Bulk RNA-seq数据的处理

1. 序列比对

参考基因组：hg19

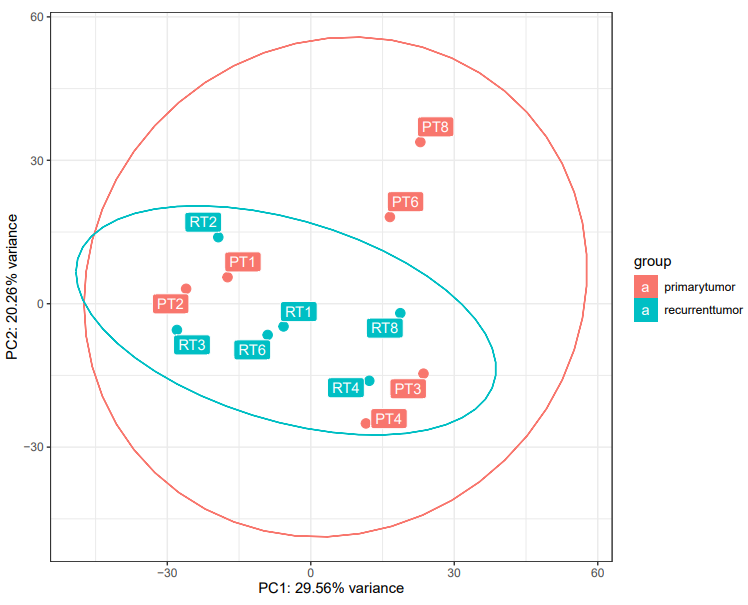
软件：序列比对hisat2 (version: 2.0.0-beta)；组装stringtie (version: 1.3.3b)；定量featureCounts (version: 1.5.3)

|  |  |  |
| --- | --- | --- |
|  | Total Reads | Mapping Rate |
| A1 | 52118490 | 97.34% |
| A2 | 53154296 | 97.04% |
| A3 | 47266533 | 96.82% |
| A4 | 54200973 | 96.99% |
| A6 | 50593123 | 97.12% |
| A8 | 49618856 | 97.02% |
| B1 | 52236646 | 97.14% |
| B2 | 48385319 | 97.05% |
| B3 | 60399844 | 98.29% |
| B4 | 47000910 | 96.91% |
| B5 | 60119552 | 97.18% |
| B6 | 40698589 | 97.14% |
| B7 | 46638087 | 96.94% |
| B8 | 48866055 | 97.39% |

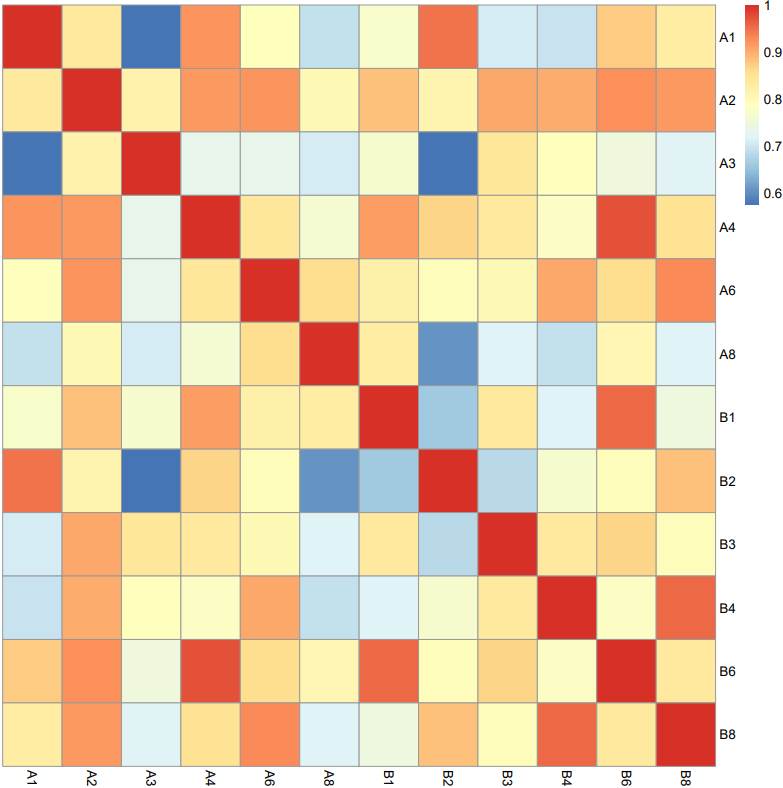
（这里，B3是由后续两次测序合并的结果：29107547(96.46%)和31292297(100.00%)我们只选择配对的样本进行后续分析，样本1,2,3,4,6和8）。

1. RNA-seq数据质量评估 （样本过滤）

标准：用主成分表示样本之间的变异，组间样本之间的距离大于组内样本之间的距离；把基因组打成10Kb为一个bin的区域，根据统计落入每个区域内的counts数目，并计算相关性，理论上组内的相似性大于组间相似性。



