Zhang-Lab 生信小课堂 第九期

和趣求真 秉实生信

（张建伟生物信息学课题组 https://zhang.hzau.edu.cn）

WGBS数据分析

2023.2.10 二综一楼C102 15:00 欢迎大家交流学习！

主讲人：刘思诗

2023/02/10

目录

CONTENTS

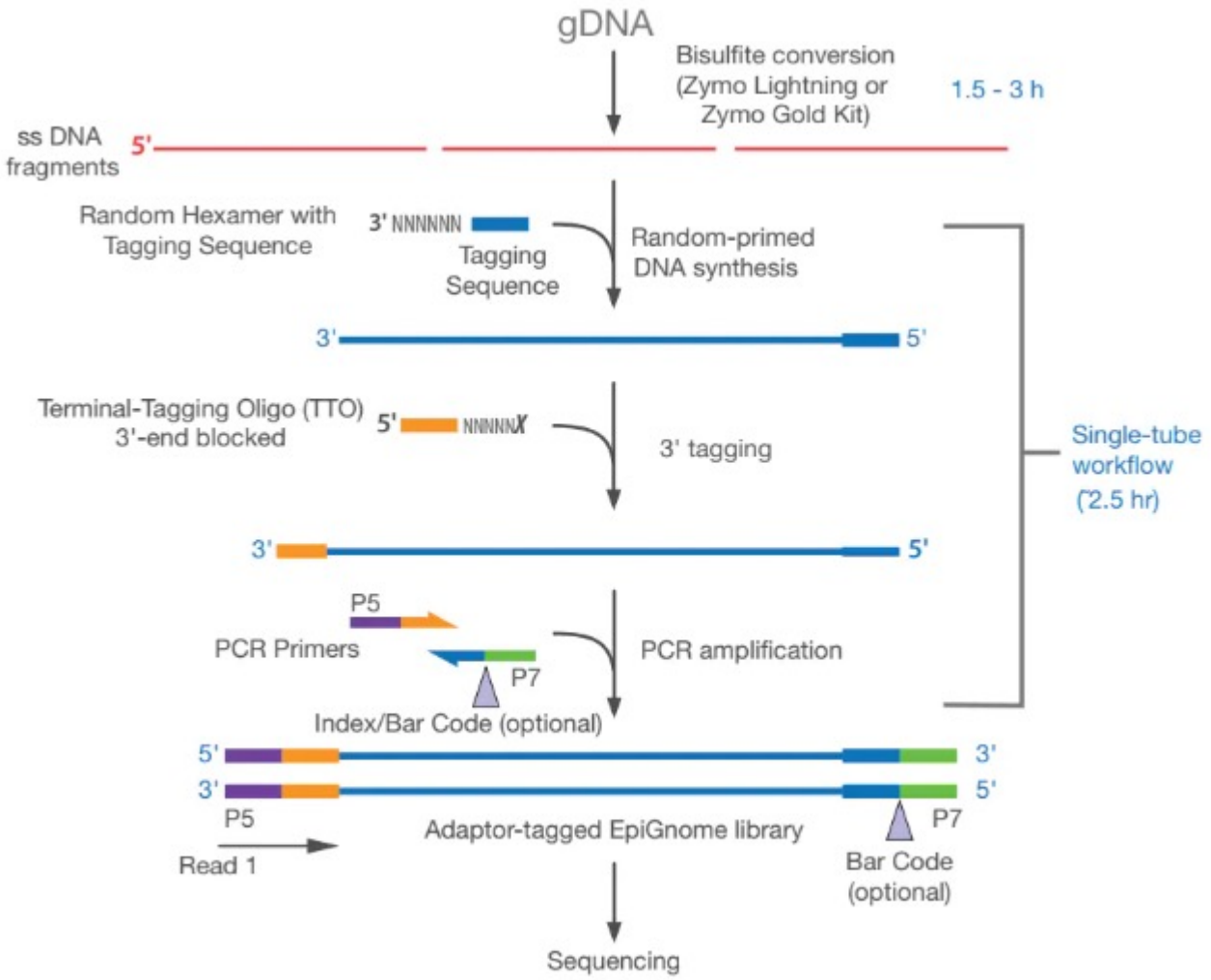
壹

什么是WGBS?

全基因组甲基化测序（WGBS）

全基因组重亚硫酸盐测序（whole-genome bisulfite sequencing，WGBS）用于描绘全基因组DNA甲基化图谱。它可以在全基因组范围内精确的检测所有单个胞嘧啶碱基（C碱基）的甲基化水平，是DNA甲基化研究的金标准。它能为基因组DNA甲基化时空特异性修饰的研究提供重要技术支持，能广泛应用在个体发育、衰老和疾病等生命过程的机制研究中，也是各物种甲基化图谱研究的首选方法。

常规全基因组甲基化测序技术通过T4-DNA连接酶，在超声波打断基因组DNA片段的两端连接接头序列，连接产物通过**重亚硫酸盐**处理将未甲基化修饰的胞嘧啶C转变为尿嘧啶U，进而通过接头序列介导的 PCR 技术将尿嘧啶U转变为胸腺嘧啶T。



Illumina建库流程

在这个过程中，亚硫酸盐处理过的单链DNA被随机引物，使用能够读取尿嘧啶核苷酸的聚合酶，合成含有特定序列标记的DNA。然后在新合成的DNA链的3 '端选择性地用第二个特定序列标记，从而得到在其5 '和3 '端具有已知序列标签的双标记DNA分子(图1)。然后用这些标签分别在原始DNA链的5 '和3 '端通过PCR添加Illumina P7和P5适配器。

