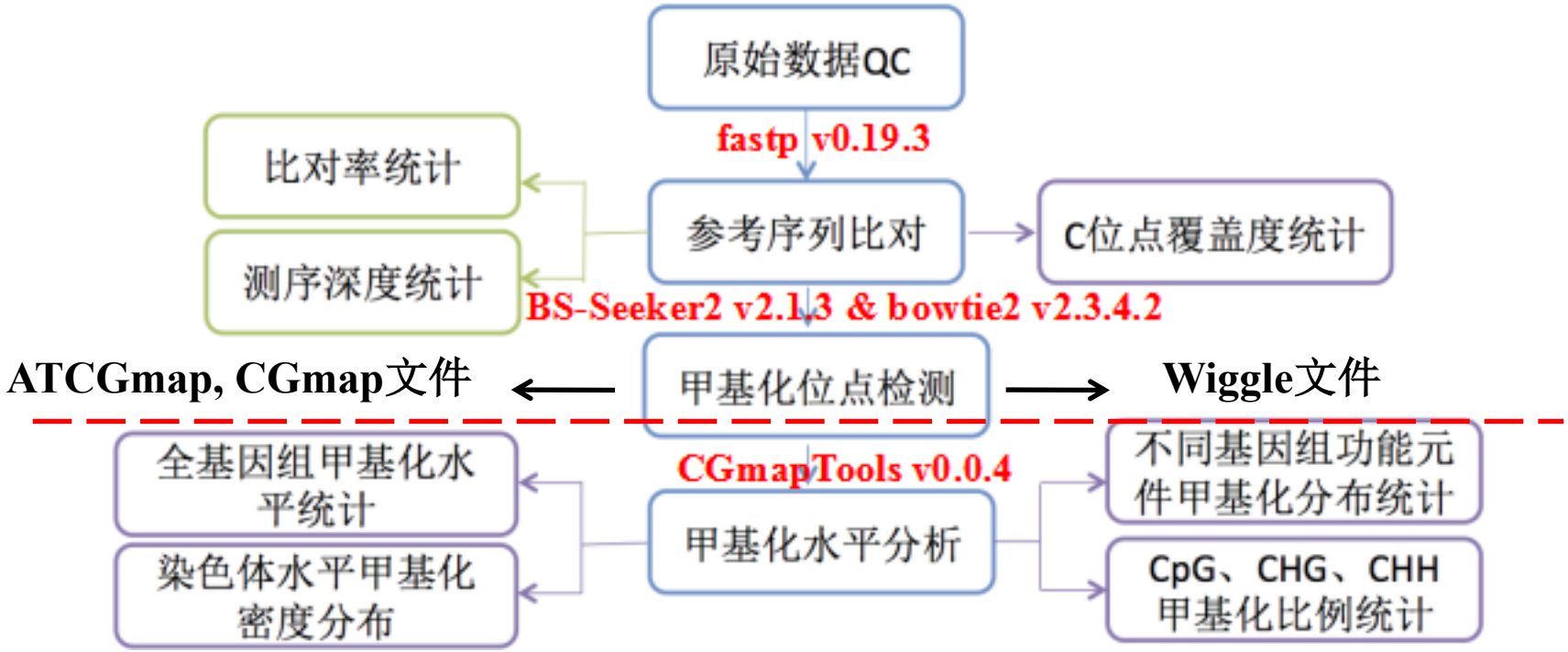
重亚硫酸盐测序数据基本分析流程代码及软件

BS-seeker2 & CGmapTools

黄湘仪
xiangyihuang@126.com



此次DNA甲基化数据分析基本流程图

BS-Seeker2

a versatile aligning pipeline for bisulfite sequencing data

SOFTWARE Open Access

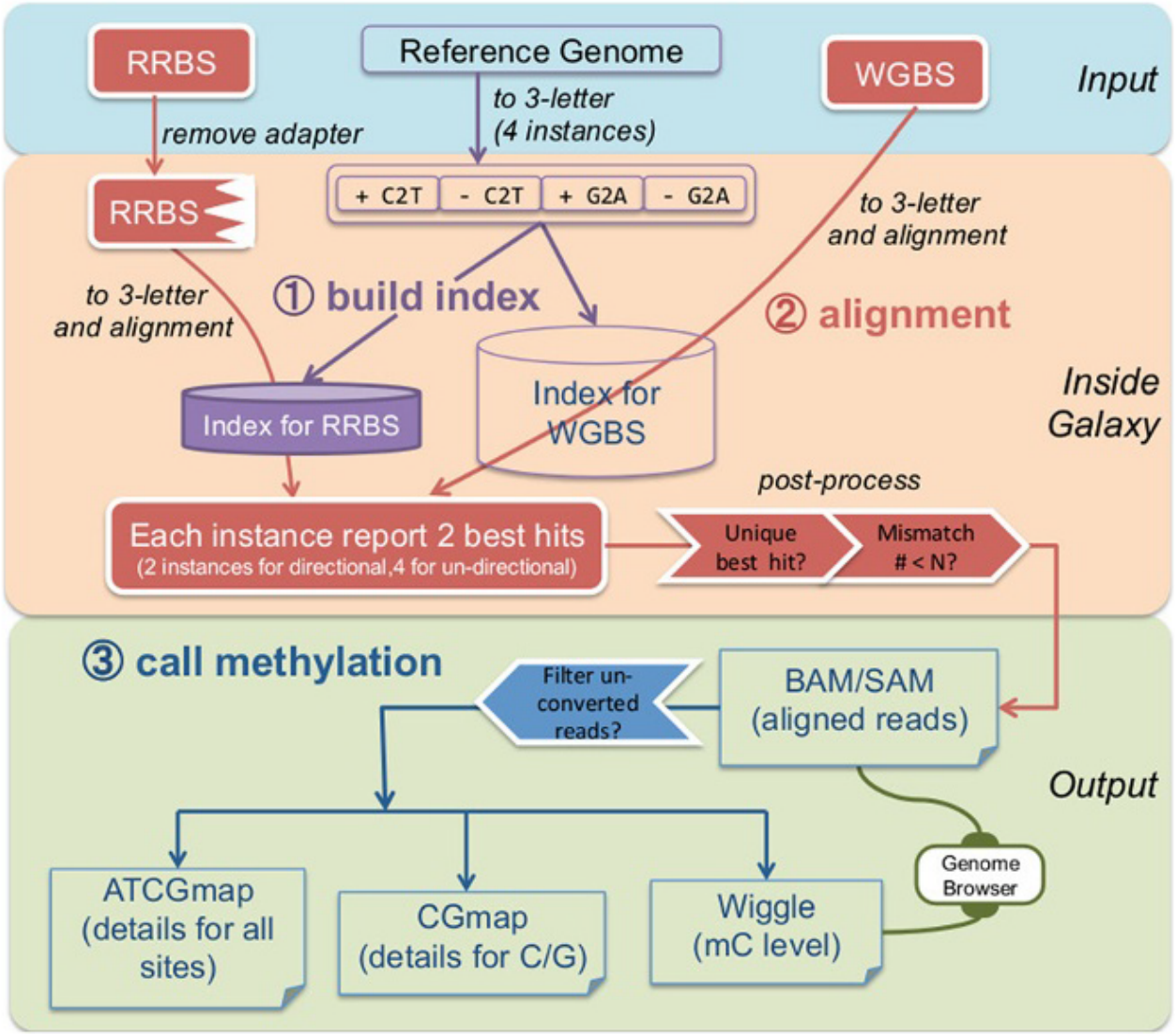
BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data

Weilong Guo[1,2], Petko Fiziev[3], Weihong Yan[4], Shawn Cokus[2], Xueguang Sun[5], Michael Q Zhang[1,6], Pao-Yang Chen[7*] and Matteo Pellegrini[2,8*]

* Corresponding authors: Pao-Yang Chen paoyang@gate.sinica.edu.tw - Matteo Pellegrini matteop@mcdb.ucla.edu

▾ Author Affiliations

文献链接：https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-774



BS-Seeker2数据分析流程图

基本流程：

 # 1 建立WGBS版索引

bs_seeker2-build.py -f genome.fa --aligner=bowtie2

参数--aligner可选择调用bowtie（默认）或bowtie2

 # 2 序列比对

2.1 单端测序reads

-i 输入可为gzip后fasta或fastq文件

bs_seeker2-align.py -i test.fq -o test.bam -g genome.fa

2.2 双端测序reads

bs_seeker2-align.py -1 R1.fq -1 R2.fq -o test.bam -g genome.fa

参数--aligner可选择调用bowtie（默认）或bowtie2来比对，

可利用--bt或--bt2来将参数传给二者，如--bt2-p设置线程，

--bt2-end-to-end和--bt2-local设置模式为全局或局部比对

 # 3 甲基化水平计算：

bs_seeker2-call_methylation.py -i test.bam -o output --db <BSseeker2_path>/bs_utils/reference_genomes/genome.fa_bowtie/ #生成 ATCGmap，CGmap 和 Wiggle文件

BS-Seeker2 基本分析流程

DNA 甲基化是最早发现的基因表观修饰方式之一，在生物界中普遍存在。DNA 甲基化通过对基因组DNA的修饰，从表观遗传水平上对生物遗传信息进行调节，在基因表达调控、发育调节、基因组印迹等方面发挥重要作用。目前，DNA甲基化研究中的主要测序文库包括：WGBS（whole genome bisulfite sequencing）和RRBS（reduced representation bisulfite sequencing）。其中，WGBS文库以MethylC-seq和PBAT（ post-bisulfite adaptor tagging）为主。

用于比对BS-seq数据并绘制DNA甲基化图谱的主流工具有：BS-Seeker2、bismark和BSMAP。BSMAP采用“wild-card”策略进行比对，而BS-Seeker2和Bismark则采用“threee-letter”策略，将基因组和测序数据看作“三碱基”再进行比对，在序列的比对过程中调用bowtie或者bowtie2。

bowtie2有local alignment和end-to-end alignment两个比对模式，而bowtie1只有end-to-end 模式，且bowtie2处理长度大于50bp的数据速度更快并支持gap形式的比对，所以此次比对选择调用bowtie2，并使用end-to-end模式。

Step 1：建立索引

```
# 建立WGBS版索引，调用bowtie2，若为RRBS版索引，添加参数 -r 即可
bs_seeker2-build.py -f ~/genome/wheat_IWGSC/IWGSCv1.fa --aligner=bowtie2
```

“ -f ” 选项表示后续参数文件为参考基因组文件
“ --aligner” 选项表示后续参数为所选用的比对工具，若不指定此参数则默认为bowtie

需要注意的是，BS-Seeker2不能直接使用bowtie2的index，需要自己创建特定的index。建好后的index默认保存在BS-Seeker2的bs_utils/reference_genomes/目录下，-d参数可重新设置目录（或-db=）。

