Whole Genome Bisulfite Sequencing Report

2019/8/26

目　录

● 分析结果

1 数据基本处理与质控

在项目 WGBS_MGI2000_RNAProj 中，我们对 2 个 hg19 样品进行了WGBS测序，平均每个样品产出96.430 Gb过滤后的clean bases。将下机数据进行过滤，包括去污染，去测序接头和低质量碱基比例过高的reads，得到clean data。表1中列出了数据产出的概况。图1显示的是测序碱基含量分布，图2显示的是碱基测序质量分布情况。

表1 数据基本处理与质控 （下载）

Sample ID	Fragment Length(bp)	Sequencing Strategy	Clean Reads Number	Clean Data Size(bp)	Clean Rate(%)
SampleB	30-600bp	PE100	966,801,754	96,680,175,400	89.72
SampleA	30-600bp	PE100	961,795,846	96,179,584,600	92.08

Clean Rate (%) = Clean Data Size (bp)/Raw Data Size (bp)

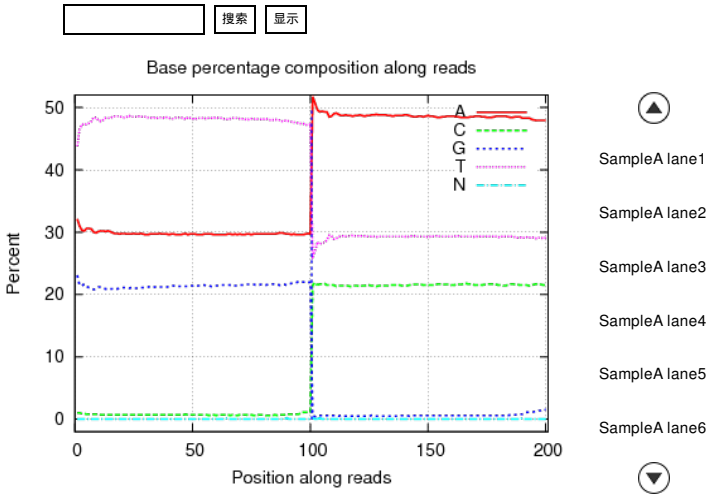


图1 Clean reads的碱基含量分布图。

横坐标表示碱基在reads上的位置，纵坐标表示碱基比例，如果图中碱基分布不平衡则说明测序过程有异常情况发生。右侧框中为样品名称，相同的样品名称出现多次是因为该样品数据来源于多个测序lane。

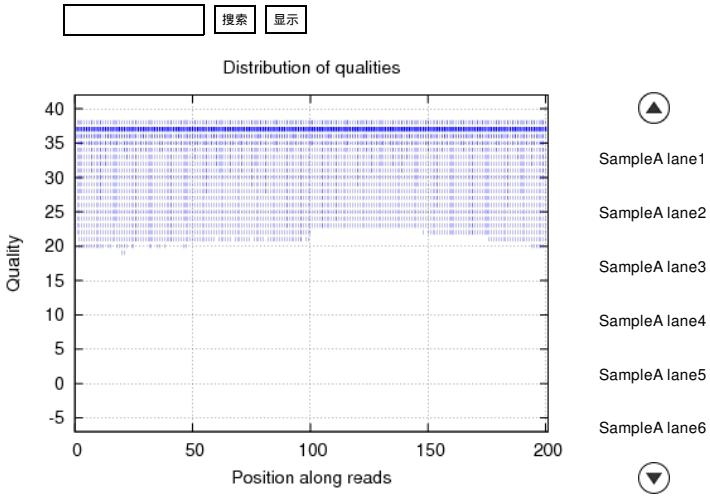


图2 Clean reads碱基质量分布图。

横坐标为reads上碱基位置；纵坐标为碱基测序质量。图中每个点表示reads中相应位置碱基的测序质量。如果低质量碱基（Q<20）的比例过多，则测序质量较差。相同的样品名称出现多次是因为该样品数据来源于多个测序lane。

在得到clean data之后，使用比对软件BSMAP[3]将reads比对到参考基因组上，比对的统计结果如表2所示；之后根据需要对各个文库的reads进行去duplication处理；然后进行质控来判断测序数据质量是否达标。

表2 比对结果统计 （下载）

Sample ID	Clean Reads	Mapped Reads	Mapping Rate (%)	Uniquely Mapped Reads	Uniquely Mapping Rate (%)	Bisulfite Conversion Rate (%)
SampleA	961,795,846	814,253,942	84.66	772,090,278	80.28	99.69
SampleB	966,801,754	824,052,387	85.23	779,747,703	80.65	99.23

Bisulfite Conversion Rate = 1 - methylation rate of Lambda DNA

下图为各样品的测序深度分布图，理论上，其最高点对应的测序深度与全基因组平均覆盖深度一致或接近，这个分布图可以用于反映测序是否均匀。

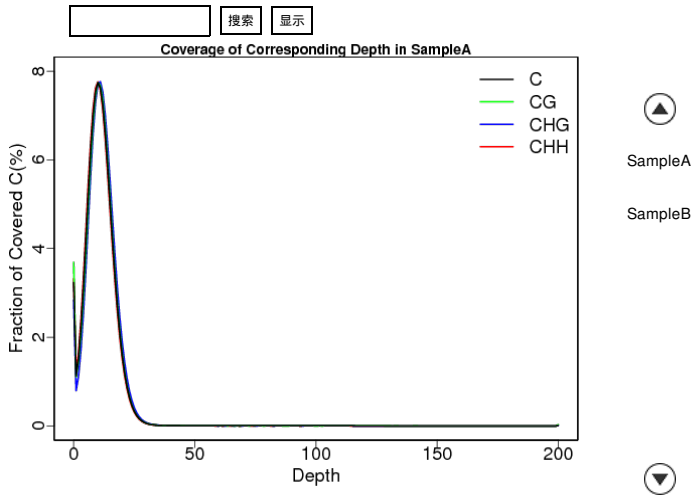


图3 测序深度分布。

X轴为测序深度，Y轴为该测序深度所占百分比。

根据胞嘧啶（C）序列特征可以将其分为三种类型CG, CHG和CHH（ H代表A或T或C碱基）[4]。下述图表中反映了不同C碱基类型有效测序深度的累积分布（基于有效数据计算）。