PCA主成分分析



PCC相关系数分析

（Fig. 1. RNA-seq数据质量评估）根据图所示，我们保留样本6和8用于后续分析

1. 识别差异表达基因

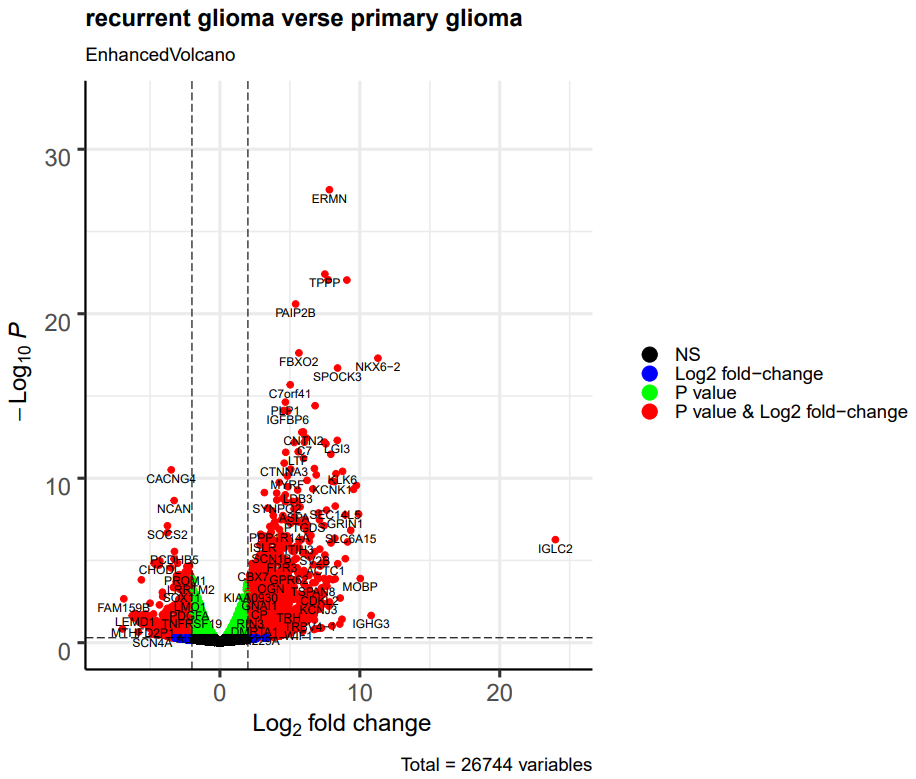
如果某个基因在所有样本中的表达都很低，我们则认为这个基因检测效果不好，需要过滤。

我们只保留在至少一个样本中的表达大于1的基因用于后续分析（基因过滤），得到了26,744个基因。

R 包: DESeq2 (version: 1.28.1)

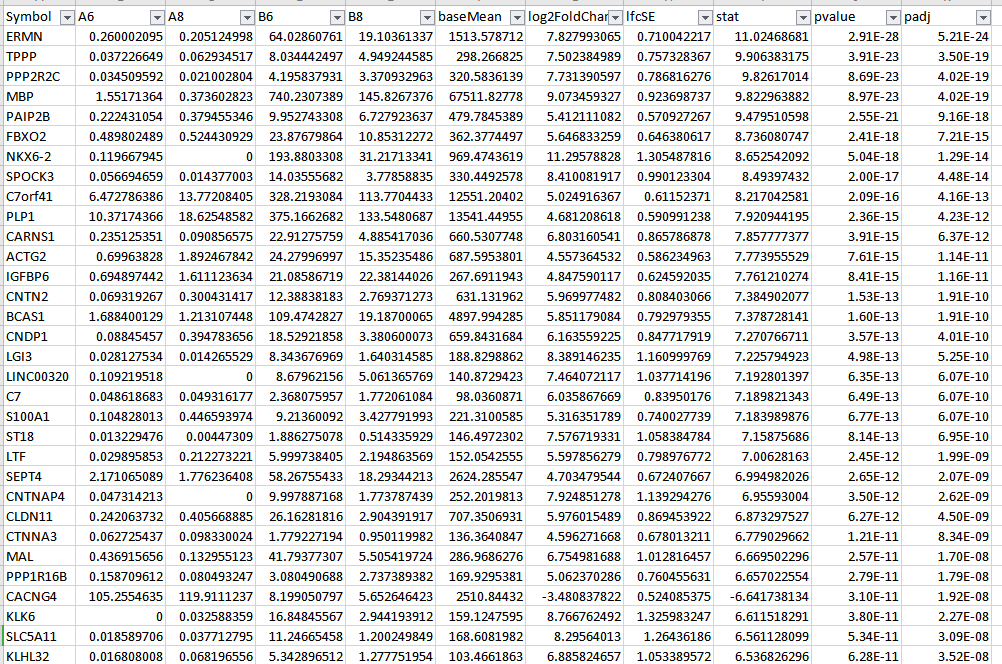
阈值：(log2FC>1或者log2FC<-1)并且P<0.05

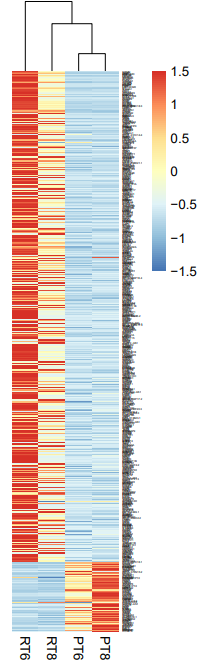
结果：相比于原发胶质瘤，复发中有538个显著上调的基因以及76个显著下调的基因。



（Fig. 2. 差异表达基因火山图）横坐标表示log2FC，纵坐标表示-log10Pvalue。

Table.1. 差异表达基因（截短）





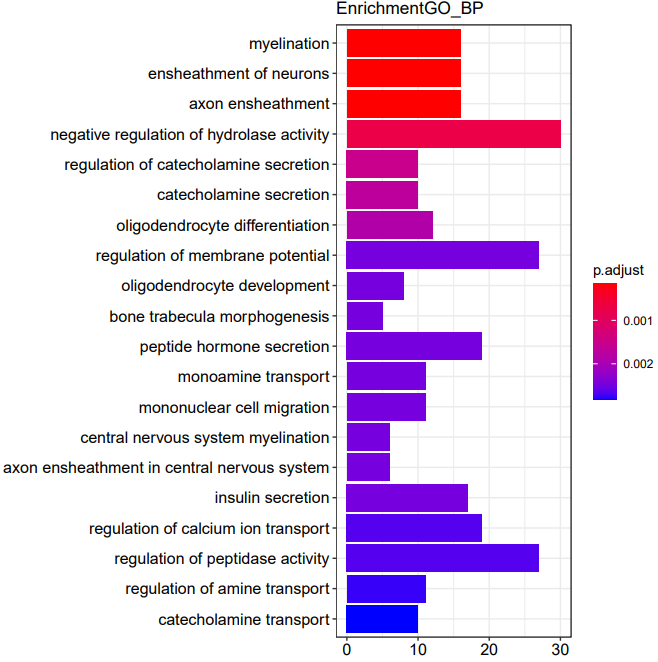
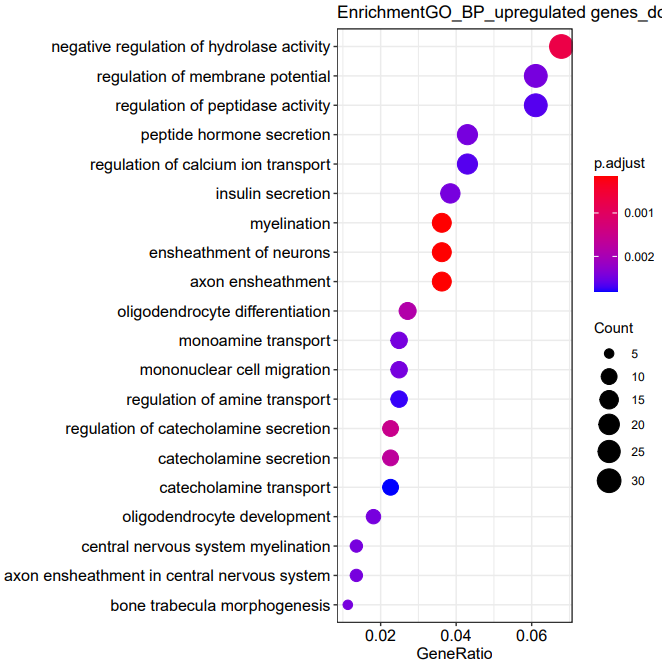
（Fig. 3. 差异表达基因热图）每一行表示一个差异表达基因，一共614行；列表示样本，其中RT代表复发，PT代表原发。把表达量标准化后用不同的颜色表示表达强度，红色代表这个基因在该样本中高表达，蓝色表示某个基因在样本中是低表达的。