Step 2：序列比对

建好index后，便可利用质控后的数据进行比对。因数据较大，可将其分成各个小块并行比对，有利于效率的提高。

```
# 将reads分成各小块（ 4M reads）
zcat clean.run_1.fastq.gz | split -l 16000000 - R1_
zcat clean.run_2.fastq.gz | split -l 16000000 - R2_
# 可将分割好的各reads压缩后再并行比对，其输入文件可为fastq、fasta、qseq或者纯序列格式及其gzip压缩文件
bs_seeker2-align.py -1 R1_aa.fq.gz -2 R2_aa.fq.gz \
-g ~/genome/wheat_IWGSC/IWGSCv1.fa \
--aligner=bowtie2 \
--bt2-p 8 --bt2--end-to-end --bt2--very-sensitive --bt2--dovetail \
--temp_dir=`pwd`  -m 0.1 --XSteve -o aa.bam
# 若为单端测序reads，则用参数为 -i
# -g 参数后的输入需与建立index时 -f 输入一致
# Bs-seeker2利用--bt2或--bt可将其后参数传递给bowtie2和bowtie，如--bt2--local则可设置为bowtie2局部比对模式
# --bt2-p 8 即执行16个线程（因其比对到两套基因组，默认四线程，即--bt2-p 2）
# 利用参数--temp_dir设置临时文件存放点，程序正常运行结束后临时文件目录自动清除
# -m 0.1 即125bp reads最多可允许12个mismatch，相当于-m 12，但去接头后reads长度不同，最好用百分比
```

BS-Seeker2指定参数 -a 便可去除adapter序列，用户只需将adapter序列写入文件即可

```
bs_seeker2-align.py -i test.fq -o test.bam -g genome.fa -a adapter.txt
```

BS-Seeker2输出的bam文件中只包含unique best hit，可在比对目录下依据生成的log文件查看比对率。

可指定参数“-M XX （或--multiple-hit=XX）”来查看被过滤掉的multiple hits的read的信息。
若需要的是没有任何比对位置的read的信息，可指定参数“-u XX (或--unmapped=XX)”。

```
# 最后把上述分成小块reads比对得到的各bam文件合并起来
samtools merge merge.bam `ls ??.bam`
# 排序好的文件可节省计算时间
samtools sort merge.bam -o sort.bam
# 标签XS:i:1代表BS-Seeker2利用CH位点的甲基化信息判断出reads failed in bisulfite-conversion
samtools view -h sort.bam | grep -v 'XS:i:1' | samtools view -bS -  > sort_rmSX.bam
# 或在call methylation步骤中，设置参数 “-x”（或“--rm-SX”）可移除被标记为 XS:i:1 的read
```

敲黑板，划重点啦！

根据经验，对于双端测序数据，比对过程中利用单端模式相对于双端模式可以比对地“更多、更准、更高效”。
虽然BS-Seeker2和其它同类软件一样，也提供了双端比对模式，但若测序读长大于80bp，建议使用单端模式对每个mate进行比对，最后合并所有bam文件。

```
# 单端模式比对mate 1
bs_seeker-align.py -i mate_1.fq -o mate_1.bam
# mate 2转为其反义序列再进行比对
Antisense.py -i mate_2.fq -o mate_2.anti.fq
bs_seeker-align.py -i mate_2.anti.fq -o mate_2.bam
# 将比对文件合并在一起
samtools merge merge.bam mate_1.bam mate_2.bam
```

Step 3：各碱基位点甲基化水平计算

将sort_rmSX.bam按染色体分离后，再分别进行call methylation，有利于提高效率

```
bs_seeker2-call_methylation.py -i chr1A.bam -o chr1A \
-d ~/xywheat3/BSseeker2-2.1.3/bs_utils/reference_genomes/IWGSCv1.fa_bowtie2/  --rm-overlap --sorted
```

“ -d ”选项表示后续参数为参考基因组索引建立后的存放路径
“--rm-overlap” 参数，确保双端测序reads重合的区域只被计算一次
“--sorted”参数，代表读入的文件是已经经过排序的，可节省计算时间。若未指定参数，BS-Seeker2读入bam文件后，首先对bam文件进行sort，并建立bai文件索引

另，对于RRBS来说，CCGG位点上的对链是后面步骤合成并添加的，计算甲基化水平的过程中应去掉该位置。参数“--rm-CCGG”可以在最终的结果中移除CCGG位点。

输出文件为ATCGmap（所有碱基），CGmap（胞嘧啶）和Wiggle文件

目前软件最新版本为 v2.1.5

软件下载地址：http://pellegrini-legacy.mcdb.ucla.edu/bs_seeker2/

软件详细说明：https://github.com/BSSeeker/BSseeker2/

上述内容部分摘自http://guoweilong.github.io/bsseeker2_doc_zh.html，更多详细内容可登陆此网站查看

CGmapTools

CGmapTools

improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data

Bioinformatics, 34(3), 2018, 381–387
doi: 10.1093/bioinformatics/btx595
Advance Access Publication Date: 18 September 2017
Original Paper

OXFORD

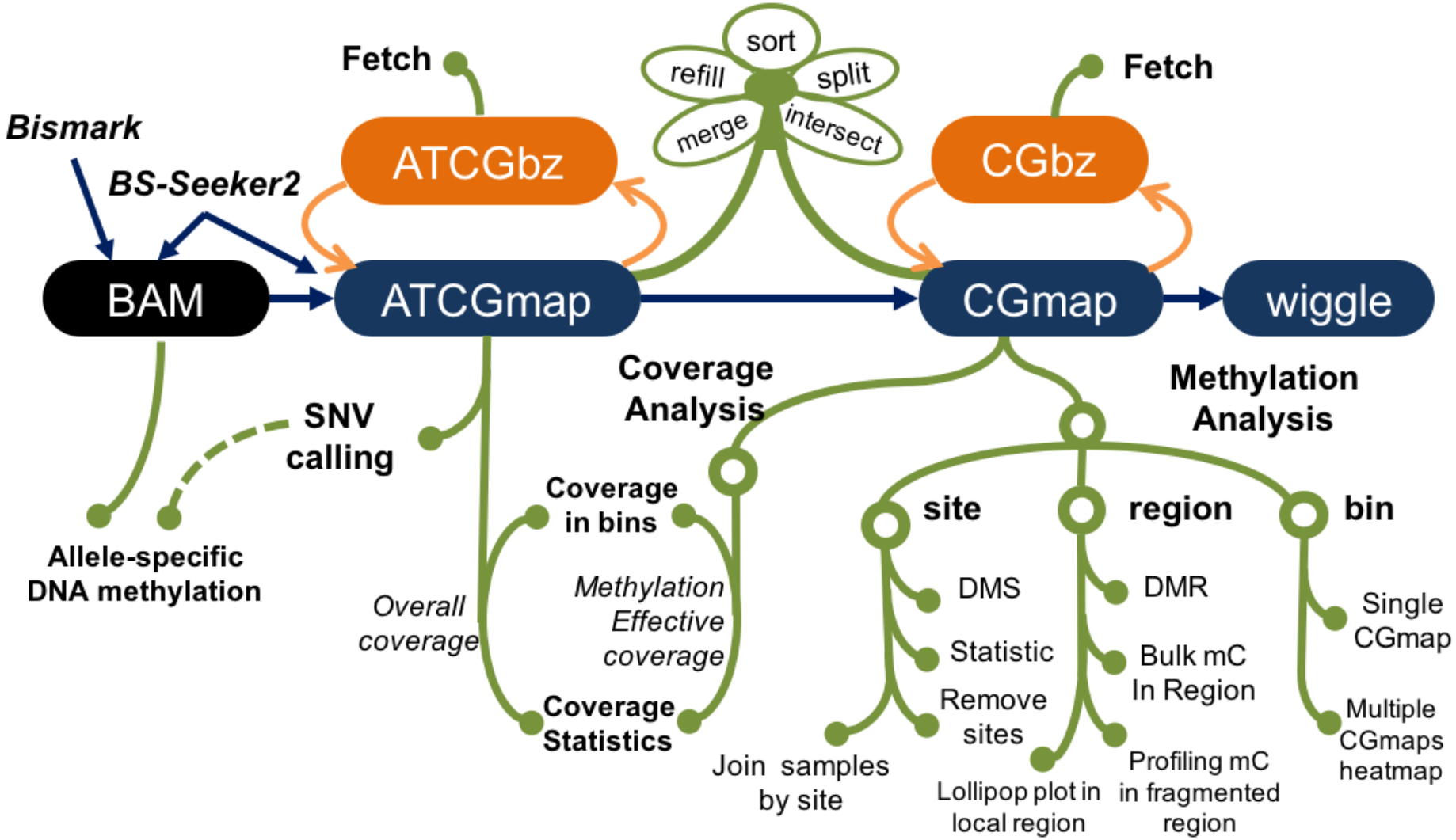
Genome analysis

CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data

Weilong Guo[1,*,†], Ping Zhu[2,3,†], Matteo Pellegrini[4], Michael Q. Zhang[5,6], Xiangfeng Wang[7] and Zhongfu Ni[1]

文章： CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data

链接：https://academic.oup.com/bioinformatics/article/34/3/381/4160682



CGmapTools数据分析流程图

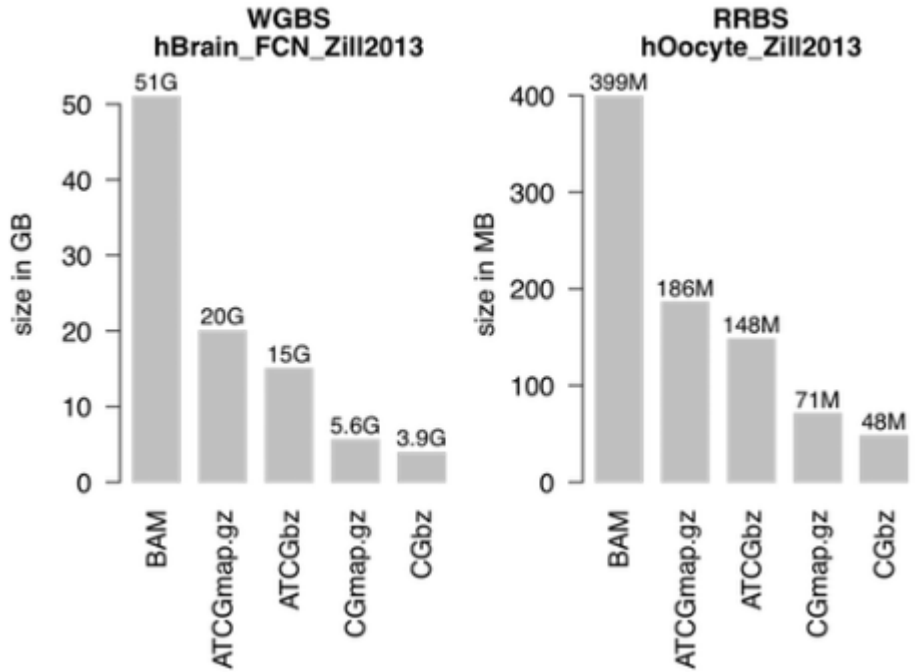
CGmapTools 基本分析流程

DNA甲基化在许多生物过程中都扮演着非常重要的角色，随着测序成本的降低以及新技术方法的开发，DNA甲基化的研究越来越多，如何高效存储、共享、分析和可视化DNA甲基化数据成为亟需解决的问题。

