME 是MACS唯一的查询参数, 其文件支持格式通过`--format`参数指定, 多个比对文件`-t A B C`, MACS将这三个文件一起处理

-c / --control control或mock数据文件. 同`-t / --treatment`

-n / --name 实验的名称. MACS使用该NAME创建输出文件名: NAME_peaks.xls, NAME_negative_peaks.xls, NAME_peaks.bed, NAME_summits.bed, NAME_modle.r

--outdir 指定输出文件夹

-f / --format FORMAT tag文件的格式, 可以为:ELAND, BED, ELANDNULTI, ELANDEXPORT, ELANDMULTIPER(for pair-end tags), SAM, BAM, BOWTIE, BAMPE, BEDPE. 默认为AUTO, MACS自动识别, 尤其当同时包含多个输入时.

当前最常用的为BED或SAM/BAM

[BED][http://genome.ucsc.edu/FAQ/FAQformat#format1]

该格式需要至少包含, 1s: chromome name, 2nd: start position(The first base in a chromosome is numbered 0), 3rd: end position, 6th: strand

[SAM/BAM][http://www.htslib.org/doc/samtools.html]

若为BAM的paired-end数据, MACS

BEDPE/BAMPE

此时, MACS2将BAM/BED文件看作paired-end数据. MACS2使用实际插入片段构建fragment pileup, 而不是根据正/负方向reads的二项式分布来预测插入片段长度. BAMPE格式为包含paired-end比对信息的BAM格式文件, 源自BWA或BOWTIE

可使用MACS2 randsample将paired-end信息的BAM文件转换为BEDPE格式:

```
macs2 randsample -i the_BAMPE_file.bam -f BAMPE -p 100 -o the_BEDPE_file.bed
```

-g / --gsize 根据实际情况设置该参数, 用于指定比对基因组大小或有效基因组大小. 由于存在重复序列, 因此实际基因组大小会比原始大小要小, 约等其基因组大小的90%或70%. 可使用k-mer工具(jellyfish)来计算有效基因组长度, 些许差异不会带来很大影响...

-s / --tsize 测序标签(sequencing tags)的大小, 若未指定, MACS使用前10个序列来判断标签长度大小 (uniquely mapped read)

-q / --qvalue 显著性区域的q-value(minimum FDR)阈值, 默认为0.05. Q-values是使用Benjamini-Hochberg处理p-values得来的

-p / --pvalue p-value阈值, 未指定, 将使用p-value而不是q-value

-m / --mfold MFOLD MFOLD 选择富集差异位于该范围内的区域用于构建model, 预测d值. 默认为5 50; 可选-m 10 30

--min-length / --max-gap 用于指定peak的最小长度和合并两个临近区域之间允许的最大gap长度. 默认采用预测的片段长度为min-length, max-gap为reads长度. 针对BROAD peak calling, 尝试大的值例如500bp, 也可以使用'--cutoff-analysis'加默认设置, 选择合理的min-length值

--nolambda 使用该参数时, MACS使用backgroud lambda作为local lambda. 这意味着MACS不会考虑peak candidate区域local bias(泊松分布,期望和方差均为λ)

--slocal, --llocal These two parameters control which two levels of regions will be checked around the peak regions to calculate the maximum lambda as local lambda. By default, MACS will considers 1000bp for small local region(--slocal), and 10000bps for large local region(--llocal) which captures the bias from a long-range like an open chromatin domain. **You can tweak these according to your project. Remember that if the region is set too small, a sharp spike in the input data may kill a significant peak.**

--nomodel 使用该参数, MACS将绕开构建shifting model 环节

--extsize 当设置--nomodel时, MACS使用该参数向reads的5'-3'方向延伸至fix-sized fragments. 例如, 针对转录因子的结合区域的大小为200bp, 同时想要取消MACS构建model, 该参数可被设置为200. 该选项只有当设置--nomodel时才有效, 或当MACS构建model失败同时--fix-bimodal选项开启

--shift 这里可以选择任意shift bp长度. 当使用非0时需要酌情处理(other than the default value, 0). 当`--nomodel`设置时, MACS使用该值moving cutting ends(5')(应该是sheared DNA位置, sequenced read), 接着使用`-extsize`从5'到3'方向延伸(when --nomodel is set, MACS will use this value to move cutting ends(5') then apply `--extsize` from 5' to 3' direction to extend them to fixed-

fragments. When this value is negative, ends will be moved toward 3'-5' direction, otherwise 5'-3' direction). 推荐在使用ChIP-Seq数据时保持默认值0), or -1 * half of EXTSIZE together with `--extsize` option for detecting enriched cutting loci such as certain DNAsel-Seq datasets. Note, you can't set values other than 0 if the format is BAMPE or BEDPE for paired-end data. The dafult is 0

> Here are some examples for combining `--shift` and `--extsize`:
>
> 1. To find enriched cutting sites such as some DNAse-Seq datasets. In this case, all 5' ends of sequenced reads should be extended in both directions to smooth the pileup signals. If the wanted smoothing window is 200bps, then use `--nomodel --shift -100 --extsize 200`.
> 2. For certain nucleosome-seq data, we need to pile up the centers of nucleosomes using a half-nucleosome size for wavelet analysis (e.g. NPS algorithm). Since the DNA wrapped on nucleosome is about 147bps, this option can be used: `--nomodel --shift 37 --extsize 73`.

--bw BW 用于选择区域计算fragment size的band宽度(**the window : roughly twice the size of the sheared chromatin across the genome**), 为用于模型构建步骤中sliding窗口长度的一半, 仅用于构建shifting model, 也就是打断的片段长度. 不建议修改该值, 默认为:300

--keep-dup 在相同位置(相同的方向和相同的链), MACS将保留的duplicate tags数目. 默认时在相同位置保留一个, default:1

--broad 当选择该选项时, 通过采用宽松阈值将附近高富集区域放进一个broad区域来尝试构成broad regions in BED12(a gene-model-like format). Broad region 通过另一个阈值`--broad-cutoff`来控制. broad region的最大长度为MACS的d的4倍. 默认为:False

--broad-off broad region的阈值. 只有设置了`--broad`时该选项才有意义. 若设置了`-p`, 为p-value阈值, 否则, 为q-value阈值. 默认为: 0.1

--scale-to <large|small> 当设置为'large'时, 将较小的数据集线性成为较大的数据集. 默认或设置为'small'时, 更大的数据集将会缩小的更小的数据集. 注意, 将小数据集扩大可能带来假阳性

-B / --bdg 若设置该选线, MACS will store the fragment pileup, control lambda in bedGraph files. The bedGraph files will be stored in the current directory names `NAME_treat_pileup.bdg` for treatment data, `NAME_control_lambda.bdg` for local lambda values from control.

--call-summits MACS will now reanalyze the shape of signal profile(p or q-score depending on the cutoff setting) to deconvolve subpeaks within each peak called from the general procedure. It's highly recommanded to detect adjacent binding events. While used, the output sub peaks of a big peak region will have the same peak boundaries, and different scores and peak summit positions

Output files

`NAME_peaks.xls`为包含检测峰的表格信息:

- 染色体名称
- 峰起点
- 峰终点
- 峰区域长度
- 峰顶点绝对位置
- 峰顶点的堆积高度
- 峰顶点的-log10(pvalue)
- 该峰顶点相对于随机泊松分布(with local lambda)的富集倍数
- 峰顶点的-log10(qvalue)

该xls文件中位置为1-based, 不同于BED文件; 当采用`--broad`参数时, 用于检出broad peak, 那么pileup, p-value, q-value, fold change将会是这个整个峰区域的平均值, 因为在检出broad peaks时不会检出峰顶点

`NAME_peaks.narrowPeak` 为BED6+4格式文件, 包含峰位置和峰顶点信息, p-value, q-value:

- 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

可直接使用[UCSC genome browser][http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.x86_64/]读取. Remove the beginning track line if you want to analyze it by other tools

`NAME_summits.bed` BED格式, 包含所有峰的峰顶点位置. 其中第5th列同`NAME_peaks.narrowPeak`. 推荐使用该文件查询结合位点的motif. 该文件也可以使用UCSC genome browser读取. Remove the beginning track line if you want to analyze it by other tools

`NAME_peaks.broadPeak` 为BED6+3格式, 类似`narrowPeak`文件,但不含邮第十行的峰顶点的注释信息. 该文件和`gappedPeak`仅当`--broad`选项使用时才会生成.

