甲基化分布图绘制

前期计划

数据获得：来自文慧丽师姐，水稻甲基化测序（亚硫酸氢盐转化法）数据：

A\_1.sorted.bam,

A\_2.sorted.bam,

B\_1.sorted.bam,

B\_2.sorted.bam,

还要获得水稻基因组数据，获得水稻各个染色体上的基因转录起始位点与转录终止位点。

利用Bismark 获得水稻基因组上胞嘧啶甲基化水平（整体的甲基化水平，CG位点的甲基化水平，CHG位点的甲基化水平，CHH位点的甲基化水平），

将水稻基因的转录起始与终止位点分成合适数量的片段（如20段），还要在转录起始终止位点前后以一定长度内进行一样的处理。

通过Bismark的结果计算这些片段内的甲基化平均水平，写入到excel文件中

获得转录起始终止位点以及其前后区段的甲基化修饰平均水平，结合这些区段的位置信息，用matplotlib绘制甲基化分布折线图。

完成记录

Bismark 环境变量设置，这个程序运行需要perl环境与bowtie1/2

先将bam文件转换为sam文件

bsub -n 16 -o B\_2.sort.bam.BSlog.txt samtools view -h -o B\_2.sort.sam B\_2.sort.bam（不必要的步骤，其实bam文件也可以作为bismark的输入文件，而且更小）

建立bismark的索引：

bsub -n 16 -e -o test.log bismark\_genome\_preparation --bowtie2 --path\_to\_bowtie /gss1/app/bowtie2-2.2.5 /gss1/home/gaohao/Bisulfite\_conversion/

--bowtie2 指定使用bowtie2

--path\_to\_bowtie<bowtie的安装位置>

最后的文件夹里需要有水稻基因组文件tigr7.fa

使用bismark的bismark\_methylation\_extractor功能:

bismark\_methylation\_extractor -p --comprehensive --no\_overlap --bedGraph --counts --report --cytosine\_report --genome\_folder /gss1/home/gaohao/Bisulfite\_conversion/ A\_1.sortBYhao.sam -o /gss1/home/gaohao/Oryza\_sativa\_methylation\_picture/newfolder/

-p：输入文件是从双端读取数据生成的Bismark结果文件。 必须指定--paired-end或--single-end。

--comprehensive: 指定此选项会将所有四个可能的特定于链的甲基化信息合并到上下文相关的输出文件中。 默认内容是：（i）CpG甲基化（ii）CHG甲基化（iii）CHH甲基化（取决于Bismark结果文件的C内容，输出文件的大小可能会达到10-30GB！）。

--no\_overlap : 对于成对的末端读取，理论上有可能read\_1和read\_2重叠。 此选项避免对重叠的甲基化调用评分两次。 尽管这消除了朝向更多的甲基化方向偏向测序片段中心的偏见，但实际上可以去除很大一部分数据。

--bedGraph: 甲基化提取完成后，将甲基化输出写入排序的bedGraph文件中，该文件使用基于0的基因组坐标报告给定胞嘧啶的位置及其甲基化状态（％，以下显示详细信息）。 甲基化提取器的输出临时分成临时文件，每个染色体一个（写到用-o / -output指定的当前目录或文件夹中）； 这些临时文件然后用于排序并随后删除。 默认情况下，仅对CpG环境中的胞嘧啶进行排序。 选项--CX\_context可以用于报告所有胞嘧啶，而与序列上下文无关（这将花费更长的时间！）。 bedGraph转换步骤由外部模块“ bismark2bedGraph”执行； 该脚本需要与bismark\_methylation\_extractor本身位于同一文件夹中。

--counts: 在输出文件中添加另外两列以进行进一步的计算：col 5：甲基化调用数col 6：未甲基化调用数如果指定了--cytosine\_report，则此选项为必需的（并在必要时自动设置）。

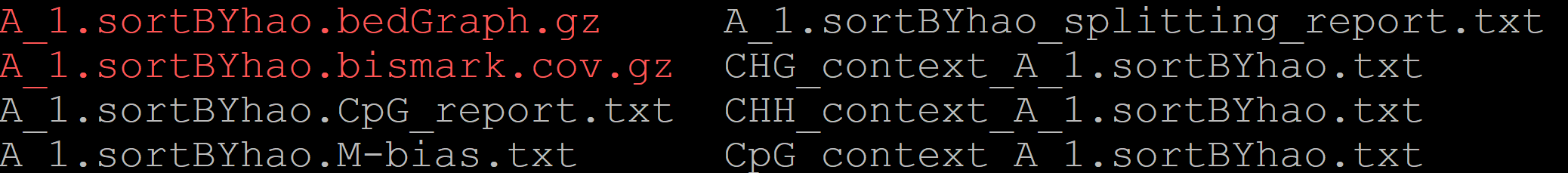
--buffer\_size <string>: 这使您可以在排序甲基化信息时指定主内存排序缓冲区。 通过附加％（例如--buffer\_size 50％）或1024字节的倍数来指定物理内存的百分比，例如 'K'乘以1024，'M'乘以1048576，以此类推，对于'T'等。（例如--buffer\_size 20G）。 有关排序的更多信息，请在命令行上键入“ info sort”。 默认为2G。

--report : 打印出简短的甲基化摘要以及用于运行此脚本的参数。

--cytosine\_report: 转换为bedGraph后，选项--cytosine\_report会生成基因组中所有胞嘧啶的全基因组甲基化报告。 默认情况下，输出使用从1开始的染色体坐标（从零开始的编码是可选的），并且仅报告CpG上下文（所有胞嘧啶上下文是可选的）。 输出考虑正向和反向链上的所有C，并报告其位置，链，三核苷酸含量和甲基化状态（如果未覆盖，则计数为0）。 半胱氨酸报告转换步骤由外部模块“ bedGraph2cytosine”执行； 该脚本需要与bismark\_methylation\_extractor本身位于同一文件夹中。