此前，已有一些针对DNA甲基化数据分析的工具被开发出来，组合现有工具能够解决我们大部分的分析需求，但由于**不同工具没有统一的输入和输出格式，给整合分析造成了较大困难，且部分工具的性能还有待较大提升。**

而CGmapTools能显著提升BS-seq数据中计算杂合SNV的精准度，支持链特异DNA甲基化等40种分析及可视化。同时，广泛使用的BS-seq比对软件BS-Seeker2定义了用于存储DNA 甲基化信息的 **CGmap和ATCGmap文件格式，CGmapTools以其作为标准文件格式接口，打破了数据格式的屏障，有助于甲基化组学领域实现数据共享互通。**

Tip 1：统一的数据格式




```
# CGmapTools 既支持对ATCGmap和CGmap文件进行直接分析
# 也支持从 BAM 格式生成 CGmap 格式文件
cgmaptools convert bam2cgmap -b x.bam -g gn.fa --rmOverlap -o WG
# 还可以将 Bismark 分析结果转换为 CGmap 格式
cgmaptools convert bismark2cgmap -i bismark.dat -o output.CGmap
```

受samtools的bam文件和sam文件的启发，CGmapTools设计了新的二进制格式：CGbz和ATCGbz，其具有以下特点：

✓ 节省硬盘空间

✓ 利用二分法从硬盘大文件中查找信息，省时、省点、不伤硬盘

```
cgmaptools fetch cgbz -h
cgmaptools fetch atcgbz -h
```

CGmap文件格式内容

染色体	核苷酸	位置	CG/CHG/CHH模式	CX模式	甲基化水平	未转化读段数	转化读段数
chr1A	G	466	CHH	CA	0.1	1	10
chr1A	C	467	CHG	CT	0.5	2	4
chr1A	G	469	CHG	CA	0.9	9	10
chr1A	C	471	CG	CG	1	5	5

……

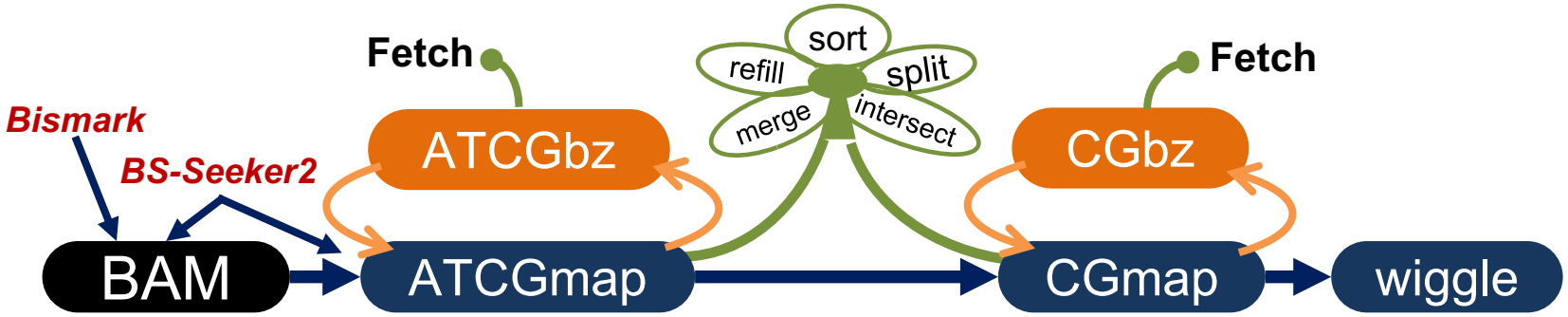
ATCGmap文件格式内容

染色体	核苷酸	位置	CG/CHG/CHH模式	CX模式	正链					负链					甲基化水平
					A	T	C	G	N	A	T	C	G	N	
chr1A	G	225	CG	CG	0	0	0	0	0	0	0	0	3	0	1
chr1A	G	226	CHG	CC	0	0	0	0	0	2	0	0	1	0	0.33
chr1A	C	227	CHH	CA	0	0	0	0	0	0	0	3	0	0	na
chr1A	A	228	--	--	0	0	0	0	0	3	0	0	0	0	na

……

注： CGmap文件共8列，包含基因组所有胞嘧啶信息，所以第二列G代表负链；ATCGmap文件共16列，包含所有碱基信息。

Tip 2：文件格式架构 & 支持标准输入(STDIN)和输出(STDOUT)，轻松实现工具的多重组合




```
cgmaptools -h
Program : cgmaptools (Tools for analysis in CGmap/ATCGmap format)
Commands:
-- File manipulation
     convert     + data format conversion tools
     fetch       + fetch a region by random accessing
     refill       refill the missing columns
     intersect     intersect two files
     merge2      + merge two files into one
     mergelist   + merge a list of files
     sort         sort lines by chromosome and position
     split       + split file by chromosomes
     select      + select lines by region/site
Note:
  Commands contain sub-commands are marked with "+"
```

Tip 3：提供多层次甲基化分析和可视化工具

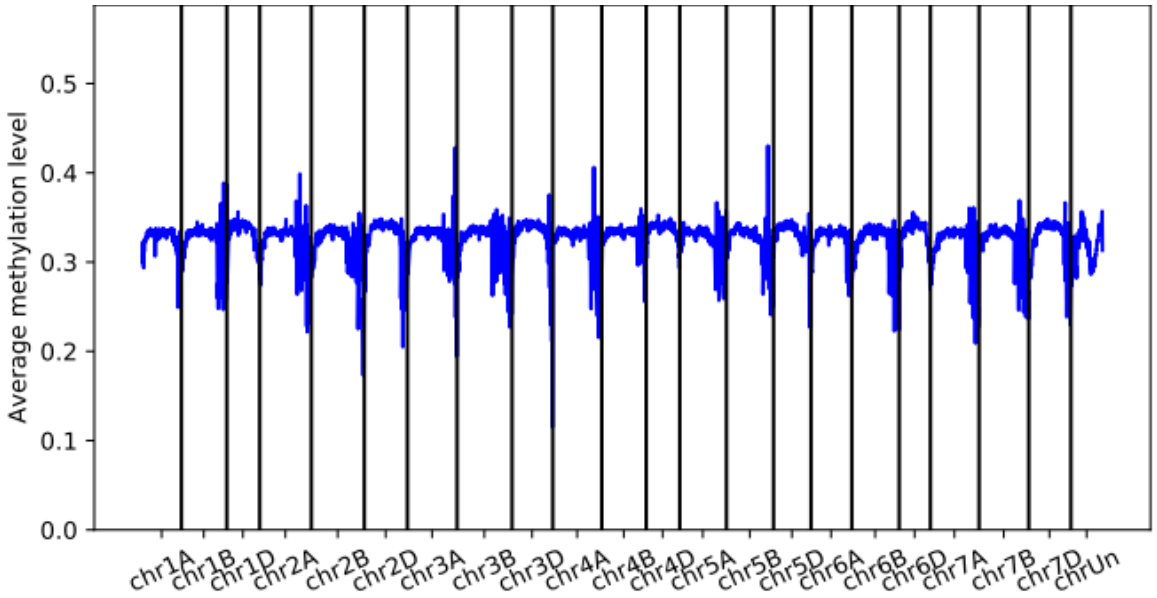
支持基于不同水平的DNA甲基化分析和可视化，如不同C的类型(CpG、非CpG等）、不同基因组区域(如基因区域、启动子区域等）、不同窗口大小、不同样本等。

```
# bin-wise DNA甲基化分析
cgmaptools mbin -h
# region-wise DNA甲基化分析
cgmaptools mfg -h
# site-wise DNA甲基化分析:“棒棒糖图”（lollipop plot）
cgmaptools lollipop -h
# read-wise DNA甲基化分析:“糖葫芦图”（tanghulu plot）
cgmaptools tanghulu -h
# sample-wise DNA甲基化分析:热图和聚类
cgmaptools heatmap -h
# context-wise DNA甲基化分析
cgmaptools mstat -h
#=============================================
cgmaptools  -h
-- Coverage analysis
    oac       +* overall coverage (for ATCGmap)
    mec       +* methylation effective coverage (for CGmap)
Note:
  Commands support figures generation are marked with "*"
  Commands contain sub-commands are marked with "+"
```

3.1 ：mbin

该工具将计算全基因组每个bin的平均甲基化水平，并生成汇总表和分布图

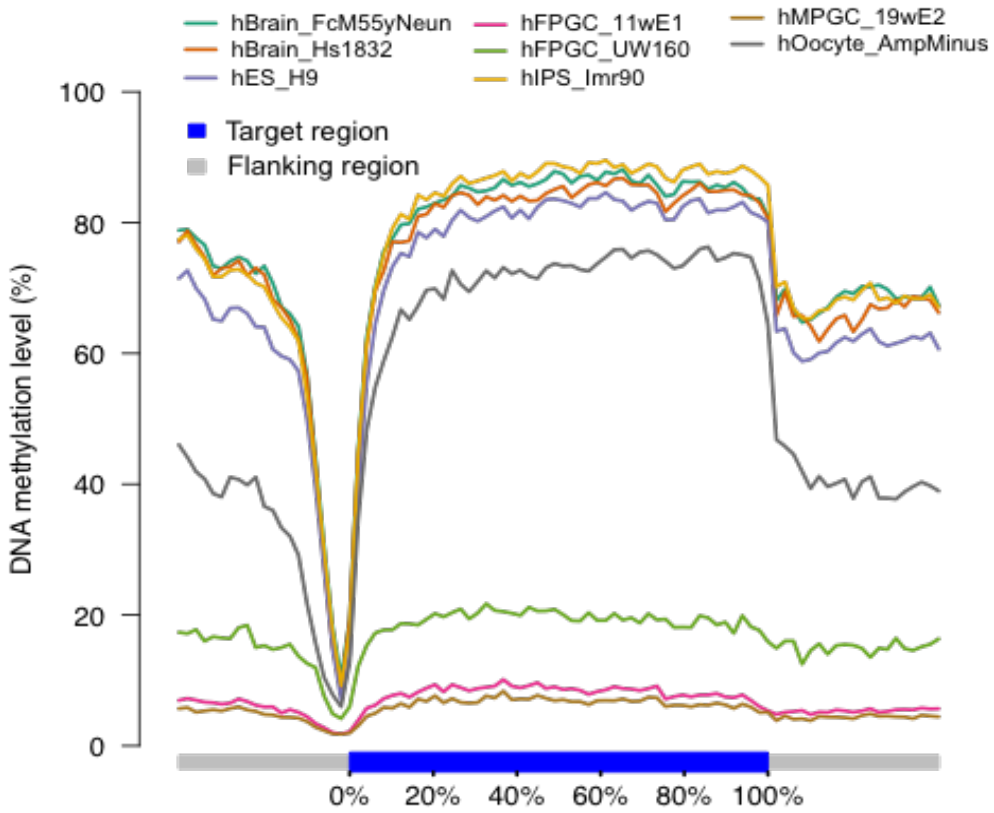
```
zcat sample.CGmap.gz | cgmaptools mbin -B 5000000 -c 10 -f pdf  -p sample.mbin -t sample.mbin  > sample.mbin.log &
```




```
 cgmaptools mbin  -h
-i FILE                 .CGmap or .CGmap.gz. 标准格式的输入文件
-B BIN_SIZE             定义bin的大小，默认： 5000000
-c COVERAGE             coverage低于此设定则不予分析，默认：10
-C CONTEXT, --context=CONTEXT
                        若不设定模式，所有位点均予以分析
                        模式设定：CG, CH, CHG, CHH, CA, CC, CT, CW
 --cXY=COVERAGEXY  对雄性只计算一条性染色体覆盖度
-f FIGTYPE, --figure-type=FIGTYPE
                        设定给出图片类型：png, pdf等
-H FLOAT                高度，单位：英寸，默认4英寸
-W FLOAT                宽度，单位：英寸，默认8英寸
-p STRING               给出图片名前缀
-t STRING, --title=STRING   图片名
```

