

**研究生学位论文开题报告**

|  |  |
| --- | --- |
| 研 究 生： | **周强伟** |
| 学 号： | **2013304110186** |
| 学 院： | **信息学院** |
| 专 业： | **生物信息学** |
| 学位级别： | **硕 士** |
| 学位类型： | **学 硕** |
| 研究方向： | **表观遗传学** |
| 课题名称： | **DNA甲基化处理流程构建** |
| 导 师： | **李国亮 教授** |
| 指导小组： |  |

华中农业大学研究生处制

2015年04月

摘要

近年来表观修饰受到人们的广泛关注，而DNA甲基化作为表观修饰中的重要组成之一，其发生机制以及作用机理成为当前研究的热点。目前已有研究表明DNA甲基化修饰能够在染色体构象、转录调控、组织分化、以及癌症等方面发挥重要作用。因此研究在不同状态下DNA甲基化的发生以及变化情况对于了解甲基化的发生过程以及作用原理具有重要意义，对一些疾病研究预防起到一定参考价值，研究DNA甲基化状态尤其是单碱基甲基化水平分布情况对于DNA甲基化的发生原理作用机制具有重要意义。而目前检测DNA甲基化的主流技术为亚硫酸测序技术，亚硫酸测序能够将未甲基化胞嘧啶转换为胸腺嘧啶因此来检测发生甲基化位点，但是目前针对该技术缺少系统简便的处理流程，因此本课题旨在针对重亚硫酸盐测序技术开发一套系统完整的处理DNA甲基化数据的软件包，该软件包的开发能够为当前DNA甲基化分析提供一个完整简便的工具，对DNA甲基化的研究起到一定的辅助作用。

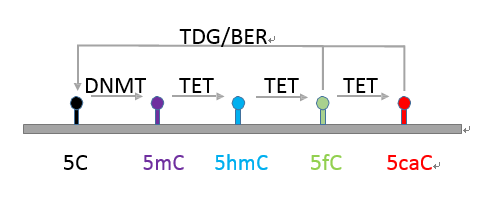
**一、实验背景**

**1.1 DNA甲基化的定义和不同形式**

DNA甲基化（DNA methylation）是最早发现的DNA修饰途径之一，是在DNA甲基化转移酶(DNA methyltransferase，DNMT)作用下，以s-腺苷甲硫氨酸(SAM)为甲基供体将甲基基团合成到胞嘧啶的第五位碳原子上形成5’-甲基胞嘧啶。早在1948年，R D Hotchkiss就报道了DNA的cytosine的5'位可以被甲基修饰 [1] ，中间对DNA甲基化的研究经历了漫长的时间，直到2000年人类表观基因组协会（Human Epigenome Consortium，HEC）启动了完成人类6号染色体完成DNA甲基化图谱计划，该计划完成后，继而在2003年提出了人类表观基因组（HEP）计划，在此之后大量对DNA甲基化修饰的研究浮出水面，一时之间DNA甲基化成为新的研究热点。迄今DNA甲基化的作用功能已经有了一定的研究基础，DNA 甲基化作为表观修饰的一种重要存在形式，在异染色质 [2] 、转录沉默 [3] 、基因表达、组织分化 [4] 、神经紊乱 [5] 、胚胎发育 [6-7]、以及对于肿瘤 [8] 发生有重要调节作用。因此研究细胞染色体DNA甲基化及其规律对理解生物功能及其调控尤为重要。

DNA甲基化的发生以及维持是一个非常复杂的过程，现在已知DNA甲基化的维持需要靠DNA甲基化以及DNA去甲基化共同起作用来维持一个平衡态，最初人们只认识到甲基化的存在，直到后来DNA甲基化在受精卵中的擦除再生现象使人们认识到DNA去甲基化的存在。而对于过程催化发生的酶一直是研究的热点，DNA甲基化酶DNMT家族的发现是由哥伦比亚大学医学院的Mary Grace Goll and Timothy H. Bestor 2005年发现 [9]。相对而言DNA去甲基化酶的发现经历了长时间的探索，DNA去甲基化酶的研究最早是在1982年，Martin实验团队在小鼠中发现一种蛋白酶敏感物具有DNA去甲基化活性 [10]。1999年Moshe Szy教授实验室鉴定到了DNA去甲基化酶可以在体外将5mC催化为胞嘧啶C [11]，但该结果难以实现重复。期间对于DNA去甲基化酶的研究一直在继续，直到2009年DNA去甲基化的研究有了大的进展，美国科学院院士Anjana Rao实验室发现Tet1酶可以将5-甲基胞嘧啶氧化成5-羟甲基胞嘧啶 [12]，而5-羟甲基胞嘧啶稳定出现在胚胎干细胞中，因此当时推测5mC可能经过氧化为5hmC进而恢复为未甲基化的胞嘧啶。然而5hmC又是如何转变的呢，这在当时成为每个表观实验室研究的热点，到2010年英国剑桥大学的Azim Surani等人在nature上提出推测可能存在一种糖苷酶参与DNA去甲基化过程 [13]，但是当时该实验室认为该酶直接作用在mC或者hmC，在此期间徐国良实验室检测到5caC甲基化的存在，该实验室针对是否去甲基化是经过5caC过程，最终在2011年在Science上与北卡莱罗纳大学的张毅实验室同时提出TDG酶可以转化5caC [14,15]，并且张毅实验室同时检测到了5fC的存在，到此位置DNA去甲基化的过程才有了一个清楚的认识。

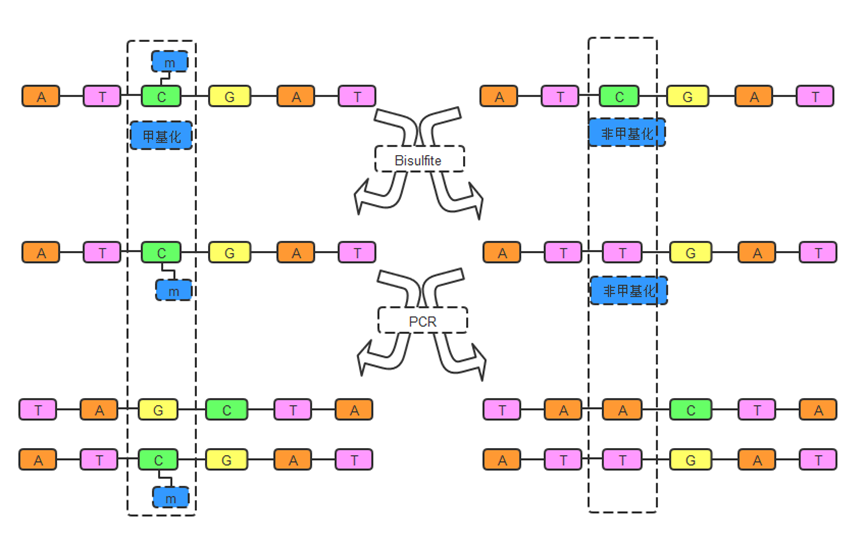
DNA甲基化以及去甲基化的过程可以总结为未甲基化胞嘧啶经DNMT酶修饰变成5mC，5mC在Tet家族双加氧酶（dioxygenases）作用下转化成另外一种修饰形式：5-羟甲基胞嘧啶（5-hydroxymethylcytosine，5hmC），进一步在TET酶作用下氧化成5- 胞嘧啶甲酰(5-formylcytosine,5fC)、5-胞嘧啶羧基（5-carboxylcytosine，5caC）[17]（如图一）。DNA甲基化在体内存在形式有5mC、5hmC [16] 、5fC [17]以及5caC [17] 。



图一：胞嘧啶在不同酶的作用下被转化为不同修饰形式

**1.2 DNA甲基化检测**

DNA甲基化检测经历了长期的发展过程，早些时候是高效液相色谱柱 [18]以及在非特异性核酸酶的基础上针对染色体易靠近区域位点检测 [19]，其中最为常用的为MNase [20]与DNaseI [21]。随着技术的发展，目前主流的DNA检测方法为亚硫酸氢盐转化测序（BS-Seq）[22]，具体流程可参见图二。亚硫酸氢盐测序技术核心在于通过亚硫酸氢盐处理将未发生甲基化的胞嘧啶转化为胸腺嘧啶，而甲基化的胞嘧啶由于甲基化保护作用并未发生变化，之后通过PCR扩增以及测序过程完成对全基因组单位点的DNA甲基化情况监测。全基因组亚硫酸氢盐测序 WGBS(whole genome bisulfite sequencing)能够在单碱基分辨率下显现碱基位点甲基化水平，构建全基因组甲基化图谱，对于研究DNA甲基化发生机理，以及不同状态下甲基化水平变化具有重要意义。



图二：亚硫酸氢测序流程图

**1.2.1 基于WGBS技术对不同甲基化形态的检测技术**

正如前面介绍DNA甲基化存在不同的形态。对于不同的DNA甲基化形态，分别有相应的技术来进行检测（如图三所示），如下具体阐述。

5hmC的检测方法为TAB-Seq(TET-assisted bisulfite seqencing) [23]技术，是2012年由加州大学的任兵(Bing Ren)教授和芝加哥大学的何川(Chuan He)教授共同合作发表在nature protacols。该技术的原理如下TET蛋白为双加氧酶，能将5mC 氧化为5hmC进一步氧化为5fC、5caC。而5hmC由于在最先加入了β-GT的保护，能够防止TET蛋白的氧化作用，因此经亚硫酸氢钠转化处理能够直接检测5hmC的存在。

