附录二

**项目研究报告**

|  |  |  |  |
| --- | --- | --- | --- |
| **项目名称：** | **全基因组DNA甲基化** | | |
|  | **及其差异分析** | | |
| **项目编号：** | **2016348** | | |
| **项目负责人：** | **董怡蕙** | **年级专业：** | **计科1403** |
| **项目组成员：** | **焦裕迪、宋琦敏** | | |
| **指导教师：** | **李国亮** | **职 称：** | **教授** |

**全基因组DNA甲基化及其差异分析项目研究报告**

董怡蕙，焦裕迪，宋琦敏

**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。

**关键词：**DNA甲基化 亚硫酸氢盐测序技术 差异分析

**Abstract：**In recent years, apparent modification has been widely concerned, and DNA methylation as an important component in the apparent modification of its occurrence mechanism and mechanism of action has become a hot research. Studies have shown that DNA methylation can play an important role in chromosome conformation, transcriptional regulation, tissue differentiation, and cancer. Therefore, the study of DNA methylation in different states of the occurrence and changes in understanding of the process of methylation and the role of the principle of great significance for the prevention of some diseases play a reference value to study the status of DNA methylation, especially single The level of methylation of DNA methylation of the mechanism of the mechanism of DNA methylation is of great significance.

**Keywords：**DNA methylation BS-seq variance analysis

1 DNA甲基化差异分析的背景与现状

DNA甲基化（DNA methylation）是最早发现的修饰途径之一，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。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。目前已有研究证明DNA甲基化对于异染色质 [1] 、转录沉默 [2]、基因表达、染色体结构、胚胎发育[3-4]、细胞全能型、以及癌症方面有重要作用。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甲基化在体内存在形式有5mC、5hmC [5] 、5fC [6]以及5caC [6] 。

目前检测DNA甲基化的金标准为BS-seq技术（亚硫酸氢盐测序技术）[7]，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。

当下DNA甲基化的研究较为火热，而检测DNA甲基化差异分析是其中必不可缺的部分，但是目前已有的检测DNA甲基化差异方法存在一定的局限性：M3D/RnBeads只能检测CG（DNA甲基化的发生有三种：CG CHG CHH）的差异分析，methylSig需要生物学重复，这些问题对于目前DNA甲基化的功能研究造成了很大的局限性。因此本课题工作旨在以batmeth（BSseq比对软件）为基础，完成DNA甲基化数据的差异分析以及hmC分析完成相应的生物学功能研究，为DNA甲基化的研究以及疾病的研究胚胎的发育等起到些微的助力。最终我们结合已有数据能够完成DNA甲基化数据分析并在此基础上解析一定的生物学问题。

2 本课题所研究的问题

运用全基因组DNA甲基化CG/CHG/CHH的差异位点（DMC）、差异区域（DMR）、差异基因（DMG）等的分析软件包进行差异分析。在此基础上结合已有的WGBS-seq数据能够对方法进行可靠性检测，并结合该生物学数据能够解析一定的生物学问题。

2.1 DNA甲基化软件包及数据库开发

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甲基化数据以及等位基因的差异甲基化进行研究。

综合分析与比较现有的DNA甲基化数据库，开发一套功能齐全，可视化界面良好的DNA甲基化数据需求平台，方便用户浏览，搜索，查询，下载后台数据库中的相关数据。

2.2 应用软件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。

3 研究实施思路

预期的目标有：完成DNA甲基化及hmC的差异检测分析，如差异位点（DMC）、差异区域（DMR）、差异基因（DMG）以及差异功能等的分析，应用相关软件包结合相应的生物学问题进行数据分析，期望能够解释一些DNA甲基化作用机制，为以后甲基化的研究提供参考。为DNA甲基化的研究在疾病、胚胎发育等研究方面起到些微的助力。

项目初期的主要任务是通过相关的的网站以及老师给的资料了解项目，和导师学长讨论交流等方式掌握生物学DNA甲基化以及羟甲基（hmC）相关背景，熟悉batmeth等已有软件的使用和数据分析。

项目中期，熟练掌握linux操作系统的基本指令，并熟悉掌握在相应的系统环境下如何安装软件；在生物相关网站ncbi等下载不同物种软件运行所需数据（比对文件，参考文件）；项目组成员熟悉DNA甲基化差异分析软件初始版本batmeth1的用法，熟练软件相关功能指令的运行，对运行结果文件进行初步的了解与分析，并在batmeth1运行的基础上，熟悉DNA甲基化差异分析软件第二版本batmeth2用法，该版本扩展了初始版本的功能，增加DNA甲基化数据比对的可视化结果文件。通过运行分析不同物种的数据，加深对软件功能的理解。