贰

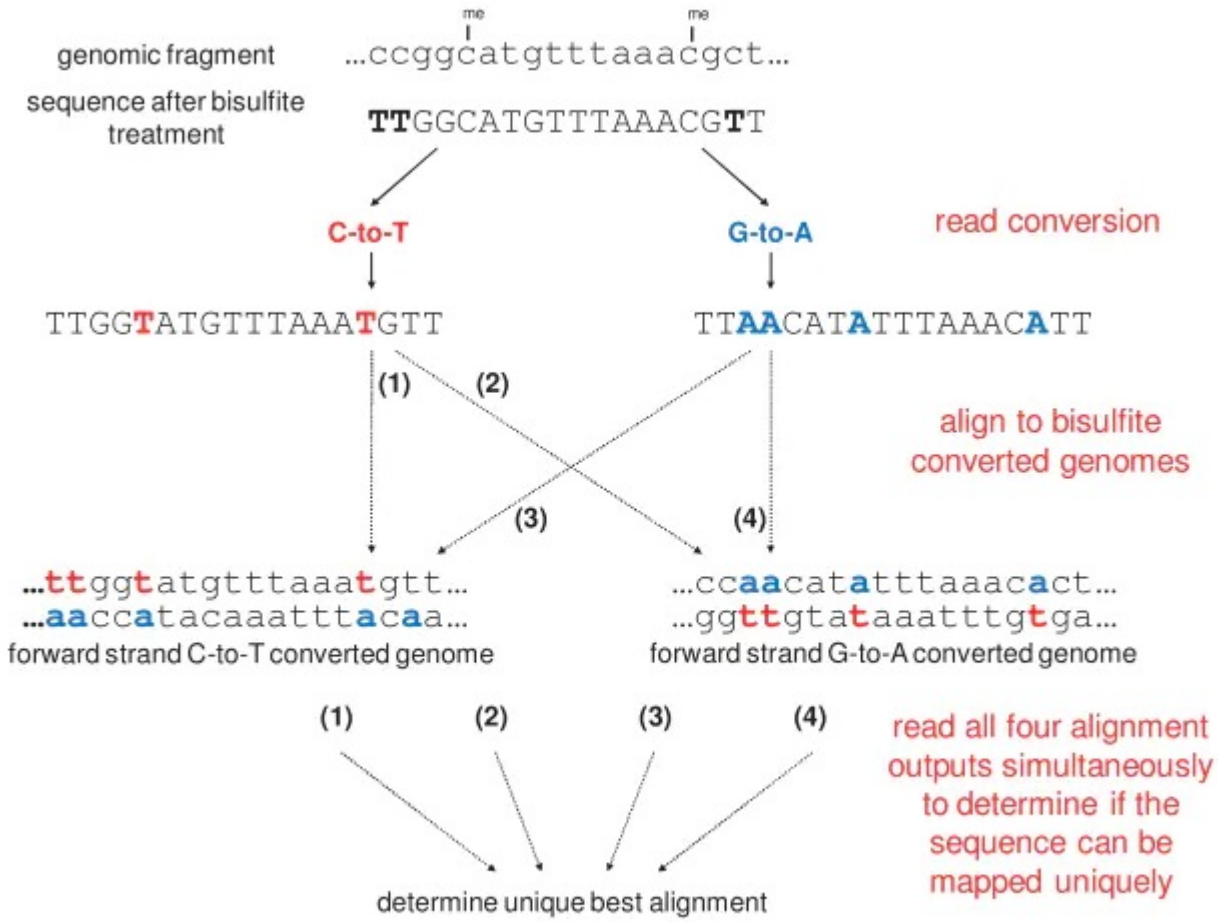
WGBS数据分析

Bismark

Bismark比对

1. Bismark 将参考基因组序列预先进行 C→T 和 G→A 2种转换。

2.比对时每一条 reads 同样进行 C→T 和 G→A 2种转换，这样组合以后每条 reads 相当于进行 4 种不同的比对，这些比对选出最佳比对，就可以确定发生甲基化的链方向和可能甲基化位点。



Bismark比对

Bismark

1. Bismark Genome Preparation（建立索引）

bismark_genome_preparation

2. Bismark Compare（进行比对）

bismark

3.Bismark Duplicate（过滤重复）

deduplicate_bismark

4.Bismark Methylation Extractor（甲基化信息提取）

bismark_methylation_extractor

Bismark

```
bismark_genome_preparation  .
```

输出文件夹
bowtie2_index

```
bismark --bowtie2 -N 0 -L 20 --quiet --un --ambiguous --bam --parallel 20 \
-o ${obj_path} bowtie2_index \
-1 test.file.R1_1.clean.fq \
-2 test.file.R1_2.clean.fq
```

输出文件
test.file.R1_1.clean_bismark_bt2_pe.bam 所有对齐和甲基化的信息
test.file.R1_1.clean_bismark_bt2_PE_report.txt 对齐和甲基化的主要信息概括

```
deduplicate_bismark -p --bam test.file.R1_1.clean_bismark_bt2_pe.bam \
--output_dir ${obj_path}
```

输出文件
test.file.R1_1.clean_bismark_bt2_pe.deduplicated.bam
test.file.R1_1.clean_bismark_bt2_pe.deduplication_report.txt 甲基化去重的主要信息概括

```
bismark_methylation_extractor -p --comprehensive --no_overlap --CX --bedGraph --counts --parallel 20 \
--buffer_size 20G --cytosine_report \
--genome_folder bowtie2_index \
test.file.R1_1.clean_bismark_bt2_pe.deduplicated.bam \
-o ${obj_path}

#可视化生成HTML 报告页面
bismark2report .
```