1. 差异表达基因功能分析

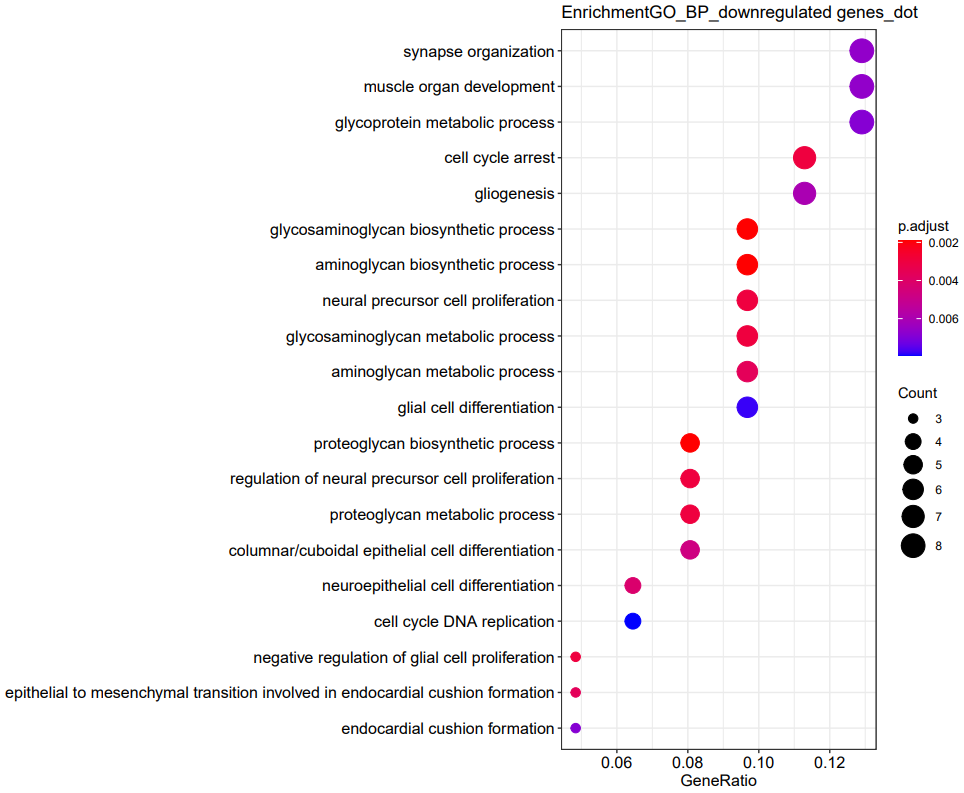
R包：clusterProfile (version: 3.16.1)

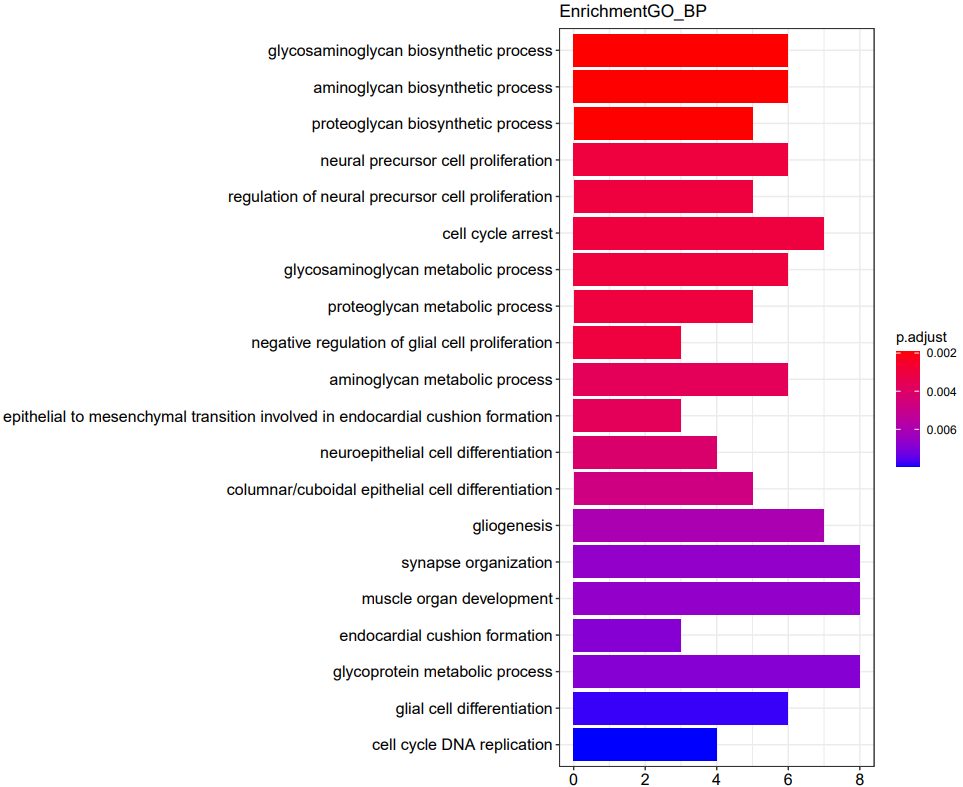
选择：GO的生物学过程biological process

上调的DEGs主要涉及到的生物学过程是：



下调的DEGs主要涉及到的生物学过程是：





Bulk ATAC-seq数据的处理

1. 质控

软件：trim\_galore (version: 0.5.0)