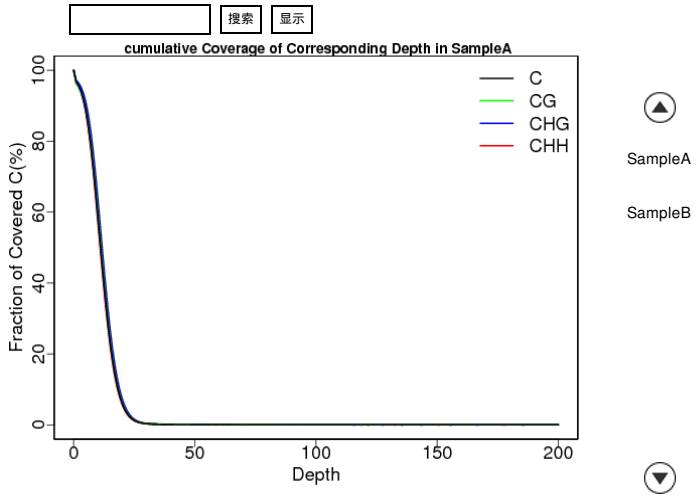


图4 C碱基测序深度的累积分布图。

横轴（x轴）表示测序深度，纵轴（y轴）表示基因组中测序深度不小于该测序深度的C碱基占全基因组全部C碱基的比例。

表3 SampleA样品在全基因组各染色体上的C位点覆盖度 （查看全部）

Chr	C (%)	CG (%)	CHG (%)	CHH (%)
chr1	97.51	97.15	97.89	97.42
chr10	97.08	96.90	97.55	96.95
chr11	98.47	98.21	98.90	98.36
chr12	98.69	98.41	99.16	98.58
chr13	98.72	98.08	99.10	98.65
chr14	98.20	97.93	98.57	98.12
chr15	95.60	95.00	95.83	95.57
chr16	96.04	95.85	96.57	95.88
chr17	97.57	97.41	97.98	97.45
chr18	98.72	98.26	99.12	98.63
chr19	97.77	97.96	98.53	97.48
chr2	97.98	97.60	98.33	97.90
chr20	98.62	98.49	99.13	98.47
chr21	98.57	98.45	99.14	98.41
chr22	96.61	96.89	97.18	96.38
chr3	98.96	98.60	99.31	98.88
chr4	98.52	97.68	98.91	98.46
chr5	97.83	97.24	98.21	97.76
chr6	98.70	98.28	99.08	98.62
chr7	97.27	96.55	97.62	97.22

表4 SampleB样品在全基因组各染色体上的C位点覆盖度 （下载）

表5 SampleA样品在全基因组各类型调控元件范围内的C位点覆盖度 （下载）

Different Genome Regions	C (%)	CG (%)	CHG (%)	CHH (%)
5-UTR	97.07	96.80	97.39	97.00
CDS	97.85	97.78	97.92	97.83
intron	97.87	97.40	98.25	97.80
3-UTR	95.72	95.33	96.03	95.65
mRNA	98.39	98.17	98.76	98.30
ncRNA	97.33	96.68	97.69	97.27
PseudoGene	79.60	79.71	80.12	79.43
transposons	96.02	95.86	96.74	95.83
tandem_repeats	55.61	75.36	77.98	53.48
CGI	95.44	95.81	95.62	95.11
CGI_shores	95.96	95.70	96.49	95.81
CGI_shelves	96.07	96.00	96.55	95.91
downstream2k	96.21	95.62	96.55	96.15
upstream2k	96.38	96.93	96.80	96.14
gene	97.76	97.34	98.11	97.69

表6 SampleB样品在全基因组各类型调控元件范围内的C位点覆盖度 （下载）

表7 各样品QC质控表 （下载）

Sample ID	Clean Reads Q20 Rate (%)	Clean Data Size (bp)	Mapping Rate (%)	Bisulfite Conversion Rate (%)	Duplication Rate (%)	Average Depth (X)	Coverage (%)
SampleA	95.90;97.47(Pass)	96,179,584,600	84.66(Pass)	99.69(Pass)	5.12(Pass)	23.41	90.401
SampleB	95.13;97.19(Pass)	96,680,175,400	85.23(Pass)	99.23(Pass)	3.64;3.23(Pass)	23.98	90.236

2 全基因组甲基化水平分析

用于分析的DNA样品为多细胞样品，因此C碱基的甲基化水平是一个0% ~ 100%范围内的数值，等于该C碱基上覆盖到的支持mC的序列数除以有效覆盖的序列总数，(详细算法请参考方法部分).通常CG甲基化存在于基因和重复序列中，在基因表达调控过程中起到非常重要的作用[1][5]。非CG类型的序列（CHG和CHH）在基因中十分少见，主要存在于基因间区和富含重复序列的区域，在沉默转座子过程中起关键作用[4]。

表8 样品 SampleA 全基因组及各染色体的平均甲基化水平 （查看全部）

Chr	C (%)	CG (%)	CHG (%)	CHH (%)
chr1	2.94	49.30	0.47	0.51
chr10	3.30	48.79	0.48	0.57
chr11	2.68	46.37	0.45	0.50
chr12	2.94	50.87	0.45	0.50
chr13	2.54	46.29	0.46	0.52
chr14	2.86	49.78	0.45	0.50
chr15	3.02	50.55	0.45	0.50
chr16	4.13	51.82	0.49	0.52
chr17	4.03	54.31	0.48	0.50
chr18	2.58	44.02	0.49	0.53
chr19	4.76	54.68	0.51	0.51
chr2	2.74	47.83	0.49	0.54
chr20	3.15	48.51	0.46	0.49
chr21	3.39	46.31	0.76	0.62
chr22	4.29	55.05	0.47	0.49
chr3	2.54	48.22	0.45	0.51
chr4	2.37	43.95	0.47	0.53
chr5	2.54	47.48	0.45	0.51
chr6	2.71	48.74	0.46	0.51
chr7	2.98	49.78	0.46	0.51

表9 样品 SampleB 全基因组及各染色体的平均甲基化水平 （下载）

表10 样品 SampleA 在全基因组各类型调控元件范围内的甲基化水平 （下载）

Different Genome Regions	C (%)	CG (%)	CHG (%)	CHH (%)
5-UTR	3.22	46.73	0.47	0.50
CDS	7.70	59.80	0.53	0.47
intron	3.29	55.76	0.46	0.51
3-UTR	3.49	55.99	0.47	0.50
mRNA	3.45	55.35	0.47	0.50
ncRNA	2.24	40.04	0.46	0.51
PseudoGene	3.81	52.81	0.55	0.53
transposons	3.18	52.23	0.49	0.54
tandem_repeats	7.93	43.06	1.24	1.94
CGI	7.10	22.48	0.84	0.66
CGI_shores	4.73	49.35	0.56	0.51
CGI_shelves	4.82	44.12	0.57	0.52
downstream2k	4.30	54.95	0.63	0.55
upstream2k	3.81	30.46	0.55	0.52
gene	3.44	55.11	0.47	0.50