3.2：mfg

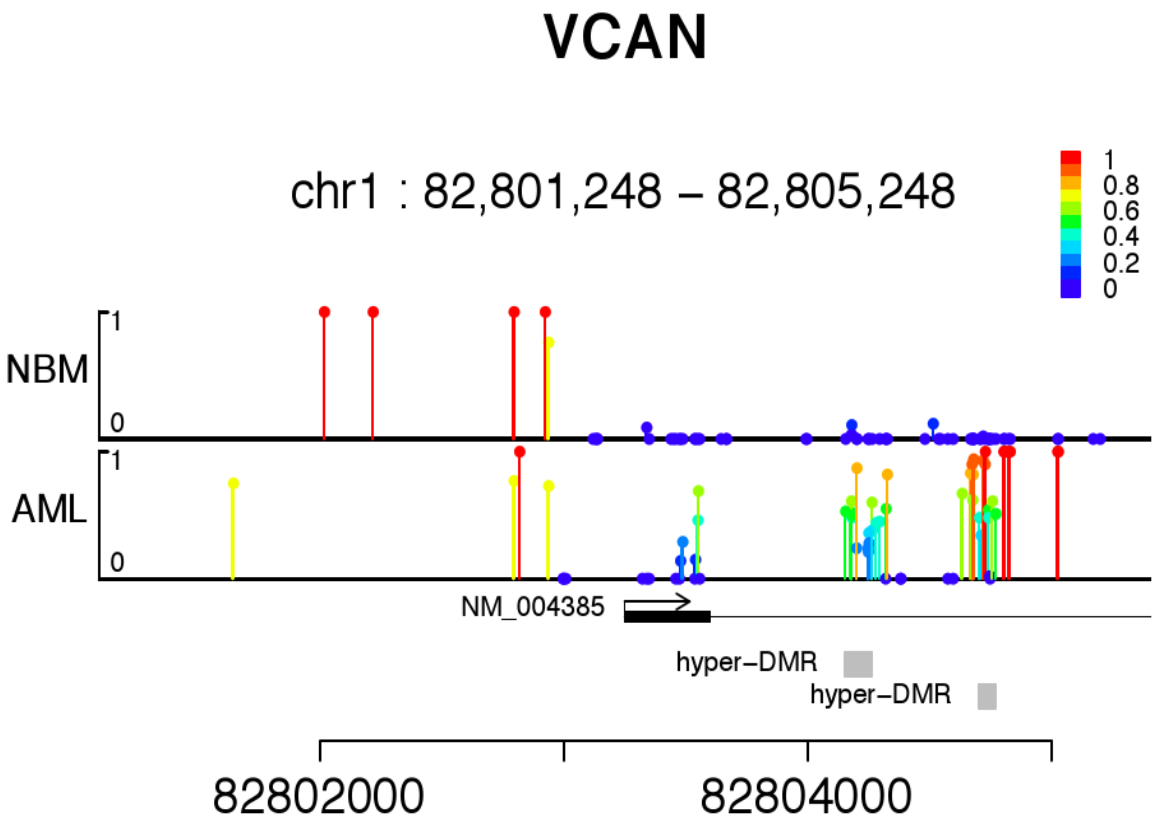
```
cgmaptools bed2fragreg
cgmaptools mfg
cgmaptools fragreg  #step by step
```



可用于绘制多样本在genebody，启动子，转座子等各元件上甲基化水平分布及CG，CHG和CHH等各context甲基化水平分布趋势。

3.3：lollipop

```
# gene-wise DNA methylation analysis
cgmaptools lollipop
# Description: Plot local mC level for multiple samples
```