1. 序列比对

软件：bowtie2 (version: 2.2.5)

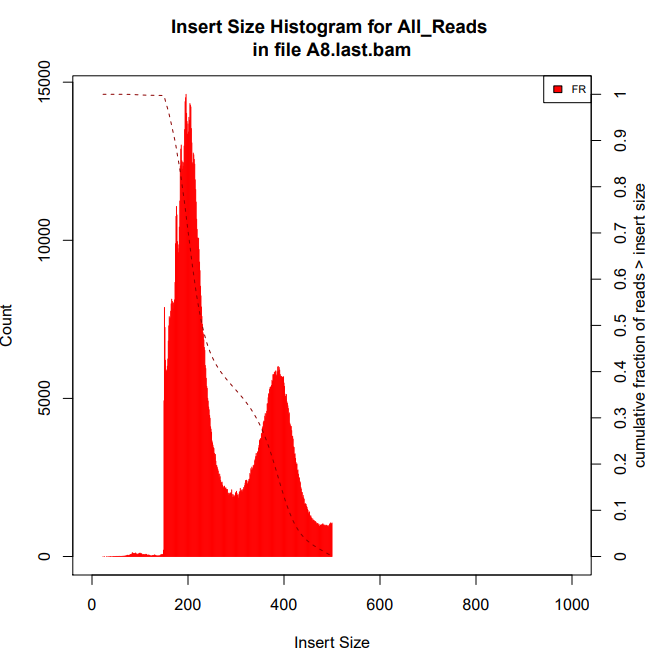
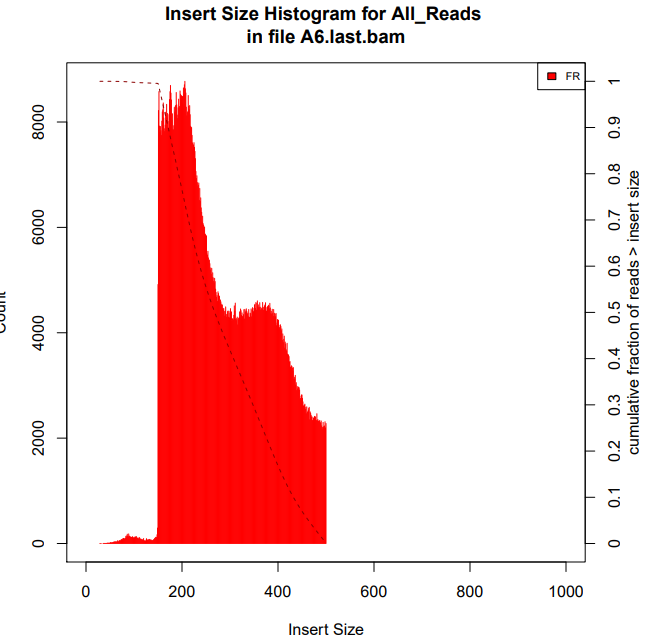
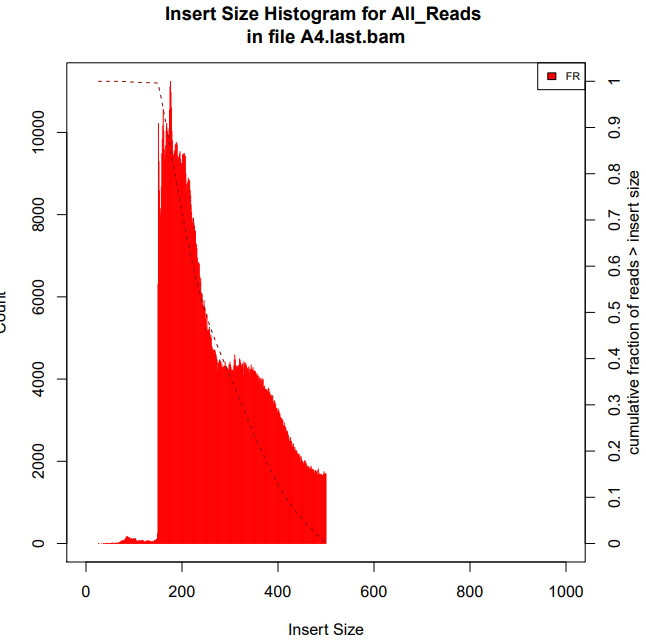
标准：期望的ATAC-seq的序列比对率一般在80%左右 实际操作上可以做调整。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Total Reads\*2 | Clean Reads\*2 | Mapped Reads | Remove Duplicate Reads | Filter or Not |
| A1 | 65328257; 36318407 | 64723398; 35990454 | 47181631(36.45%); 37366703(51.91%) | 27493486; 22694448 | Remove (插入片段长度) |
| A2 | 52891435 | 52321895 | 55043749(52.60%) | 34099300 | Remove (没有对应的复发样本) |
| A3 | 5697621; 13867758; | 5125819; 13755376 | 7754747(75.64%); 15641209(56.85%) | 4642186; 8168654 | Remove (插入片段长度) |
| A4 | 58302183 | 57466501 | 53081852(46.19%) | 32379190 | Keep |
| A6 | 36549716; 26229104; 13737972 | 35484583; 25973889; 13471770 | 38989636(54.94%); 26210558(50.46%); 12332056(45.77%) | 21729248; 15231906; 6830402 | Keep |
| A8 | 47354454 | 46610193 | 52466180(56.28%) | 32127524 | Keep |
| B1 | 20312337; 45462355 | 19959501; 44984079 | 16959199(42.48%); 30396517(33.79%) | 9048844; 17492824 | Keep |
| B3 | 51234949 | 50662229 | 46284657(45.68%) | 25801192 | Remove (插入片段长度) |
| B4 | 44691086; 18868132; 14380817 | 44219264; 18690791; 14036179 | 31725710(43.31%); 16412626(43.91%); 12130300(43.21%) | 18412056; 8426658; 5970122 | Keep |
| B6 | 23040154; 14743976 | 22711637; 14406654 | 20468871(45.06%);  12223521(42.42%) | 10926114; 6357288 | Keep |
| B8 | 32173687; 12699574 | 31716951; 12228224 | 27952816(44.07%); 12797794(52.33%) | 14376050; 6390074 | Keep第一个样本; Remove第二个样本 (插入片段长度) |