`NAME_peaks.gappedPeak` 为BED12+3格式, 包含broad region和narrow peaks.:

5. `NAME_peaks.gappedPeak` is in BED12+3 format which contains both the broad region and narrow peaks. The 5th column is the score for showing grey levels on the UCSC browser as in `narrowPeak`. The 7th is the start of the first narrow peak in the region, and the 8th column is the end. The 9th column should be RGB color key, however, we keep 0 here to use the default color, so change it if you want. The 10th column tells how many blocks including the starting 1bp and ending 1bp of broad regions. The 11th column shows the length of each block and 12th for the start of each block. 13th: fold-change, 14th: *-log10pvalue*, 15th: *-log10qvalue*. The file can be loaded directly to the UCSC genome browser. Refer to `narrowPeak` if you want to fix the value issue in the 5th column.

`NAME_model.r` 为R脚本, 用于生成PDF图像文件展示model: `R NAME_model.r`

`NAME_treat_pileup.bdg`和`NAME_control_lambda.bdg`文件为bedGraph格式, 可导入UCSC genome browser或转换为更小的bigWig文件:

7. The `NAME_treat_pileup.bdg` and `NAME_control_lambda.bdg` files are in bedGraph format which can be imported to the UCSC genome browser or be converted into even smaller bigWig files. The `NAME_treat_pielup.bdg` contains the pileup signals (normalized according to `--scale-to` option) from ChIP/treatment sample. The `NAME_control_lambda.bdg` contains local biases estimated for each genomic location from the control sample, or from treatment sample when the control sample is absent. The subcommand `bdgcmp` can be used to compare these two files and make a bedGraph file of scores such as p-value, q-value, log-likelihood, and log fold changes.

Tips of fine-tuning peak calling

1. `bdgcmp`可用于`*_treat_pileup.bdg`和`*_control_lambda.bdg`或其他来源的bedGraph文件, 计算score track
2. `bdgpeakcall`可用于`*_treat_pvalue.bdg`或其他bdgcmp/bedGraph文件生成的文件, 根据指定阈值, 最大gap距离, 最小峰长度来检出峰. bdgbroadcall和bdgpeakcall用法类似, 只是输出BED12格式的`_broad_peaks.bed`
3. 差异检出工具`bdgdiff`, 可用于4个bedGraph文件, 包含treatment1/control1,

treatment2/control2, treatment1/treatment2, treatment2/treatment1 比值. 根据最短长度, 最大gap和阈值输出一致性和唯一位置

4. You can combine subcommands to do a step-by-step peak calling. Read detail at [MACS2 wikipage][https://github.com/taoliu/MACS/wiki/Advanced%3A-Call-peaks-using-MACS2-subcommands]

[Advanced: Call peaks using MACS2 subcommands] [https://github.com/taoliu/MACS/wiki/Advanced%3A-Call-peaks-using-MACS2-subcommands]

`CTCF_ChIP_200K.bed.gz`, `CTCF_Control_200K.bed.gz` 可在MACS2 github仓库下载

下面流程不针对paired-end reads

Step1: Filter duplicates

默认最大允许的duplicated reads为1, `--keep-dup=1`

```
macs2 filterdup -i CTCF_ChIP_200k.bed.gz --keep-dup=1 -o CTCF_ChIP_200K_filterdup.bed
```

```
macs2 filterdup -i CTCF_Control_200K.bed.gz --keep-dup=1 -o CTCF_Control_200K_filterdup.bed
```

Step2: Decide the fragment length d

测序reads长度仅是DNA片段长度(插入序列)的一个末端, 需要评估该DNA长度来获得实际的富集情况. 该例子中, 预测的fragment length d为254bp

```
macs2 predicted -i CTCF_ChIP_200K_filterdup.bed -g hs -m 5 50
```

如果不想通过该延伸reads的方法获得DNA长度, 或对DNA长度有很好的估计, 可以跳过这一步

Step3: Extend ChIP sample to get ChIP coverage track

针对ChIP样本, 根据第二步评估的片段长度, 使用`pileup`生成BEDGRAPH格式的pileup track.

`pileup`步骤默认从5'到3'方向延伸reads, 如果处理一些DNAse-Seq data, 或者认为cutting site(应该是sheared DNA位置, sequenced read), 被短序列reads检测到的, 正好位于感兴趣片段的中间, 可以使用`-B`选项将reads向两边延伸

```
macs2 pileup -i CTCF_ChIP_200K_filterdup.bed -o CTCF_ChIP_200K_filterdup.pileup.bdg --extsize 254
```

文件`CTCF_ChIP_200K_filterdup.pileup.bdg`文件包含ChIP样本fragment pileup信号

Step4: Build local bias track from control

默认, MACS2 `callpeak` 函数通过选取围绕的1kb(set by --slocal), 10kb(set by --llocal), fragment length d的大小来自之前预测结果(`predicted`), 整个基因组背景选取最大的bias来计算local bias.

The d backgroud

一般而言, 为构建背景信息噪音track, 需要使用`pileup`函数向两个方向延伸control reads. 因为, 来自control 样本的cutting site包含的噪音代表了围绕它的一个区域(The idea is that the cutting site from control sample contains the noise representign a region surrounding it). 选取d/2长度

```
macs2 pileup -i CTCF_Control_200K_filterdup.bed -B --extsize 127 -o d_bg.bdg
```

文件`d_bg.bdg`包含来自control样本的d 背景

The slocal background

默认选取1kb窗口构建构建背景噪音track. 简单想象每一个测序的reads代表一个1kb围绕的噪音

```
macs2 pileup -i CTCF_Control_200K_filterdup.bed -B --extsize 500 -o 1k_bg.bdg
```

这里500位1k的一半. 然而ChIP信号track是通过将read延伸d 片段长度构建, 这里将1kb噪音乘以d/slocal, 254/1000=0.254

```
macs2 bdgopt -i 1k_bd.bdg -m multiple -p 0.254 -o 1k_bg_norm.bdg
```

The llocal background

来自更大区域的背景噪音可使用同上步相同步骤完成, llocal大小为10kb

```
macs2 pileup -i CTCF_Control_200K_filterdup.bed -B --extsize 5000 -o 10k_bg.bdg
```

```
macs2 dbgopt -i 10k_bg.bdg -m multiply -p 0.0254 -o 10k_bg_norm.bdg
```

The genome background

全基因组背景可计算:

`the_number_of_control_reads * fragment_length/genome_size`, 该例子中为: `199867 * 254 / 2700000000 ~= 0.0188023`

Combine and generate the maximum background noise

选取slocal(1k)和llocal(10k)背景间的最大(maximum)

```
macs2 bdgcmp -m max -t 1k_bg_norm.bdg -c 10k_bg_norm.bdg -o 1k_10k_bg_norm.bdg
```

接着, 通过和d背景比较选取最大(maximum)

```
macs2 bdgcmp -m max -t 1k_10k_bg_norm.bdg -c d_bg.bdg -o d_1k_10k_db_norm.bdg
```

最后, 使用`bdgopt`合并基因组背景

```
macs2 bdgopt -i d_1k_10k_bg_norm.bdg -m max -p 0.0188023 -o local_bias_raw.bdg
```

这里得到的文件`local_bias_raw.bdg`为BEDGRAPH文件, 包含了来自control样本的原始local bias

Step5: Scale the ChIP and control to the same sequencing depth

为比较ChIP和control信号, ChIP pileup和control lambda需要scale到相同的测序深度. `callpeak`默认是将大的样本scale到小的样本深度. 这样可保证噪音不会扩大, 提高最终结果的特异性.