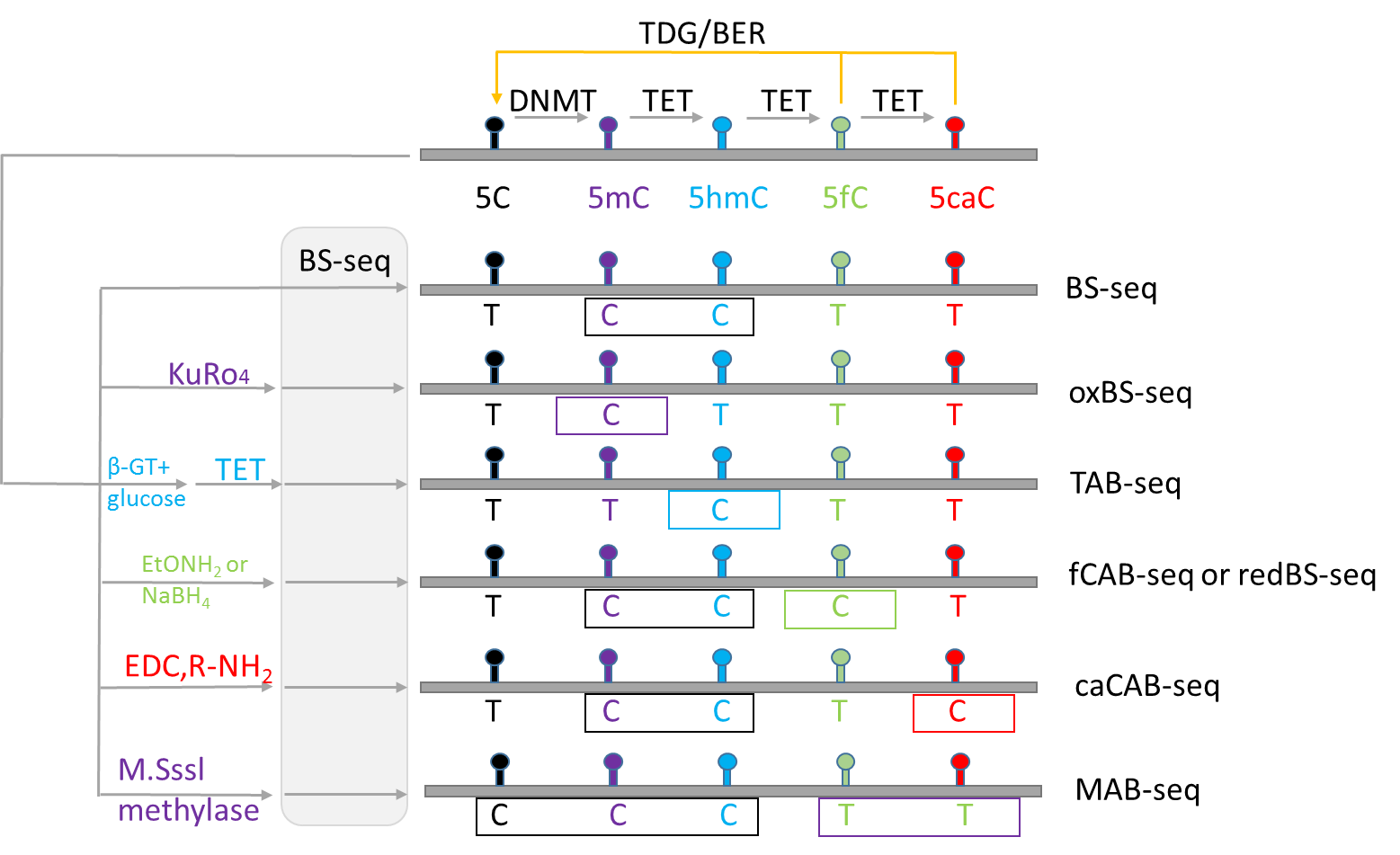
5mC的检测技术是在2012年剑桥大学的Michael J. Booth等发表在nature期刊上。该技术为oxBS-Seq [24]，首先加入kuRo4将5hmC氧化为5fC，进一步进行亚硫酸氢钠处理，因此可以直接检测5mC的甲基化水平以及分布情况。

fCAB [25] or redBS-Seq [26] ：可以用于检测5fC的甲基化水平。fCAB的原理在于加入EtoNH2保护5fC防止在加入亚硫酸氢盐时转化为T，可以检测到mC、hmC以及fC的甲基化水平，然后进一步结合BS-Seq（mC与hmC）技术可以检测到5fC的甲基化分布。redBS-Seq是在开始加入NaBH4可以将5fC还原为5hmC，同样可以检测三种甲基化状态（mC，hmC以及fC），结合BS-Seq（检测到mC与hmC）能够检测到5fC的甲基化。

caCAB-seq [27] 用于检测5caC，该技术是有芝加哥大学的何川教授发表在2013年。技术原理为通过加入EDC，R-NH2保护5caC不被亚硫酸氢钠转换为T，因此最终可以检测到mC、hmC以及caC甲基化水平，之后进一步结合BS-Seq（检测mC、hmC）最终得到5caC的甲基化水平分布。

MAB-seq [28]技术是由美国霍华德休斯医学院的Li Shen和Yi Zhang教授在2014年发表在Nature Biotechnology。该技术可以同时检测5caC以及5fC，具体原理为在MSssl methylase酶作用下能够保护未发生甲基化的胞嘧啶、5mC以及5hmC不被亚硫酸氢盐转换变为T，而5fC与5caC被亚硫酸氢钠转换为U经过PCR扩增为胸腺嘧啶，因此最终可以同时检测到5caC以及5fC甲基化分布情况。

综上所述，目前针对不同状态DNA甲基化形态已有对应的检测技术，多种检测技术之间能够相互结合使用，检测所有不同甲基化情况，完成不同的目的要求，具体各项技术的处理流程图参见图三。



图三：BS检测不同状态DNA甲基化示意图

**1.2.2 其他检测DNA甲基化技术**

MeDIP-Seq(Methylated DNA Immunoprecipitation sequencing) [29]甲基化DNA免疫共沉淀测序，该技术开发于2012年，是一种高效的富集DNA甲基化的方法。主要原理是通过抗体富集进行全基因组上发生甲基化的DNA片段富集检测。该测序方法覆盖较广性价比高，可以在全基因组水平上进行CpG密集的高甲基化区域研究，但是不能精确到单碱基分辨率水平，分辨率在100bp，只局限于一段区域的甲基化水平，主要通过peak分析进行比较甲基化变化。

MRE-Seq(Methylation-sensitive Restriction Enzyme digestion followed by sequencing ) [30-31]甲基化敏感限制性酶测序，是由美国加利福尼亚大学的Raman P Nagarajan和and Joseph F Costello开发于2010年，该方法同样存在一些检测的局限性，但是性价比较高。

MIRA-seq （Methylated-CpG Island Recovery Assay sequencing）[32，33]是在2010年开发，主要用于检测CpG岛区域的甲基化情况，通过加入MBD2B/MBD3L1 复合物，能够富集甲基化的CpG岛区域。MIRA-seq技术主要应用于多种疾病分析，多用于医学方面的研究。

RRBS（Reduced Representation Biuslfite Sequenceing）[34] 该方法是在酶切的基础上，前期使用MspI酶切，该酶切位点为（CCGG），在此基础上将CpG位点进行富集,之后处理分析甲基化水平。该技术是由麻省理工学院的Alexander Meissner在2005年发表，主要用于在多样本间研究启动子或者转座元件的甲基化变化情况。该方法能够检测包含多CpG位点的甲基化位点，序列靶向性好，单碱基分辨率，但是酶切位点有限。

MBD-Seq( methyl-CpG binding domain protein-enriched genome sequencing) [35] 技术是2013年由约翰霍普金斯大学医学院Shannath L Merbs课题组完成。原理通过特异性结合甲基化DNA的蛋白MDB2b富集甲基化的DNA片段，然后结合测序技术实现高甲基化CpG区域区域检测，可以用于在样本间全基因组范围内检测DNA甲基化变化模式。

综上可以看出对于DNA甲基化的研究技术层出不穷，这也表明目前各种技术都存在一定的缺陷，而且很多技术是针对动物高CpG区域检测，是由于动物的甲基化一般发生在CpG胞嘧啶，因此一些对于CpG的富集检测能够在节约成本基础上检测到大部分甲基化分布。但是在分析过程中发现植物（水稻中）CHG以及CHH占到了一定的比例 [36]，因此目前只针对CG的甲基化变化研究存在较大的局限性。并且上述不同的检测技术对于后期的数据分析有不同的要求，对数据的处理提高了难度要求。

**二、研究的问题**

**2.1 DNA甲基化软件包开发**

正如前面介绍目前DNA检测技术的多样性，以及DNA甲基化的重要性，因此对于研究DNA甲基化的功能机制具有重要意义，但是目前对DNA甲基化数据的处理缺少系统化的流程软件，工欲善其事必先利其器，因此我们目前亟需一套容易操作功能齐全的处理DNA甲基化数据软件包对当前BS-seq甲基化数据进行分析研究。

**2.2应用软件包，进行相关的生物学研究**

在完成DNA甲基化处理软件的基础上，结合相应的生物学问题进行生物学数据分析，希望能够揭示DNA甲基化的一些作用机制，为以后甲基化的研究能有一些帮助，并针对一些有意义数据能够进行甲基化分析，建立DNA甲基化全基因组图谱，希望对DNA甲基化的分析研究起一定参考作用。

**二、研究目的与意义**

**DNA甲基化分析的重要性**

我们都知道DNA甲基化对于生命体具有重要意义，根据生物意义和功能的不同，DNA甲基化的分析目前涉及多个方面，有对于分化过程中DNA修饰的动态过程的分析，以及在肿瘤细胞中甲基化变化的研究。另外DNA甲基化的变化对于胚胎干细胞的全能性有一定的影响，并且研究发现DNA甲基化的状态能够区分人类胚胎干细胞以及诱导胚胎干细胞 [6] 。因此对于DNA甲基化的研究对于进一步了解生命机制，以及心血管疾病肿瘤等疾病的治疗提供了一个新的解决手段。

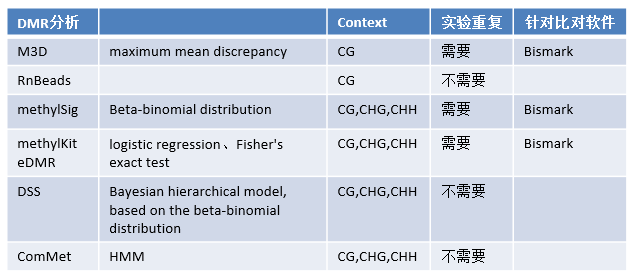
DNA甲基化对于基因表达以及发育具有关键作用，因此呈现部分基因或者全基因组的甲基化状态，对于进一步了解DNA甲基化在细胞中对基因的调控，以及表观机制的研究有重要意义。