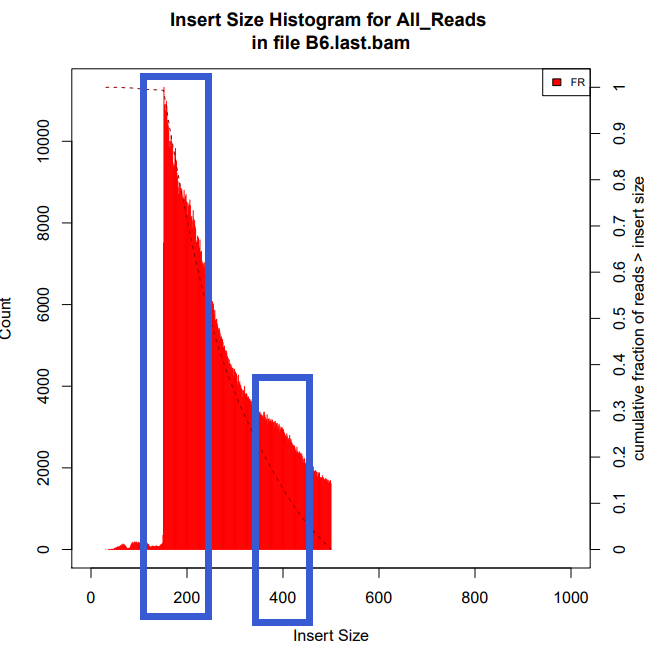
1. 合并对应的样本并计算插入片段的长度

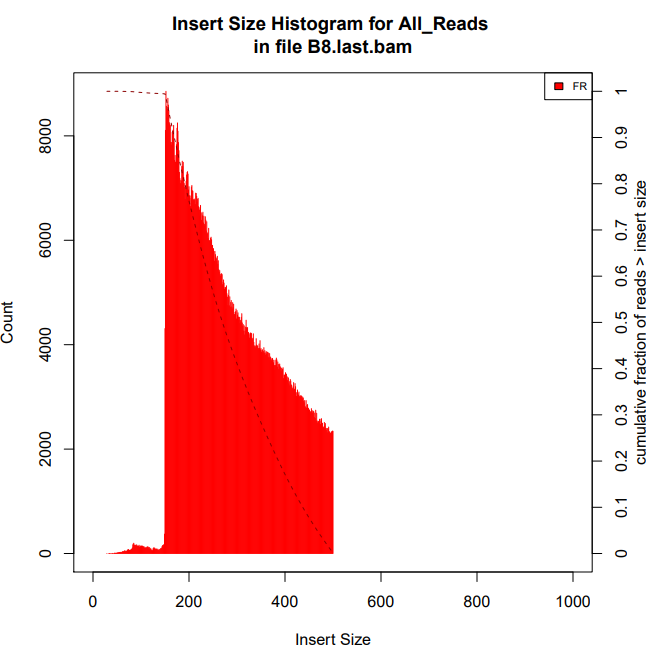
标准：Tn5 transposase可以随机插入到打开的chromatin中，因为Tn5 transposome中设计有PCR primer，当我们随后进行PCR扩增以后，每两个Tn5插入位点之间的序列就会被扩增出来。因为大多数的Linker DNA的大小介于10-80bp之间，所有得到的大多数片段都会是小于100bp的。而每个Nucleosome的DNA大小为180bp左右，加上两边插入进的冗余，我们会得到大约200bp长度是mono-nucleosome的DNA。如果是两个Nucleosome之间的片段的话，就是400bp左右。因此，成功的ATAC-seq实验会生成具有逐级递减和周期性的峰，对应无核小体区域（NFR，100bp），单核小体区域（mono-nucleosome，200bp），双核小体区域（400bp），三核小体区域（600bp）。通过质量控制的样本为：