该例子中, ChIP和control过滤后duplication得到的reads数目为199583和199867. 因此, control bias需要scale down, 199583/199867=0.99858

```
macs2 bdgopt -i local_bias_raw.bdg -m multiply -p 0.99858 -o local_lambda.bdg
```

输出文件命名为`local_lambda.bdg`, 因为该文件中的值可被认为是lambda(期待值), 可通过泊松分布和ChIP信号比较

Compare ChIP and local lambda to get the scores in pvalue or qvalue

使用某一统计模型, 根据ChIP信号和local lambda, 检测富集区域, 预测peaks

```
macs2 bdgcmp -t CTCF_ChIP_200k_filterdup.pileup.bdg -c local_lambda.bdg -m qpois -o CTCF_ChIP_200K_qvalue.bdg
```

或者

```
macs2 bdgcmp -t CTCF_ChIP_200k_filterdup.pileup.bdg -c local_bias.bdg -m ppois -o CTCF_ChIP_200k_pvalue.bdg
```

文件`CTCF_ChIP_200K_pvalue.bdg`和`CTCF_ChIP_200K_qvalue.bdg`包含了来自local 泊松检测的`-log10(p-values)`和`-log10(q-values)`. 这意味着, 每个碱基位置的ChIP信号将会使用泊松模型比较对应的local lambda.

Step7: Call peaks on score track using a cutoff

根据明确阈值选择peak区域. 这里使用`bdgpeakcall`函数检出narrow peak, 使用`bdgbroadcall`函数检出broad peak.

首先, 假如两个区域都超过了阈值, 但是区域间值较低, 假如间隔区域足够小, 则应该将两个相邻区域合并成为一个更大的区域. 该值设置为read长度, 因为read长度代表了数据集的分辨率. 针对`bdgpeakcall`, 需要使用`-g`选项设置第一步获得read长度或查看原始数据

第二, 无需检出太多小的peaks, 因此设置了最小的peak长度. 自动使用片段长度d最为最小的peak长度, 这里通过指定`-l`来表示d 长度

最后, 设置cutoff值. 输出文件中的值时`-log10格式`, 因此若需cutoff为0.05, 那么-log10转换后值约为1.3.

```
macs2 bdgpeakcall -i CTCF_ChIP_200K_qvalue.bdg -c 1.301 -l 245 -g 100 -o CTCF_ChIP_200K_peaks.bed
```

输出为narrowPeak 格式文件(典型的BED文件), 包含peak区域和峰尖位置

[github issue][https://github.com/taoliu/MACS/issues/353]

@huizhen2014 My two comments: 1) When the genome size is small, to keep only 1 unique pair may dilute the real signals. You may consider subsample your ChIP data or set a higher number for `keep-dup` option. You currently have 5 million 50bps reads which provide, assuming uniformly distributed in the bacterial genome, a 50 fold coverage. I would suggest you subsample only 1/50 of your data and redo the analysis. 2) The complexity in the input data is strangely low, with 99% redudant rate. You may need to check if the experiment/data really works. In general, an input is like whole genome DNA sequencing with a more uniform distribution than ChIP data. However only 37.2K unique 50bps reads were kept which means only small proportion of bacteria genome is covered in the input.

1. subsample the ChIP data/ set a higher number for `keep-up` option

With the current situtation, 5M/50bp, about 50 fold coverage, subsample only 1/50 of data and redo

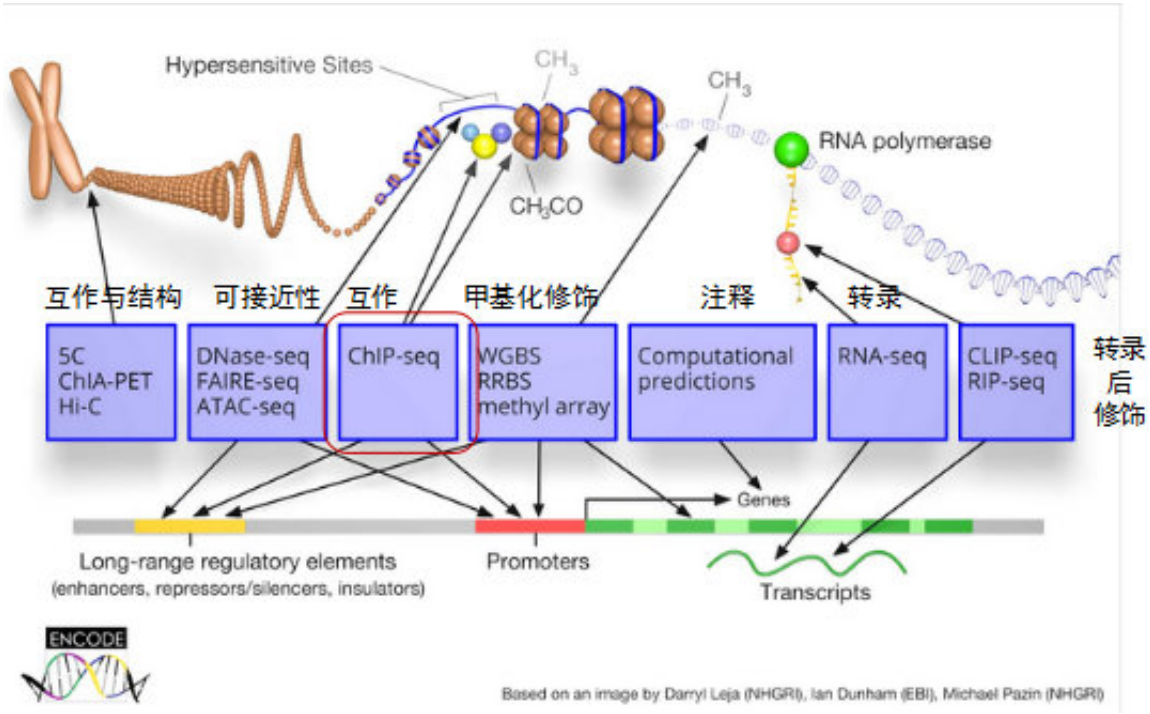
2. check the experiment/data works, since the complexity is strangely low

1. Regarding `--bw` : this parameter is for pre-scanning candidate peaks to build the shifting model (or more accurately to find the parameter to smooth signals as `--extsize`). The prescan is mainly controlled by `--bw` and `--mfold` options. The accuracy of the candidate peaks is not our goal since we only want to find "some enrichment across the genome to estimate the size of DNA fragment in the ChIP sample". I would only use the estimated fragment size from MACS2 to evaluate the data quality -- e.g. is my library over-/under-digested or the ChIP is not successful so too many noises are included. Even if the result is not what you expect, my suggestion is to always skip model building and set the `--extsize` a fixed value across all your samples (in many of our pipelines, we set 147bps or 200bps), so that we can compare the enrichment across different conditions.

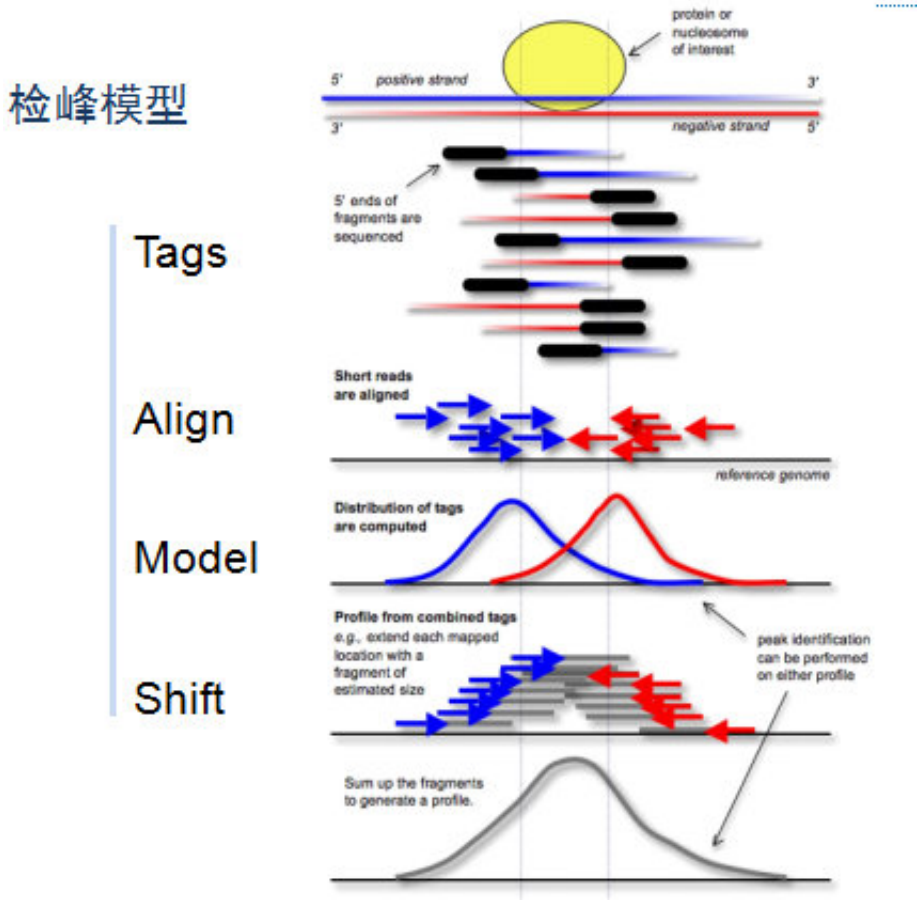
2. Yes. If you believe the fragment size is 300bps, then set `--extsize 300` . But keep in mind that, first `extsize` is for analyzing single-end ChIP data; second, ideally, the `extsize` should also be affected by the actual size of DNA bound/protected by the crosslinked protein that you are ChIPing, so the actual value may be different from the shearing size; third, `extsize` in fact is a parameter to smooth the data, so if your data is in good quality, a slightly off `extsize` shouldn't change your overall biological discovery.