表11 样品 SampleB 在全基因组各类型调控元件范围内的甲基化水平 （下载）

3 甲基化C碱基中CG, CHG 与CHH的分布比例

mCG，mCHG和mCHH三种碱基类型的构成比例在不同物种中，甚至在同一物种不同样品中都存在很大差异。因此，不同时间、空间、生理条件下的样品会表现出不同的甲基化图谱，各类型mC(mCG、mCHG和mCHH)的数目，及其在全部mC的位点中所占的比例，在一定程度上反映了特定物种的全基因组甲基化图谱的特征。mCG、mCHG和mCHH分别表示表示甲基化CG、甲基化CHG和甲基化CHH。三种碱基类型占比总和为100%，甲基化C鉴定方法依据Lister的文章描述进行[1]。

表12 mCG、mCHG和mCHH三种类型甲基化胞嘧啶的比例 （下载）

		mCG	mCHG	mCHH
SampleA	mC number	35,550,303	191,494	717,721
	proportion (%)	97.506	0.525	1.969
SampleB	mC number	34,673,596	308,025	1,015,871
	proportion (%)	96.322	0.856	2.822

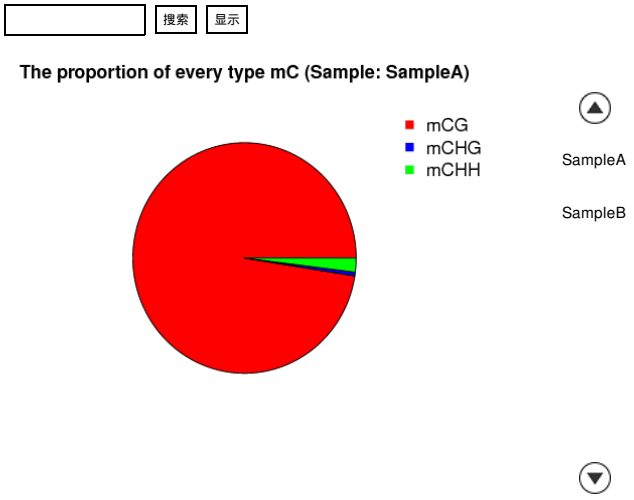


图5 不同序列类型甲基化C碱基的分布比例。

红色部分为mCG，蓝色为mCHG，绿色为mCHH。三者之和等于全基因组所有mC（100%），即构成一个整圆。

4 甲基化CG、CHG和CHH的甲基化水平分布

不同类型的C碱基(mCG、mCHG和mCHH)，其甲基化水平在不同物种间，甚至同一物种不同细胞类型不同条件下其甲基化水平都存在差异[1][6]。此图统计每种类型(CG、CHG和CHH)甲基化C的甲基化水平分布，反映了该物种DNA甲基化特征。

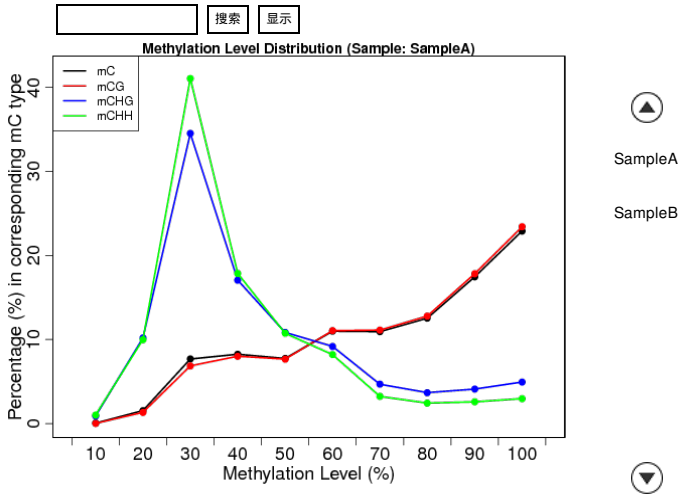


图6 甲基化C位点的甲基化水平分布。

图形横轴（x轴）表示甲基化水平，从左往右为0% ~ 100%，每10%为一档，纵轴（y轴）表示特定甲基化水平的mC在全部mC中所占比例，其中C碱基的甲基化水平等于该C碱基位点上有效覆盖序列中支持该位点为甲基化位点的序列所占的比例值。

5 甲基化的CG，CHG，CHH附近碱基的序列特征分析

在一些真核生物中，甲基化位点附近碱基的序列特征，对反映甲基化发生的序列偏向有指导意义[4]。为了研究序列特征与甲基化偏向性之间的联系，我们计算了甲基化位点上下游9个碱基 （mC位于第四个碱基）的甲基化百分比。

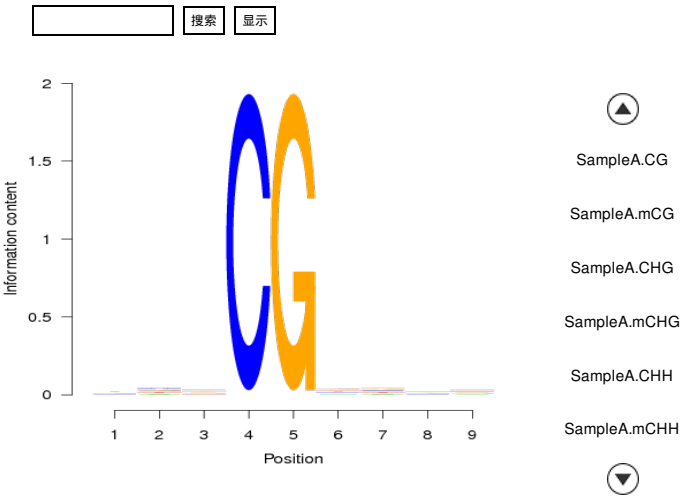


图7 C位点临近碱基的序列特征。

图形横轴（x轴）表示碱基位置，其中第四位上为用于分析的C碱基。纵轴（y轴）为熵值（0为最小值，表示四种碱基比例均匀，都为25%，2为最大值，表示四种碱基分布最不均匀，即只有一种特定碱基出现，如第四位的C与第五位的G）。

6 染色体水平的甲基化C碱基密度分布

有研究证明非CG型的甲基化与CG型甲基化C的密度有很大的差异[1]。染色体亚端粒区域DNA甲基化水平通常较高，这一现象与端粒长度以及重组有十分重要的作用，此外还有基因表达和蛋白与DNA的相互作用都有紧密联系[1]。如下图的显示，整体的甲基化水平图谱显示DNA甲基化的密度在整条染色体上变化很大。