A4, A6和A8；B1, B4, B6（勉强）和部分B8（勉强）。

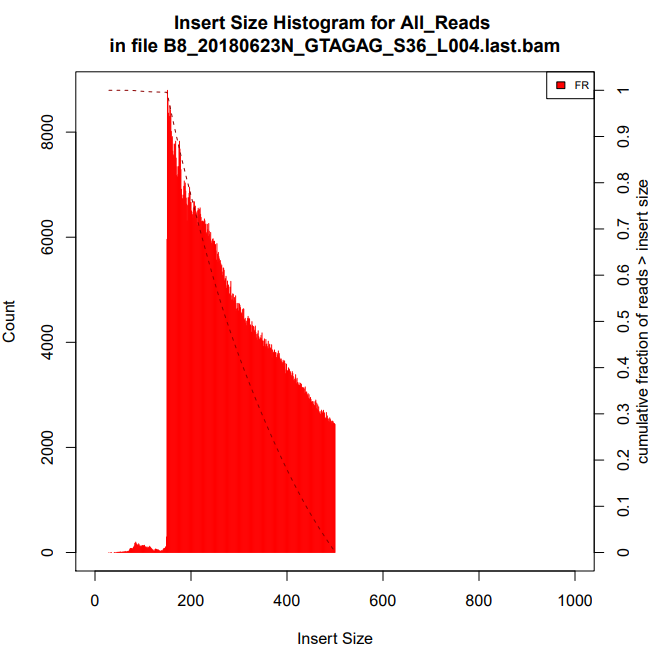




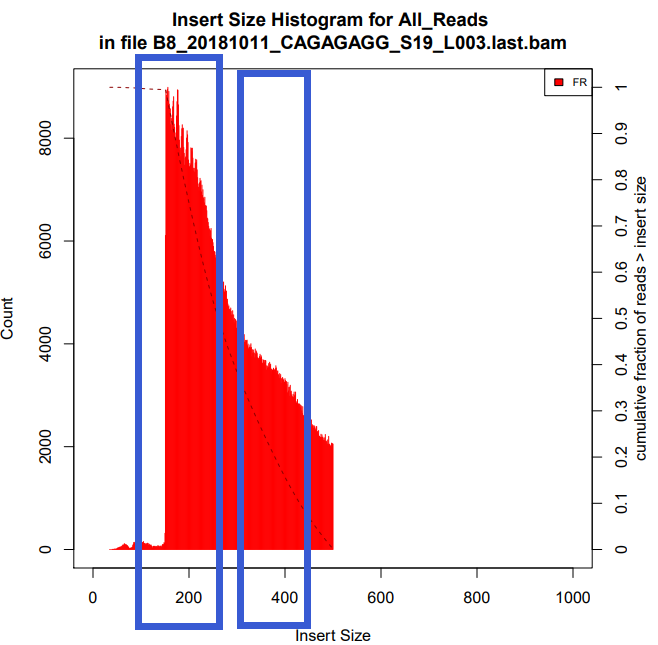




(B8来自两次测序数据的合并，这里由合并后的B8在200bp处的peak不明显)



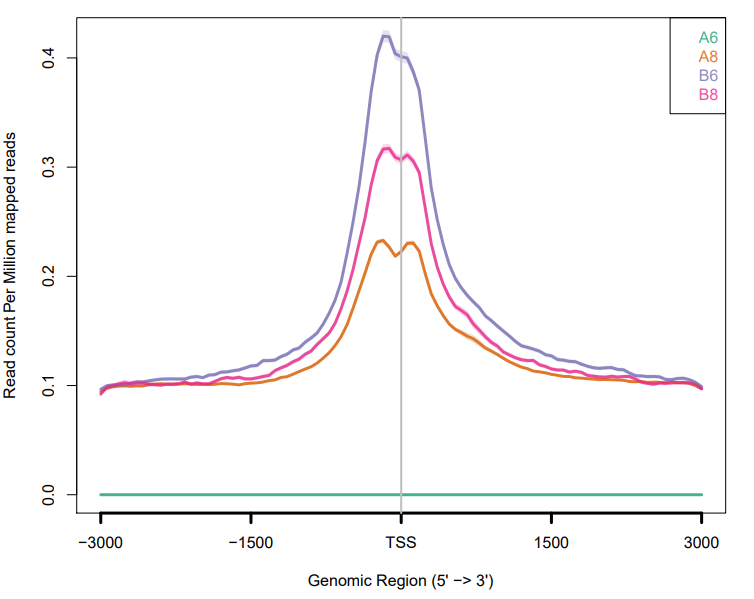
（B8的一个样本）



（B8的另一个样本）

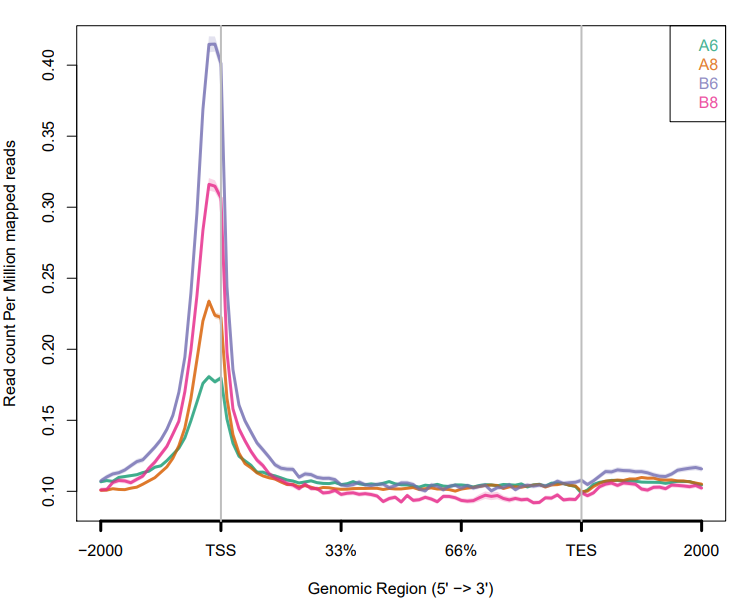
（B8是由两个测序样本合并的，其中一个效果不好，所以我们仅选择B8中的一个样本）

