1 简介

本流程为Chip-seq项目标准信息分析流程，可处理HiSeq各平台及BGISEQ500（Zebra）平台数据，利用soap/bwa/bowtie/ELAND等软件进行比对，然后采用MACS/MACS2/SICER等软件对比对结果进行peak calling，再进行peak统计，相关基因及GO、pathway注释等分析；流程使用方法简单，对脚本的并行化处理，又最大限度压缩信息分析周期，对于多样品项目极为便利，可达到快速、高效、准确交付给客户结果的目的。

2 适用范围

测序平台：HiSeq 2000，HiSeq 2500，HiSeq 4000，BGISEQ-500（Zebra）

测序策略：PE/SE

物种：具有参考基因组的物种

**3 流程使用方法和注意事项**

1. 首先配置config文件,参数说明见下面5。例子/ifs1/ST\_EPI/USER/limeiyan/project/CG\_project/chip-sbs/chip-kit/shengmao\_6monthstability/config.在配置config文件时候Analysis = clean align peakcall report这个参数可以根据自己分析需求来写，这里给的是我对于BGISEQ-500数据分析需求，这个参数全部是Analysis = clean align peakcall peakanno motif diffpeak diffanno upload report。
2. 生成脚本：perl /ifs1/ST\_EPI/PMO/Anaysis\_pipline\_database/analysis\_pipline/Chip-seq2015/pipeline/Chip-sep\_2015a\_zebra.pl -i config。这里程序也可以用科服脚本程序/ifs4/BC\_PUB/biosoft/pipeline/EPI/Chip-Seq/Chip-seq\_2016a/pipeline/Chip-seq\_2016a.pl。因委科服这个程序很多R都是在程序里直接输入了，我们集群可能不存在，而且也没有num\_proc,建议用前面我目录下的程序。
3. 进入shell目录运行:nohup sh main.shi & 在登录节点可以直接运行，自动投递任务。
4. 全部脚本跑完以后，sh report.sh outdir(和config里输出目录一样就可以) 就可以生成本地报告。这一步是我最近刚弄，我这边没有问题，大家先试看有没有问题啊。如果有问题可能就要改一下arf里面东西。
5. 给大家测试的数据在/ifs1/ST\_EPI/USER/limeiyan/project/CG\_project/chip-sbs/chip-kit/exercise\_Chip-seq，需要的相关文件我都放在目录下了。我提供给大家的数据用的是H3K4me3 是属于sharp peak的。

**注意：**目前本流程提供3个软件进行peak calling，该怎么选择？目前流程中，**macs14**主要是用于转录因子或sharp peak数据的处理，如CTCF，H3K4me3等；**SICER**主要用于组蛋白修饰或者结合区域较长的结合蛋白进行分析，如H3k9me3、H3K27me3、RNA Polymerase II等；而最新版本的**macs2**即能处理转录因子数据也能处理组蛋白修饰等长结合区域数据的处理，需要设置好对应的参数并加上–broad。所以，在项目开始前，需要提前向客户了解蛋白信息，是转录因子还是组蛋白修饰？是宽峰还是窄峰？从而选择不同的Peak calling 软件及参数；得到结果后可以从peakstat.xls结果中大致看出peak是否合理。

**4 功能列表**

（1）下机数据QC：主要根据config文件中数据目录中的report文件，对Q20, Q30, Error Rate, Adapter Rate, 测序数据量等基本QC项进行统计；

（2）过滤：使用SOAPnuke去除原始数据中低质量、接头污染、N含量过高的reads，同时可在这一步进行数据量截取；

（3）比对：比对模块可使用soap（默认）/bwa/bowtie/ELAND来完成，需在config文件中进行相应参数配置；

（4）Peak检测： 使用MACS（默认）/MACS2/SICER等软件对处理后的比对结果进行计算统计，得到peak信息。分析peak长度、深度及其在基因元件上的分布；

（5）Peak注释：注释出peak相关基因并进行GO和pathway富集分析；（6）计算样品间差异Peak： 基于MAnorm工具，对两个样品进行差异分析，确定存在样品间差异修饰的区间，并分析这些差异peak在基因元件上的分布；