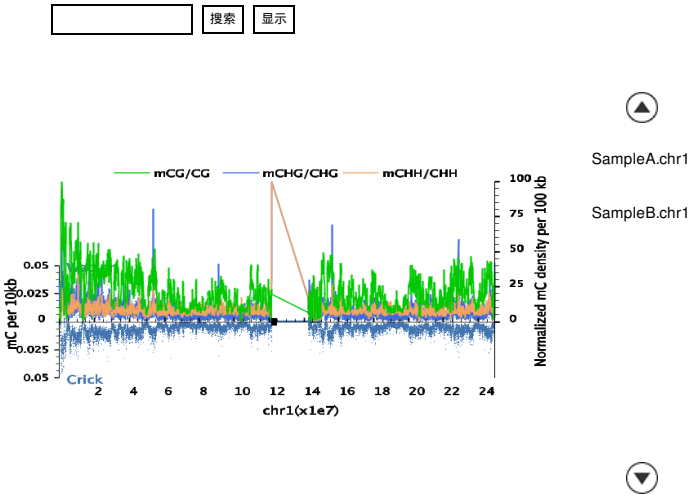


图8 各染色体上甲基化C的密度分布。

图形横轴表示 chr1 染色体，从左往右为染色体起点到终点。左边纵轴表示10kb为窗口计算得到的mC密度，以蓝点表示mC密度在染色体上的分布情况，右边纵轴表示标准化的mC比例，光滑曲线则表示不同类型甲基化C碱基(CG、CHG和CHH)的密度分布。横轴上黑色部分表示着丝粒。

7 基因组的不同区域的甲基化分布特征

基因组的不同区域具有各自不同的甲基化模式，也行使着不同的生物学功能[4]，下图中以热度图的形式表示基因组各特征区域的甲基化水平情况[7]，有助于进一步了解这些区域的甲基化特征。

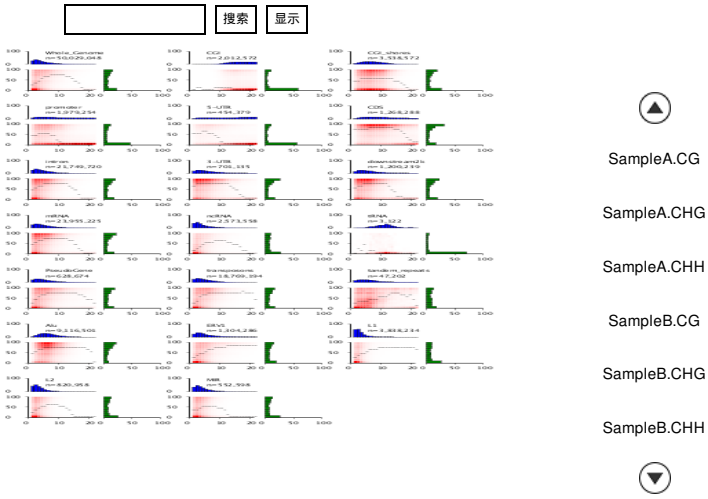


图9 基因组不同区域的甲基化分布，以及相应的CpG密度分布。

左下热度图中，CpG密度（X轴）的定义为在200bp的窗口中CpG的个数，纵轴表示CpG的平均甲基化水平。图中的黑色细线表示在特定CpG密度下，这类窗口甲基化水平的中位数值。红色区域从浅到深表示特定甲基化水平和CpG密度下的CpG个数。上方的蓝色柱状图表示CpG密度的分布,映射到横轴上，右侧的绿色柱形图表示甲基化水平的分布，并且映射到纵轴上。

8 基因组不同转录元件中的DNA平均甲基化水平

为了深入地揭示DNA甲基化与基因表达的内在联系，所有编码基因序列被分成7种不同的转录元件区域，在此基础上对不同转录元件区域的平均甲基化水平进行统计。DNA甲基化水平在不同功能区的分布特点有助于从全基因组水平去了解不同区域的DNA甲基化修饰的作用[7]。

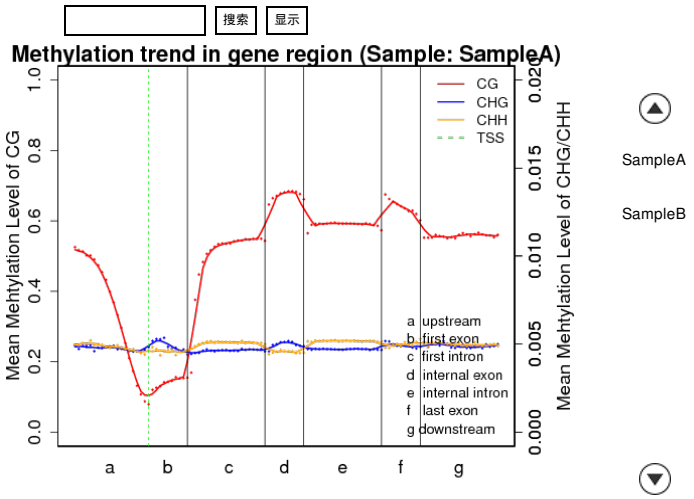


图10 全基因组不同功能元件区域的甲基化平均水平分布。

X轴为整个基因被分成的7个不同的转录元件区，每个转录元件区域的长度等分为相等数量的bin区（指包含一定数量碱基的分档区间），每一个点即为一个bin区的平均甲基化水平，各色曲线表示的是各bin区的甲基化水平的五点均值，纵轴表示甲基化水平（数值从0 ~ 1），a与b间的绿色虚线为TSS（转录起始位点）位置。

9 DMR的检测

差异甲基化区域（DMRs）是指不同样品中基因组表现出不同的甲基化模式的某些DNA片段。 DMR与遗传印记相关，在个体中表现为与父本或母本的甲基化状态一致。甲基化的等位基因经常表现为沉默状态。亲本与子代甲基化模式的差异常常导致表观遗传缺陷[7]，而人工繁殖技术可能会导致异常甲基化的比例升高，并导致疾病的发生。

我们用滑动窗口的方法检测DMR。在两个样品基因组相同位置上寻找包含至少5个CG（或CHG，或CHH）的窗口，比较该窗口在两个样品数据中甲基化水平的差异，寻找在两个样品中甲基化水平有差异的区域[2]。