Bismark report html

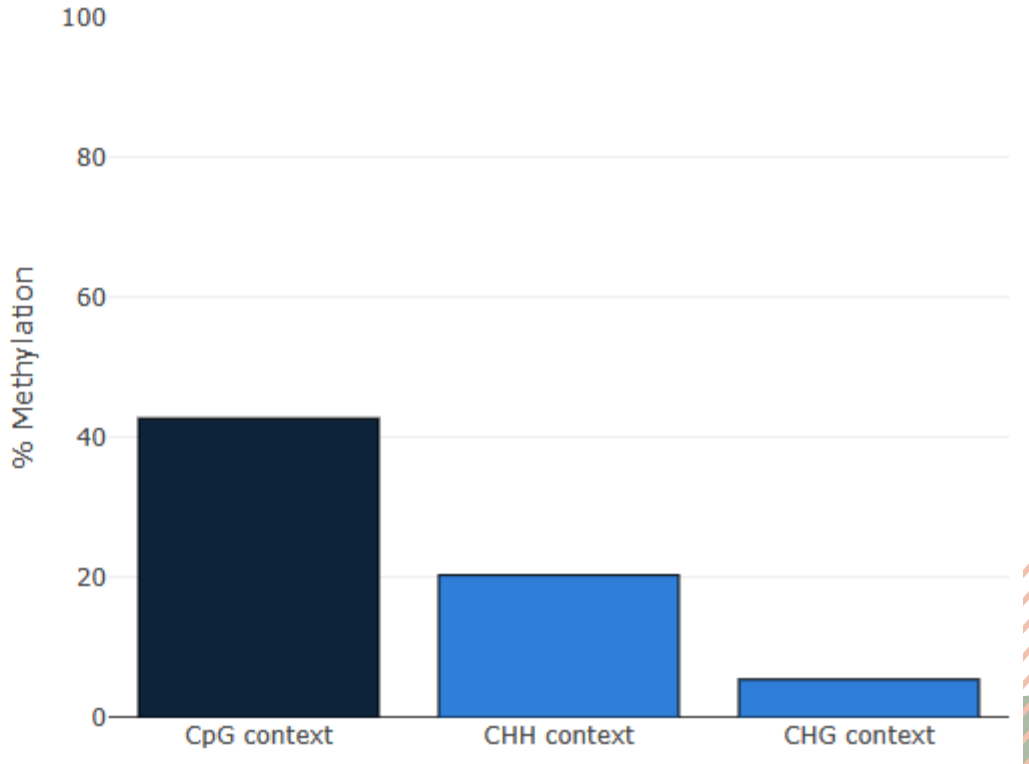
Alignment Stats

Cytosine Methylation

Deduplication

Cytosine Methylation after Extraction

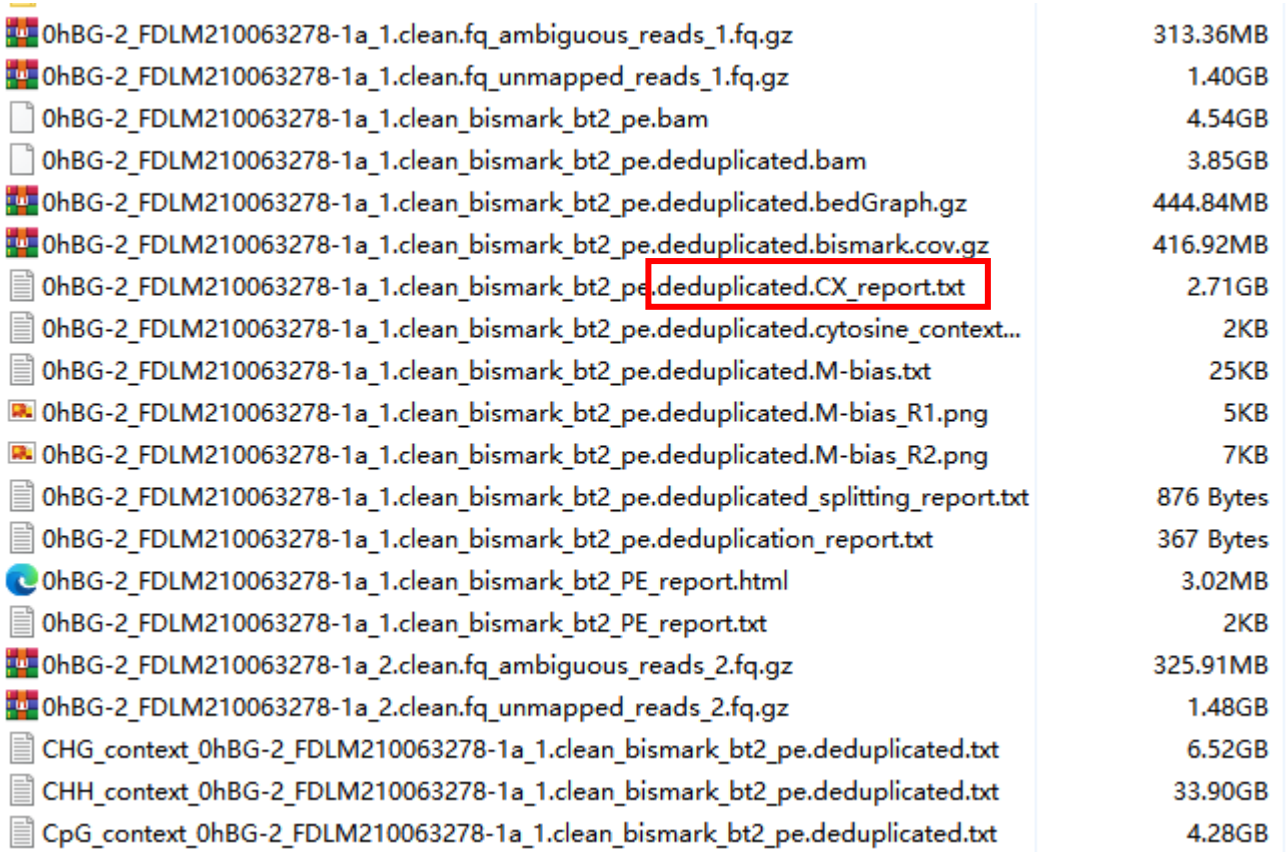
Total C's analysed	591822890
Methylated C's in CpG context	24216991
Methylated C's in CHG context	17584254
Methylated C's in CHH context	24691550
Unmethylated C's in CpG context	32419452
Unmethylated C's in CHG context	68737272
Unmethylated C's in CHH context	424173371
Percentage methylation (CpG context)	42.8%
Percentage methylation (CHG context)	20.4%
Percentage methylation (CHH context)	5.5%