1. `--bw` is used for shiftin model, such the DNAse-Seq, and exerts the similar function as `--extsize` in nomodel fashion.
2. It's trustful to adopt the nomodel and `--extsize` a fixed value across the analysis by knowing the approximate shared fragment length.

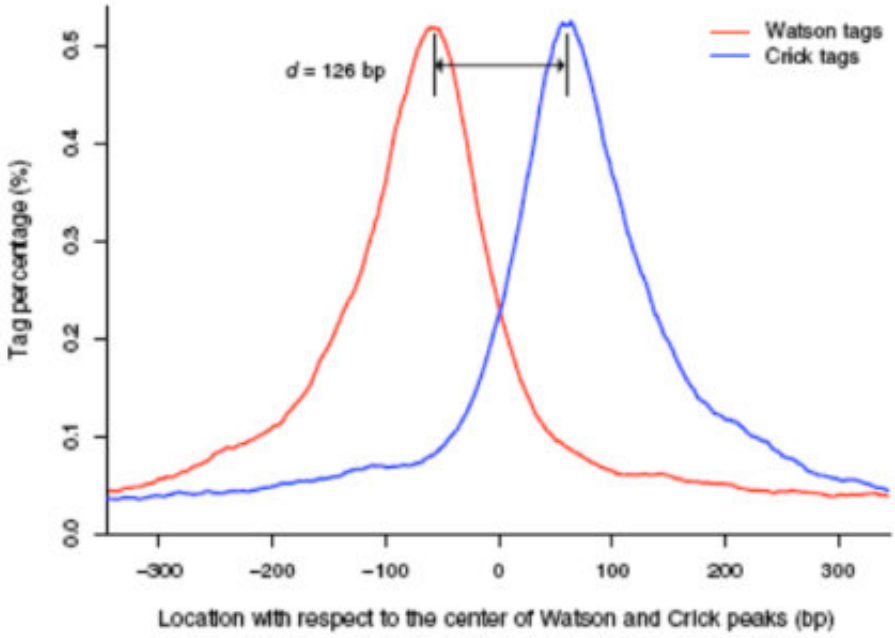
Miscellaneous



Call peaks



Peaks / d



[Evaluating ChIP-seq data][ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia]

Browser inspection and previously known sites

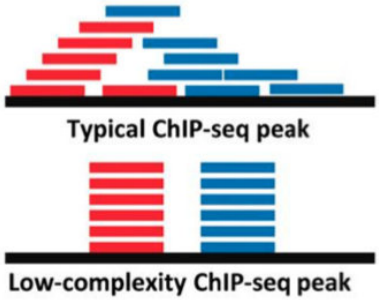
使用IGV查看, 尽管无法定量, 但是对于一个已知的结合位点, 相对于对照的read分布可用来检查. 真实的信号将显示一个非对称的reads比对分布. 可利用该方法检查有限个最强的信号位点.

Library complexity

NRF: 数据集中nonredundant mapped reads的比例

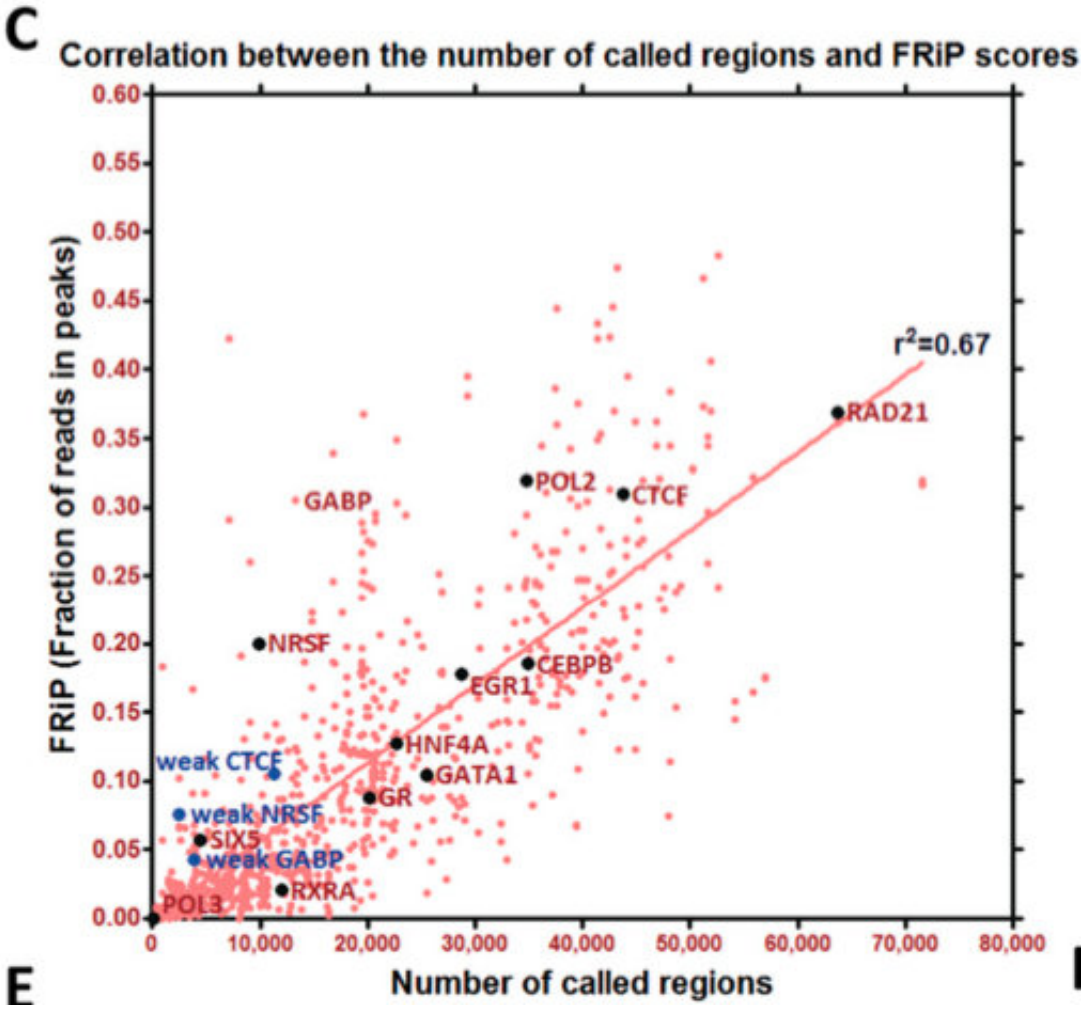
$$NRF = \frac{N_{nonred}}{N_{all}}$$

- NRF decreases with sequencing depth
- ENCODE: NRF ≥0.8 for 10M uniquely mapped reads