--genome\_folder <path> : 输入您要用来提取序列的基因组文件夹（仅全路径）。 可接受的格式是以“ .fa”或“ .fasta”结尾的FastA文件。 必须指定基因组文件夹路径。

-o Dir: 允许指定不同的输出目录（绝对或相对路径）。 如果未明确指定，则输出将被写入当前目录。

Bismark输出以下文件：

CHG\_context CHH\_context CpG\_context 三个文件每列含义：

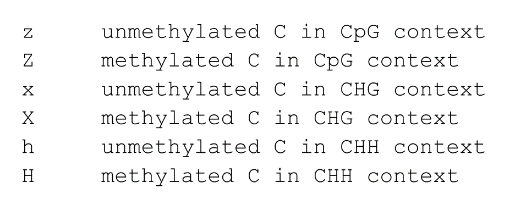
(1) seq-ID

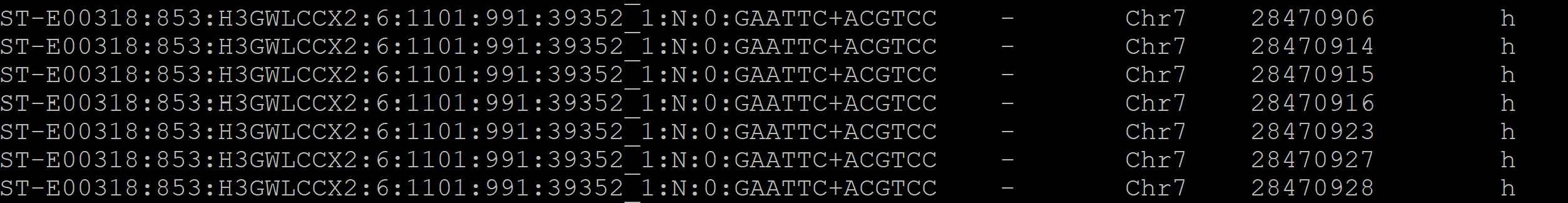
(2) methylation state 甲基化的胞嘧啶将接受“ +”方向，未甲基化的胞嘧啶将接受“-”方向。

(3) chromosome 染色体号码

(4) start position (= end position) 起位点也即终位点

(5) methylation call



Eg：

A\_1.sortBYhao.bedGraph文件里面：

<chromosome> <start position> <end position> <methylation percentage>

A\_1.sortBYhao.bismark.cov 文件里面：

<chromosome> <start position> <end position> <methylation percentage> <count methylated> <count unmethylated>

A\_1.sortBYhao.CpG\_report.txt 文件里面：

<chromosome> <position> <strand> <count methylated> <count unmethylated> <C-context> <trinucleotide context>

至此已经获得了水稻染色体

Chr9

Chr5

Chr10

Chr12

Chr8

Chr3

Chr1

Chr2

Chr7

Chr11

Chr6

Chr4

ChrM|11706

ChrC|11562上的5mc状态，可以计算某一区段内的甲基化程度。

还需要获得水稻全基因组范围内基因的起始与终止位点。

Chr9

Chr5

Chr10

Chr12

Chr8

Chr3

Chr1

Chr2

Chr7

Chr11

Chr6

Chr4

ChrUn

ChrSy

标灰色数据不参加绘图

Python 代码绘制出CpG在基因内部以及上下游的甲基化图片绘制

./bismark\_v0.22.1/bismark2bedGraph --CX -o output\_name

上面的A\_1.sortBYhao.CpG\_report.txt文件中只有CG甲基化的信息，要想知道CHH、CHG的信息需要在运行上面的命令当然也可以在一开始就加—CX

bismark2bedGraph --CX -o A\_1.sortBYhao.CG\_CHH\_CHG\_report.txt CHG\_context\_A\_1.sortBYhao.txt CHH\_context\_A\_1.sortBYhao.txt CpG\_context\_A\_1.sortBYhao.txt(没得到想要的文件，应该是缺参数)

bsub -n 16 -o test.log bismark\_methylation\_extractor -p --comprehensive --no\_overlap --bedGraph --counts --report --cytosine\_report --CX --genome\_folder /gss1/home/gaohao/Bisulfite\_conversion/ A\_1.sortBYhao.sam -o /gss1/home/gaohao/Oryza\_sativa\_methylation\_picture/A\_1.methylation/

bsub -n 16 -o test.log bismark\_methylation\_extractor -p --comprehensive --no\_overlap --bedGraph --counts --report --cytosine\_report --CX --genome\_folder /gss1/home/gaohao/Bisulfite\_conversion/ A\_2.sortBYhao.sam -o /gss1/home/gaohao/Oryza\_sativa\_methylation\_picture/A\_2.methylation/

bsub -n 16 -o test.log bismark\_methylation\_extractor -p --comprehensive --no\_overlap --bedGraph --counts --report --cytosine\_report --CX --genome\_folder /gss1/home/gaohao/Bisulfite\_conversion/ B\_1.sortBYhao.sam -o /gss1/home/gaohao/Oryza\_sativa\_methylation\_picture/B\_1.methylation/

bsub -n 16 -o test.log bismark\_methylation\_extractor -p --comprehensive --no\_overlap --bedGraph --counts --report --cytosine\_report --CX --genome\_folder /gss1/home/gaohao/Bisulfite\_conversion/ B\_2.sortBYhao.bam -o /gss1/home/gaohao/Oryza\_sativa\_methylation\_picture/B\_2.methylation/

编写python脚本，结合matplotlib绘图得：