表13 SampleA_vs_SampleB间DMRs (CG pattern)的统计结果 （查看全部）

#Chr	DMR number	DMR length
chr1	1,189	392,566
chr10	830	265,270
chr11	701	223,723
chr12	680	226,364
chr13	383	120,171
chr14	403	126,185
chr15	455	144,509
chr16	633	195,840
chr17	758	247,446
chr18	389	120,218
chr19	660	201,770
chr2	1,155	376,047
chr20	494	160,412
chr21	246	75,434
chr22	399	126,916
chr3	769	254,726
chr4	740	237,070
chr5	725	226,626
chr6	809	257,599
chr7	796	245,801

表14 SampleA_vs_SampleB间DMRs (CHG pattern)的统计结果 （下载）

表15 SampleA_vs_SampleB间DMRs (CHH pattern)的统计结果 （下载）

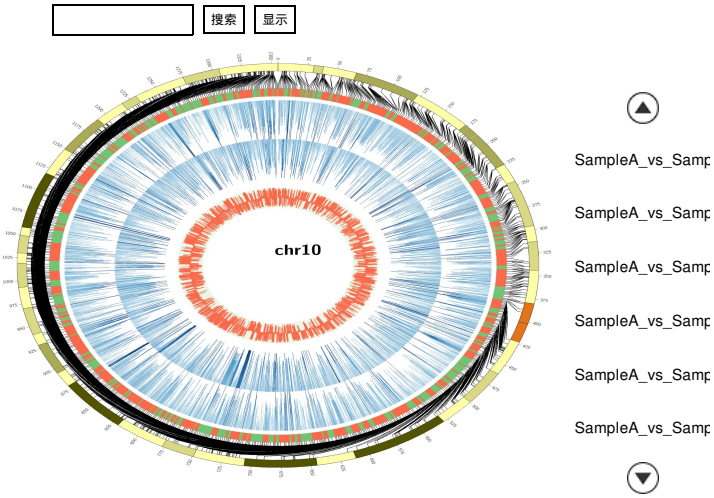


图11 DMRs中的甲基化水平分析。

最外圈表示基因组染色体的位置；第二圈为DMR区域，红色区域代表样品1的甲基化水平高于样品2，绿色区域表示样品1的甲基化水平低于样品2；第三圈表示样品1各个位点的甲基化率；第四圈表示样品2各个位点的甲基化率；第五圈表示甲基化率的差异程度。

10 DMR相关基因的GO和Pathway分析

基因本体论（Gene ontology，GO）是所有物种中最主要的了解基因和基因产物属性的生物信息学分析手段，GO分析能够用于鉴定基因产物的性能，它包含了三类基因功能信息：细胞组分（Cellular Component），

分子功能（Molecular Function）和生物学过程（Biological Process）。为了探讨表观遗传变异在通路和生物学过程中起到的作用，我们对DMR相关的基因进行了GO和Pathway分析。

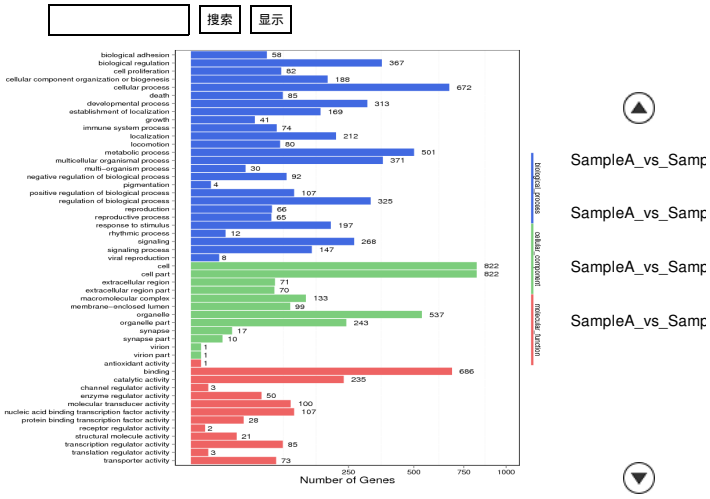


图12 DMR相关基因的GO聚类分析。

图形横轴表示DMR相关基因的数量，纵轴表示各种GO term，所有GO term分位三类，蓝色为生物学过程，绿色为细胞组分，红色为分子功能。

KEGG (Kyoto Encyclopedia of Genes and Genomes)是有关Pathway的主要公共数据库，该数据库整合了基因组、化学以及系统功能信息，特别是测序得到的基因集与细胞、生物体以及生态环境的系统性功能相关联。所有样品中的DMR相关基因均用KEGG数据库进行分析。

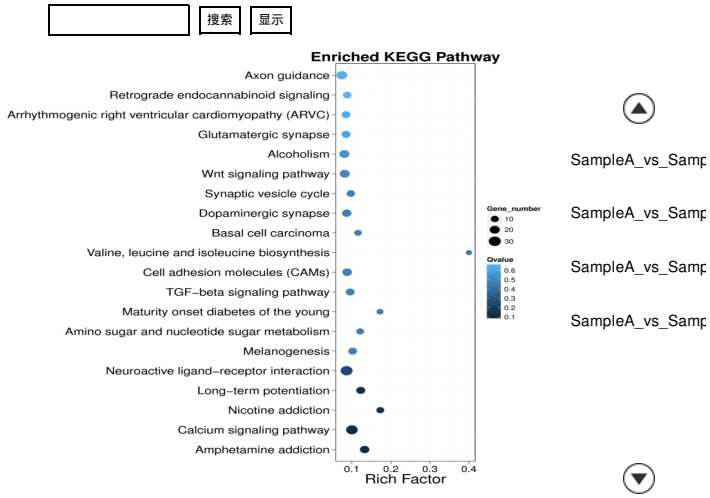


图13 DMR相关基因的Pathway功能显著性富集分析。

横轴（rich factor）是在通路中DMR相关基因占所有基因总数的比例，Rich-Factor值越高则在该pathway中越富集。Q为修正的p value，其值为0-1，Q值越小，富集度越高。在此图中仅显示前20个富集通路。