叁

MethylKit差异分析

差异分析+注释(genomation)



染色体编号;位置;正负链信息;甲基化碱基数目;非甲基化碱基数目;类型;具体背景

chr1	4	+	0	0	CHH	CCC
chr1	5	+	0	0	CHH	CCT
chr1	6	+	0	0	CHH	CTA
chr1	11	+	0	0	CHH	CCC
chr1	12	+	0	0	CHH	CCT
chr1	13	+	0	0	CHH	CTA
chr1	18	+	0	0	CHH	CCC
chr1	19	+	0	0	CHH	CCT
chr1	20	+	0	0	CHH	CTA
chr1	25	+	0	0	CHH	CCC
chr1	26	+	0	0	CHH	CCT
chr1	27	+	0	0	CHH	CTA
chr1	32	+	0	0	CHH	CCC
chr1	33	+	0	0	CHH	CCT
chr1	34	+	0	0	CHH	CTA
chr1	39	+	0	0	CHH	CCC
chr1	40	+	0	0	CHH	CCT
chr1	41	+	0	0	CHH	CTA
chr1	46	+	0	0	CHH	CCC
chr1	47	+	0	0	CHH	CCT
chr1	48	+	0	0	CHH	CTA
chr1	53	+	0	0	CHH	CCC
chr1	54	+	0	0	CHH	CCT
chr1	55	+	0	0	CHH	CTA
chr1	60	+	0	0	CHH	CCC
chr1	61	+	0	0	CHH	CCT
chr1	62	+	0	0	CHH	CTA
chr1	67	+	0	0	CHH	CCC
chr1	68	+	0	0	CHH	CCT
chr1	69	+	0	0	CHH	CTA

准备输入文件

test.file.R1_1.clean_bismark_bt2_pe.**deduplicated.CX_report.txt**

test.CG.txt

chrBase	chr	base	strand	coverage	freqC	freqT
chr1.48	chr1	48	R	2	100.0	0.0
chr1.49	chr1	49	F	2	100.0	0.0
chr1.232	chr1	232	R	6	100.0	0.0
chr1.233	chr1	233	F	4	100.0	0.0
chr1.246	chr1	246	R	6	100.0	0.0
chr1.247	chr1	247	F	6	100.0	0.0
chr1.258	chr1	258	R	5	100.0	0.0
chr1.259	chr1	259	F	7	100.0	0.0
chr1.366	chr1	366	R	7	100.0	0.0
chr1.367	chr1	367	F	8	100.0	0.0
chr1.432	chr1	432	R	8	100.0	0.0
chr1.433	chr1	433	F	5	100.0	0.0
chr1.435	chr1	435	R	7	100.0	0.0
chr1.436	chr1	436	F	5	100.0	0.0
chr1.557	chr1	557	R	4	100.0	0.0
.....						

test.CHG.txt

test.CHH.txt

读取过滤

```
library(methylKit)
setwd("D:/methlkit/")
file.list1 <- list("BG-2.CG.txt","BG-4.CG.txt","VA-1.CG.txt","VA-5.CG.txt")
m1 = methRead(file.list1,assembly = "csi",sample.id = list("BG-2","BG-4","VA-1","VA-5"),treatment = c(1,1,0,0),context = " CpG")
filtered.m1 = filterByCoverage(m1,lo.count = 4,lo.perc = NULL,hi.count = NULL,hi.perc = 99.9)
write.table(getData(filtered.m1[[1]]),file = "BG-VA-CG/BG-2.CG.filter.txt",row.names = F,quote=F,sep="\t")
write.table(getData(filtered.m1[[2]]),file = "BG-VA-CG/BG-4.CG.filter.txt",row.names = F,quote=F,sep="\t")
write.table(getData(filtered.m1[[3]]),file = "BG-VA-CG/VA-1.CG.filter.txt",row.names = F,quote=F,sep="\t")
write.table(getData(filtered.m1[[4]]),file = "BG-VA-CG/VA-5.CG.filter.txt",row.names = F,quote=F,sep="\t")
```

chr	start	end	strand	coverage	numCs	numTs
chr1	1360	1360	-	10	10	0
chr1	1505	1505	+	11	10	1
chr1	3933	3933	+	10	8	2
chr1	4352	4352	-	10	10	0
chr1	4517	4517	+	10	8	2
chr1	4547	4547	+	10	8	2
chr1	4560	4560	+	10	10	0
chr1	4700	4700	+	11	11	0
chr1	4708	4708	+	11	11	0
chr1	4714	4714	+	10	10	0
chr1	5756	5756	-	11	11	0
chr1	5835	5835	-	10	10	0
chr1	5870	5870	-	10	10	0
chr1	6221	6221	+	10	9	1