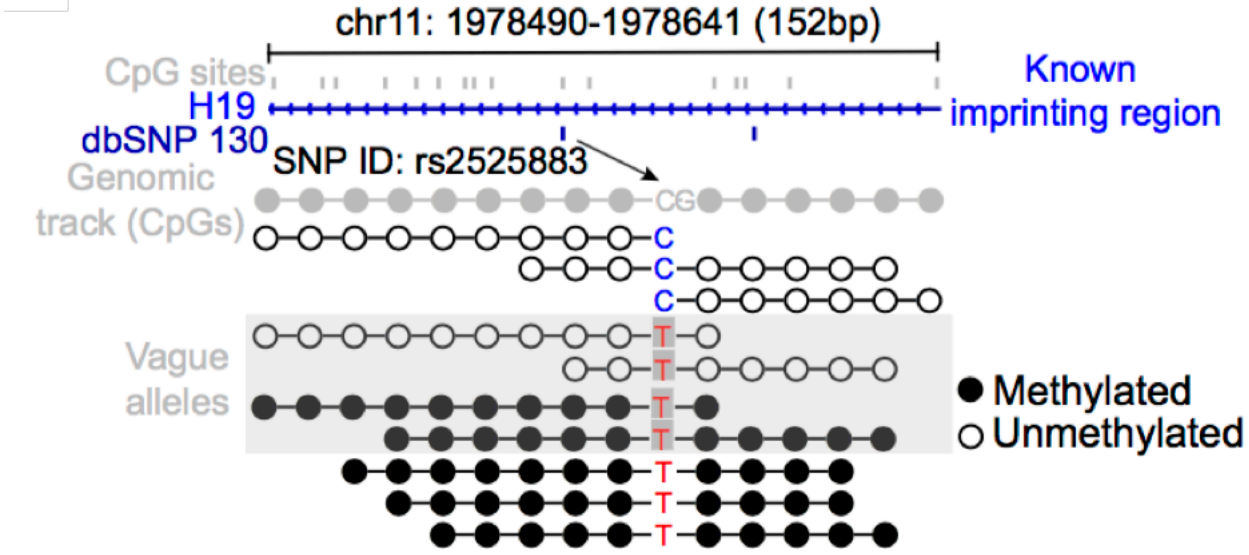
每个圆点代表一个被检测（被覆盖）到的胞嘧啶，其高度和颜色代表甲基化水平。

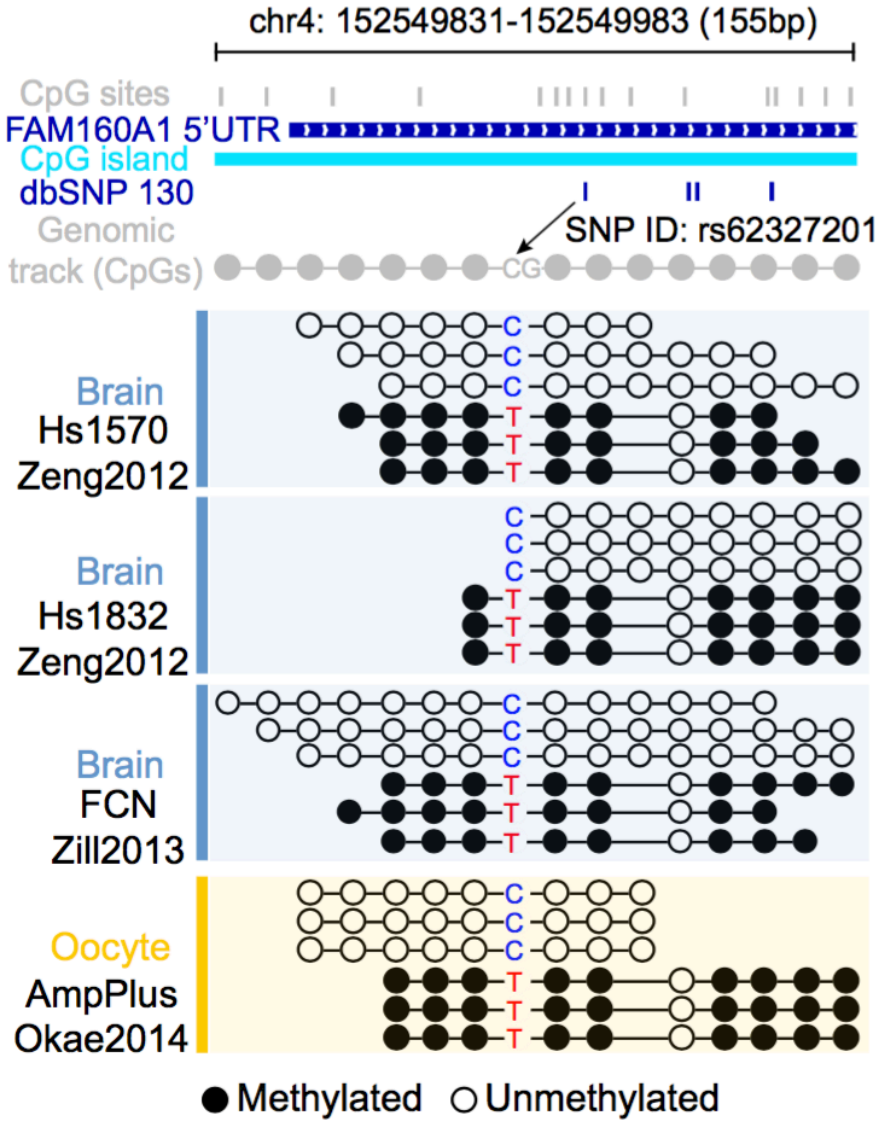
可用于区分未甲基化和未被覆盖的位点。

3.4：tanghulu

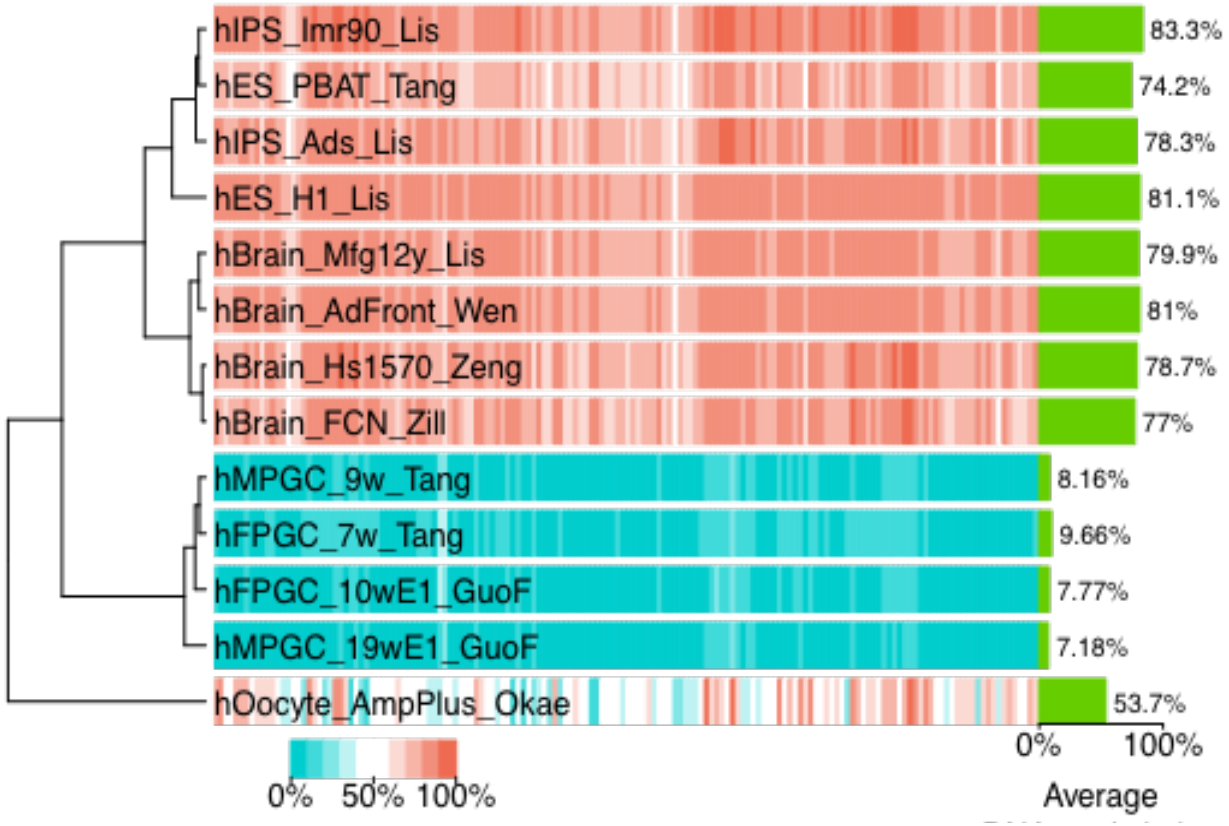
CGmapTools中ASM分析中考虑vague allele

cgmaptools tanghulu





3.5：heatmap



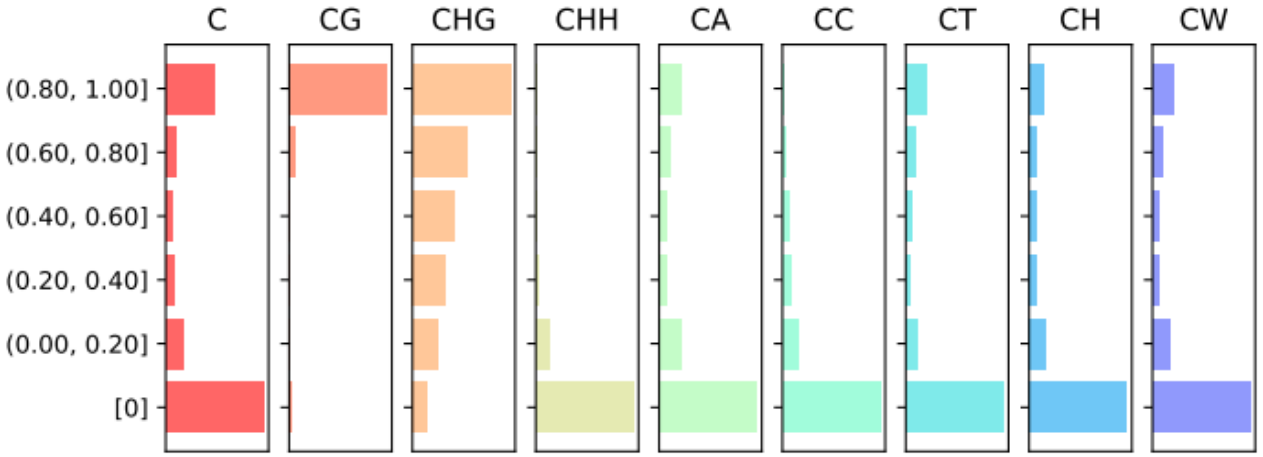
Chromosome-wise DNA甲基化分析

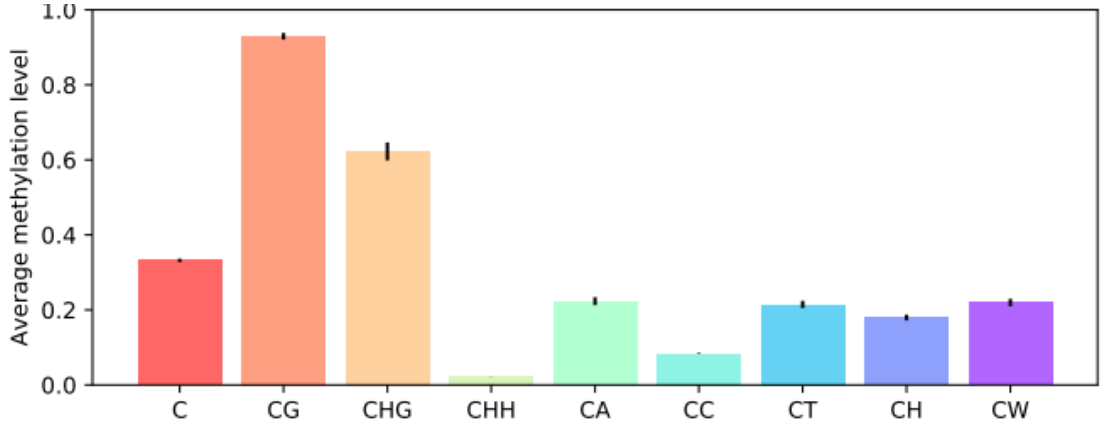
cgmaptools heatmap
Description: Plot methylation dynamics of target
region for multiple samples

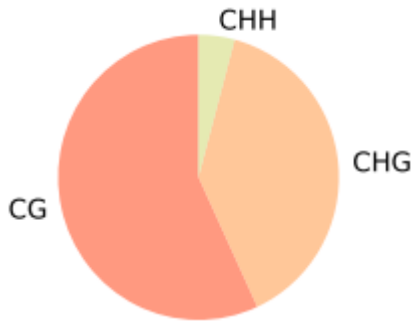
3.6：mstat

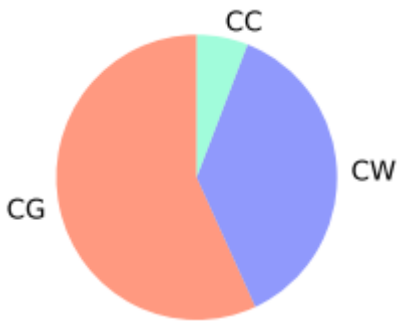
该工具将计算全基因组各context的甲基化水平，即context在甲基化胞嘧啶中所占比例

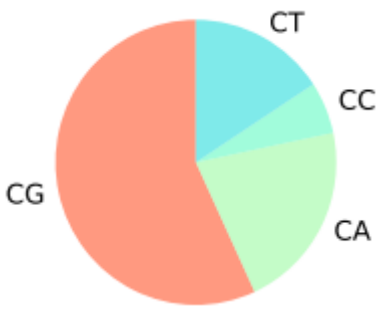
```
zcat sample.CGmap.gz | cgmaptools mstat -c 10 -f pdf -p sample.mstat -t sample.mstat > sample.mstat.log &
```











