**ChIP-seq(version2.0)**

**Gui tuantuan & Zoukeke**

[**guituantuan@sibs.ac.cn**](mailto:guituantuan@sibs.ac.cn)

**2015.5.20**

1.背景介绍：

CHIP-seq: 将染色质免疫共沉淀技术（Chromatin immunoprecipitation, ChIP）与下一代高通量测序技术相结合，是全基因组水平研究DNA与蛋白质相互作用的有力工具和标准方法。

在生理状态下，把细胞内的DNA与蛋白质交联（Crosslink）后裂解细胞，分离染色体，通过超声或酶处理将染色质随机切割，利用抗原抗体的特异性识别反应，将与目的蛋白相结合的DNA片段沉淀下来，再通过反交联（Reverse crosslink）释放结合蛋白的DNA片段，最后通过多种技术（定量PCR、芯片、测序等）获得DNA片段的序列。

最常见的两种ChIP实验技术是N-ChIP[12]和X-ChIP[10]。前者用来研究DNA与高结合力蛋白的互作，采用核酸酶消化染色质，适用于组蛋白及其异构体方面的研究；后者用来研究DNA与低结合力蛋白的互作，采用甲醛或紫外线进行DNA和蛋白交联，超声波片段化染色质，适用于多数非组蛋白方面的蛋白的研究。