DNA甲基化作为主要的表观修饰之一，对生命体具有重要的功能作用，高通量测序对于研究DNA甲基化在基因组上的分布情况对于了解甲基化对机体的调控作用机制是非常必要的，但是目前DNA甲基化数据分析缺乏系统完整的分析流程以及标准，尤其是在目前DNA甲基化数据量较大的情况下系统简便的分析流程更是必不可少的。因此，我们旨在结合目前已有的分析软件并整合DNA甲基化数据分析需要实现的功能，形成一套完整的DNA甲基化数据分析流程。对DNA甲基化数据的分析以及功能机理探索提供有力的工具。

**三、目前已有软件介绍（研究进展）**

目前DNA甲基化分析已有的比对软件主要有BSmap [37] 、Bismark [38] 、BatMeth [39] ，后续处理软件包包括COHCAP [40]、M3D [41] 、methy-Pipe [42] 、RnBeads [43] 、methylKit[44]、methylSig[45]、eDMR[45]等，信息介绍见表一。

但是目前已有的程序软件存在一定的不足，首先M3D软件包需要数据重复这对目前DNA甲基化检测来说是严重的问题，并且差异检验只针对CG context，因此只适合于动物DNA甲基化分析。RnBeads同样只能针对CG context，并且后续的一些分析只针对人与小鼠，这对DNA甲基化数据分析是很大的局限。methySig首先需要重复，并且使用过程中会出现随机位点的变动，这会导致后期具体位置分析的误差，而且对DMR的分析并不完善，以及缺少后续统计分析。methyKit与methySig的功能是一致的但是结合eDMR能够实现后续统计分析。DSS由于DMR的识别是基于DML（差异甲基化位点）因此可能存在一定误差，并且该软件不能区分正负链这对后期DMG分析具有很大影响。ComMet程序在DML、DMR等差异分析缺少显著性判断，很难确切筛选差异。综上所有软件包在处理方面存在各种不足，尤其缺少后期的统计分析，给DNA甲基化的数据分析工作提供了很大的难道。

表一：目前已有软件包信息表

**3.1使用不同的软件，我们需要完成以下基本的分析内容：**

**3.1.1全基因组甲基化图谱**

计算单位点甲基化水平可以得到全基因组甲基化图谱。

MethRatio = effective C/(#C + #T)

**3.1.2甲基化context注释**。

根据甲基化CG/CHG/CHH区分注释，便于对三种不同的甲基化形态标注以及研究不同甲基化环境发生的机理以及作用。

**3.1.3 差异甲基化位点 DMCs**

计算基因组上不同样本位点之间甲基化差异。

**3.1.4差异甲基化区域 DMRs 以及差异甲基化基因DMG**

根据不同区域甲基化水平可以计算得到差异甲基化区域，差异甲基化基因，对照差异区域在基因组上的具体位置集合RNA-seq以及CHIP-seq数据具体分析变化区域的功能或者探索变化的原因，已有计算方法主要有fisher精确检验 [44]、二项式检验 [45]、方差检验 [4] 、最大均值差异[41]、秩和检验[42]等。

**3.1.5差异甲基化DMC、DMR在不同功能元件上的分布**

统计在不同功能组分上的数目分布情况（Promoter，genebody，TE），对于研究基因以及转座元件的功能作用具有重要意义。

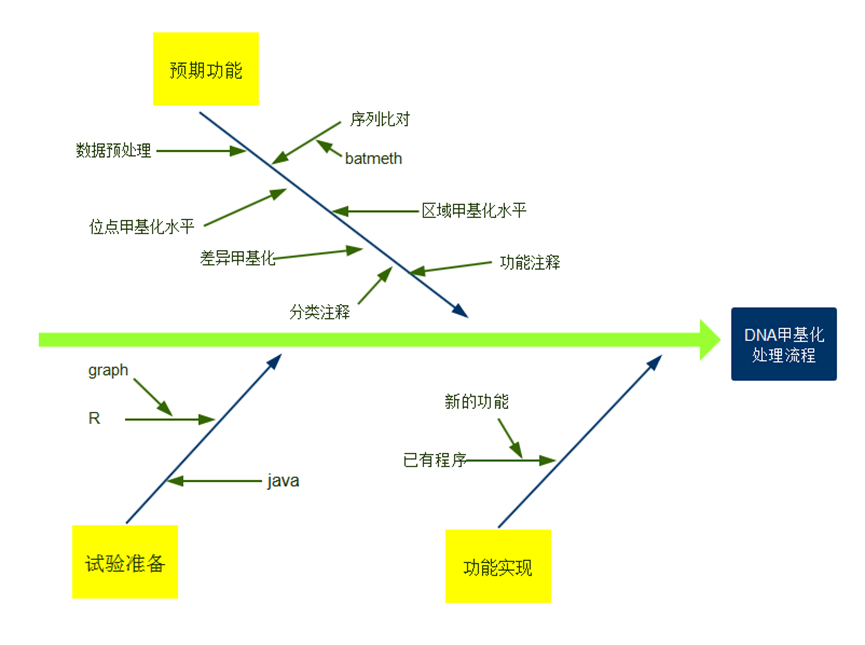
**3.1.5差异甲基化基因功能分析**

对于差异甲基化基因可以进行GO分析，以及KEGG分析，根据功能注释结果，检测突变组导致DNA甲基化的变化对于生命体的功能结果具体变化。

**四、技术路线**

比对软件C++完成，准备阶段程序实现需要学习java语言，一些图形显示部分是需要借助R来完成，后期结合目前已经完成程序以及分析过程中需要添加的新的内容，最终完成分析DNA甲基化数据的流程包，为以后DNA甲基化数据的分析以及功能的研究提供新的方法。

对于DNA甲基化程序包的开发主要分三步工作，1）准备工作，首先理论方面需要学习DNA甲基化数据分析需要注意的问题当前已有软件存在的不足，并且熟悉DNA甲基化数据的分析过程以及分析过程中存在的问题和解决方案。另外对于编程上学习c++ /java编程语言以及R语言，这对后期软件包的开发编辑是必不可少的。2）设计预期需要实现的功能，目前来说主要包括数据预处理、序列比对、位点甲基化水平计算、区域甲基化水平统计计算、差异甲基化分析（DML/DMR/DMG）、根据gene以及TE进行甲基化水平计算和分类注释、差异基因功能分析。3）预期功能的实现，完成预期实现功能以及结合后期数据分析中遇到的问题或需求最终实现DNA甲基化程序软件包。



图四：实验技术流程图

**五、目前DNA甲基化分析中存在的问题，和需要提高的部分**

目前DNA甲基化数据分析仍然存在一些问题，对于全基因组亚硫酸测序，目前的测序平台，对测序样本要求4种碱基比较平均，但是亚硫酸氢盐处理过后， C基本都变成了T，这导致样本碱基基本变为A,T,G的情况，复杂度降低。测序的效果就下降，并且分析过程更容易出现比对的复杂性以及误差。并且BS-seq测序对于单碱基位点甲基化水平需要较高的测序深度，导致所需要的费用是昂贵的。最后分析过程中没有统一的标准、数据分析缺乏系统简便的分析流程，对于一些较细的需求难以满足，而且目前缺乏对于不同甲基化状态的分析软件。

其次基于亚硫酸转化衍生的其他检测技术RRBS以及特定区域DNA甲基化测定MBD-Seq等虽然可以有效降低成本，但是只能研究部分特定区域的甲基化水平，无法实现整个基因组甲基化图谱，并且存在重亚硫酸盐处理不完全导致的假阳性问题。

实验不足如果针对特定区域或者问题可以选择RRBS等检测技术，而要实现全基因组甲基化水平测序，随着实验技术的不断开发以及测序平台的提高一定会逐渐得到改善。

对于分析不足，该实验目的正在于结合目前已有的分析软件，能够开发出一套实现序列比对，获取位点甲基化水平，不同context的差异甲基化分析（DMCs、DMRs、DMGs），分类注释，功能分析等功能，使用简便的软件包。

**六、已有的工作基础以及分析结果**

完成修改比对软件部分，并根据比对结果计算位点甲基化水平，以及区域甲基化水平，基因区甲基化水平，均值方差法检测甲基化差异，基因和TE区甲基化变化趋势，差异区域分类注释等功能已用java完成。

目前对于BS-Seq数据分析主要处理过水稻以及人类的DNA甲基化数据。

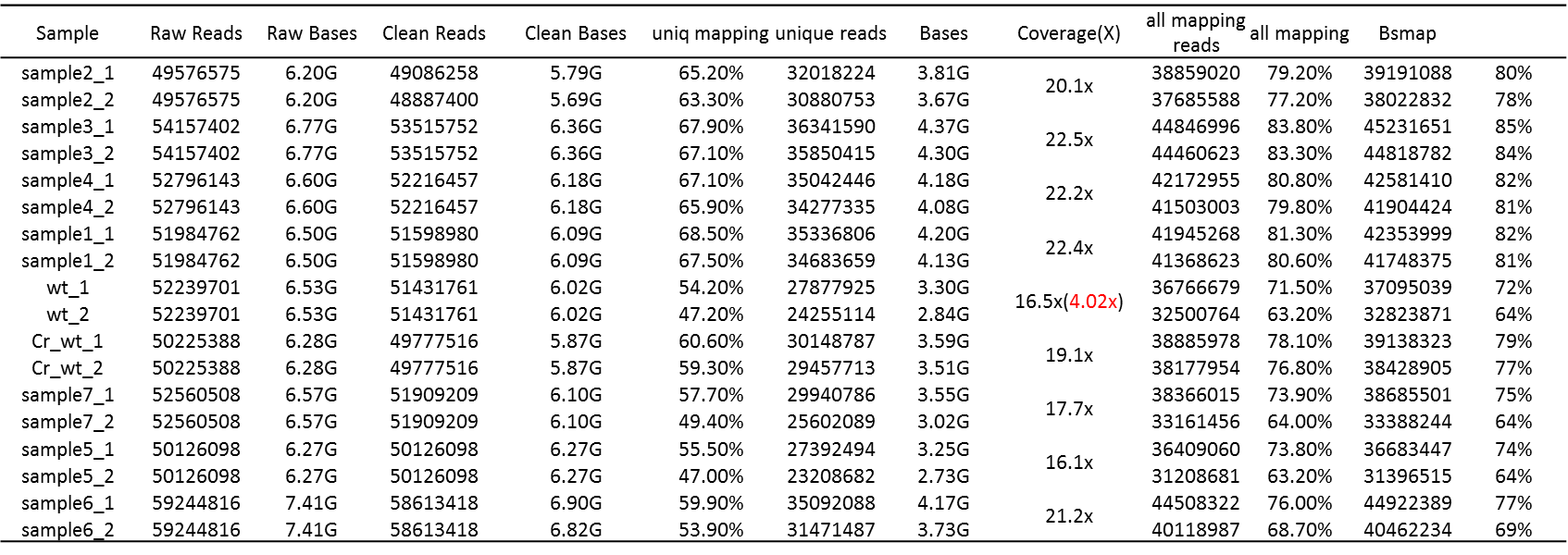
**6.1 水稻BS-seq 数据分析**

水稻的S706 是一个DNA 甲基转移酶基因，对于从头甲基化以及CHH的维持具有重要作用 [46-47]。水稻DNMT主要参与DNA甲基化的维持 [48-50]。