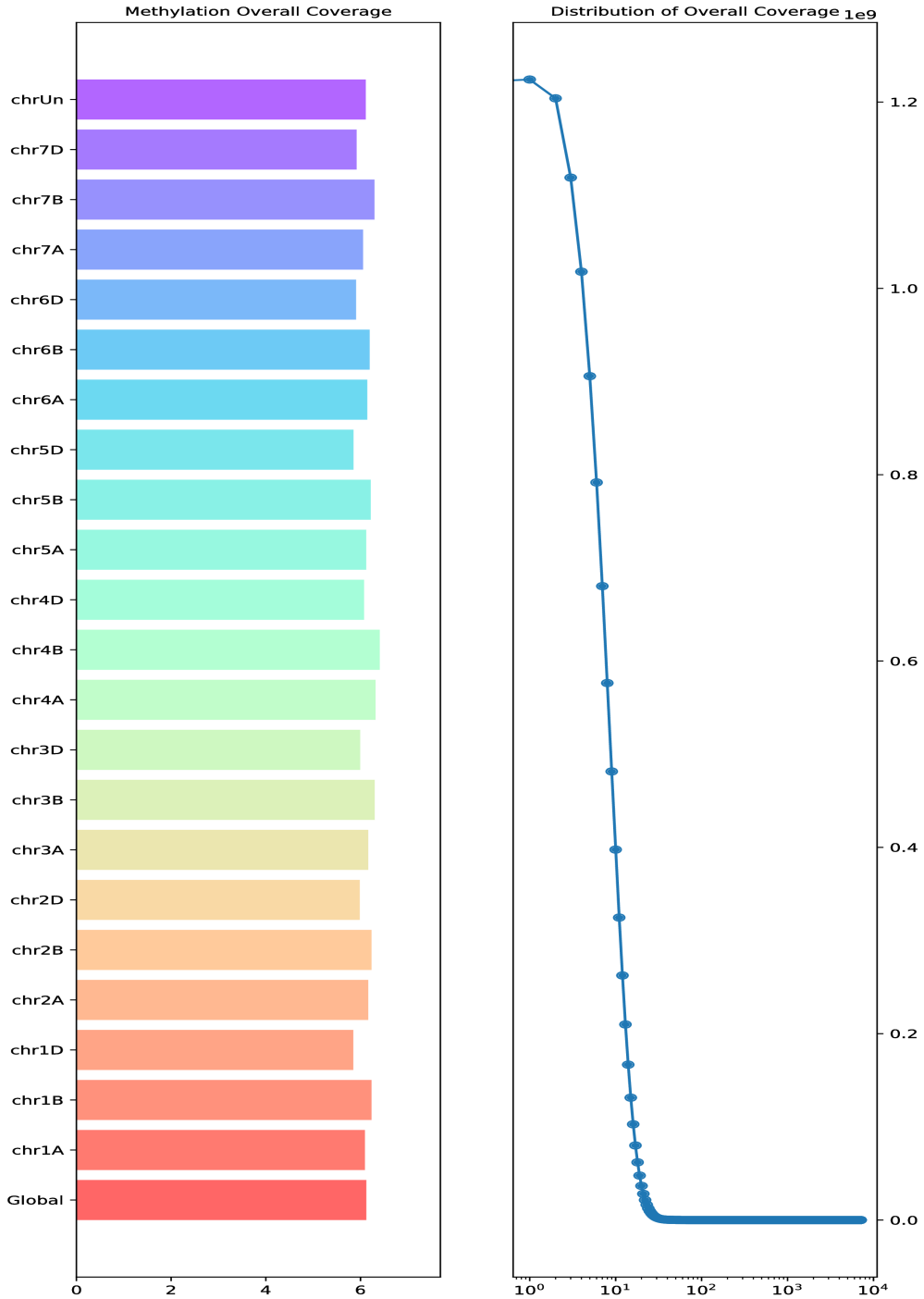
植物

动物研究中关注的context

左上，各context中甲基化胞嘧啶的分布。
如含CG context的片段，
其CG中的胞嘧啶绝大部分全发生了甲基化。

右上，胞嘧啶及各context平均甲基化水平。

左下，甲基化胞嘧啶中，各context所占比例。



3.7：oac　该工具将统计全基因组所有碱基覆盖度信息

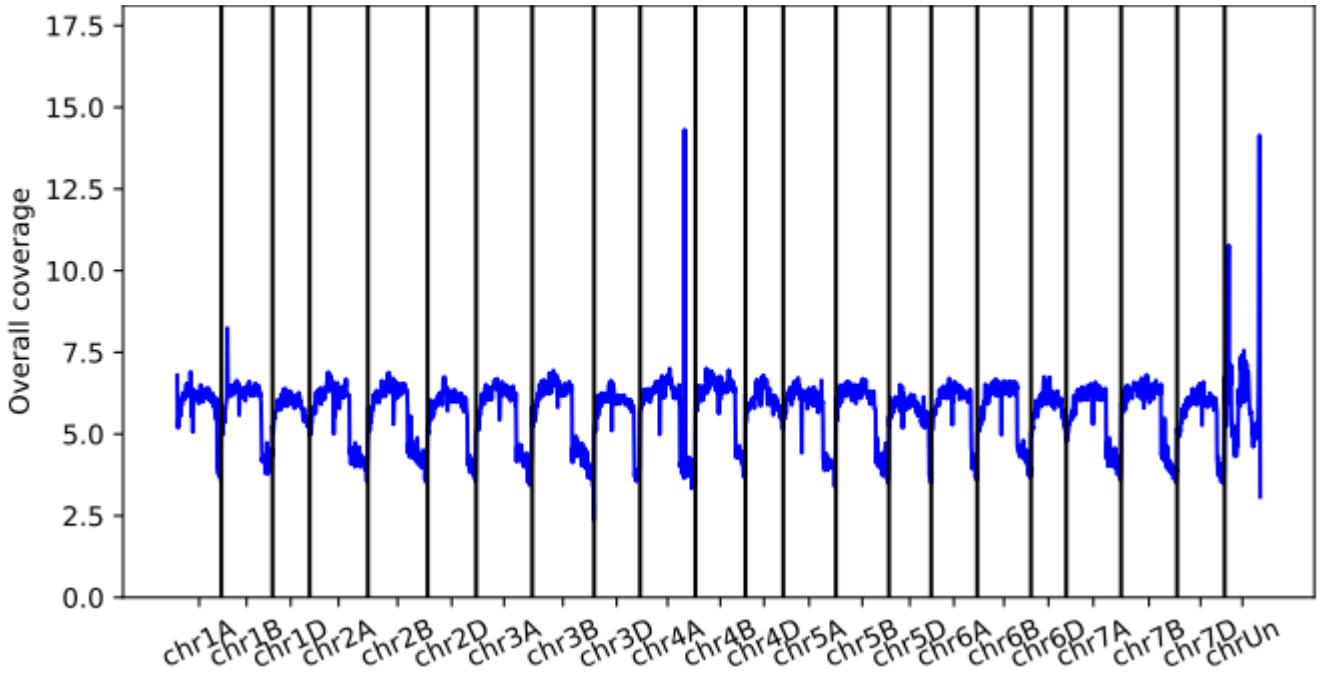
左图：oac stat

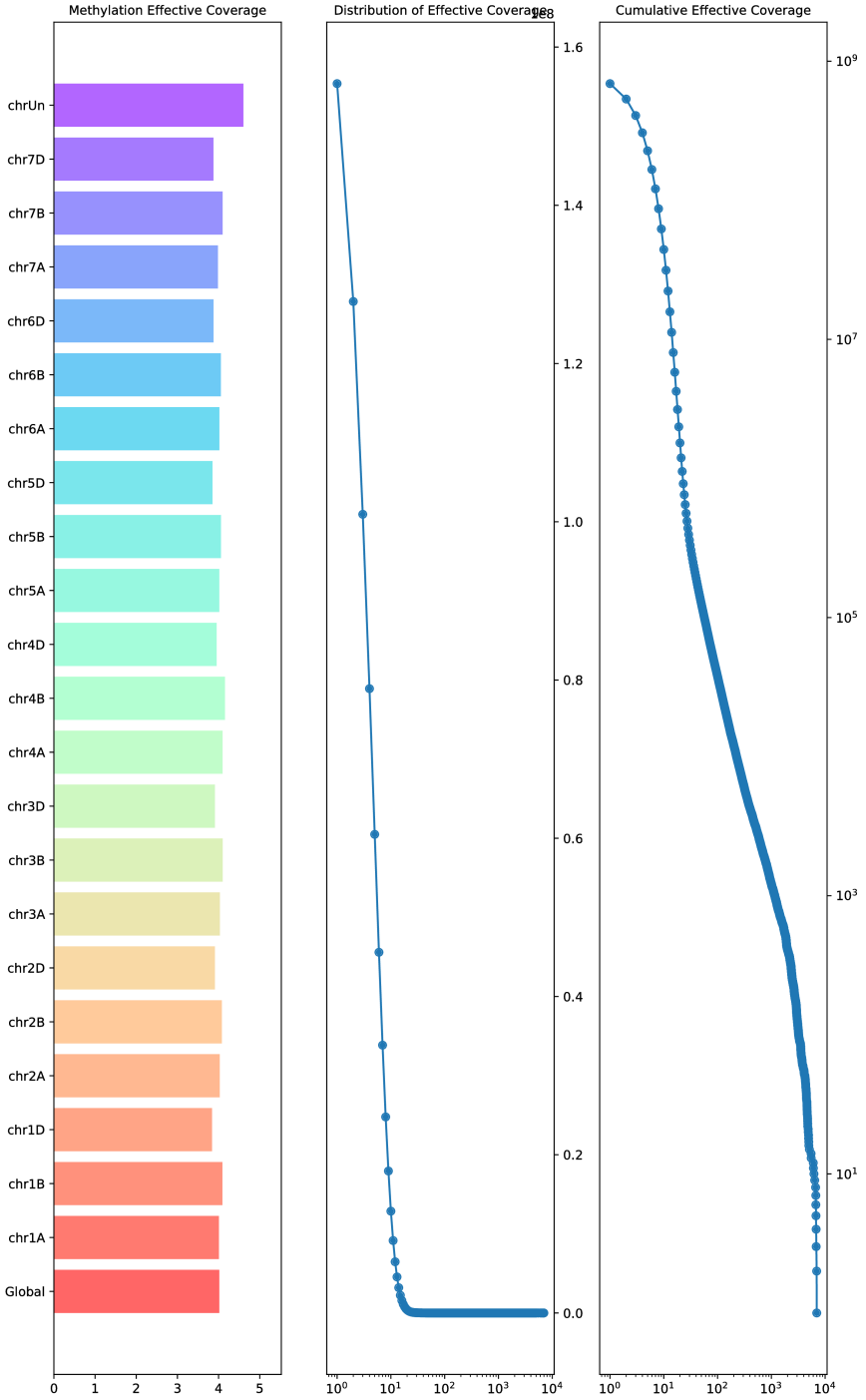
```
zcat sample.ATCGmap.gz | cgmaptools oac stat -f pdf  -p \
sample.oac_stat > sample.oac_stat.log &
```

其左半边图为各染色体所有碱基覆盖度信息，右半边图横轴为覆盖度，纵轴为reads数，单位为1e9（1G）。

右下图：oac bin

```
zcat sample.ATCGmap.gz | cgmaptools oac bin -B 5000000 -f pdf \
 -p sample.oac_bin -t sample.oac_bin > sample.oac_bin.log &
```





3.8：mec　该工具将统计全基因组所有胞嘧啶覆盖度信息

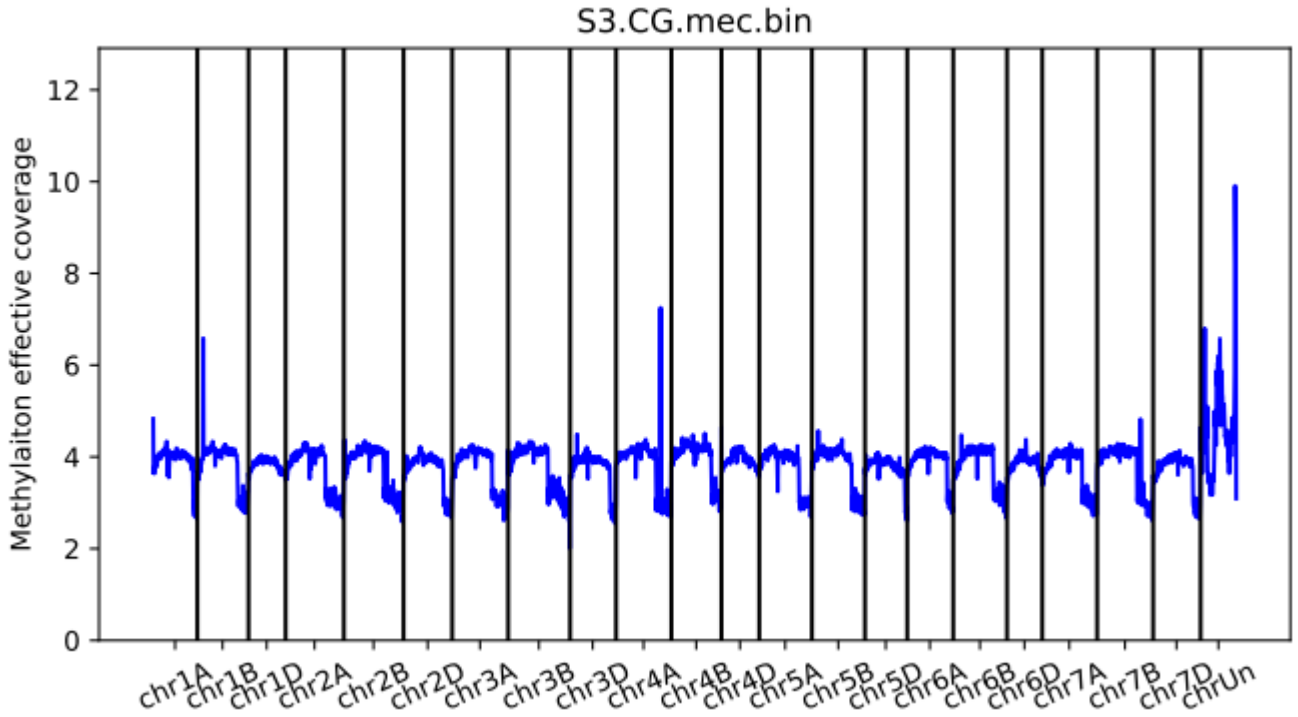
左图：mec stat

```
zcat sample.CGmap.gz | cgmaptools mec stat  -f pdf  -p S.CG.mec.stat \
-C CG > S.mec_stat_CG.log    # -C 设定 context
```