ChIP-seq的一般实验流程（见下图）：

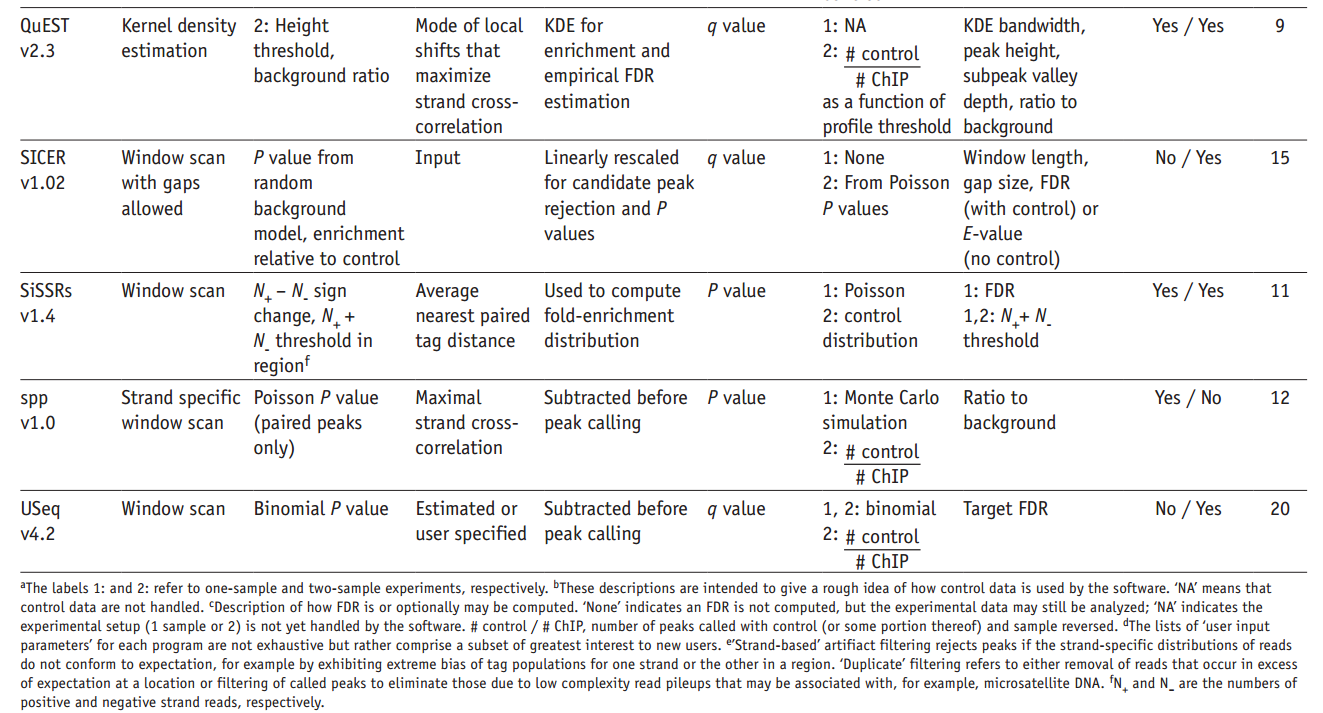
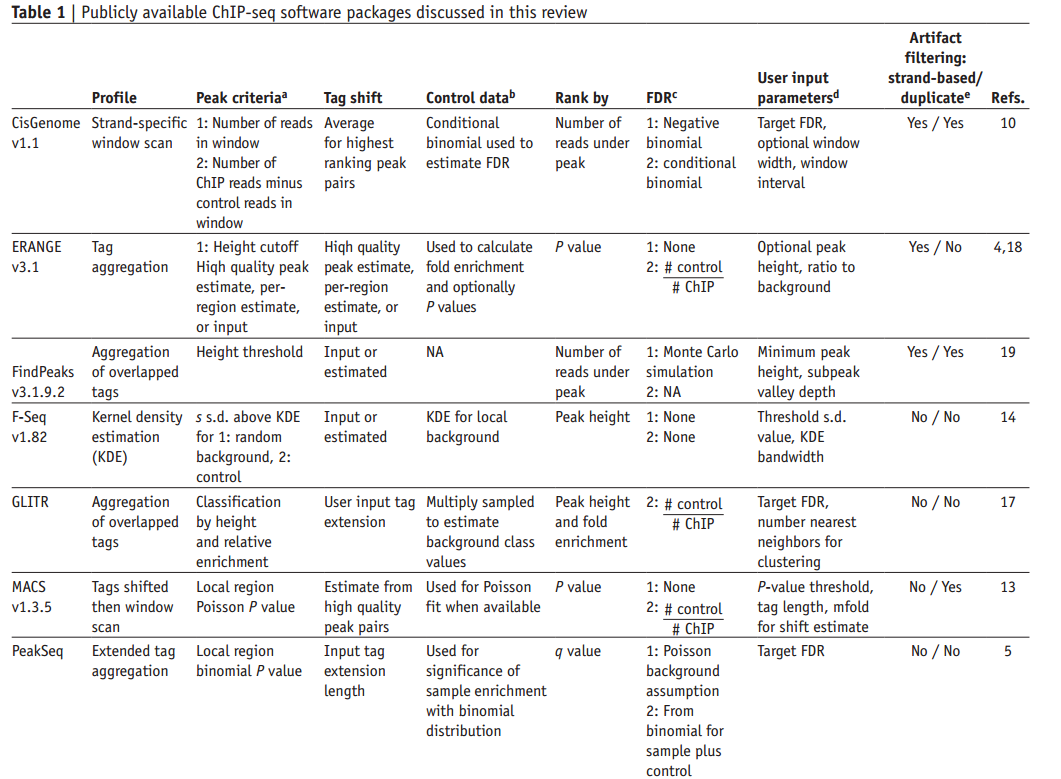
1. 解剖/收集样本
2. 交联
3. 溶解细胞和降解
4. 免疫沉淀
5. 测序

因此我们需将得到的与蛋白相互作用的那部分DNA序列信息进行分析，主要步骤见下：

1. 对测序质量进行控制 —> 去掉较短和质量较差的序列
2. 将reads mapping到基因组上
3. Peak-calling —> 找到真实的蛋白结合位点
4. Peak-calling后期分析