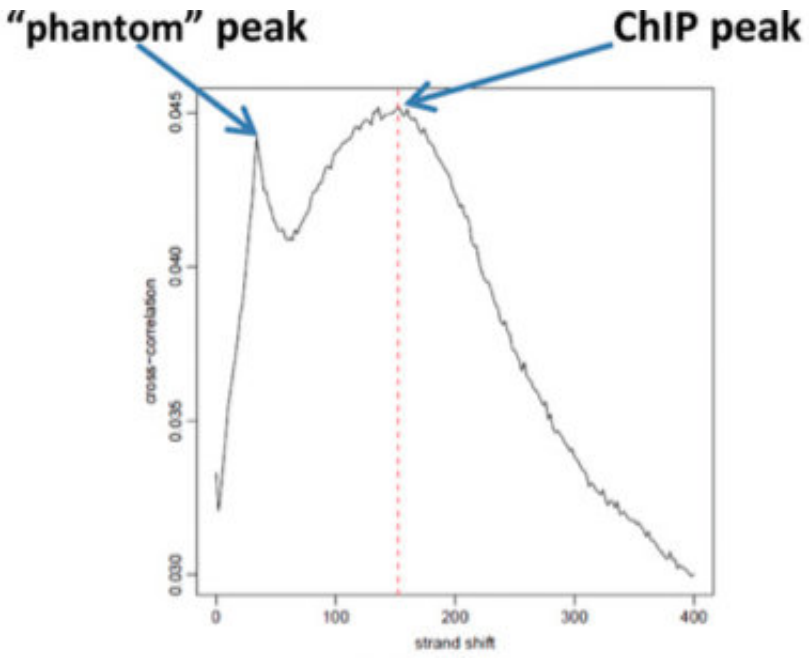
Measuring global ChIP enrichment(FRiP)

一般而言, 只有少部分reads比对到了显著性富集区域. 因此, 比对到峰区域的reads比例为一简单的指控指标, 称为: fraction of reads in peaks, FRiP. **一般而言, FRiP值和所检出区域正比且线性相关, 使用MACS默认参数检验哺乳动物时, FRiP富集比例在1%以上.** 但是, 高于该阈值并不意味着实验成功, 低于该阈值也不意味着实验失败.



Cross-correlation analysis

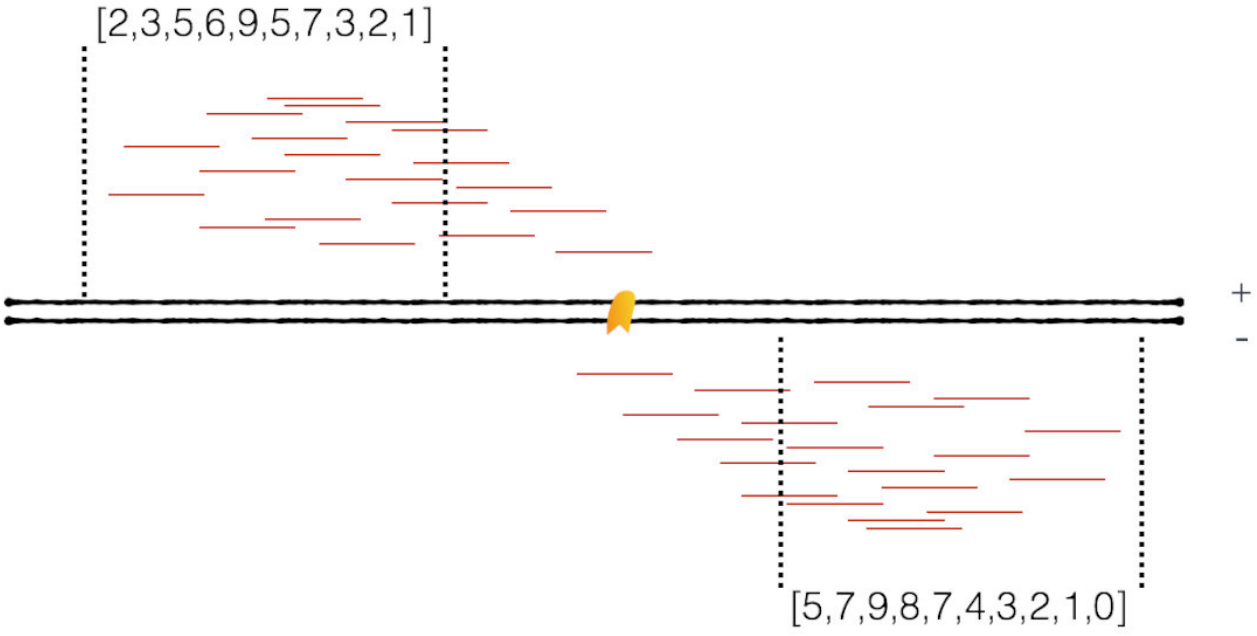
ChIP-seq一个有用的质控指标是检出峰具有独立性且是链交叉相关的(strand cross-correlation). 也就是说, 显著性富集位点富集的DNA序列tags标签(富集区域的reads)是以结合位点为中心, 同时分别比对到正负链的. 因此, 该质控标准基于基因组范围上tags的密度来量化其片段的聚集(IP clustering). 其计算为Crick链和Watson链的Person线性相关性.



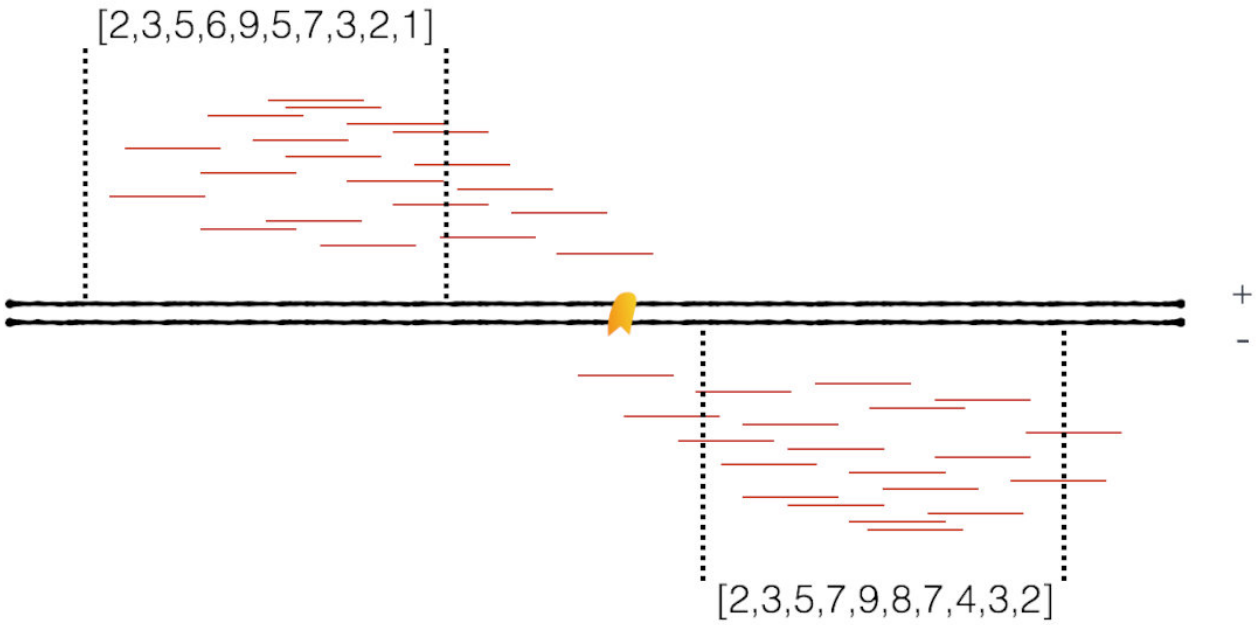
It is computated as the Person linear correlation between the Crick and the Waston strand, after shifting Waston by k base pairs. **This typically produces two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length("phantom" peak).**

Reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Person correlaiton between the per-position read count vectors for each strand is calculated. 也就是每个链每个位置的read count的向量之间的相关性系数:

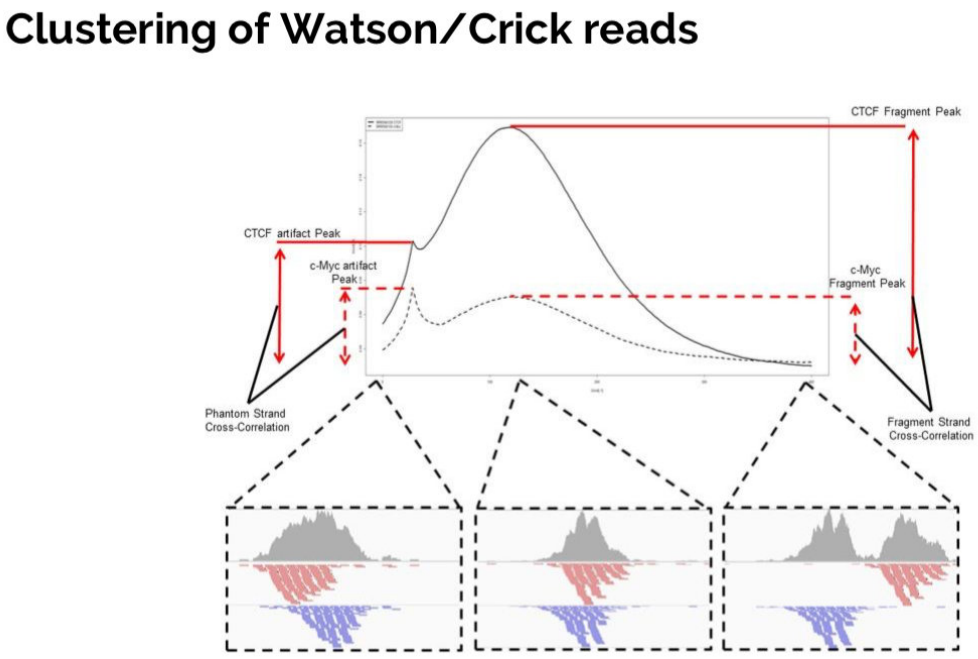
链移动为0时: Person=0.539



链位移5bp时: Person=0.931

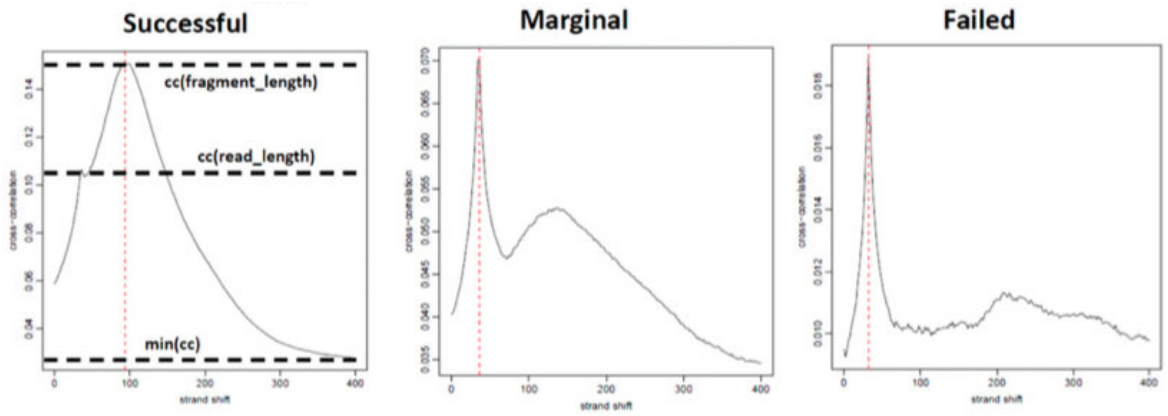


fragment-length cross-correlation peak 和 background cross-correlation (normalized strand coefficient, NSC)标准化后的比例, fragment-length peak and the read-length peak(relative strand correlation, RSC)标准化后比例, 是ChIP-Seq有力的信噪比指标. 高质量的测序数据集倾向于较大的 fragment-length peak 比 read-length peak.



NSC是最大交叉相关值除以背景交叉相关的比率(所有可能的链转移的最小交叉相关值)

RSC是片段长度相关值减去背景相关值除以phantom-peak相关值减去背景相关值



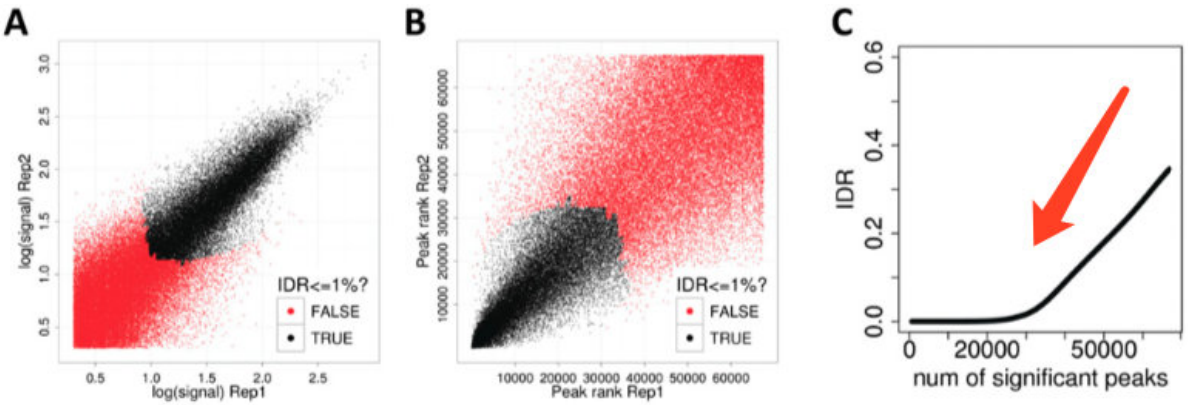
$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

ENCODE: NSC > 1.05 and RSC > 0.8 for point source TFs.

Consistency of replicates: Analysis using IDR

DR, irreproducible discovery rate

给定一对重复数据集, 它们检出的峰可以根据显著性(p-value, q-value), ChIP-to-input enrichment, 或者每个峰的reads覆盖度排序(be ranked). 假如两个重复样本位于相同的生物条件下, 最显著性的峰(likely to be genuine signals) 应该具有最高的一致性, 而低显著性的峰(likely to be noise), 应表现出低的一致性.



This consistency transition provides an internal indicator of the change from signal to noise and suggests how many peaks have been reliably detected.

Increased consistency comes from the fact that IDR uses information from replicates, whereas the FDR is computed on each replicate independently.

[phantompeakqualtools] [https://github.com/kundajelab/phantompeakqualtools]

该package用于计算来自ChIP-seq/DNase-seq/FAIRE-seq/MNase-seq数据的富集信息和质量值. 同时也可用于获得稳定的predominant fragment长度评估或者特征性tag shift值(or characteristic tag shift values).

该套程序用于处理tagAlign或BAM格式的illumina单端测序read数据，可用于:

1. 基于strand cross-correlation peak, 计算片段长度(prdominant insert-size length)

2. 基于相对phantom峰计算数据质量
3. Call Peaks and regions for punctate binding datasets

Run run_spp.R

`Rscript run_spp.R <options>`

参数

用于质控必须参数:

`-c=<ChIP_alignFile>` tagAlign/BAM 文件的全路径和名称(文件名后缀必须为tagAlign.gz, tagAlign, bam, bam.gz)

用于peak calling的必须参数:

`-i=<Input_alignFile>` 同上

可选参数:

`-s=<min>:<step>:<max>` 用于评估cross-correlation的链shift值, 默认为: -500:5:1500

`-speak=<strPeak>` 用户定义的cross-correlation peak strandshift

`-x=<min>:<max>` 排除的strand shifts(主要为了避免phantom peak的区域), 默认为: 10(readlen+10)

`-p=<nodes>` 并行处理节点数目, 默认为:0

`-fdr=<falseDiscoveryRate>` 用于peak calling的错误检出率

`-npeak=<numPeaks>` 检出peaks数目阈值

`-tmpdir=<tmpdir>` 临时文件夹

`-filtchr=<chrnamePattern>` 用于去除比对到特异染色体的tags的patterns e.g._will remove all tags that map to chromosomes with _ in their name

输出选项:

`-odir=<outputDirectory>` 输出文件夹目录

`-savn=<narrowpeakfilename>` NarrowPeak 文件名称(固定峰宽)

`-savn`

`-savr=<regionpeakfilename>` RegionPeak 文件名(围绕峰尖的富集区域的不同宽度的峰)