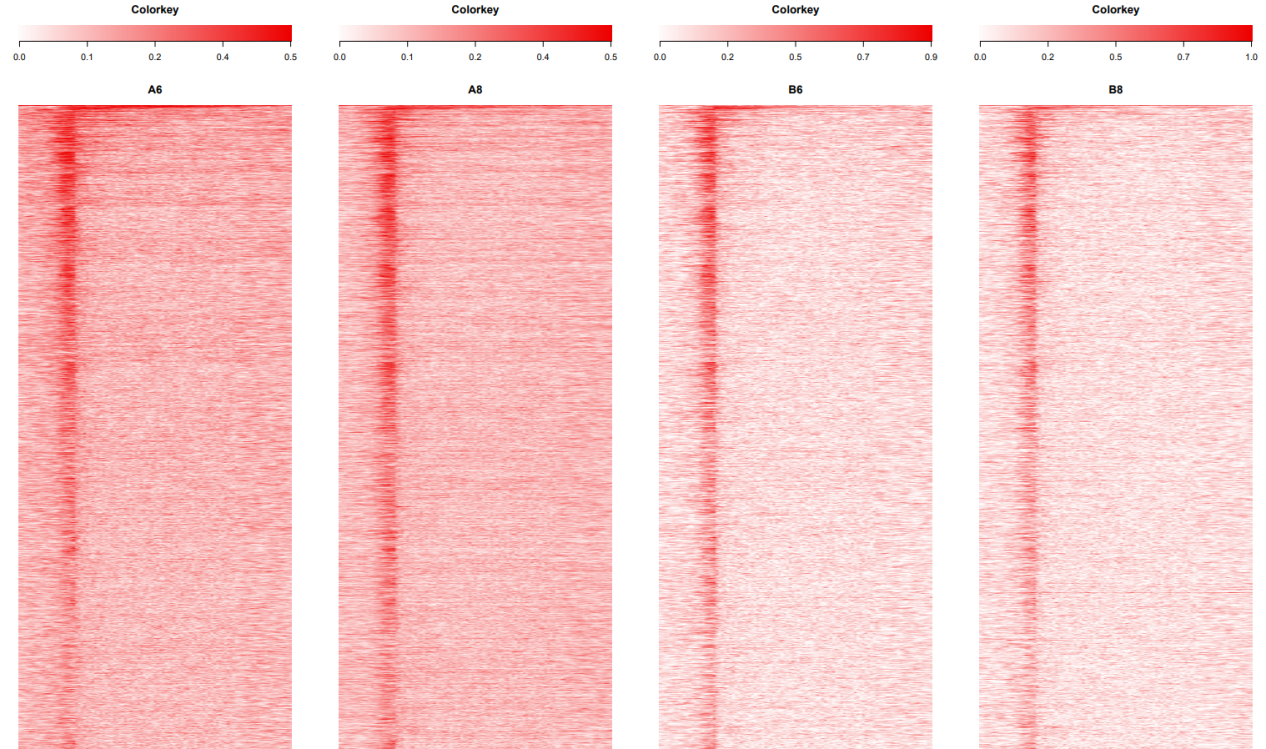
1. 可视化peaks



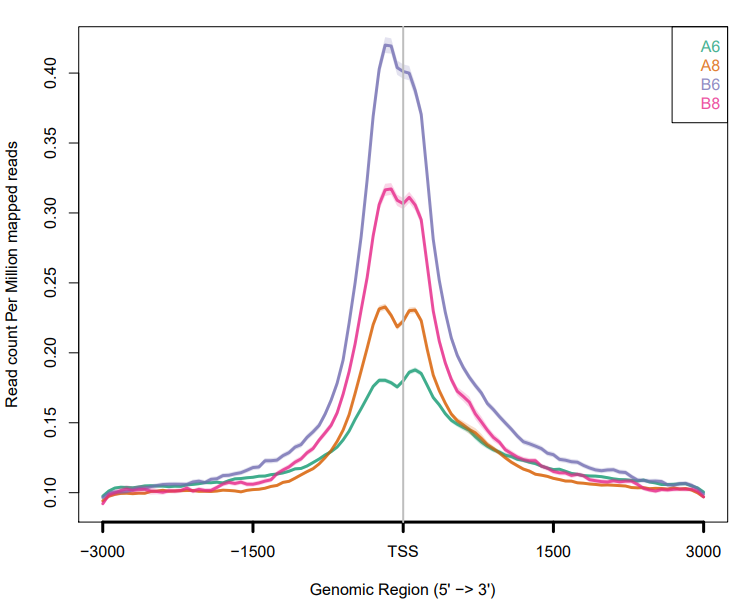
由图可知：A6样本在TSS附近的peak信号强度不高，我们发现A6是由三个测序样本合并而来的，这里我们分别计算，最后我们只保留其中两个样本（A6\_20180803N\_CGTACT\_S67\_L002和A6\_20181011\_CGTACTAG\_S16\_L003）即去掉样本A6\_20180623N\_GGACTC\_S29\_L004。