前两歩是一些常规的步骤，不做详细介绍。在这里主要介绍下Peak-calling的方法，

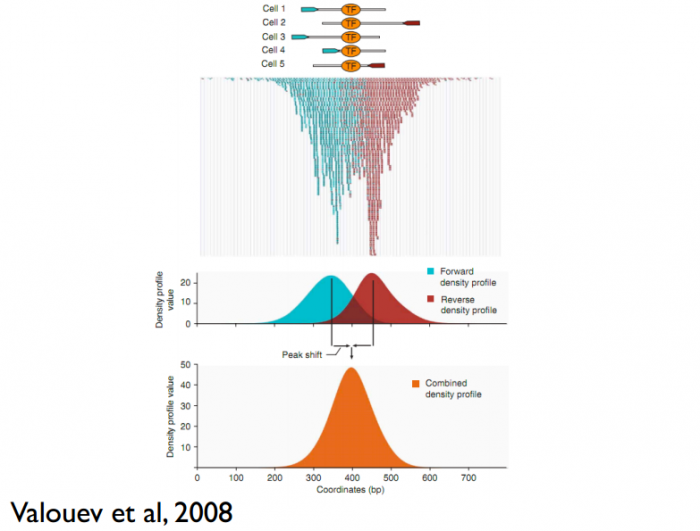
常见的Peak-calling的方法及软件见下图[5]：



这里主要对软件MACS方法进行讨论：

MACS背景：

read只是跟随着TF一起沉淀下来的DNA fragment的末端，read的位置并不是真实的TF结合的位置。所以在peak-calling之前，延伸read是必须的。不同TF大小不一样， 对read延伸的长度也理应不同。我们知道，测得的read最终其实会近似地平均分配到正负链上，这样，对于一个TF结合热点而言，read在附近正负链 上会近似地形成“双峰”。[MACS](http://www.plob.org/tag/macs)会 以某个window size扫描基因组，统计每个window里面read的富集程度，然后抽取（比如1000个）合适的（read富集程度适中，过少，无法建立模型，过 大，可能反映的只是某种偏好性）window作样本，建立“双峰模型”。最后，两个峰之间的距离就被认为是TF的长度D，每个read将延伸D/2的长 度。如下图：

[](http://upload.plob.ybzhao.com/wp-content/uploads/2014/05/2.png)

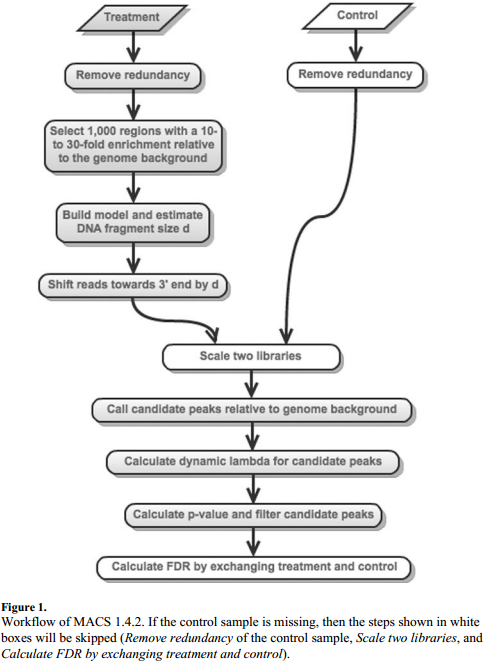
在某些情况下，比如，对组蛋白修饰的ChIP-seq数据peak-calling时，“双峰模型”会建立失败，这是因为组蛋白修饰往往并不是孤立存在的，可能很长一段染色质区间都被同一个组蛋白修饰占据，换句话说，组蛋白修饰的peak并不典型。这时，只要多加一个参数：

–nomodel –shiftsize=number

–nomodel将略过“双峰模型”建立的过程，而–shiftsize将人为指定reads延伸的长度。因为我们知道一个核小体上大概缠绕着147bp长的DNA，在对组蛋白修饰做peak-calling时我们可以指定：

–nomodel –shiftsize=73

MACS 1.4.2流程表见下图[6]：



CHIP-seq的后期分析：

TF ChIP-seq，为了找到TF调控的相关基因，一般会在call出peak之后分析：

1. TF-DNA binding site所覆盖的基因
2. TF-DNA binding site富集区域（TSS,gene,promoter）的比例
3. TF-DNA binding site的motif

Histome modification的位点，例如H3K4me3, K3K36me3,H3K27me3,H3K27ac, etc通常不具有motif，因此通常多他们所富集的区域感兴趣。