目前已有结果主要有：

**6.1.1各个样本的比对结果**

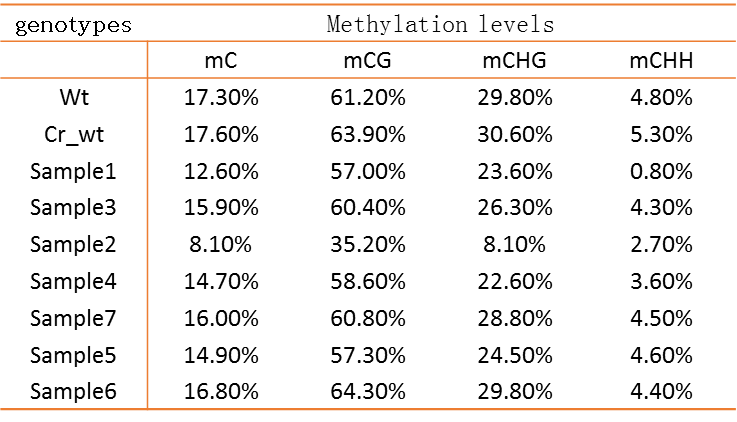
统计每组样本的数据原始数据、预处理结果clean reads以及比对信息、覆盖度等（见表二）。



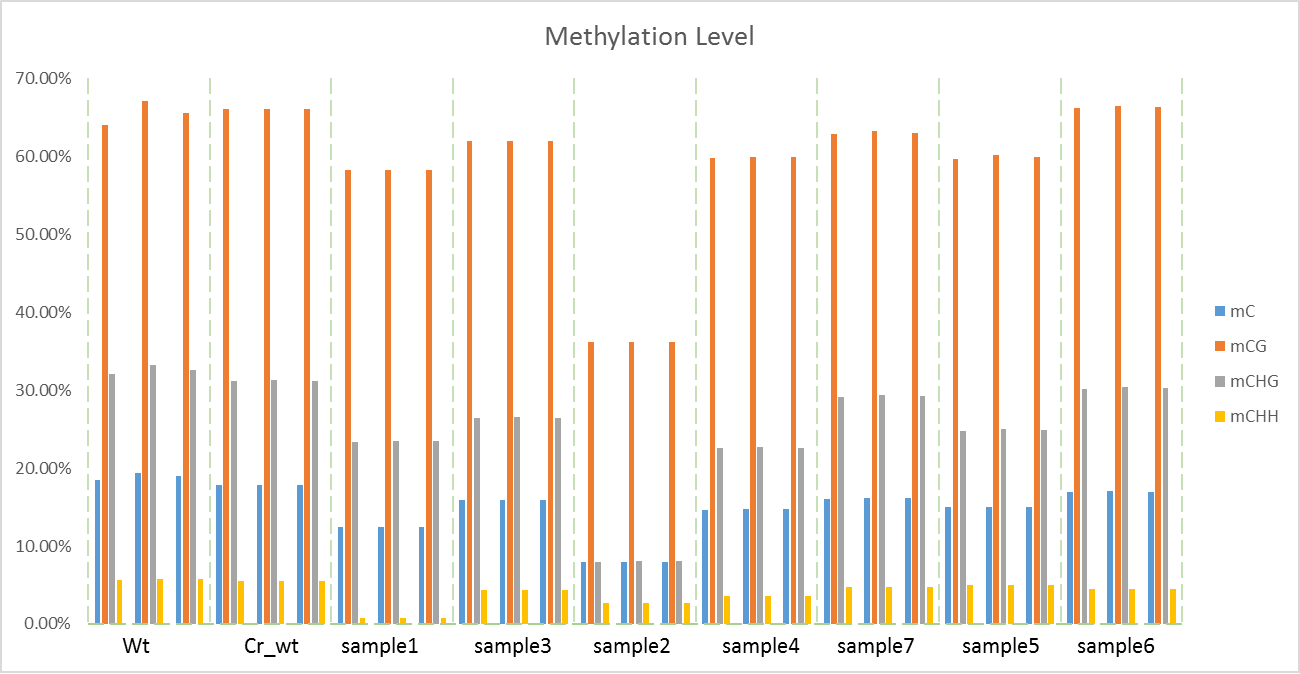
表二：各个样本的预处理以及比对结果。

**6.1.2不同样本平均甲基化水平**

计算样本在全基因组上不同甲基化context DNA甲基化水平（参见表二、图二），可以比较不同样本间DNA甲基化水平整体变化情况。



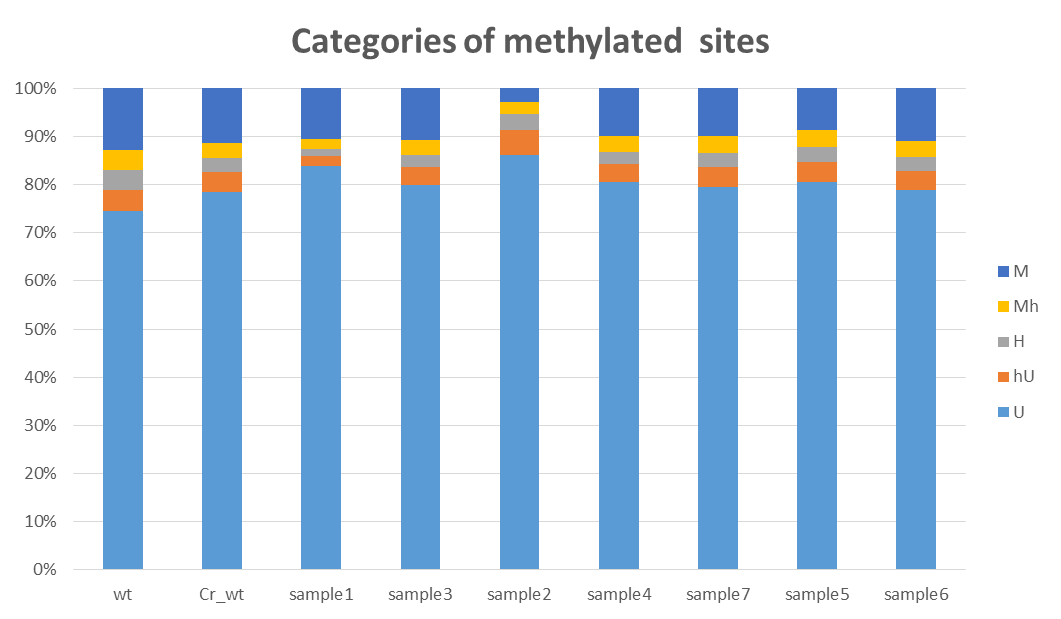
表三，不同表型平均甲基化水平