● 分析方法

1 实验流程

1）基因组DNA用Bioruptor (Diagenode, Belgium) 打断成平均大小为250 bp的片段；

2）DNA片段末端修复、3’端加A碱基，连接甲基化接头；

3)采用EZ DNA Methylation-Gold kit（ZYMO）进行Bisulfite处理；

4)2%的琼脂糖凝胶电泳，片段选择；

5)用QIAquick Gel Extraction kit (Qiagen)回收DNA片段；

6)PCR扩增，合格的文库上机测序。

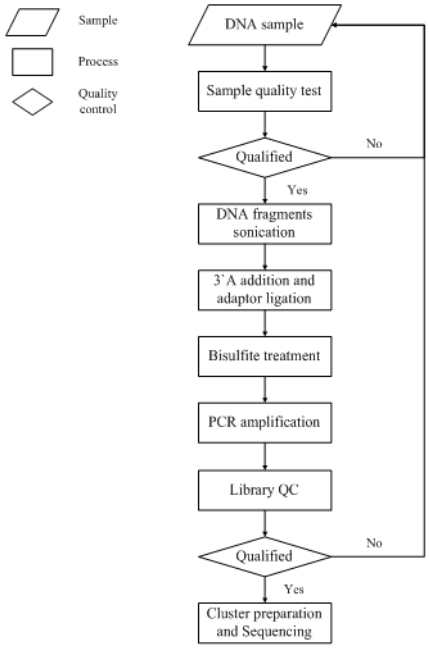


图1 BS实验流程示意图。

得到DNA样品后，首先对样品进行质量检测。在确定样品质量合格后，使用样品进行BS文库构建工作。文库构建完成后需对文库进行质量检测。质检合格的文库将用于上机测序。

2 信息分析流程

测序所得的下机数根据客户需求进行各项后续分析。信息分析流程如图2：

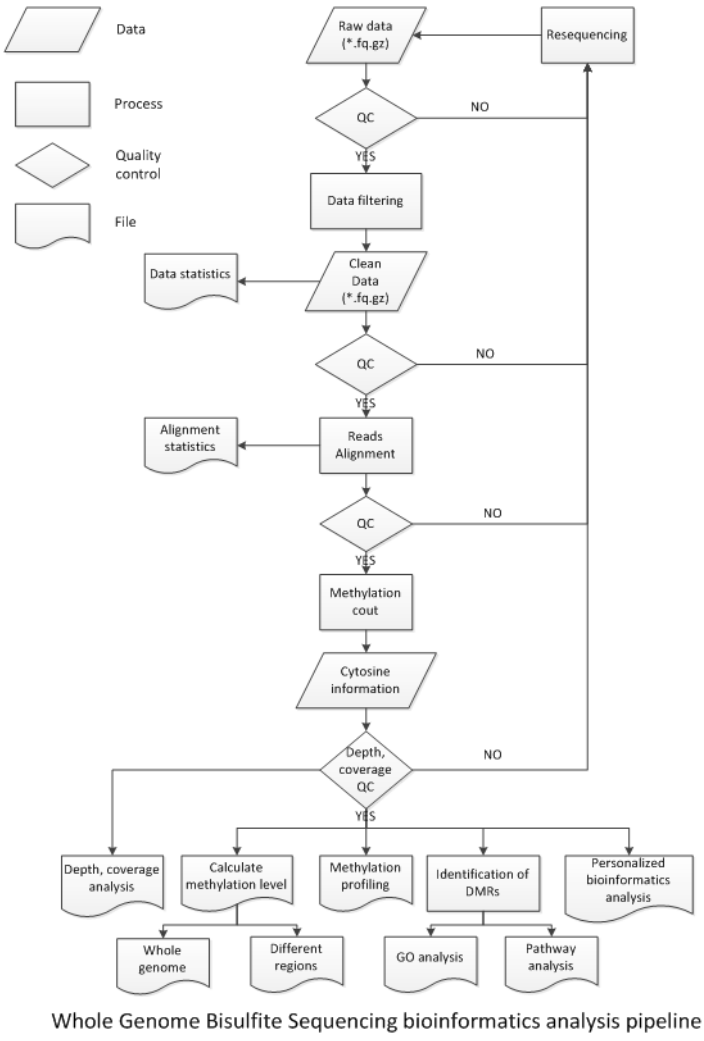


图2 BS信息分析流程示意图。

得到下机数据后，首先进行数据过滤，去掉低质量数据，得到可用数据。完成数据过滤后，需检测可用数据量是否符合合同要求。检测合格后，将可用数据与参考基因组进行比对，得到比对结果。在确认比对质量合格后，使用唯一比对数据计算得到全基因组C碱基甲基化信息，进行信息分析处理，得到标准信息分析结果和个性化分析结果。

3 Data Filtering

数据过滤包括去 、污染以及低质量序列。数据过滤分析使用华自主的分析软件，低质量的reads包括以下两类，符合任意一条的都会被剔除：

1) N > 10%；

2) 质量值小于20的碱基>10%。

完成过滤后的reads称为clean reads，这些数据存储为FASTQ格式（参见帮助页中的FASTQ格式）。

4 序列比对

过滤完成后，clean data与参考基因组进行比对（BSMAP），并计算每个样品的比对率和bisulfite转化率等统计信息。

序列比对过程

BSMAP parameters for PE reads: bsmap -a filename_1.clean.fq.gz -b filename_2.clean.fq.gz -o filename.sam -d ref.fa -u -v 9 -z 33 -p 8 -n 0 -w 20 -s 16 -f 10 -L 100
If use Hiseq2000, -z should be set 64;
samtools parameters:
samtools view -S -b -o filename.bam filename.sam
samtools sort -m 2000000000 filename.bam filename.sort
samtools index filename.sort.bam

5 甲基化水平

甲基化水平是支持甲基化的reads数占所有覆盖该位点的reads数的比例[1]。计算公式如下：

$$Rm_{average} = \frac{\mathrm{Nm_{all}}}{\mathrm{Nm_{all}+Nnm_{all}}} * 100\%$$

Nm为支持该位点是甲基化C的reads数，Nnm为支持该位点是非甲基化C的reads数。

6 DMR检测

在两个样品基因组相同位置上寻找包含至少5个CG（CHG或CHH）的窗口，比较该窗口在两个样品数据中CG甲基化水平的差异，寻找在两个样品中甲基化有显著差异（2倍差异，且fisher检验P value <= 0.05）的区域即为DMR。如果两个相邻的DMR形成的连续区域在两个样品中甲基化水平有明显差异，则这两个DMR将被合并为一个连续的DMR，否则为两个独立的DMR。

7 甲基化水平程度差异

我们用CIRCOS比较样品间DMR的甲基化水平差异来计算两个样品之间甲基化程度的差异，两样品间某位点的甲基化水平的差异程度可以用下面的公式来计算：

$$\text{degree of difference} = \frac{\log_2 \mathrm{Rm1}}{\log_2 \mathrm{Rm2}}$$