项目末期主要是运用batmeth2进行数据（包括栽培稻，拟南芥，人类肝癌基因等)分析，包括使用fastqc、trimmomatic等软件进行数据质量分析和修剪等预处理。batmeth2处理流程包括：建立参考基因序列索引、映射比对甲基化、获取甲基化、甲基化水平和密度计算、甲基化可视化处理等。同时项目组搜索收集现有的关于DNA甲基化数据库的相关资料，整理出现有数据库的异同点，为将要形成的数据库做需求分析。

4 研究结果

4.1 背景知识掌握

掌握了生物学DNA甲基化以及羟甲基（hmC）相关背景，熟悉batmeth等已有软件的使用和数据分析。

4.2 环境搭建与相关软件使用

项目成员通过u盘引导安装的方式安装了Ubuntu系统，并自学了 linux下的相关指令及其系统操作。下载了methpipe的软件包， 并根据methpipe里README的说明安装了meth pipe，通过阅读methpipe的pdf指令说明书了解其用法。然后在服务器上熟悉其相关指令运行测试了batmeth1版本。

熟练掌握linux操作系统的基本指令，并熟悉掌握在相应的系统环境下如何安装软件；在生物相关网站ncbi等下载不同物种软件运行所需数据（比对文件，参考文件）；项目组成员熟悉DNA甲基化差异分析软件初始版本batmeth1的用法，通过Xshell与学校服务器连接后，下载batmeth1软件包，熟悉软件开发环境，掌握软件安装过程，熟练软件相关功能指令的运行，对运行结果文件进行初步的了解与分析。了解指令相关的生物背景知识；在batmeth1运行的基础上，熟悉DNA甲基化差异分析软件第二版本batmeth2用法，该版本扩展了初始版本的功能，增加DNA甲基化数据比对的可视化结果文件。通过运行分析不同物种的数据，加深对软件功能的理解。

4.3软件测试数据与分析

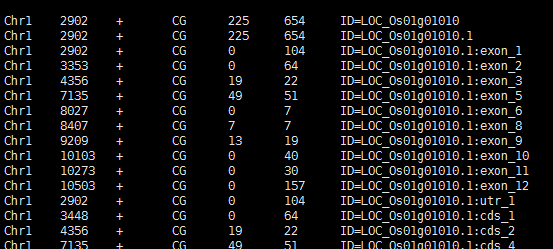
运用batmeth2进行数据（包括栽培稻，拟南芥，人类肝癌基因等)分析，包括使用fastqc、trimmomatic等软件进行数据质量分析和修剪等预处理。batmeth2处理流程包括：建立参考基因序列索引、映射比对甲基化、获取甲基化、甲基化水平和密度计算、甲基化可视化处理等。具体流程如下：

(1) 建立索引，像建立索引目录一样，可以实现基因的快速比对。

(2) 映射比对（Map）。比对之后得到了甲基化。包括单端比对和双端比对，比对序列长度大于150，fastqc 中reads的长度(.fq文件，AGCT也可以看长度),我们一般使用120,125的长度。

(3) 获取甲基化,得到甲基化水平相关的文件。

(4) 甲基化水平和密度计算，获得更多的甲基化信息为后续进行可视化的处理提供便利。图4.1中第一行Chr1表示染色体名称，2902 表示基因在染色体上的位置，225为其中支持甲基化的基因，654为基因总数，从而我们可以得到为甲基化水平：225/654。