图五：样本甲基化水平分布

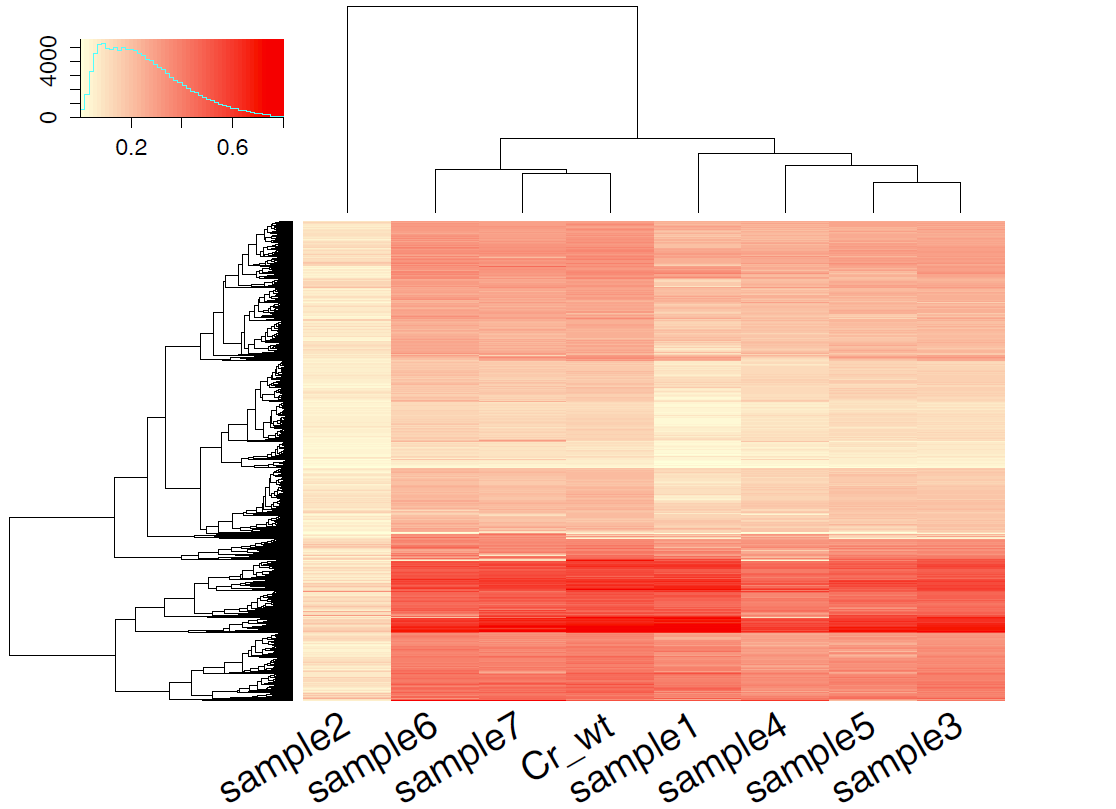
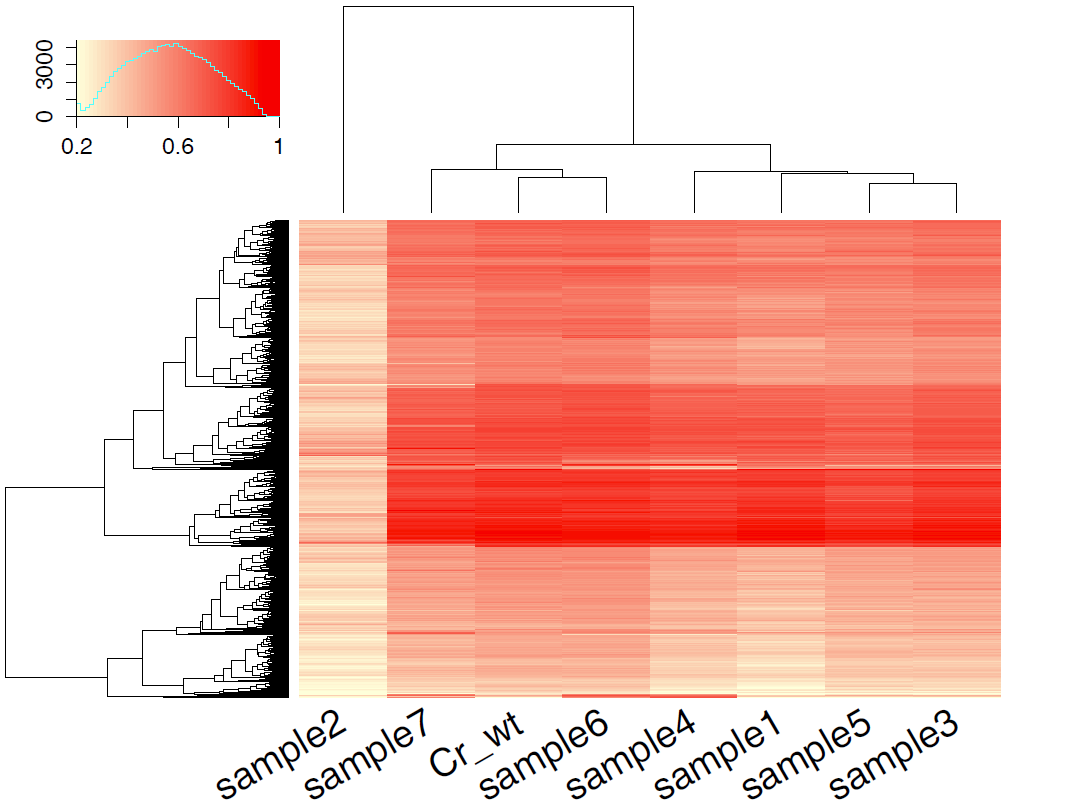
**6.1.3 根据甲基化水平划分得到不同样本的甲基化水平比重图**

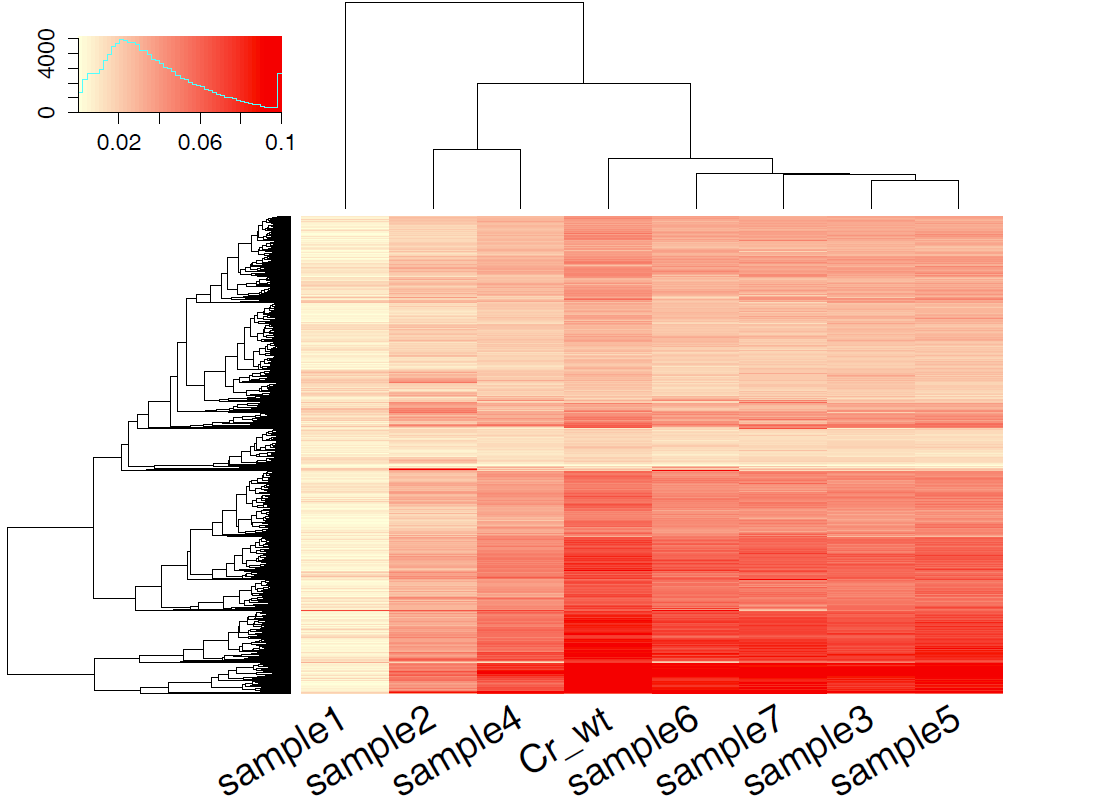
将DNA甲基化水平划分为五个等级（M: more than 80% Mh: between 60%, 80% H: between 40% and 60% hU: between 20% and 40% U: less than 20%），并分别统计不同级别DNA甲基化分布情况，因此我们可以看出不同样本DNA甲基化的变化（见图六）。