划分窗口与组合

```
tiles=tileMethylCounts(filtered.m1,win.size=100,step.size=100)   #100bp窗口
#tiles2=tileMethylCounts(filtered.m1,win.size=1000000,step.size=1000000)   #1MB窗口
tiles.1=getData(tiles[[1]])
write.csv(tiles.1,file = "BG-VA-CG/tiles.BG-2.CG100bp.csv",row.names = F)
tiles.1=getData(tiles[[2]])
write.csv(tiles.1,file = "BG-VA-CG/tiles.BG-4.CG100bp.csv",row.names = F)
tiles.1=getData(tiles[[3]])
write.csv(tiles.1,file = "BG-VA-CG/tiles.VA-1.CG100bp.csv",row.names = F)
tiles.1=getData(tiles[[4]])
write.csv(tiles.1,file = "BG-VA-CG/tiles.VA-5.CG100bp.csv",row.names = F)
```

chr	start	end	strand	coverage	numCs	numTs
chr1	1301	1400	*	10	10	0
chr1	1501	1600	*	11	10	1
chr1	3901	4000	*	10	8	2
chr1	4301	4400	*	10	10	0
chr1	4501	4600	*	30	26	4
chr1	4601	4700	*	11	11	0
chr1	4701	4800	*	21	21	0
chr1	5701	5800	*	11	11	0
chr1	5801	5900	*	20	20	0
chr1	6201	6300	*	10	9	1
chr1	7101	7200	*	10	9	1
chr1	7201	7300	*	10	10	0
chr1	7401	7500	*	10	9	1
chr1	9201	9300	*	60	46	14
chr1	9301	9400	*	113	106	7
chr1	9501	9600	*	68	58	10

```
meth1 = unite(tiles,destrand =F)
write.table(getData(meth1),file = "BG-VA-CG/meth100bp.txt",row.names = F,quote=F,sep="\t")
```

chr	start	end	strand	coverage1	numCs1	numTs1	coverage2	numCs2	numTs2	coverage3	numCs3	numTs3	coverage4	numCs4	numTs4
chr1	7101	7200	*	10	9	1	12	10	2	10	10	0	12	10	2
chr1	9201	9300	*	60	46	14	50	38	12	44	39	5	31	24	7
chr1	9301	9400	*	113	106	7	121	116	5	82	76	6	55	54	1
chr1	9501	9600	*	68	58	10	77	71	6	58	54	4	70	64	6
chr1	10601	10700	*	11	10	1	28	23	5	29	19	10	10	10	0
chr1	16501	16600	*	18	17	1	19	19	0	17	17	0	11	11	0
chr1	23201	23300	*	43	37	6	47	35	12	36	30	6	10	9	1
chr1	23301	23400	*	58	56	2	72	69	3	47	47	0	33	33	0
chr1	29901	30000	*	65	58	7	30	28	2	32	30	2	24	24	0
chr1	31701	31800	*	41	36	5	11	9	2	11	11	0	44	43	1
chr1	31901	32000	*	93	66	27	10	10	0	10	4	6	76	67	9
chr1	32001	32100	*	19	15	4	10	7	3	11	2	9	13	4	9
chr1	32901	33000	*	37	18	19	33	12	21	28	0	28	39	0	39
chr1	35101	35200	*	46	1	45	36	0	36	31	0	31	71	1	70
chr1	35201	35300	*	35	0	35	39	0	39	42	0	42	59	0	59

DMRs

getCorrelation(meth1,plot = TRUE)
clusterSamples(meth1,dist = "correlation",method = "ward",plot = TRUE)
PCASamples(meth1,adj.lim = c(0.5,0.5))

myDiff = **calculateDiffMeth**(meth1)
m1Diff50p.all = **getMethylDiff**(myDiff,difference = **50**,qvalue = 0.05,type = "**all**")
m1Diff50p.hyper = **getMethylDiff**(myDiff,difference = **50**,qvalue = 0.05,type = "**hyper**")
m1Diff50p.hypo = **getMethylDiff**(myDiff,difference = **50**,qvalue = 0.05,type = "**hypo**")
CpGallFrame1 <- getData(m1Diff50p.all)
CpGallFrame2 <- getData(m1Diff50p.hyper)
CpGallFrame3 <- getData(m1Diff50p.hypo)
write.csv(CpGallFrame1,file="BG-VA-CG/CGallFrame_all.csv")
write.csv(CpGallFrame2,file="BG-VA-CG/CGallFrame_hyper.csv")
write.csv(CpGallFrame3,file="BG-VA-CG/CGallFrame_hypo.csv")

	chr	start	end	strand	pvalue	qvalue	meth.diff
12	chr1	32001	32100	*	0.000159	0.001864	50.86207
208	chr1	112101	112200	*	1.63E-12	1.10E-10	53.65766
209	chr1	112201	112300	*	5.31E-12	3.28E-10	53.125
211	chr1	112401	112500	*	6.72E-13	4.78E-11	69.23077
212	chr1	112501	112600	*	4.12E-14	3.56E-12	75.67568
213	chr1	112801	112900	*	1.42E-45	2.25E-42	76.92308
214	chr1	112901	113000	*	9.20E-14	7.50E-12	61.37681
414	chr1	249101	249200	*	1.20E-11	6.99E-10	-53.1778
451	chr1	265901	266000	*	6.91E-13	4.91E-11	55.81738
660	chr1	354801	354900	*	5.74E-36	4.61E-33	-66.6667
661	chr1	354901	355000	*	2.14E-20	4.15E-18	-80.3571
662	chr1	355001	355100	*	6.56E-21	1.33E-18	-57.3626
663	chr1	355101	355200	*	2.22E-28	9.74E-26	-80.8842
664	chr1	355201	355300	*	8.94E-40	9.68E-37	-72.638
710	chr1	378701	378800	*	3.80E-10	1.71E-08	59.35551
855	chr1	484801	484900	*	8.01E-17	1.00E-14	84.50226
900	chr1	512101	512200	*	2.18E-10	1.03E-08	73.32029
934	chr1	524601	524700	*	4.20E-20	7.87E-18	89.58333
949	chr1	529401	529500	*	4.40E-08	1.30E-06	-54.2842
950	chr1	529501	529600	*	2.01E-20	3.91E-18	-74.4186