（7）样品间差异Peak注释：注释出样品间差异peak相关基因并进行GO和pathway富集分析；

（8）RNA-seq差异表达基因（DEGs）与ChIP-Seq peaks关联分析；

（9）Motif分析：  基于MEME SUITE工具，对各样品检测出的peak进行motif分析，预测motif序列及位置信息。

### 5参数说明

各参数详细信息描述如下:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 参数 | 参数类型 | 参数说明 |
| ProjectType = | <STRING> | 项目类型 |
| ProjectNo = | <STRING> | 大项目名 |
| SubProjectCode = | <STRING> | 子项目代码，-P参数 |
| DPD = | <STRING> | 部门简写 |
| Query = | <STRING> | -q参数 |
| CDTS\_usr = | <STRING> | 本项目CDTS账号，用于上传数据到CDTS |
| CDTS\_pwd = | <STRING> | 本项目CDTS密码，用于上传数据到CDTS |
| CDTS\_loc = | <STRING> | 上传结题报告的地点（大写），SZ、HK、WH |
| Pipeline\_Name = | <STRING> | 流程类型，Chip-seq |
| Pipeline\_Version = | <STRING> | 流程版本，2016a |
| Analyst\_Email = | <STRING> | 信息分析人员邮箱 |
| SeqPlatform = | <STRING> | 测序平台，HiSeq 或者 Zebra |
| FqType = | <STRING> | 测序类型，决定后续的过滤及比对分析，PE or SE |
| DataSize = | <STRING> | 这个可以是G、M、K的单位，用于raw data的质控 |
| Library\_Size = | <STRING> | 文库长度，用于生成质控表 |
| Readlength = | <int> | reads长度 |
| Directory = | <STRING> | 结果输出路径 |
| DataBase = | <STRING> | 数据库路径 |
| Species = | <STRING> | 物种参考序列名称，如hg19，mm9 |
| SpeType = | <STRING> | 物种类别，animal，plant等，DB中Pathway/需配置SpeType\_ko\_map.tab |
| Analysis = | <STRING> | 分析项，可以根据具体情况选择（空格符隔开），全部包括clean align peakcall peakanno  diffpeak diffanno Peak\_DEGs motif |
| CleanSoft = | <STRING> | 过滤软件 |
| rmAdapter = | <STRING> | 过滤时是否去除adapter，默认是去除 |
| CleanParameter = | <STRING> | 过滤参数，具体设置参考SOAPnuke参数说明 |
| Aligner = | <STRING> | 比对软件路径：BWA、SOAP、bowtie、ELAND |
| Samtools = | <STRING> | samtools路径 |
| AlignMem = | <STRING> | 比对内存设置 |
| AlignParameter = | <STRING> | 比对参数，可参考比对软件参数说明 |
| AlignIndex = | <STRING> | Reference的index文件:BWA、SOAP、ELAND |
| ReadElement = | <STRING> | 基因的bed文件，用于reads的TSS、TES计算  TSS的上游2k、TES的下游2k。可以改变，不过需要在\*/GeneInfo下放置对应的文件 |
| DepthMax = | <int> |
| Extend = | <int> |  |
| Peakcaller = | <STRING> | Peakcaller软件路径，请注意根据抗体类型选择合理的软件，对于转录因子请用macs14；  组蛋白修饰分析或抗体类型为长peak类型请用sicer或macs2 |
| Pythonpath = | <STRING> | MACS python路径 |
| Python = | <STRING> | python路径 |
| Peakparameter = | <STRING> | peak calling参数，请根据选用的软件正确设置相应的参数。值得注意的是，请根据测序物  种设置好macs的-g 参数，用macs2进行长peak calling时注意加上–broad |
| RegionElement = | <STRING> | peak注释的bed文件，可以设置成5K或其它，不过需在database中准备好相关文件 |
| MotifTopnum = | <int> | 用MEME或DREME进行motif分析时，设计的挑选enrichment值较大的peak |
| MotifSoft = | <STRING> | meme可以获取长度较长的motif，dreme可以获取6-8bp的motif，且dreme运行效率更高 |
| MotifParameter = | <STRING> | meme的参数设置，具体可见motif分析软件的参数说明 |
| OverlapCutoff = | <int> | 用bedtools计算两个文件overlap时的参数设置 |
| Diff = | <STRING> | 差异peak分析的样品分组 |
| Diffparameter = | <STRING> | 差异peak分析的参数 |
| DEGs = | <STRING> | 进行联合分析的RNA-seq样品对，如JJJ\_zebra1&JJJ\_zebra2 |
| DEGs\_Path = | <STRING> | DEGs文件路径(path)，path路径下对应着存放上述联合分析的RNA-seq样品对的DEGs文件，  如path/JJJ\_zebra1-vs-JJJ\_zebra2/JJJ\_zebra1-vs-JJJ\_zebra2.xls。该xls文件格式为6列：差异  基因名\t差异表达倍数（fold change）\t染色体编号\t正负链\t基因在染色体上的起点\t基因  在染色体上的终点 |
| Rscript= | <STRING> | R路径 |
| CaseSample = | <STRING> | case样品 |
| ContSample = | <STRING> | control样品 |
| Fq= | <STRING> | 样品下机路径（Hiseq为下机路径，如/nas/fqdata025/data/F15FTSEUHT0433\_HUMhevC  /WHCDNAPEP00003703/150628\_I137\_FCH2TVVBBXX\_L2\_wHAPPI019628-90 ；Zebra为  Fq文件绝对路径，如/nas/fqdata035A/Zebra/TEST/BGI500CHIP8-N/160607\_I2\_CL300001  027\_L2\_BGI-500Test-C7/CL300001027\_L02\_47.fq.gz） |
| Control = | <STRING> | 样品的对照品名称 |

### 6****分析流程图****