其中间图纵轴为reads数，单位为1e8；右半边图横轴为覆盖度，横轴从右往左看，纵轴为reads累积数，可依此设定cut-off值。

右下图：mec bin

```
zcat sample.CGmap.gz | cgmaptools mec bin -B 5000000 -f pdf  -p  \
sample.CG.mec.bin -t S3.CG.mec.bin -C CG > S.mec_bin_CG.log
```



coverage and methyletion level

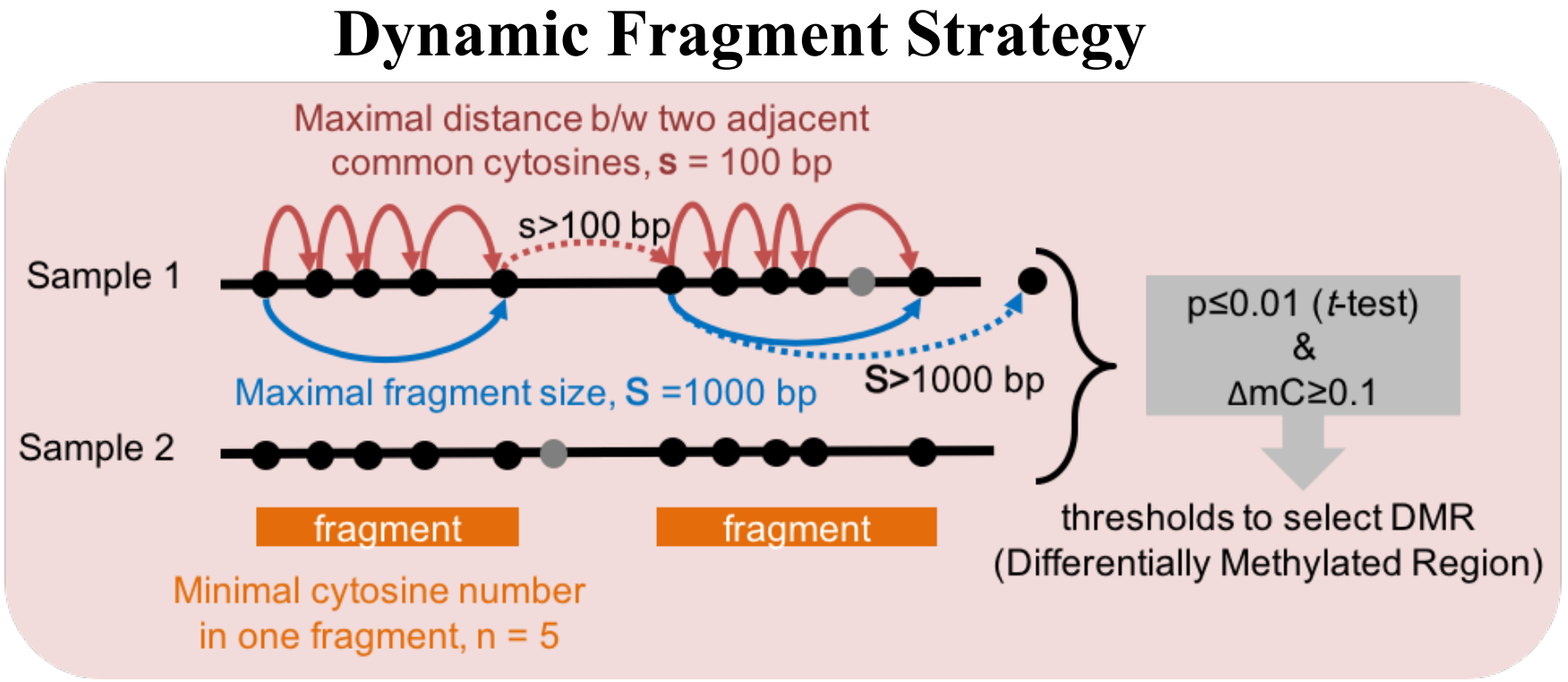
sample	global coverage				mC levels of different contexts			mC contributions of different contexts			mC
	oac	mec									
		CG	CHG	CHH	CG	CHG	CHH	CG	CHG	CHH	
cs_leaf2		4.2809	4.3752	3.8224	0.9449	0.6310	0.0267	0.5825	0.3752	0.0423	0.3583
cs_leaf6	8.0879	5.0275	5.1391	4.4660	0.9440	0.6292	0.0280	0.5853	0.3689	0.0458	0.3515
cs_leaf8	6.1224	4.0165	4.1703	3.7402	0.9295	0.6226	0.0234	0.5688	0.3897	0.0415	0.3321
cs_leaf9	7.1110	4.4376	4.6160	4.1591	0.9270	0.6293	0.0226	0.5685	0.3896	0.0419	0.3246
cs_leaf10	6.5924	4.2749	4.4393	3.9792	0.9355	0.6428	0.0226	0.5678	0.3925	0.0397	0.3362

注：覆盖度分析包括所有碱基覆盖信息的overall coverage（oac）和只包含胞嘧啶覆盖信息的methylation effective coverage（mec）。6至8列代表CG、CHG和CHH各context甲基化水平。9至11列代表所有甲基化胞嘧啶中各context所占比例。12列代表甲基化的胞嘧啶占所有胞嘧啶的比例。

Tip 4：优化差异甲基化区域分析

针对基因组**覆盖度较低的全基因组甲基化测序数据**(WGBS)和**覆盖不连续的简化甲基化测序数据**(RRBS)，该工作提出了**动态片段化(dynamic fragment)的策略**，实现数据的有效比较。

```
# 计算 DMR
cgmaptools intersect -1 A.CGmap.gz -2 B.CGmap.gz | cgmaptools dmr -o DMR_A_vs_B.gz
# 计算 DMS
cgmaptools intersect -1 A.CGmap.gz -2 B.CGmap.gz | cgmaptools dms -o DMS_A_vs_B.gz
```

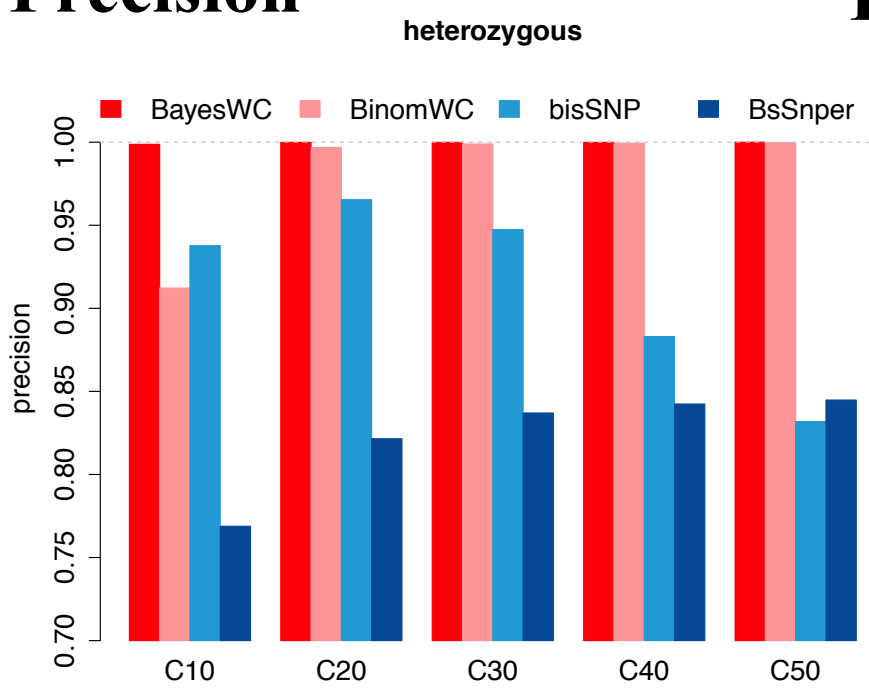


Tip 5：BS-Seq数据准确计算SNV新方法

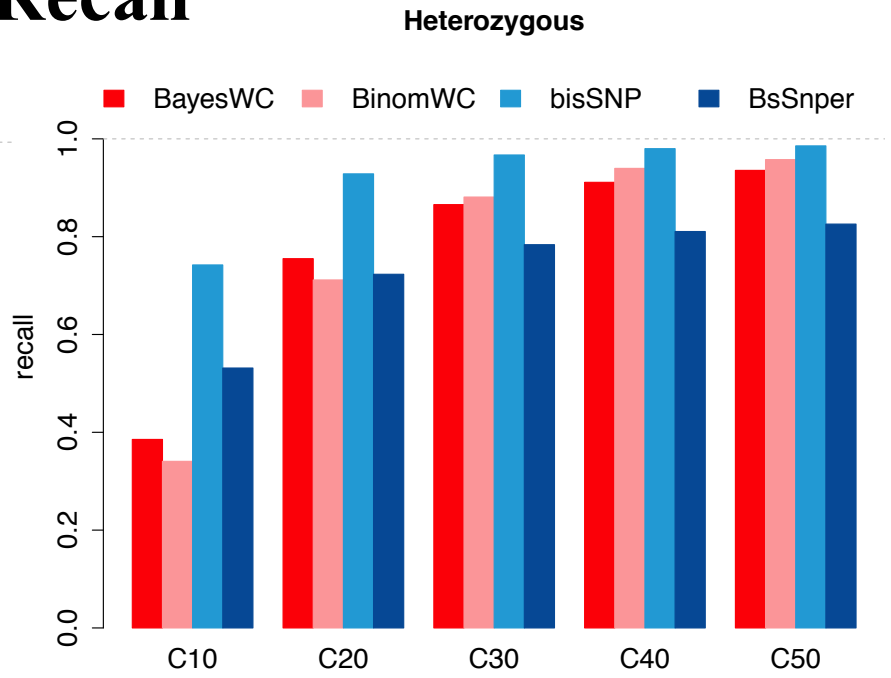
利用ATCGmap文件的信息，通过引入wildcard的基因型，结合贝叶斯模型和二项分布模型设计了BayesWC和BinomWC两种SNV calling策略，使得从DNA甲基化数据计算heterozygous SNV的precision从之前80%提高到99%。

```
# BayesWC模型
cgmaptools snv -i WG.ATCGmap.gz -m bayes -v bayes.vcf -o bayes.snv --bayes-dynamicP
# BinomWC模型
cgmaptools snv -i WG.ATCGmap.gz -m binom -o binom.snv
```