图六：样品间甲基化比重分布

**6.1.4 样本间甲基化水平热图**



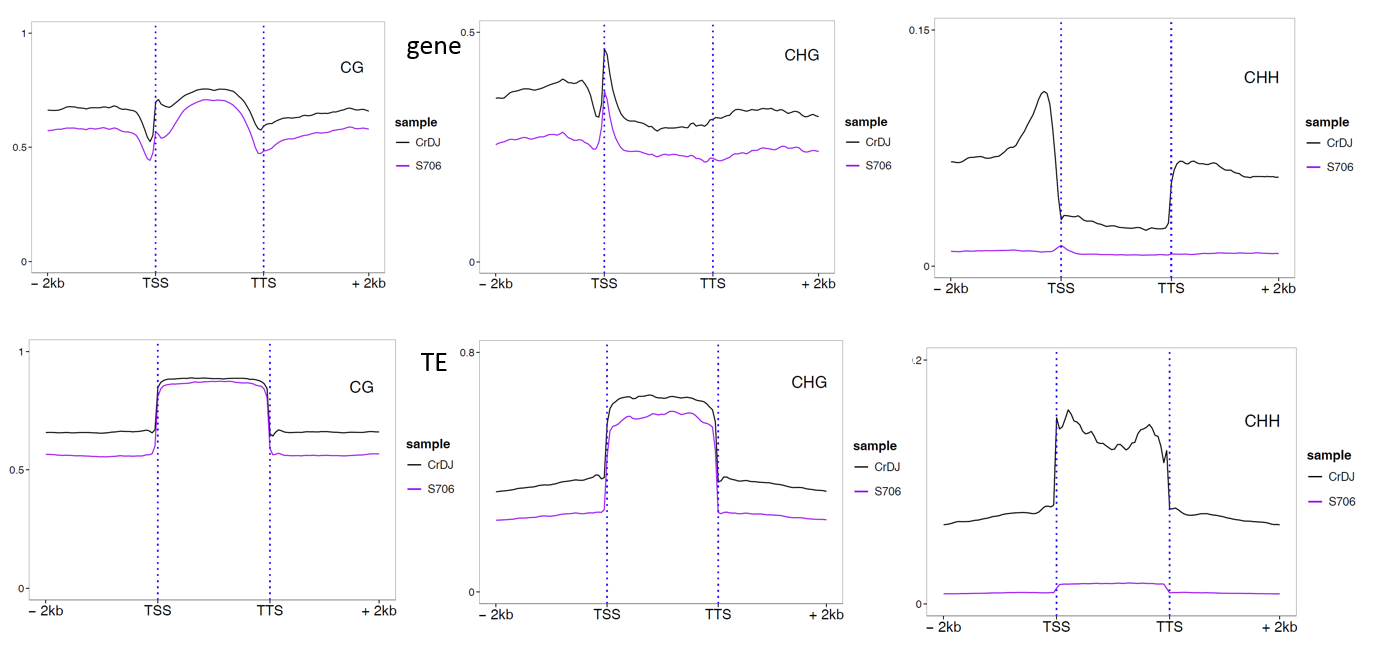


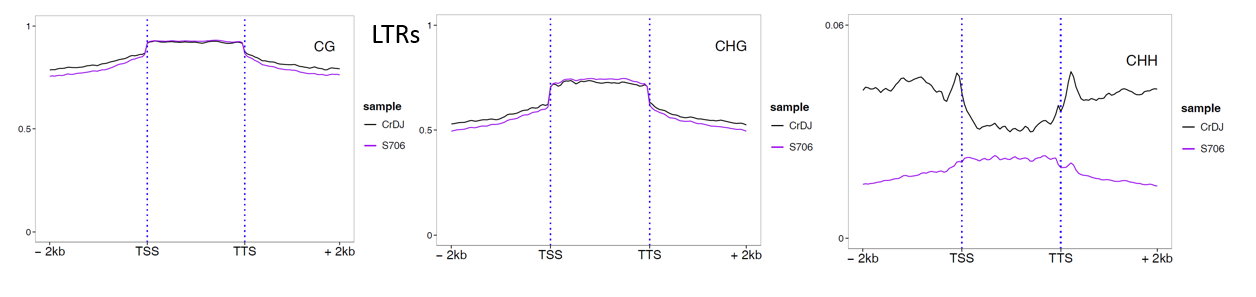
图七：甲基化水平分布cg、chg、chh。

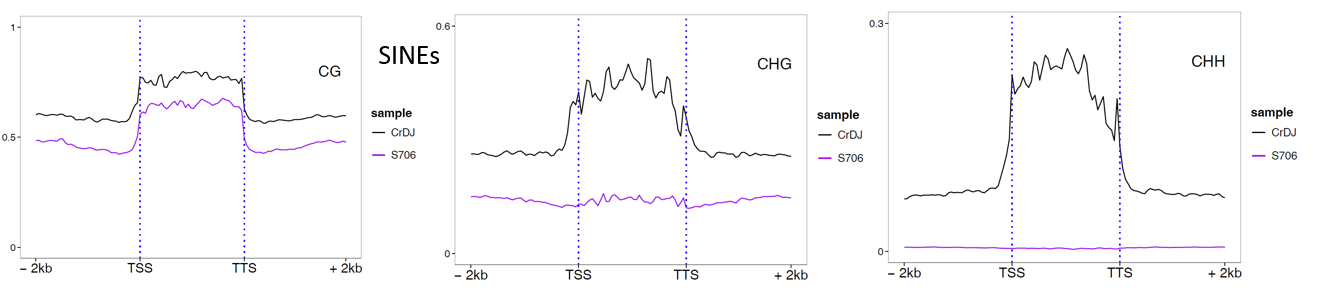
对所有样本分别根据甲基化context进行DNA甲基化聚类分析（如图七），根据聚类结果，我们可以得到不同甲基化context下与野生型相比变化最大的样本，这对我们具体分析提供参考。

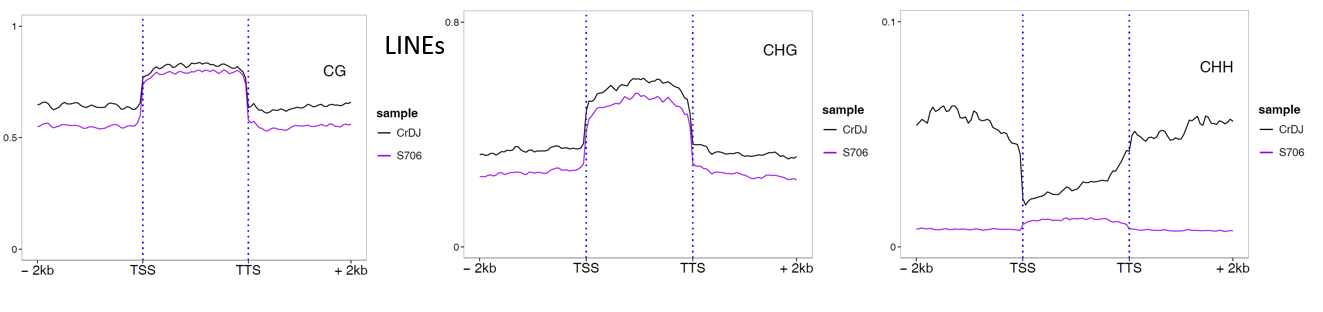
**6.1.5野生型组培样本Cr\_DJ与DNMT突变组相比在Gene以及转座子区域甲基化水平分布。**

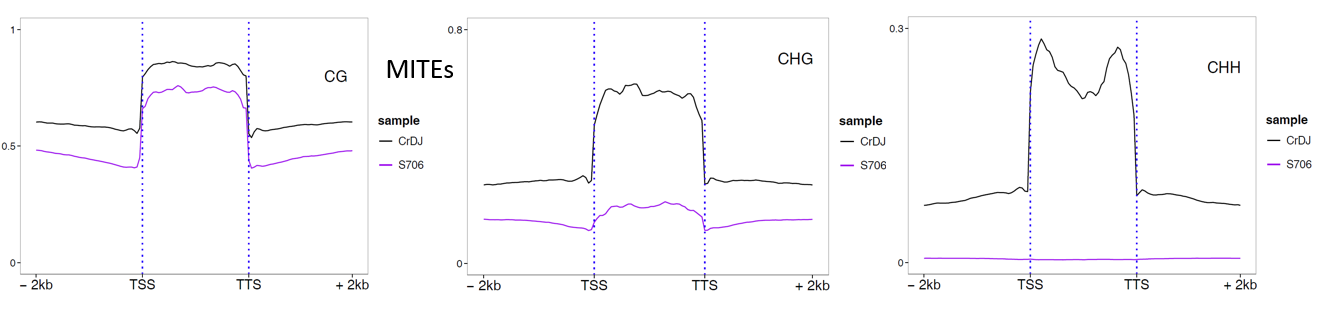
对Gene以及TE进行甲基化水平统计，并绘制其上下游甲基化变化情况（如图八），根据得到结果我们可以看出在gene或TE上下游甲基化变化，以及gene TSS甲基化变化情况，这对我们了解物种在gene以及TE上DNA甲基化的分布以及预测研究DNA甲基化的作用机制具有重要作用。







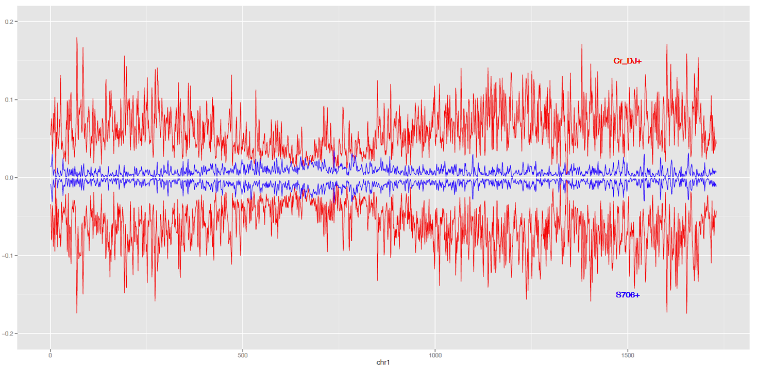




图八：野生型样本Cr\_DJ与OsDRM2突变组相比在Gene以及转座子区域甲基化水平分布。

**6.1.6 野生型样本与突变组相比Chr1染色体CHH甲基化水平分布**

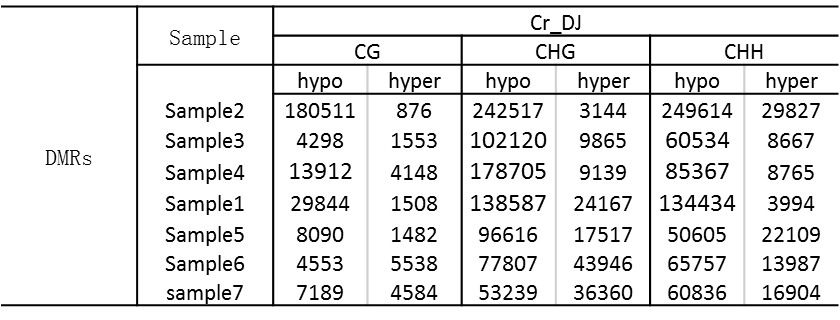
进行不同样本间染色体甲基化水平分布图，由图五野生型样本（红色）与突变组（蓝色）甲基化水平变化情况，可以看出该基因对于CHH的维持起着决定性作用。



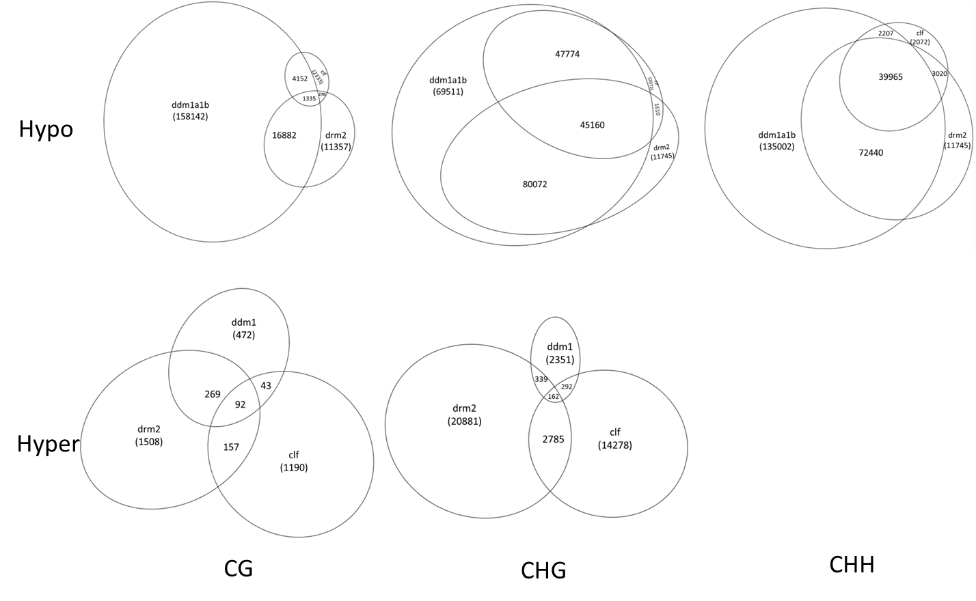
图九：野生型样本Cr\_DJ与S706突变组相比Chr1染色体CHH甲基化水平分布