`-savr`

`-savd=<rdatafile>` 保存Rdata文件

`-savd`

`-savp=<plotdatafile>` 保存cross-correlaton图

`-savp`

`-out<resultfile>` 将peakshift/phantomPeak结果附加到该文件

`-rf` 如果plot/rdata/narrowPeak文件存在, 覆盖之; 否则会报错终止

`-clean` 在读取原始chip和质控文件后删除. CAUTION: Use only if the script calling `run_spp.R` is creating temporary files

用法

1. 评估strand cross-correlation peak/显著片段长度或质量值. `-out=<outFile>`将会输出, 同时包含多个重要特征值的数据集将附加其中

```
Rscript run_spp.R -c=<tagAlign/BAMfile> -savp -out=<outFile>
```

The file contains 11 tab delimited columns.

col.	abbreviation	description
1	Filename	tagAlign/BAM filename
2	numReads	effective sequencing depth i.e. total number of mapped reads in input file
3	estFragLen	comma separated strand cross-correlation peak(s) in decreasing order of correlation.
4	corr_estFragLen	comma separated strand cross-correlation value(s) in decreasing order (COL2 follows the same order)
5	phantomPeak	Read length/phantom peak strand shift
6	corr_phantomPeak	Correlation value at phantom peak
7	argmin_corr	strand shift at which cross-correlation is lowest
8	min_corr	minimum value of cross-correlation
9	NSC	Normalized strand cross-correlation coefficient (NSC) = COL4 / COL8
10	RSC	Relative strand cross-correlation coefficient (RSC) = (COL4 - COL8) / (COL6 - COL8)
11	QualityTag	Quality tag based on thresholded RSC (codes= -2:veryLow, -1:Low, 0:Medium, 1:High, 2:veryHigh)

输出处于最大cross-correlation值90%内的三个最大local maximal locations. 在几乎所有的情况, 列表中的top(第一个)值代表显著性片段长度. If you want to keep only the top value simply run:

```
sed -r 's/,[^\t]+//g' <outFile> > <newOutFile>
```

NSC值范围为最小的1到一个更大的数值. 1.1为一个重要阈值. NSC值远小于1.1(<1.05)的数据倾向于拥有偏低的信号/噪音值, 或更少的峰数目(this could be biological eg. a factor that truly binds only a few sites in a particular tissue type OR it could be due to poor quality)

RSC值范围从0到较大的正数值. 1为重要的阈值. RSC值显著低于1(<0.8)倾向于拥有低的信号/噪音值. 偏低值可能是由于ChIP失败或差的质量值, 低的read序列质量以因此带来很多错配, 较低的测序深度(significantly below saturation). 像NSC一样, 更少的结合位点的生物上可证实的数据也会带来更低的RSC值.

Qtag为RSC的阈值版本.

2. Peak calling

```
Rscript run_spp.R -c=<ChIP_tagalign/BAM_file> -i=<control_tagalign/BAM_file> -
fdr=<fdr> -odir=<peak_call_output_dir> -savr -savp -savd -rf
Rscript run_spp.R -c=<ChIP_tagalign/BAM_file> -i=<control_tagalign/BAM_file> -
npeak=<npeaks> -odir=<peak_call_output_dir> -savr -savp -savd -rf
```

3. For IDR analysis you want to call a large number of peaks(relaxed threshold) so that the IDR model has access to a sufficient noise component.

```
Rscript run_spp.R -c=<ChIP_tagalign/BAM_file> -i=<control_tagalign/BAM_file> -
npeak=300000 -odir=<peak_call_output_dir> -savr -savp -rf -out=<resultFile>
```

Notes:

- 务必过滤掉多重比对reads, 大量数目该reads将会严重影响phantom峰的系数和峰检出结果
- For the IDR(Irreproducible Discovery Rate) rescue strategy, one needs to pool reads from replicates and then shuffle and subsample the mapped reads to create two balanced pseudoReplicates. This is much easier to implement on tagAlign/BED read-mapping files using the unix 'shuf' command. So it is recommended to use the tagAlign format.
- 大多数情况, 可简单使用最大报告的strand correlation peak作为显著片段长度. 然而, 建议手动查看cross-correlation plot确保所选的最大峰不是错误的.
- 如果文库片段的选择存在问题, 那么数据的cross-correlation轮廓会有多个强cross-correlation峰.

输入文件

TagAlign files: 为文本格式的BED3+3比对格式, 包含6个tab分隔的列:

col.	abbrv.	type	description
1	chrom	string	Name of the chromosome
2	chromStart	int	The starting position of the feature in the chromosome. The first base in a chromosome is numbered 0.
3	chromEnd	int	The ending position of the feature in the chromosome or scaffold. The chromEnd base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as chromStart=0, chromEnd=100, and span the bases numbered 0-99.
4	sequence	string	Sequence of this read
5	score	int	Indicates uniqueness or quality (preferably 1000/alignmentCount).
6	strand	char	Orientation of this read (+ or -)

针对IDR rescue 分析, 需要使用shuffled比对文件并且subsample. 因此在使用TagAlign格式时会简单, 使用unix的shuf命令. 因此, 推荐使用TagAlign格式.

转换BAM为TAGALIGN文件

去除未必对, 低质量和多重比对reads后:

```
samtools view -F 0x0204 -o - <bamFile> | awk 'BEGIN{OFS="\t"}{if (and($2,16) > 0) {print $3,($4-1),($4-1+length($10)),"N","1000","-"} else {print $3,($4-1),($4-1+length($10)),"N","1000","+"} }' | gzip -c > <gzip_TagAlignFileName>
```

[ChIPQC] [http://www.bioconductor.org/packages/release/bioc/html/ChIPQC.html]

简单生成ChIP-seq实验或者样本的QC报告:

`samples` 包含样本sheet

```
> experiment = ChIPQC(samples)
> experiment
> ChIPQCreport(experiment)
```

默认条件下, HTML报告以及相关图片将会输出至子目录, `ChIPQCreport`, [例如] [http://chipqc.starkhome.com/Reports/tamoxifen/ChIPQC.html]

或者单个bam文件

```
> sample = ChIPQCsample("chip.bam")
> sample
> ChIPQCreport(sample)
```

1. Experiment sample sheet

第一步时构建样本sheet, 描述ChIP-seq实验. 可以是一个数据框或保存为一个csv文件. 同样该实验可以使用DiffBind包构建为DBA对象

```
> samples
  SampleID  Tissue Factor Replicate           bamReads                          Peaks
1   CTCF_1    A549   CTCF         1 reads/SRR568129.bam peaks/SRR568129_chr22_peaks.bed
2   CTCF_2    A549   CTCF         2 reads/SRR568130.bam peaks/SRR568130_chr22_peaks.bed
3   cMYC_1    A549   cMYC         1 reads/SRR568131.bam peaks/SRR568131_chr22_peaks.bed
4   cMYC_2    A549   cMYC         2 reads/SRR568132.bam peaks/SRR568132_chr22_peaks.bed
5   E2F1_1 HeLa-S3   E2F1         1 reads/SRR502355.bam peaks/SRR502355_chr22_peaks.bed
6   E2F1_2 HeLa-S3   E2F1         2 reads/SRR502356.bam peaks/SRR502356_chr22_peaks.bed
>
```

该样本sheet详述了实验数据, 同时也包含文件路径, 比对reads和called peaks. 另外, 若为csv格式文件, 可直接传递給`ChIPQC`

2. Constructing a `ChIPQCexperiment` object

`ChIPQC`接受一个样本sheet和一些可选参数, 针对每个样本计算质量值. It does this using the BiocParallel package, which by default will run in parallel, using all available cores on your machine.