Rm1、Rm2分别代表样品1和样品2的mC的甲基化水平。如果Rm1或Rm2的值为0则用0.001 代替[2]。

8 GO注释

GO（Gene Ontology，基因本体论）数据库是目前对基因功能分析的一个重要工具，GO富集分析提供所有在DMR相关基因中有明显富集的GO term，并过滤特定生物学功能的DMR相关基因。这个方法主要是基于GO TermFinder（http://www.yeastgenome.org/help/analyze/go-term-finder），首先将DMR相关基因比对到GO term的数据库中（http://www.geneontology.org/），计算每个term的基因数量，然后应用超几何检验，找出与整个基因组背景相比，在DMR相关基因中显著性富集的GO term。我们研发了十分严格的分析方法，主要计算方法如下：

$$\mathrm{P} = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N为GO注释的所有基因数，n为所有基因中与DMR相关的基因数，M为注释的某特定GO term的所有基因数，m是该特定GO term中与DMR相关的基因数。算出的p值 通过Bonferroni检验，阈值设定为p≤0.05。满足这

些条件的GO term则为显著富集。 该分析可以识别DMR相关基因行使的主要的生物学功能。

9 KEGG通路富集

在生物体内，由于不同基因通过翻译、表达、调控、相互协调使其发挥特定的生物学功能，基于 Pathway的分析有助于更进一步了解某些基因所参与的代谢通路。KEGG[9]是有关Pathway的主要公共数据库，Pathway显著性富集分析以KEGG Pathway为单位找出与整个基因组背景相比在DMR相关基因中显著性富集的Pathway。主要的计算方法与GO分析相同。

● 帮助

1 华大科技在线知识库

更多帮助信息，请查阅：华大科技在线知识库 http://www.techdoc.bgi.com/#/share/82c7d17e-8c23-4ef8-b846-cb5e0c90de02/2af248f6-443d-4794-bf24-f7cba575489b/

2 FASTQ 格式说明

对于Illumina测序，通过base calling将原始图像数据转化成序列信息，以FASTQ格式存储。FASTQ文件为用户得到的最原始文件，里面存储reads的序列以及reads的测序质量。在fastq格式文件中每个read由四行描述：

```
@A80GVTABXX:4:1:2587:1979#ACAGTGAT/1
NTTTGATATGTGTGAGGACGTCTGCAGCGTCACCTTTATCGGCCATGGT
+
BMMTKZXUUUddddddddddddddddddddddddddddddddddddd^WYYU
```

第一行与第三行为测序仪生成的序列名称，第二行为碱基序列，第四行为质量值。质量值中的每个字母代表第二行对应位置的碱基质量值，具体的质量值等于字符的ASCII码值减去64或33。HiSeq2000通常-64，HiSeq4000通常-33，此处以HiSeq2000 为例，如c字符的ASCII码值等于99，代表碱基质量值为35。质量值的范围通常为2到41，根据仪器的具体型号会有一些区别。表1展示了测序错误率与测序质量的对应关系。若以E表示测序错误率，Q表示碱基质量值，以下公式描述了他们之间的关系。

$$SQ = -10 \times (\log \frac{E}{1-E}) / (\log 10)$$

$$E = \frac{Y}{1+Y}$$

$$Y = \frac{SQ}{e^{-10 \times log\,10}}$$

表1 测序错误率与测序质量的对应关系 （下载）

Sequencing Error Rate(%)	Sequencing Quality Value	Character
1.00	20	5
0.10	30	?
0.01	40	I

FASTQ格式的详细说明请参见 http://en.wikipedia.org/wiki/FASTQ_format.

3 BAM 格式说明

将测序序列比对到参考基因组上得到的比对结果以bam文件进行存储，BAM实际上是SAM（Sequencing Alignment/Map）的二进制格式，SAM是一种文本文件格式，各列说明如表2所示。其中FLAG列的每一个bit的

意义如表3所示。samtools可以进行SAM和BAM格式的相互转换，也支持进行其他许多复杂的操作，比如变异检测、比对结果查看、排序、索引以及提取数据等。关于samtools的更多资料请见 http://samtools.sourceforge.net/samtools.shtml#5

表2 SAM格式各列表头描述 （下载）

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

表3 SAM文件中FLAG一列的定义 （下载）

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

4 Cout 格式说明

通过对比对结果的计算可以确定甲基化胞嘧啶的位置并将它们的信息储存在cout格式文件中，它包括了如下图所示的9列信息，

chr1	10004	+	CHH	AACCC	1	0	0	1
chr1	10005	+	CHH	ACCCT	1	0	0	1
chr1	10006	+	CHH	CCCTA	1	0	0	1
chr1	10010	+	CHH	AACCC	1	0	0	1
chr1	10011	+	CHH	ACCCT	1	0	0	1
chr1	10012	+	CHH	CCCTA	1	0	0	1
chr1	10016	+	CHH	AACCC	1	0	0	1
chr1	10017	+	CHH	ACCCT	1	0	0	1
chr1	10018	+	CHH	CCCTA	1	0	0	1
chr1	10022	+	CHH	AACCC	1	0	0	1

表4 cout格式文件每列的定义 （下载）

Col	Description
1	Chromosome ID
2	Position
3	+/- chain
4	cytosine pattern (CG/CHH/CHG)
5	cytosine sequence context
6	copy number : the same value as rep_rate from qmap file colum 11
7	methyl-reads num
8	non-methyl-reads num
9	Binomial distribution expected values