**6.1.7 不同样品间差异甲基化分析（差异区域DMR、差异碱基DMC、差异基因DMG）**

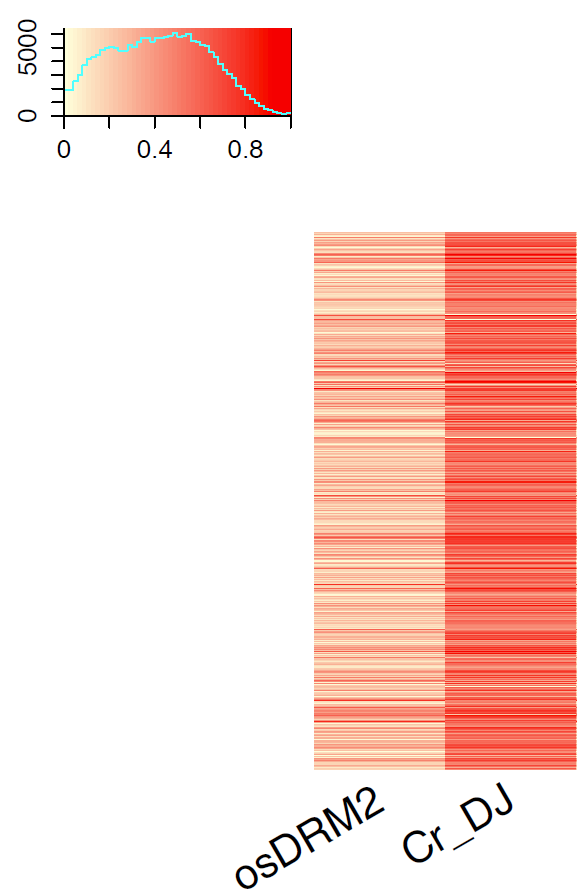
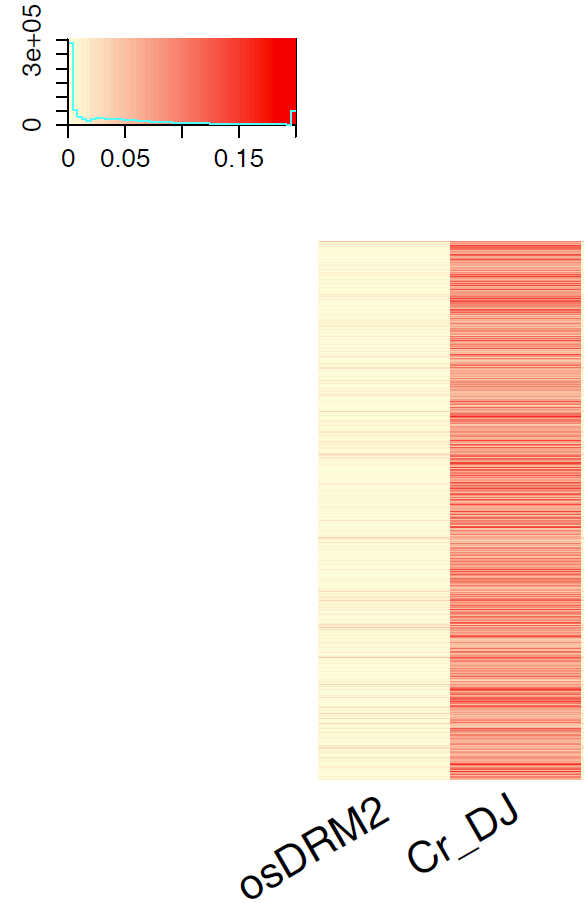
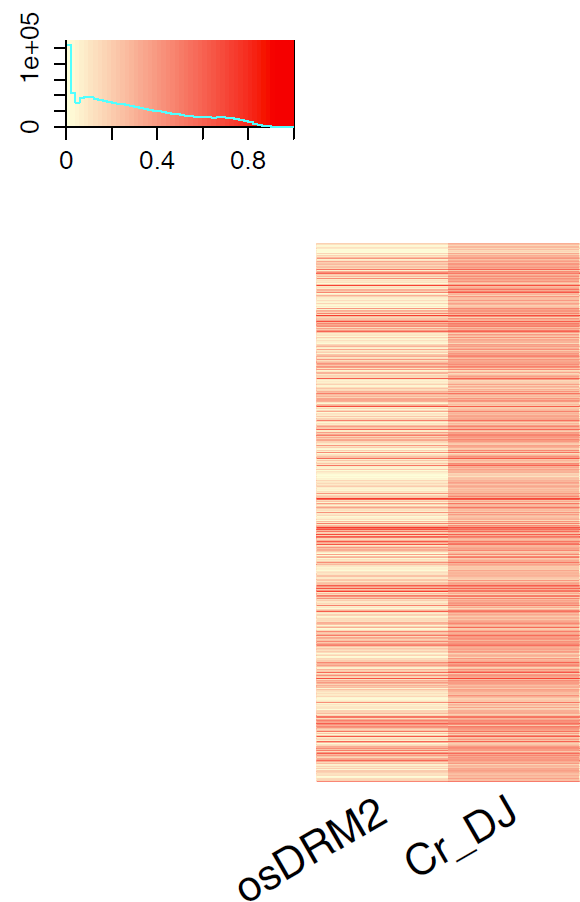
统计样本间差异甲基化位点（图十三），差异甲基化区域（图十），差异甲基化基因（图十四），以及差异甲基化区域或者差异甲基化基因在不同样本间的分布情况（图十二）以及相交情况（图十一、图十五）。之后根据统计结果进行具体问题分析。



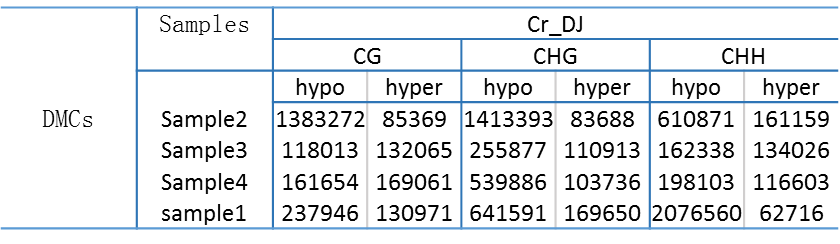
图十：DNA甲基化差异区域



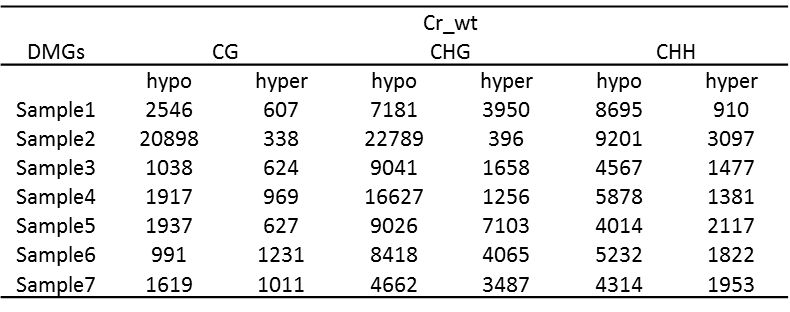
图十一：不同样本间DNA甲基化差异区域交集图

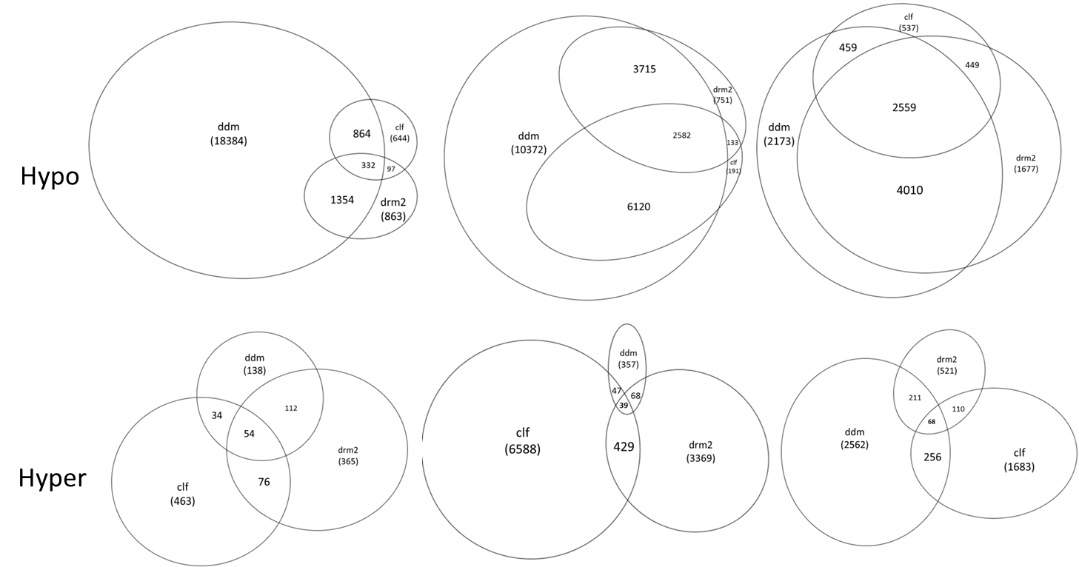
图十二：差异区域cg、chg、chh甲基化水平显示



图十三：DNA甲基化差异位点



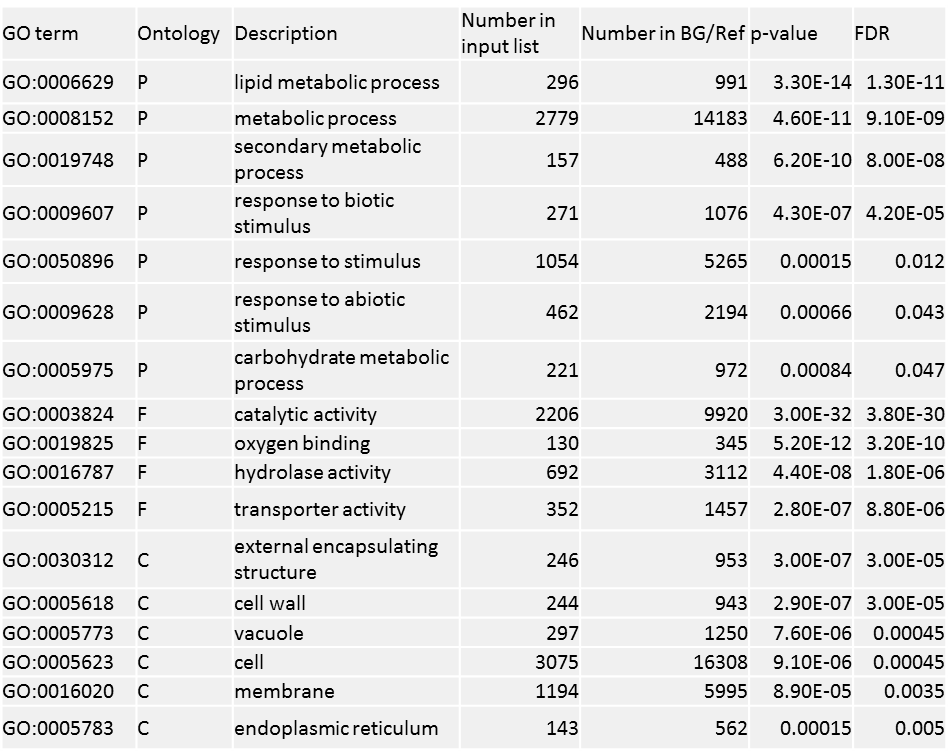
图十四：DNA甲基化差异基因



图十五：不同样本间DNA甲基化差异基因交集图

**6.1.8差异基因功能注释**

对DNA甲基化差异基因进行GO分析KEGG分析，找到不同样本间DNA甲基化变化导致的具体基因变化可能的相关功能（如同十六）。



图十六：差异基因GO分析

**七、未来的工作计划**

未来的工作计划分为两部分：1）DNA甲基化数据分析软件包开发，分析水稻DNA甲基化数据过程对甲基化数据分析有了整体的认识，进一步分别比照已经使用过的软件包和其他软件对DNA甲基化数据分析进行比较，结合目前已有的结果以及在数据分析过程中遇到的问题和需要添加的功能，最终完成DNA甲基化数据分析流程化的软件包；2）应用该软件包，进行具体的生物学研究，发现在水稻中DNA甲基化变化对生物功能的调控，对当前DNA甲基化数据进行分类注释以及后续分类注释，并结合已有CHIP-seq数据研究DNA甲基化与组蛋白甲基化之间的关系，以及结合RNA-seq分析DNA甲基化与表达量之间的关系。重点熟悉DNA甲基化的分析流程以及需要得到的结果，为后期软件包的流程以及功能提供一定的背景参照。