DMGs

#准备genomation注释需要的bed文件
module load **TransDecoder**/5.5.0
gff3_file_to_bed.pl protein-coding.genes.gff >bed/protein-coding.genes_raw.bed
sed 's/;LOC\S*\t/\t/g' protein-coding.genes._raw.bed |sed 's/ID=//g' >protein-coding.genes.bed

library(genomation)
gene.obj = **readTranscriptFeatures**("protein-coding.genes.bed",up.flank=**2000**,down.flank=**2000**)

diffAnn1 = **annotateWithGeneParts**(as(m1Diff50p.all,"GRanges"),gene.obj)
getTargetAnnotationStats(diffAnn1,percentage=TRUE,precedence=TRUE)
write.csv(diffAnn1@dist.to.TSS,"BG-VA-CG/CG_all_dist.to.TSS.csv")
write.csv(getMembers(diffAnn1),"BG-VA-CG/CG_all_members.csv")

diffAnn2 = **annotateWithGeneParts**(as(m1Diff50p.hyper,"GRanges"),gene.obj)
getTargetAnnotationStats(diffAnn2,percentage=TRUE,precedence=TRUE)
write.csv(diffAnn2@dist.to.TSS,"BG-VA-CG/CG_hyper_dist.to.TSS.csv")
write.csv(getMembers(diffAnn2),"BG-VA-CG/CG_hyper_members.csv")

diffAnn3 = **annotateWithGeneParts**(as(m1Diff50p.hypo,"GRanges"),gene.obj)
getTargetAnnotationStats(diffAnn3,percentage=TRUE,precedence=TRUE)
write.csv(diffAnn3@dist.to.TSS,"BG-VA-CG/CG_hypo_dist.to.TSS.csv")
write.csv(getMembers(diffAnn3),"BG-VA-CG/CG_hypo_members.csv")

*dist.to.TSS.csv

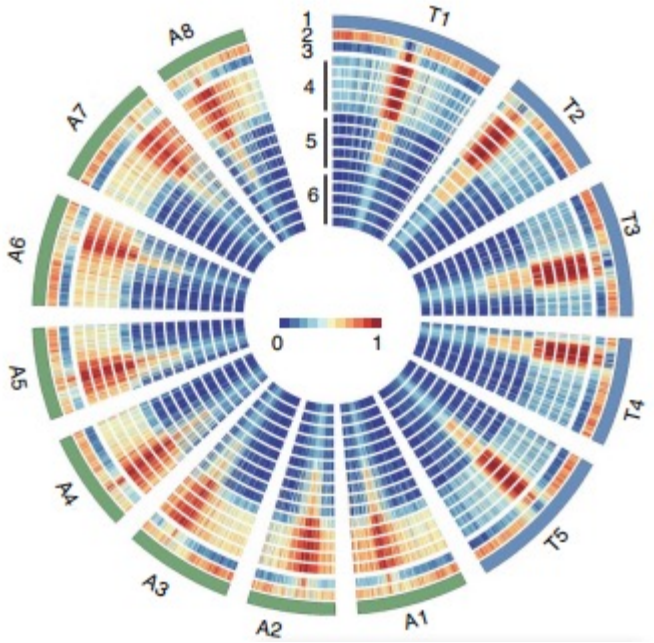
	prom	exon	intron
1	0	0	1
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	1	0
9	1	0	1
10	1	0	0
11	1	0	0
12	1	0	0
13	1	0	0
14	1	0	0
15	0	0	1
16	1	0	0
17	0	0	0
18	1	0	0
19	0	0	0