5 dmr格式说明

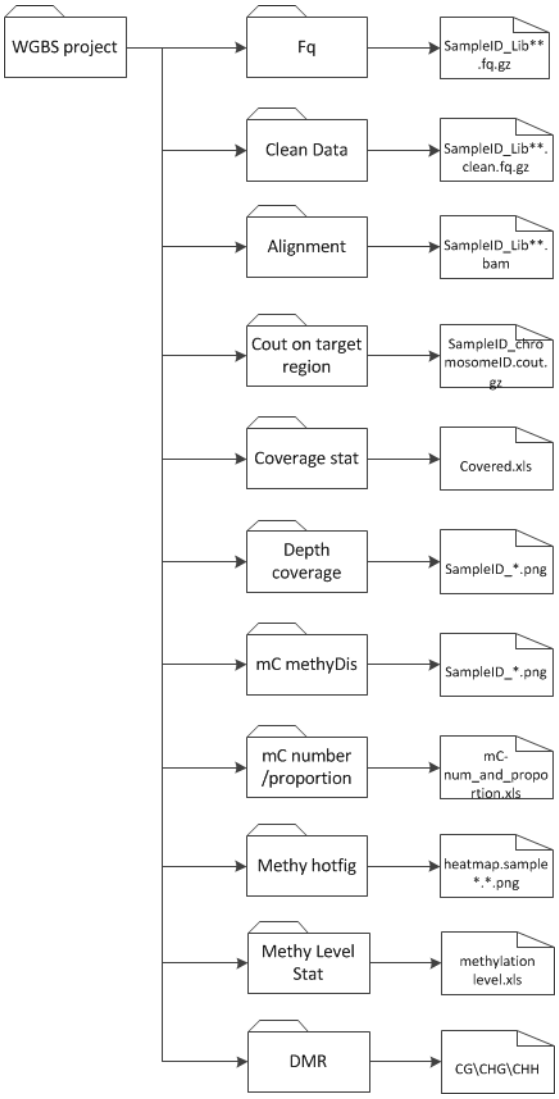
我们在两个样品基因组相同位置上寻找包含至少5个CG的窗口，比较该窗口在两个样品数据中CG甲基化水平的差异，寻找在两个样品中甲基化有差异的区域。计算出来的DMRs信息存储在dmr格式文件中，它包含了如下图所示的15列信息，

```
chr1    870425  870703  5.2918303E-18   0.13    0.35    0.01    0.03    582     557     2471.89 2047.98 15
chr1    906363  906603  3.0920712E-21   0.44    0.19    0.07    0.06    628     644     3022.12 3446.21 15
chr1    906395  906603  1.4762550E-20   0.44    0.19    0.08    0.07    603     627     3359.92 3792.53 13
chr1    933589  933836  1.1518100E-31   0.12    0.33    0.02    0.02    1013    1240    1270.18 2632.92 25
chr1    1072469 1072879 1.3948890E-43   0.30    0.14    0.02    0.01    2413    3077    2400.64 3419.37 35
chr1    1093359 1093603 1.5177954E-09   0.25    0.02    0.02    0.01    153     146     63.92   100.11  14
chr1    1108890 1109192 1.6596868E-09   0.13    0.26    0.03    0.03    598     628     3892.33 5174.61 12
chr1    1109375 1109694 4.0971980E-08   0.12    0.34    0.02    0.08    227     195     112.80  78.22   22
chr1    1120692 1121046 7.0695243E-13   0.39    0.81    0.08    0.06    138     145     86.82   108.17  18
chr1    1182031 1182304 6.6505283E-16   0.26    0.08    0.04    0.00    504     613     315.82  701.63  30
```

表5 dmr格式文件每一列的定义 （下载）

Col	Description
1	Chromosome ID
2	Starting position
3	End position
4	p-value(Fisher exact test)
5	All CpG average methylation rate for sample1
6	All CpG average methylation rate for the sample2
7	var-methy1
8	var-methy2
9	Depth for the for sample1
10	Depth for the for sample2
11	var-depth1
12	var-depth2
13	Numbers of CpG
14	Uncovered CpG number in sample 1
15.	Uncovered CpG number in sample 2

6 FTP 文件结构说明



7 联系我们

如果您有任何问题，请第一时间联系我们，我们将为您提供优质的服务。

服务热线:400-706-6615

售后服务:customer@genomics.cn

技术支持:tech@genomics.cn

投诉专线:010-80481175(Beijing) 0755-25273291(Shenzhen)

● 常见问题

Bisulfite-Seq在项目开始之前需要考虑哪些因素？

是否为低甲基化率的物种；该物种的基因组完成情况如何（影响BS-SEQ的比对）；基因组是否存在复杂因素：GC含量偏高、杂合度偏高、转座子、重复区域等。

可以对无参考基因组的物种进行Bisulfite研究吗？

Bisulfite-Seq强烈依赖基因组的完成程度，基因质量的好坏直接影响后续的分析结果，因此更适合有完整基因组信息的物种。

Bisulfite-Seq的数据量？

Bisulfite-Seq需要计算每一个C碱基的甲基化率，因此必须保证一定的数据量，以得到相对精确的分析结果。建议为20X-40X基因组大小，对于疾病研究的人或者小鼠，我们建议90G的数据；若平均甲基化水平很低或者基因组结构复杂的的物种，考虑增加测序深度或是增加一个生物学重复的样本，可参考家蚕的研究思路；对于微生物的甲基化研究，建议先测1G的数据。

Bisulfite-Seq建库需要多长时间？

建库需要4个工作日。

Bisulfite-Seq有什么特殊接头？

Bisulfite的甲基化接头中的所有C都是甲基化的，避免在后续的处理中被改变。

Bisulfite-Seq微量建库样品量最低是多少？

100 ng的起始样本可以得到1个文库，每个文库可以得到7G (50PE)的原始数据。根据BGI的研发初步结果来看，100ng建库的数据比对结果，与常规建库的比对结果相关性很好。

Bisulfite的转化率？

一般可达到99%以上；如果样品的DNA不存在不发生甲基化的DNA作为对照（例如：拟南芥的叶绿体的DNA都是不发生甲基化），都会在样品中混入control DNA来验证Bisulfite的转化率。

Bisulfite-Seq中如何定义支持甲基化的reads？

Reads是否支持甲基化是针对某一个位点而言的，在这个位点上，若read1（正链），C与C比对上（即：甲基化的C，没有发生转化），若read2（负链），G与G比对上，那么这个read在这个位点上就是支持甲基化的read，反之就是支持非甲基化的read。

Bisulfite-Seq中如何定义C是一个甲基化的C？

如果该C位点支持甲基化的reads数大于等于期望值，则判断为甲基化的C。

Bisulfite-Seq是否可以统计甲基化频率？

可以，但是并不包含在标准信息分析流程中。甲基化频率，就是指相隔多少个碱基（ATCG），会出现一个甲基化的C，模式生物拟蓝芥和人中都做过相关统计（Nature）。

● 参考文献

[1] Lister R., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271):315-22

[2] Mortazavi, A., et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5(7): 621-8.

[3] Xi Y. , et al.(2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics,10:232.

[4] Xiang, H. ,et al.(2010). Doi, A., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. BMC Bioinformatics , 41(12): 1350-1353.

[5] Lister, R., et al. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell, 133(3), 523-536.

[6] Xiang H., et al. (2010). Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nature Biotechnology, 28(5):516-20.

[7] Li Y., et al. (2010). The DNA methylome of human peripheral blood mononuclear cells.PLoS Biol. 8(11).

Technical Support E-mail:info@bgitechsolutions.com
Website: www.bgitechsolutions.com

华大基因
BGI

基因科技造福人类

联系我们

服务热线：400-706-6615　　网址：www.bgitechsolutions.com

邮　箱：info@bgitechsolutions.com　　地址：广东省深圳市盐田区洪安三街华大综合园7栋（邮编：518083）

2017年01月