**八、预期的目标**

结合目前已有的工作以及分析过程中需求的功能，最终能够实现DNA甲基化数据预处理、序列比对、甲基化位点计算、差异甲基化分析、分类注释、功能分析等功能，并且对不同形态的DNA甲基化（mC/hmC）数据都能够进行分析处理，期望对现今DNA甲基化的机理研究起到一定的辅助作用。

**参考文献**

1. J Biol Chem. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. HOTCHKISS RD, 1948 Aug;175(1):315-32.

2. Keshet I, Lieman-Hurwitz J, Cedar H: DNA methylation affects the formation of active chromatin. Cell 1986, 44:535-543.

3. Reik W, Dean W, Walter J: Epigenetic reprogramming in mammalian development. Science 2001, 293:1089-1093.

4. Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. 20, 320–331.

5. Feng, J., and Fan, G. (2009). The role of DNA methylation in the central nervous system and neuropsychiatric disorders. Int. Rev. Neurobiol. 89, 67–84.

6. Huang, K., Shen, Y., Xue, Z., Bibikova, M., April, C., Liu, Z., Cheng, L., Nagy, A., Pellegrini, M., Fan, J.-B., et al. (2014). A Panel of CpG Methylation Sites Distinguishes Human Embryonic Stem Cells and Induced Pluripotent Stem Cells. Stem Cell Rep. 2, 36–43.

7. Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A.A., Kol, N., Salmon-Divon, M., Hershkovitz, V., Peer, E., Mor, N., Manor, Y.S., et al. (2015). m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. Science.

8. Takahashi, T., Matsuda, Y., Yamashita, S., Hattori, N., Kushima, R., Lee, Y.-C., Igaki, H., Tachimori, Y., Nagino, M., and Ushijima, T. (2013). Estimation of the Fraction of Cancer Cells in a Tumor DNA

9. Goll, M.G., and Bestor, T.H. (2005). EUKARYOTIC CYTOSINE METHYLTRANSFERASES. Annu. Rev. Biochem. 74, 481–514.

10. Gjerset RA, Martin DW. Presence of a DNA demethylating activity in the nucleus of murine erythroleukemic cells. J Biol Chem 1982; 257(15): 8581-83.

11. Bhattacharya SK, Ramchandani S, Cervoni N, et al. A mammalian protein with specific demethylase activity for mCpG DNA. Nature, 1999, 397: 579-83

12. Tahiliani M, Koh KP, Shen Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mamma-lian DNA by MLL partner TET1. Science, 2009, 324: 930-5

13. Hajkova P, Jeffries SJ, Lee C, et al. Genome-wide reprogram-ming in the mouse germ line entails the base excision repair pathway. Science, 2010, 329: 78-82

14. Ito S, Shen L, Dai Q, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxy-lcytosine. Science, 2011, 333: 1300-3

15. Gu, T.-P., Guo, F., Yang, H., Wu, H.-P., Xu, G.-F., Liu, W., Xie, Z.-G., Shi, L., He, X., Jin, S., et al. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature 477, 606–610.

16. Penn, N.W., Suwalski, R., O’riley, C., Bojanowski, K., and Yura, R. (1972). The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. Biochem J 126, 781–790.

17. Zhang, L., Lu, X., Lu, J., Liang, H., Dai, Q., Xu, G.-L., Luo, C., Jiang, H., and He, C. (2012). Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. Nat Chem Biol 8, 328–330.

18. Kuo K C, McCune R A, Gehrke C W, et al. Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA [J]. Nucleic Acids Res, 1980, 8: 4763-4776.

19. Zentner, G.E., and Henikoff, S. (2014). High-resolution digital profiling of the epigenome. Nat Rev Genet 15, 814–827.

20. Reeves, R. & Jones, A. Genomic transcriptional activity and the structure of chromatin. Nature 260, 495–500 (1976).

21. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. Science 193, 848–856 (1976).

22. Frommer, M., E. Marianne, L. , S. Millar, D. ,etc.(1991)A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Genetics . 89, 1827-1831.

23. Yu, M., Hon, G.C., Szulwach, K.E., Song, C.-X., Jin, P., Ren, B., and He, C. (2012). Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine.Nat. Protoc. 7, 2159–2170.

24. Booth, M.J., Branco, M.R., Reik W, Balasubramanian, S, et al. (2012). Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. Science 336, 931–934.

25. Song, C.-X., Szulwach, K.E., Dai, Q., Fu, Y., Mao, S.-Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., et al. (2013). Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. Cell 153, 678–691. Sample Using DNA Methylation. PLoS ONE 8, e82302.

26. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D., and Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nat. Chem. 6, 435–440.

27. Lu, X., Song, C.-X., Szulwach, K., Wang, Z., Weidenbacher, P., Jin, P., and He, C. (2013). Chemical Modification-Assisted Bisulfite Sequencing (CAB-Seq) for 5-Carboxylcytosine Detection in DNA. J. Am. Chem. Soc. 135, 9315–9317.

28. Wu, H., Wu, X., Shen, L., and Zhang, Y. (2014). Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. Nat. Biotechnol.

29. Taiwo, O., Wilson, G.A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S., and Butcher, L.M. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. Nat Protoc. 7, 617–636.

30. Fouse, S.D., Nagarajan, R.P., and Costello, J.F. (2010). Genome-scale DNA methylation analysis. Epigenomics 2, 105–117.

31. Li, D., Zhang, B., Xing, X., and Wang, T. (2015). Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods 72, 29–40.

32, Rauch, T.A. et. al. DNA Methylation Profiling using the Methylated-CpG Island Recovery Assay (MIRA). Methods. Volume 53 Issue 3. November 2010. pp. 213-217.

33, Mitchell, N. et. al. Methylated-CpG Island Recovery Assay. Methods Mol Bio. Volume 791, 2011. pp. 125-133.

34. Meissner, A. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 33, 5868–5877.

35. Oliver, V. F., Wan, J., Agarwal, S., Zack, D. J., Qian, J., & Merbs, S. L. (2013). A novel methyl-binding domain protein enrichment method for identifying genome-wide tissue-specific DNA methylation from nanogram DNA samples. Epigenetics & Chromatin, 6, 17. doi:10.1186/1756-8935-6-17.

36. Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. 11, 204–220.

37. Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10, 232.

38. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572.

39. Lim, J.-Q., Tennakoon, C., Li, G., Wong, E., Ruan, Y., Wei, C.-L., and Sung, W.-K. (2012). BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. Genome Biol 13, R82.

40. Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. 41, e117–e117.

41. Mayo, T.R., Schweikert, G., and Sanguinetti, G. (2014). M3D: a kernel-based test for spatially correlated changes in methylation profiles. Bioinformatics.

42. Jiang, P., Sun, K., Lun, F.M.F., Guo, A.M., Wang, H., Chan, K.C.A., Chiu, R.W.K., Lo, Y.M.D., and Sun, H. (2014). Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis. PLoS ONE 9, e100360.

43. Assenov, Y., Muller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. Nat Meth 11, 1138–1140.

44. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., Mason, C.E., and others (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13, R87.

45. Park, Y., Figueroa, M.E., Rozek, L.S., and Sartor, M.A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics 30, 2414–2422.

46. Zhong, X., Hale, C.J., Nguyen, M., Ausin, I., Groth, M., Hetzel, J., Vashisht, A.A., Henderson, I.R., Wohlschlegel, J.A., and Jacobsen, S.E. (2015). DOMAINS REARRANGED METHYLTRANSFERASE3 controls DNA methylation and regulates RNA polymerase V transcript abundance in Arabidopsis. Proc. Natl. Acad. Sci. 112, 911–916.

47. Hu, L., Li, N., Xu, C., Zhong, S., Lin, X., Yang, J., Zhou, T., Yuliang, A., Wu, Y., Chen, Y.-R., et al. (2014). Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. Proc. Natl. Acad. Sci. 111, 10642–10647.

48. Dangwal, M., Malik, G., Kapoor, S., and Kapoor, M. (2013). De Novo Methyltransferase, OsDRM2, Interacts with the ATP-Dependent RNA Helicase, OseIF4A, in Rice. J. Mol. Biol. 425, 2853–2866.

49. Moritoh, S., Eun, C.-H., Ono, A., Asao, H., Okano, Y., Yamaguchi, K., Shimatani, Z., Koizumi, A., and Terada, R. (2012). Targeted disruption of an orthologue of DOMAINS REARRANGED METHYLASE 2, OsDRM2, impairs the growth of rice plants by abnormal DNA methylation: osdrm2 affects DNA methylation and development. Plant J. 71, 85–98.

50. Higo, H., Tahir, M., Takashima, K., Miura, A., Watanabe, K., Tagiri, A., Ugaki, M., Ishikawa, R., Eiguchi, M., Kurata, N., et al. (2012). DDM1 (Decrease in DNA Methylation) genes in rice (Oryza sativa). Mol. Genet. Genomics 287, 785–792.