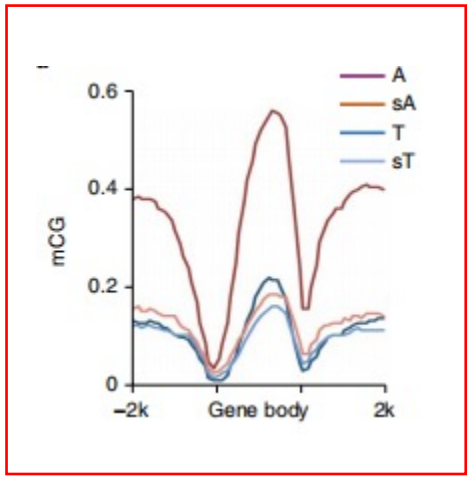
*members.csv

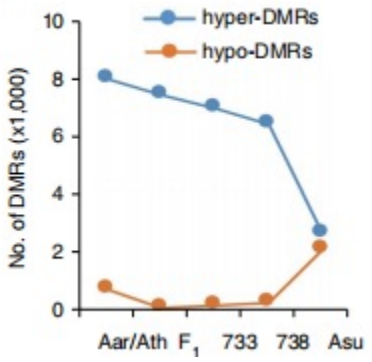
	target.row	dist.to.feature	feature.name	feature.strand
3778	1	4425	Cs_ont_1g000020.1	-
5228	2	-5156	Cs_ont_1g000100.1	+
5228.1	3	-5056	Cs_ont_1g000100.1	+
5228.2	4	-4856	Cs_ont_1g000100.1	+
5228.3	5	-4756	Cs_ont_1g000100.1	+
5228.4	6	-4456	Cs_ont_1g000100.1	+
5228.5	7	-4356	Cs_ont_1g000100.1	+
3086	8	3579	Cs_ont_1g000240.1	-
6177	9	1045	Cs_ont_1g000270.1	-
3102	10	-1337	Cs_ont_1g000410.1	+
3102.1	11	-1237	Cs_ont_1g000410.1	+
3102.2	12	-1137	Cs_ont_1g000410.1	+
3102.3	13	-1037	Cs_ont_1g000410.1	+
3102.4	14	-937	Cs_ont_1g000410.1	+
85	15	4814	Cs_ont_1g000450.1	+
3362	16	-973	Cs_ont_1g000520.1	+
2810	17	4460	Cs_ont_1g000560.1	+
1211	18	-860	Cs_ont_1g000570.1	-
1211.1	19	-5660	Cs_ont_1g000570.1	-

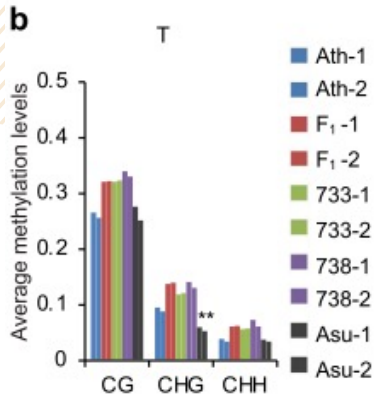
肆

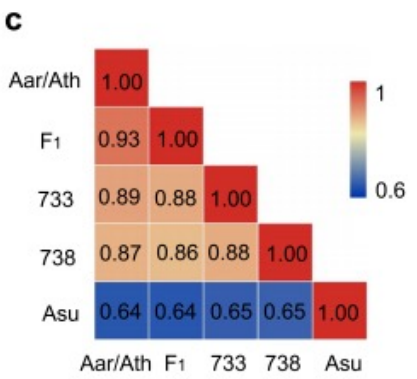
可视化

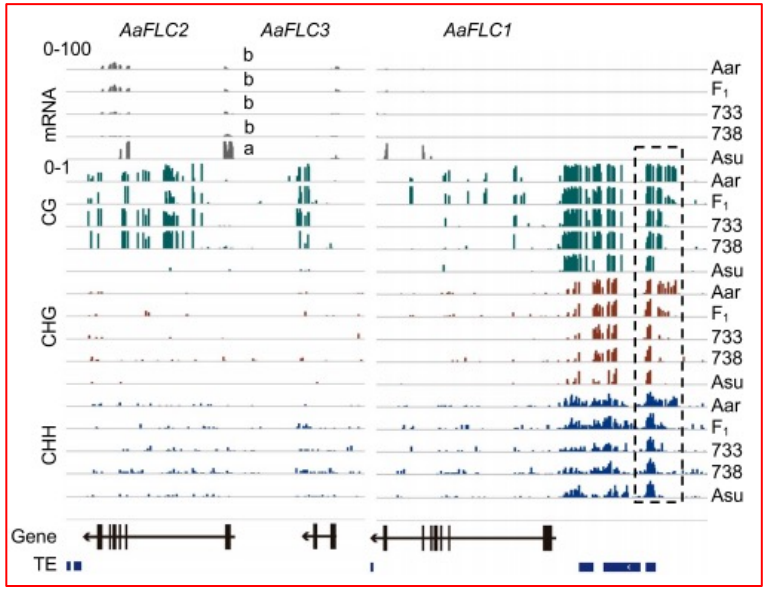


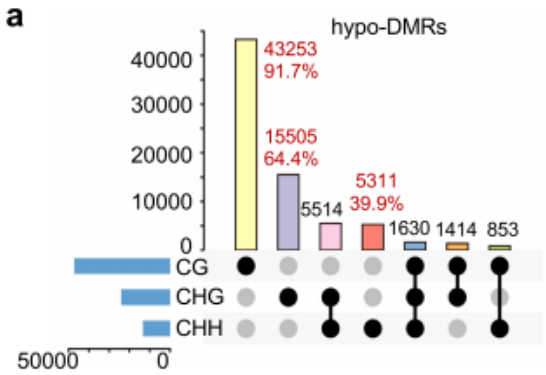










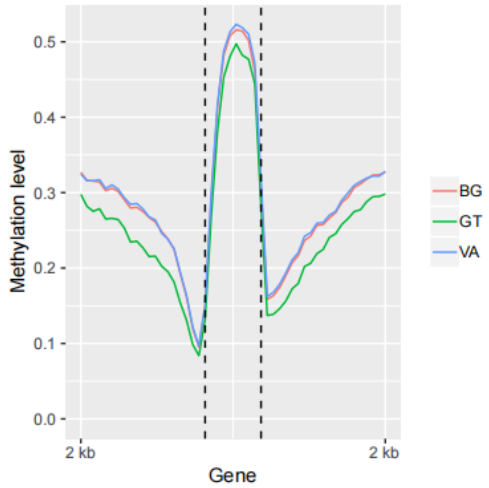


Jiang, X., Song, Q., Ye, W. et al. Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. Nat Ecol Evol 5, 1382–1393 (2021). https://doi.org/10.1038/s41559-021-01523-y