Precision



Recall



CGmapTools
预测SNV的精准度更高

Bayes + Wildcard = **BayesWC**
Binom + Wildcard = **BinomWC**

即使不能明确 genotype，依然可以判断是否为 heterozygous SNV

Wildcard symbol table for ambiguous genotypes

Ambiguous GN symbol	Possible genotypes	Hete- or Homo-zygous	sure to be SNV if reference is
Y	TT / TC / CC	not sure	A, G
R	AA / AG / GG	not sure	T, C
A,Y	AT / AC	heterozygous	A, T, C, G
C,Y	CT / CC	not sure	A, T, G
G,Y	GT / GC	heterozygous	A, T, C, G
T,Y	TT / TC	not sure	A, C, G
A,R	AA / AG	not sure	T, C, G
C,R	CA / CG	heterozygous	A, T, C, G
G,R	GA / GG	not sure	A, T, C
T,R	TA / TG	heterozygous	A, T, C, G

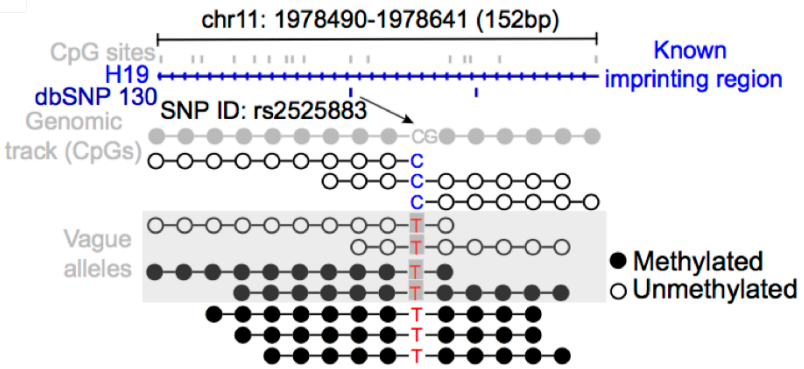
注：表格来自文献Table1，通配符定义： Y=T/C ， R=A/G。

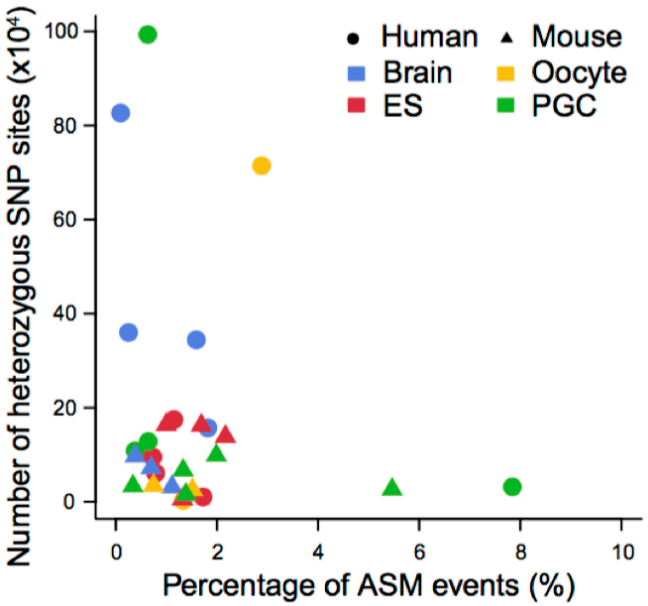
Tip 6：支持Allele特异的甲基化分析和可视化

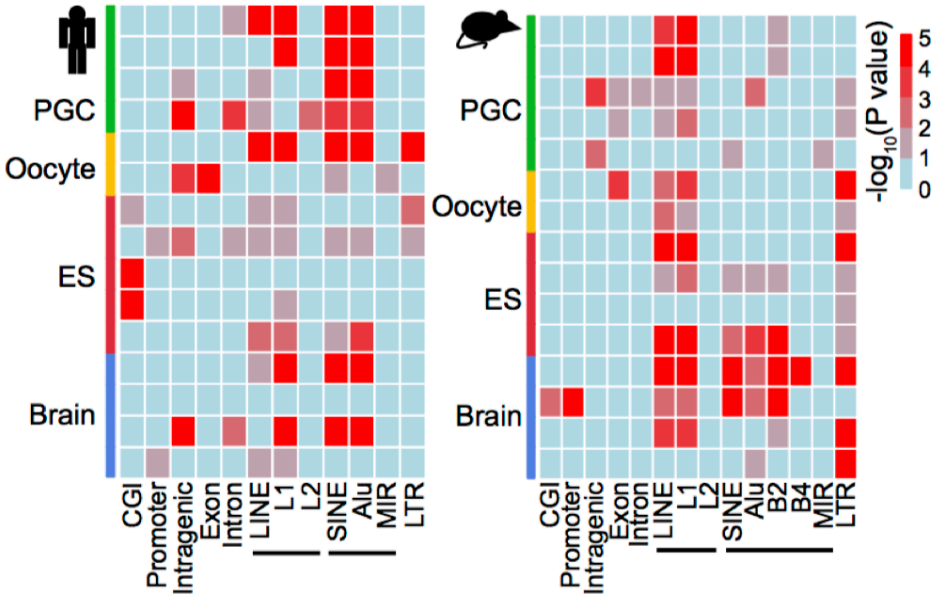
利用高精确度的杂合SNV作为输入，分析Allele-specific DNA methylation (ASM)，并通过Tanghulu(“糖葫芦”)图对ASM区域的读段上DNA甲基化的状态进行直观展示。

```
# 利用预测的杂合SNV信息计算链特异的DNA甲基化（ASM）
gawk '{if(/^#/){print}else{print "chr"$0;}}' bayes.vcf > bayes2.vcf
cgmaptools asm -r genome.fa -b WG.bam -l bayes2.vcf > WG.asm
```

左：糖葫芦图







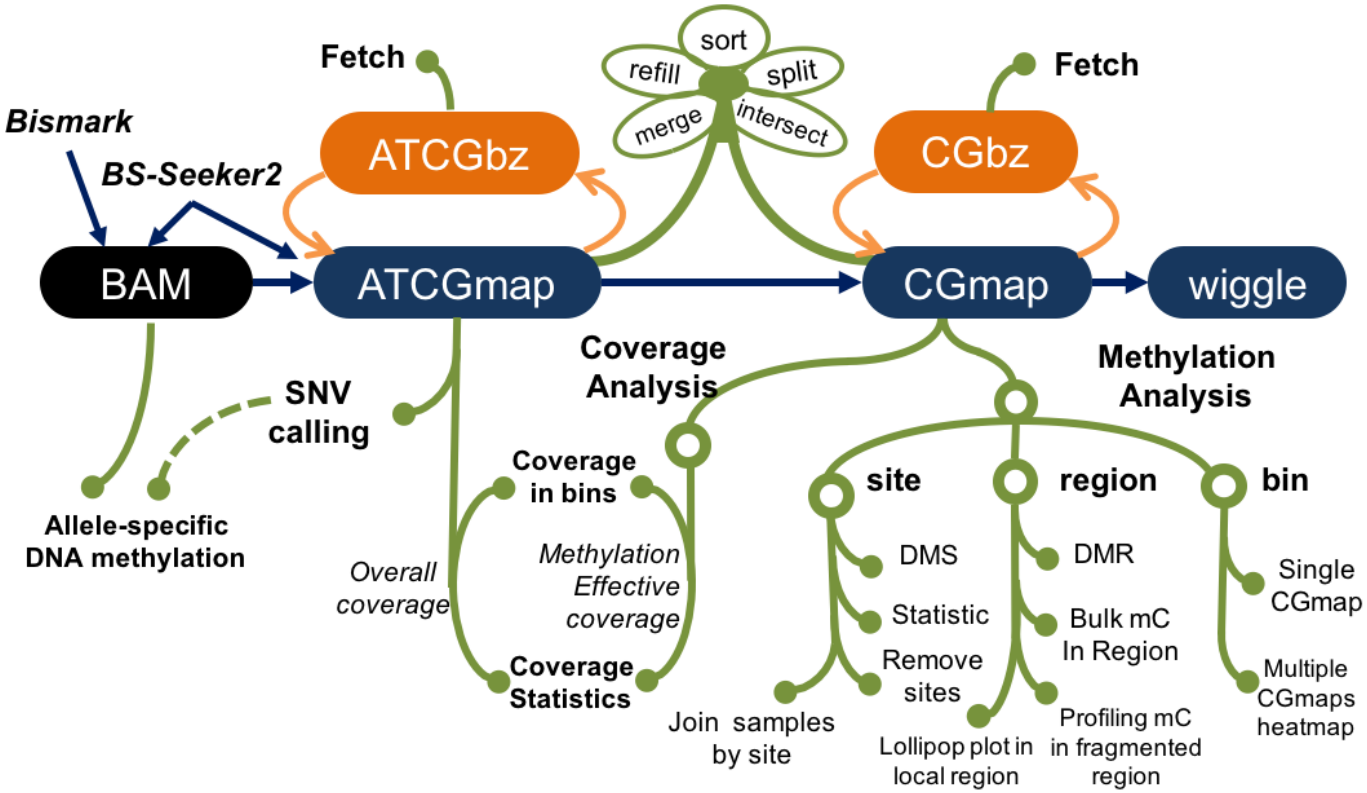
CGmapTools中ASM分析中考虑vague allele

大于1.5%的区域是ASM region，且主要富集在一些转座子上

目前软件最新版本为 v0.1.1

软件下载地址：https://github.com/guoweilong/cgmaptools

软件详细说明：https://cgmaptools.github.io/



推荐CGmapTools用作BS-Seeker2比对后分析

CGmapTools: DNA甲基化数据分析和可视化工具

显著提升BS-seq数据中计算杂合SNV的精准度，支持链特异DNA甲基化等40种分析及可视化。

1 统一的数据格式
2 命令行模式轻松支持功能扩展
3 支持快速检索
4 优化差异甲基化区域分析
5 BS-Seq数据准确计算SNV新方法
6 支持Allele特异的甲基化分析和可视化
7 提供多层次甲基化分析和可视化工具

超过1.3万行代码，40个子程序

上述内容摘自https://cgmaptools.github.io/zh/CGmapTools_zh.html，更多详细内容可登陆此网站查看