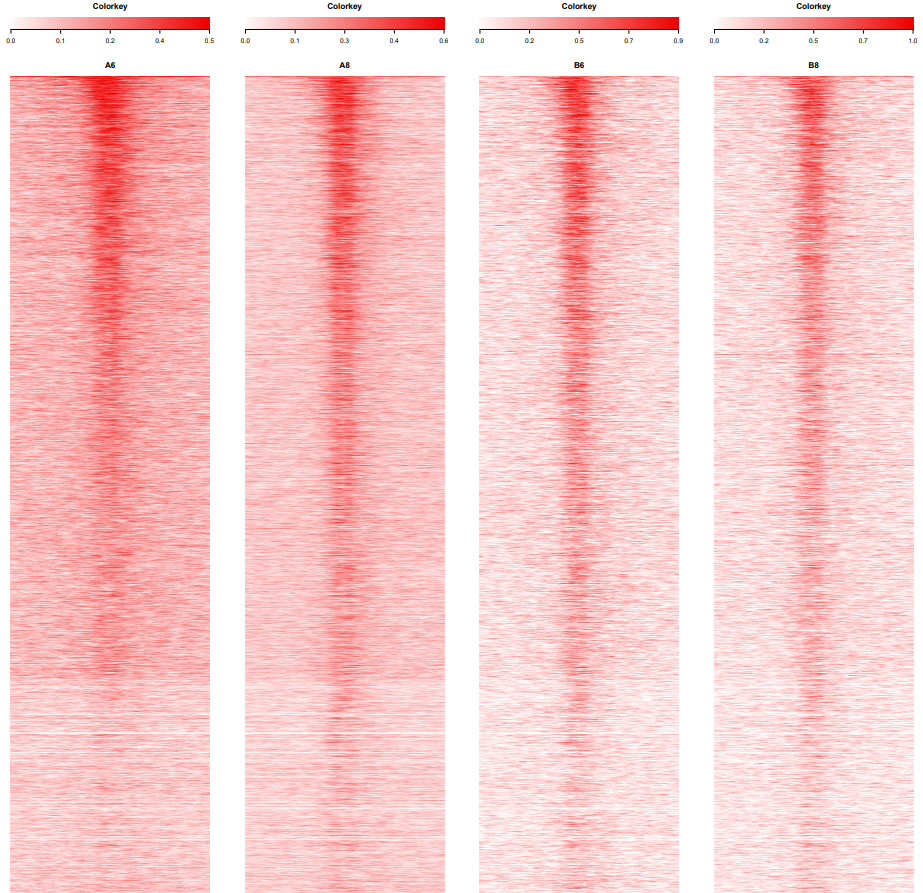
最终得到TSS附近和gene body区域的atac-seq信号





每一行代表ATAC-seq信号在一个基因上面的peaks强度。从图中可以看出atac-seq在整个基因区间内 都是结合在上游TSS附近的。



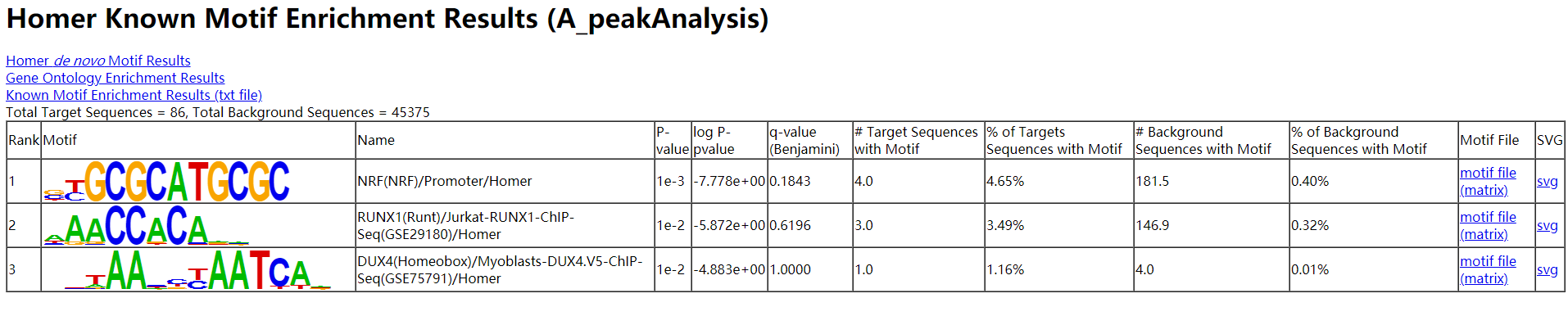


每一行代表ATAC-seq信号在一个基因上面的peaks强度。从图中可以看出atac-seq在所有基因的TSS上的富集强度更大。

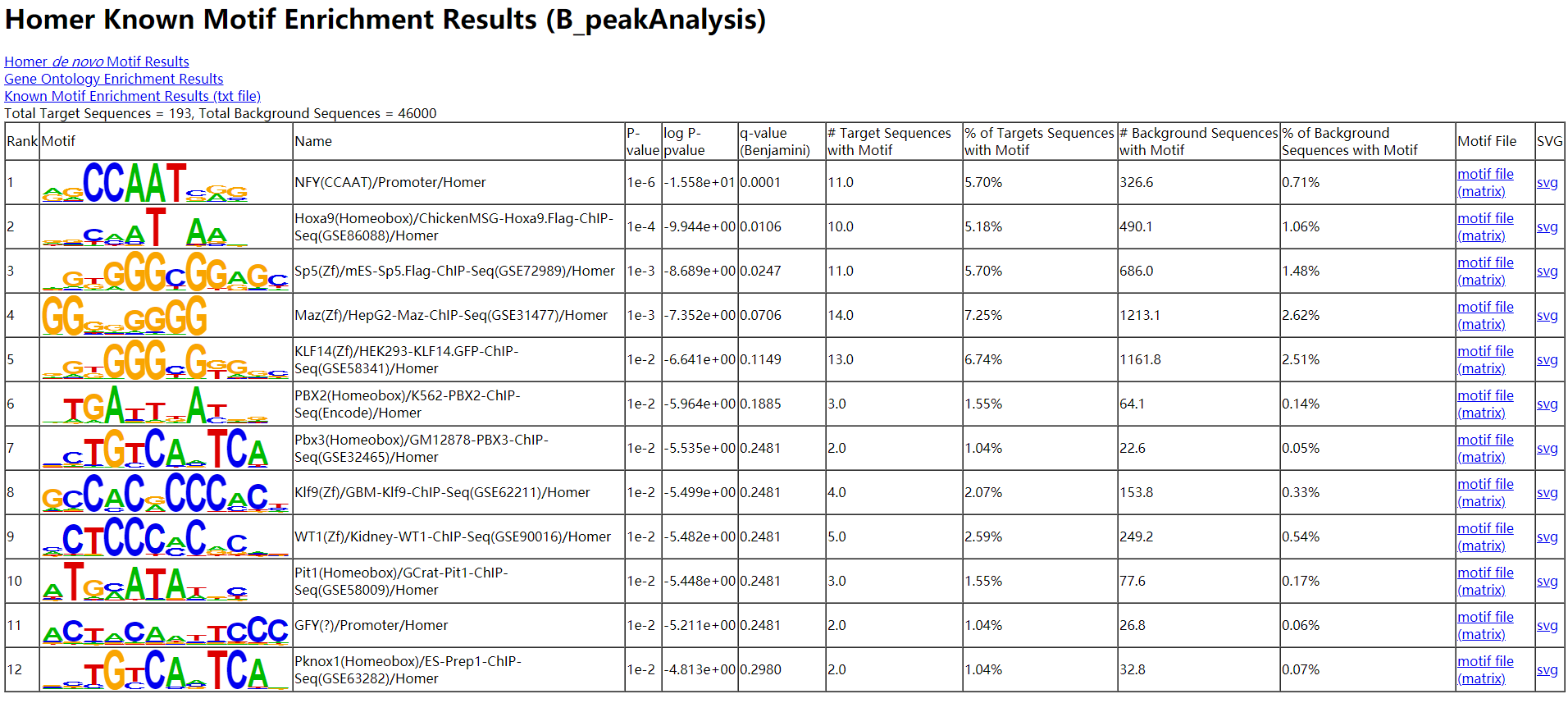
数据进行到这里，我们合并A6和A8：53,917,348 (约等于54 million) reads 而B6和B8：12,692,394 (约等于13 million) reads。原发和复发样本量差距大，所以这里我们又查看了复发里面的其他样本，最终我们选择了样本B4：32,808,836 (约等于33 million) reads，最终得到的数据量为reads。

1. motif富集分析

原发：



复发：



1. Motif印记分析