2.软件描述、所需其他软件、脚本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **软件/数据库** | **版本号** | **机器配置** | **下载地址** | **参考文献** | **功能** |
| Fastqc |  | Linux | http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ |  | 质量控制 |
| fastx |  | Linux | http://hannonlab.cshl.edu/fastx\_toolkit/index.html |  | 质量控制 |
| Bowtie2 |  | Linux | http://bowtie-bio.sourceforge.net/bowtie2/index.shtml | Langmead B, Salzberg S. [Fast gapped-read alignment with Bowtie 2](http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html). [*Nature Methods*](http://www.nature.com/nmeth). 2012, 9:357-359. | 映射到基因组上 |
| samtools |  | Linux | http://samtools.sourceforge.net/ | Li H.\*, Handsaker B.\*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: [19505943](http://www.ncbi.nlm.nih.gov/pubmed/19505943)] |  |
| MACS |  | Python2.6 | http://liulab.dfci.harvard.edu/MACS/ | Zhang et al. Model-based Analysis of ChIP-Seq ([MACS](http://liulab.dfci.harvard.edu/MACS/index.html)). Genome Biol (2008) vol. 9 (9) pp. R137 | Peak-Calling |
| UCSC |  |  | http://genome.ucsc.edu/cgi-bin/hgCustom |  | 可视化 |
| bedtools |  | Linux | https://github.com/arq5x/bedtools2/releases | Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842 | 后期分析 |

3.软件安装

Wget http…

Tar zxvf .gz

Cd filename

./configure

Make

Make install

4.参数说明及使用建议：

重要的单参数：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **软件** | **命令** | **参数** | **说明** | **默认值** |
| MACS | Macs14 | -t | ChIP-seq treatment files. REQUIRED. When ELANDMULTIPET is selected, you must provide two files separated by comma, e.g. s\_1\_1\_eland\_multi.txt,s\_1\_2\_eland\_multi.txt |  |
|  |  | -c | Control files. When ELANDMULTIPET is selected, you must provide two files separated by comma, e.g.  s\_2\_1\_eland\_multi.txt,s\_2\_2\_eland\_multi.txt |  |
|  |  | -n | Experiment name, which will be used to generate output file names. | NA |
|  |  | -f | Format of tag file, "AUTO", "BED" or "ELAND" or"ELANDMULTI" or "ELANDMULTIPET" or "ELANDEXPORT" or"SAM" or "BAM" or "BOWTIE". The default AUTO option will let MACS decide which format the file is. Please check the definition in 00README file if you choose EL AND/ELANDMULTI/ELANDMULTIPET/ELANDEXPORT/SAM/BAM/BOWTI  E. | AUTO |
|  |  | -g | Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts:'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for fruitfly (1.2e8) | hs |
|  |  | --nomodel | Whether or not to build the shifting model. If True, MACS will not build model. by default it means shifting size = 100, try to set shiftsize to change it. | False |
|  |  | --shiftsize | The arbitrary shift size in bp. When nomodel is true, MACS will use this value as 1/2 of fragment size. | 100 |
|  |  | -B | Whether or not to save extended fragment pileup at every bp into a bedGraph file. When it's on, -w, --space and --call-subpeaks will be ignored. When--single-profile is on, only one file for the whole genome is saved. WARNING: this process is time/space consuming!!When --single-profile is on, only one file for the whole genome issaved. WARNING: this process is time/space consuming!! |  |
|  |  | -S | When set, a single wiggle file will be saved for treatment and input. | False |
|  |  | --call-subpeaks | If set, MACS will invoke Mali Salmon's PeakSplitter soft through system call. If PeakSplitter can't be found, an instruction will be shown for downloading and installing the PeakSplitter package. -w option needs to be on and -B should be off to let it work. | False |
|  |  | -p | Pvalue cutoff for peak detection. | 1e-5 |

常用的参数组合：

1.组蛋白的基本参数组合：

|  |  |  |  |
| --- | --- | --- | --- |
| **软件** | **命令** | **参数组合** | **文献** |
| MACS | Macs14 | macs14 -t BROAD\_GM12878\_H3K36me3.bam -c  BROAD\_GM12878\_H3K36me3\_Control.bam -g hs -n  BROAD\_GM12878\_H3K36me3 --nomodel --shiftsize 73 -B -S –-pvalue 1e-3 --call-subpeaks | **Identifying ChIP-seq enrichment using MACS** |