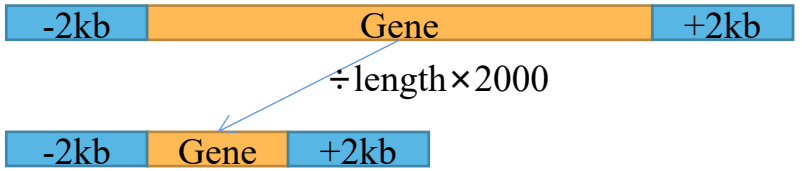
Gene±2kb

ViewBS

#数据准备
samtools faidx genome.fasta
#Bismark结果需bgzip压缩
bgzip ../A.1.clean_bismark_bt2_pe.**deduplicated.CX_report.txt** ./
#生成tbi结尾index文件
tabix -p vcf A.1_bismark_bt2_pe.deduplicated.CX_report.txt.gz
#可视化
ViewBS **MethCoverage** --reference reference.fa \
--sample VA.clean_bismark_bt2_pe.deduplicated.CX_report.txt.gz,VA \
--sample GT.clean_bismark_bt2_pe.deduplicated.CX_report.txt.gz,GT \
--sample BG.clean_bismark_bt2_pe.deduplicated.CX_report.txt.gz,BG \
--outdir MethCoverage --prefix BS_seq_allsam



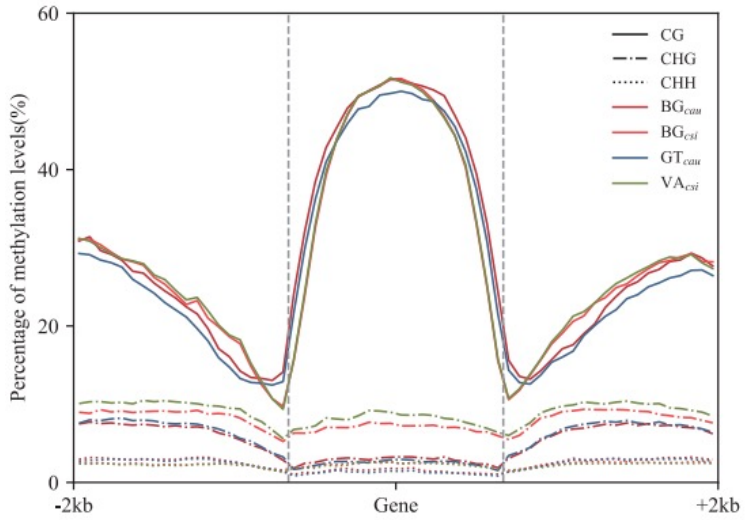
自己写脚本



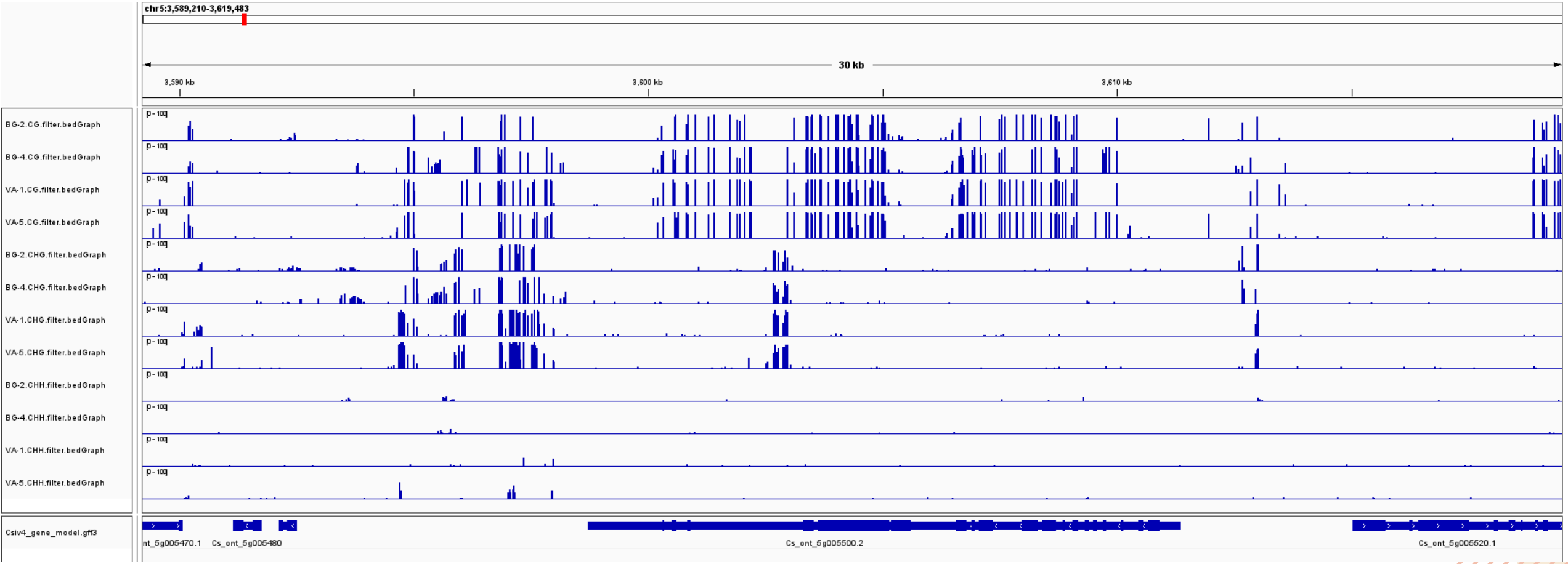
(x,y)

winLen=100
winNum=20

python groupby包



Gene 甲基化位点窗口



感谢大家交流学习！