`annotation` 指明分析的基因组

`chromosomes`指明需分析计算的染色体, 默认第一个; 选择`NULL`表示所有

`mapQCth` 表明过滤比对质量的阈值, 默认为15

`blacklist` 为一个文件或`GRanges`对象, 表示将这些区域的reads过滤掉

`exampleExp <- ChIPQC(samples, annotation="hg19")`

3. Quality metrics summary

```
> exampleExp
Samples: 6 : CTCF_1 CTCF_2 ... E2F1_1 E2F1_2
        Tissue Factor Replicate Peaks
CTCF_1    A549   CTCF         1  1118
CTCF_2    A549   CTCF         2   648
cMYC_1    A549   cMYC         1   253
cMYC_2    A549   cMYC         2   159
E2F1_1 HeLa-S3   E2F1         1   325
E2F1_2 HeLa-S3   E2F1         2   249
        Reads Map% Filt%   Dup% ReadL FragL RelCC  SSD  RiP% RiBL%
CTCF_1 341055  100  26.3 16.600    28   131 2.740 2.53 31.20  1.33
CTCF_2 303856  100  28.4  7.320    28   131 2.690 1.43 12.80  0.00
cMYC_1 287462  100  31.0 13.600    28    97 0.347 1.47  6.59  1.78
cMYC_2 317537  100  29.9  4.540    28   129 0.386 1.09  2.79  0.00
E2F1_1 223580  100  31.7  1.010    28   101 0.701 1.29  7.80  2.00
E2F1_2 194919  100  31.8  0.663    28   107 0.303 1.40  5.36  2.79
```

第一行描述样本数量, 以下信息可通过函数`QCmetrics(exampleExp)`获得

```
> QCmetrics(exampleExp)
        Reads Map% Filt%   Dup% ReadL FragL RelCC  SSD  RiP% RiBL%
CTCF_1 341055  100  26.3 16.600    28   131 2.740 2.53 31.20  1.33
CTCF_2 303856  100  28.4  7.320    28   131 2.690 1.43 12.80  0.00
cMYC_1 287462  100  31.0 13.600    28    97 0.347 1.47  6.59  1.78
cMYC_2 317537  100  29.9  4.540    28   129 0.386 1.09  2.79  0.00
E2F1_1 223580  100  31.7  1.010    28   101 0.701 1.29  7.80  2.00
E2F1_2 194919  100  31.8  0.663    28   107 0.303 1.40  5.36  2.79
```

`Dup%`: 至少包含一个其他read比对到基因组确切位置的read比率(The percentage of reads that map to the exact position in the genome as at least one other read is the reported). 上图, 可见重复率具有很大的变异, **好的ChIPs质量数据,期待狭窄地结合的转录因子拥有很高的富集度的区域, 该区域将包含来自相同位置的片段. 当因子具有很高的结合性时, 就会产生具有生物学意义的'duplicate'片段.**

期待看到指控样本表现出更低的duplicate率(5%), ChIP样本拥有较高的比率(15-20%), 但是不会过高(>50%). 过高的duplicate率可能由于不均一的PCR扩增.

read长度, 源于bam文件数据, 随后为评估的平均片段长度. 片段长度是通过chipseq包, 系统性相互朝向地shift每条链上的reads直到实现最小的基因组覆盖度(estimated by methods in implemented in the chipseq package by systematically shifting the reads on each strand towards each other until minimum genome coverage is achived).

`RelCC`为`RelativeCC`, 通过比较maximum cross coverage peak(at the shift size corresponding to the fragment length)和 the cross coverage at a shift size corresponding to the read length, 较高的值代表好的实验(一般大于等于1)

The fragment length is estimated from the data by systematically shifting the reads on each strand towards each other until the highest degree of cross-vocerage is obtained

RelativeCC一般较高的值(一般为2或更大)表明好的富集结果

`SSD` 为htSeq Tools采用的另一个富集证据. It is computed by looking at the standard deviation of signal pile-up along the genome normalised to the total number of reads. 富集的样本典型性拥有显著性pile-up区域, 因此更高的SSD值越能代表好的富集结果.

SSD值接近1一般对应较差的富集样本, 然而成功的ChIPs值一般为1.5, 较高的富集样本的SSD值可以为2或更高

`RiP%` 代表了跨越called peak的reads的百分率(也称为FRIP). 可认为是'signal-to-noise' 比值, 表示来自结合位置的片段reads比上背景reads. `RiP%`值针对ChIPs一般在5%或更高表示富集成功.

RiP%针对ChIPs为15%或更高方反应成功的富集

`RiBL%` 为reads落在blacklist区域的reads比率. 来自blacklisted的信号能够混淆peak callers和片段长度评估...

- **ID** - Unique sample ID.
- **Tissue/Factor/Condition** - Metadata associated to sample.
- **Replicate** - Number of replicate within sample group
- **Reads** - Number of sample reads within analysed chromosomes.
- **Dup%** - Percentage of MapQ filter passing reads marked as duplicates
- **FragLen** - Estimated fragment length by cross-coverage method
- **SSD** - SSD score (htSeqTools)
- **FragLenCC** - Cross-Coverage score at the fragment length
- **RelativeCC** - Cross-coverage score at the fragment length over Cross-coverage at the read length
- **RIP%** - Percentage of reads wthin peaks
- **RIBL%** - Percentage of reads wthin Blacklist regions

4. Generating a summary QC report for experimental sample groups

`ChIPQCreport(exampleExp)`

[Report][http://ChIPQC.starkhome.com/ Reports/exampleExp/ChIPQC.html.]

[Detail][https://rdrr.io/bioc/ChIPQC/man/ChIPQC.html]

ChIPQC(experiment, annotation, chromosomes, samples, consensus=FALSE, bCount=FALSE, mapQCth=15, blacklist=NULL, profileWin=400, fragmentLength=125, shifts=1:300,...)

`experiment`	A specification of the ChIP-seq experiment to evaluate. This can either be a dataframe, a filename for a .csv file, or a `DBA` object as defined in the `DiffBind` package. Columns names in sample sheet may include: • SampleID: Identifier string for sample • Tissue: Identifier string for tissue type • Factor: Identifier string for factor • Condition: Identifier string for condition • Treatment: Identifier string for treatment • Replicate: Replicate number of sample • bamReads: file path for bam file containing aligned reads for ChIP sample • bamControl: file path for bam file containing aligned reads for control sample • ControlID: Identifier string for control sample • Peaks: path for file containing peaks for sample. Format determined by PeakCaller field or caller parameter • PeakCaller: Identifier string for peak caller used. If Peaks is not a bed file, this will determine how the Peaks file is parsed. If missing, will use default peak caller specified in caller parameter. Possible values: ◦ “raw”: text file file; peak score is in fourth column ◦ “bed”: .bed file; peak score is in fifth column ◦ “narrow”: default peak.format: narrowPeaks file ◦ “macs”: MACS .xls file ◦ “swembl”: SWEMBL .peaks file ◦ “bayes”: bayesPeak file ◦ “fp4”: FindPeaks v4 • PeakFormat: string indicating format for peak files; see PeakCaller and `dba.peakset` • ScoreCol: column in peak files that contains peak scores • LowerBetter: logical indicating that lower scores signify better peaks See the documentation for the `sampleSheet` parameter of `dba` for details.

- annotation 基因组及版本，或之前定义的(by QCannotation)

  - “hg19”: Human, version 19
  - “hg18”: Human, version 18
  - “mm10”: Mouse, version 10
  - “mm9” : Mouse, version 19
  - “rn4” : Rat, version 4
  - “ce6” : C. Elgans, version 6
  - “dm3” : D. Melanogaster, version 3

  Alternatively, you can construct your own annotation; see the package vignette for more information.

或构建自己的注释信息

- chromosomes 需计算QC的染色体, 若未指明, 则仅计算第一条染色体，如果为NULL, 则计算所有. ("chr18")
- samples list或ChIPsampe对象
- consensus 若consensus为GRanges对象, 那么在构建peak-based metrics, 所有样本均使用该peakset. 如果consensus=TRUE, 那么一个consensus peakset将会生成且用于所有样本, consensus peakset通过合并所有提供的peaksets中重叠的peaks, 保留了至少两个样本包含重叠的peaks, 取消该选项, 设置consensus=FALSE. 此时, 只会计算提供了peaksets的样本, controls没有提供peakset将不会计算

- bCount if TRUE, the peak scores for all samples will be based on read counts using dba.count using a consensus peakset.
- mapQCth 整数, 表示比对质量阈值
- blacklist GRangs对象或指定的bed文件包含不进行分析的基因组区域
- profileWin 整数, 表明碱基单位的宽度, 用于peak profiles. Peaks will be centered on their summits, and include half the window size upstream and half downstream of this point
- fragmnetLength 整数, 期待的文库片段长度
- shifts 当计算最佳的shift大小时所尝试一个向量值