2.转录因子的基本参数组合

|  |  |  |  |
| --- | --- | --- | --- |
| **软件** | **命令** | **参数组合** | **文献** |
| MACS | Macs14 | macs14 -t HAIB\_T47D\_FoxA1.sam -n HAIB\_T47D\_FoxA1 –g hs -B -S -  -call-subpeaks | **Identifying ChIP-seq enrichment using MACS** |

5.结果处理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **作者** | **脚本** | **目的** | **输入** | **输出** |
|  |  |  |  |  |

6.结果图形化展示：

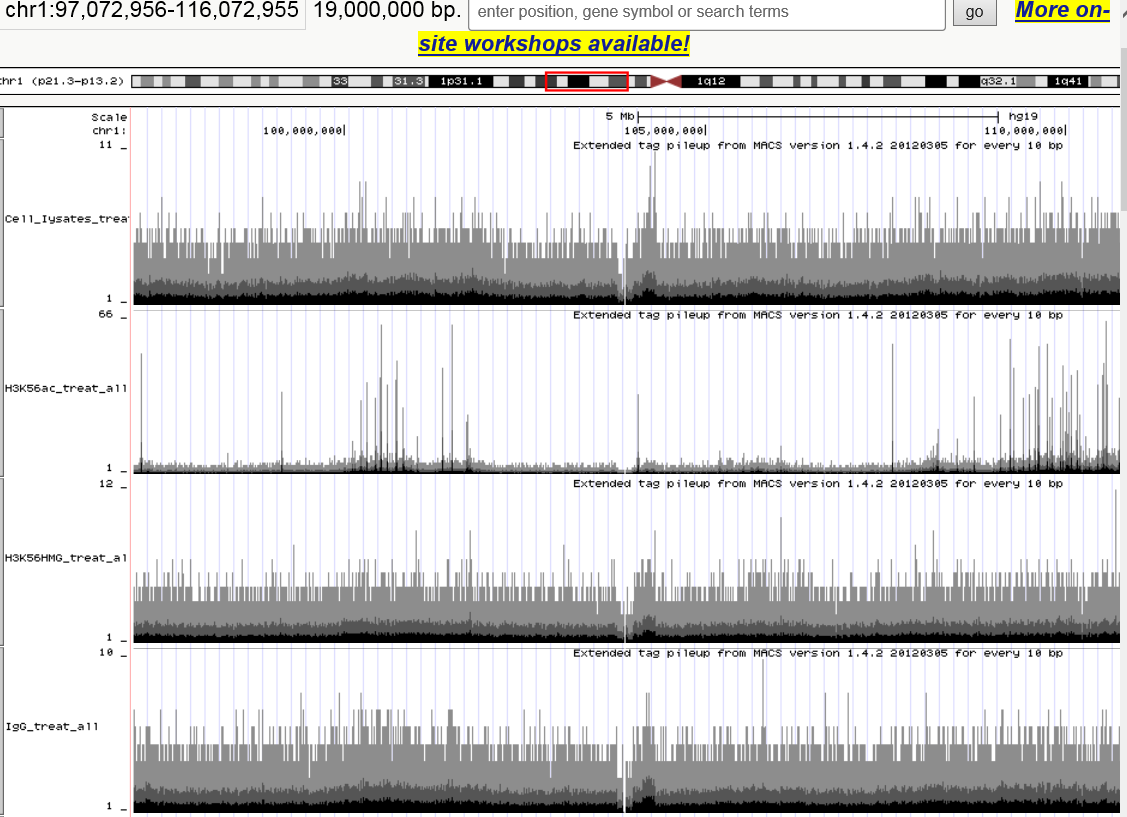
7.实例：

脚本见文件：

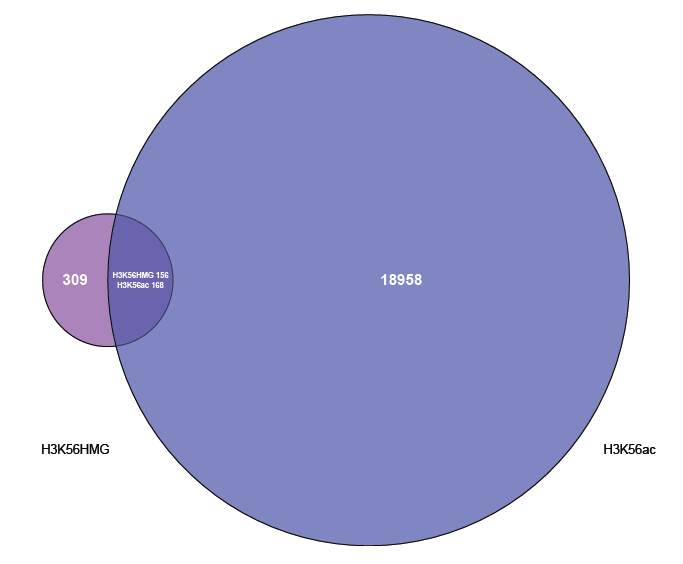
\10.10.155.251\incoming\Personal\Protocol\新流程构建\Epigenetics\ChIP-seq.pipeline.sh

结果见下：

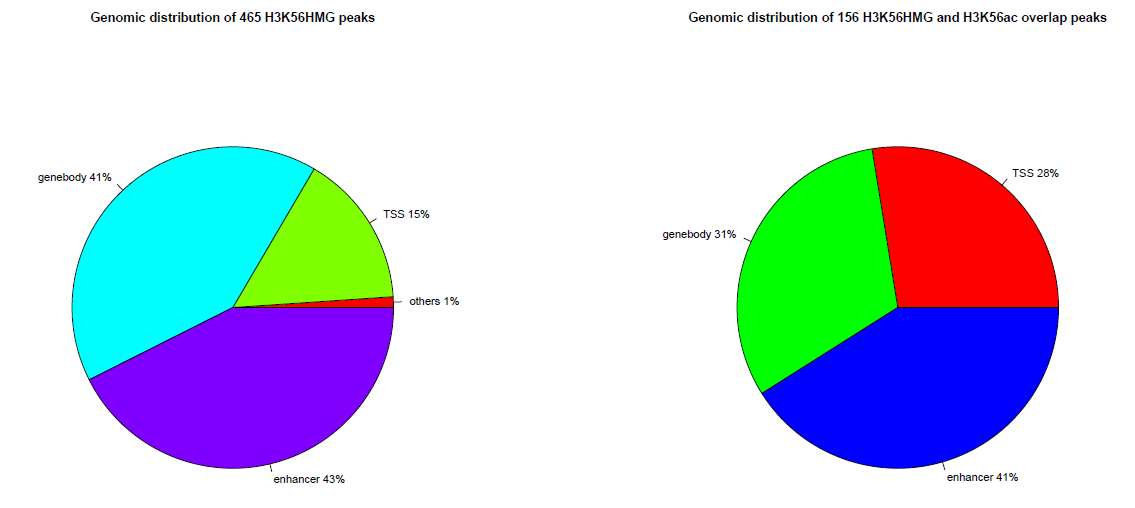
Call peaks的UCSC展示如下图所示：



H3K56ac和H3K56HMG的Peaks如下图所示：



Peak在染色体上的分布如下图所示：



8.参考文献：

1.Langmead B, Salzberg S. [Fast gapped-read alignment with Bowtie 2](http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html). [*Nature Methods*](http://www.nature.com/nmeth). 2012, 9:357-359.

2. Li H.\*, Handsaker B.\*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: [19505943](http://www.ncbi.nlm.nih.gov/pubmed/19505943)]

3. Zhang et al. Model-based Analysis of ChIP-Seq ([MACS](http://liulab.dfci.harvard.edu/MACS/index.html)). Genome Biol (2008) vol. 9 (9) pp. R137

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

5. Jianxing Feng1\*, Tao Liu2\*, Bo Qin1, Yong Zhang1, and Xiaole Shirley Liu2. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012 September ; 7(9): . doi:10.1038/nprot.2012.101

6. Shirley Pepke, Barbara Wold,Ali Mortazavi2. computation for chIP-seq and rNA-seq studies. nature methodS SuPPLement. 2009 NOVEMBER ,VOL.6 NO.11s