图4.1 甲基化水平文件输出

1. DNA甲基化数据可视化。图4.2默认使用0.025基因的步长。

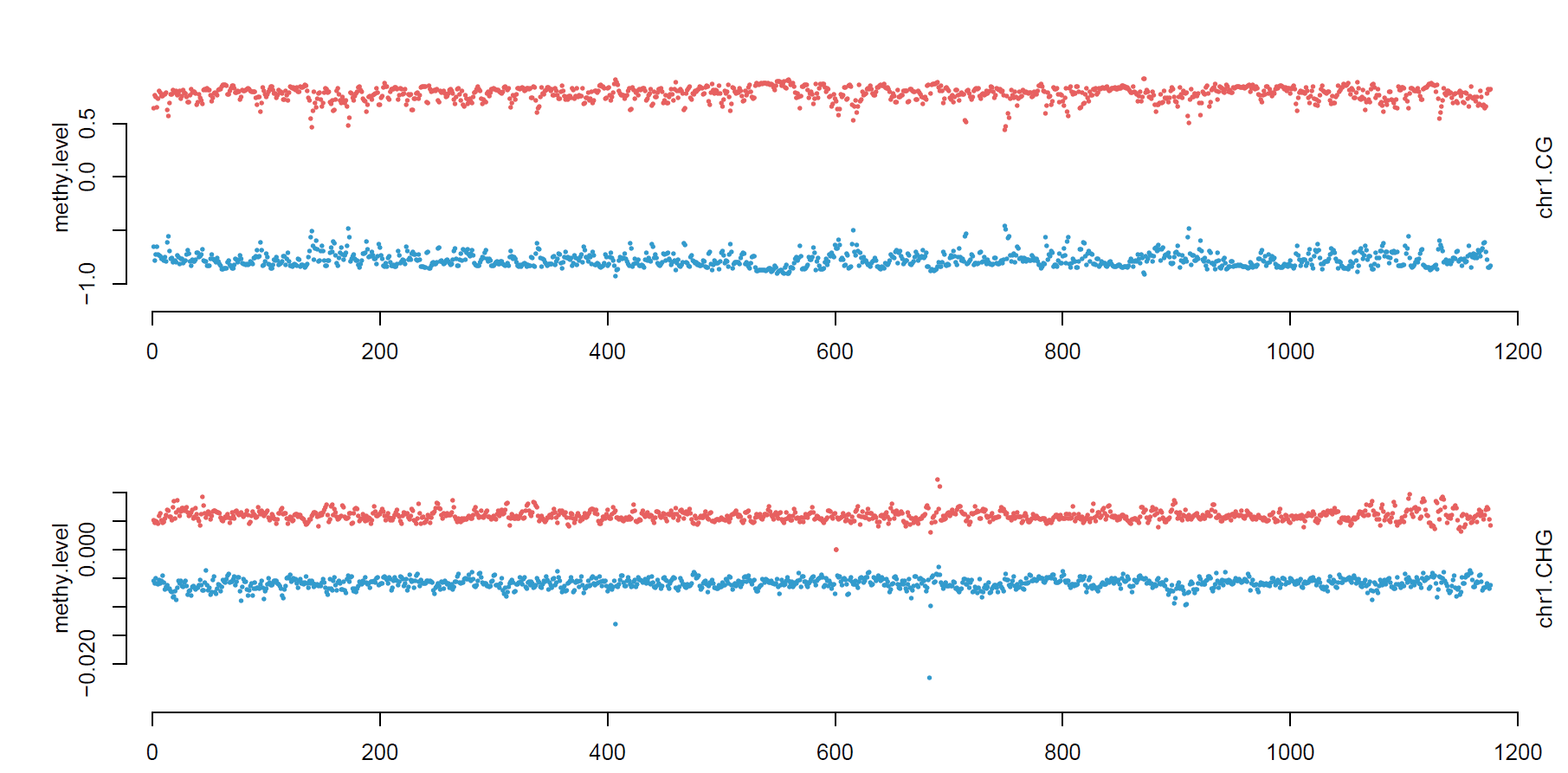


图4.2甲基化水平可视化

4.4数据库需求分析

搜索收集现有的关于DNA甲基化数据库的相关资料，整理出现有数据库的异同点，形成表格为以后数据库开发做需求分析,提供一定的助力。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据库名称 | 物种及所研究疾病、组织 | 浏览 | 搜索 | 下载 | 特点 |
| methBank | 斑马鱼，小鼠；  不同生物配子，早期胚胎。 | 高分辨率的DNA甲基化谱以及导频序列，基因的表达水平，SNPs，CpG岛。显示甲基化水平，创电子表达谱，DMRs和单核苷酸多态性 | 1.对于每一个物种，可搜索CpG岛，DMR、基因表达、甲基化水平，参考基因参考序列和SNP，2.特定基因或区域的生物多样性信息3.可选择感兴趣和轨道缩放和滚动基因组中的任何区域。当点击一个在一个特定的轨道基因/地区，显示相应细节。4.指定一个基因符号，可以获得其在启动子和基因本体在多个发育阶段的甲基化状态，以及它的基本信息，和基因表达。5.对于一个给定的基因符号或特定基因组区域，可以在两个发育阶段之间提供所有相关的区域 | 浏览器中当点击一个特定的轨道的基因或区域，相应的细节信息可下载 | DMR、等位基因信息 |
| methHC | 人；  癌症，肿瘤；  人的正常和肿瘤组织。 | DNA甲基化、mRNA基因表达，RNA基因和microRNA表达，DNA甲基化和mRNA/microRNA表达的不同关系，可视化图像界面。基因区域，CGI地区和DNA甲基化水平的方法估计 | 1.选择一个特定的KEGG通路和确定在一组的肿瘤基因甲基化的差异2.甲基化和去甲基化的基因，产生分别在肿瘤基因甲基化和去甲基化 | 6000个样本18个人的can-cers，6548个microarrays和12567的RNA测序数据。 | 1.microRNA小的非编码RNA，RNA和基因表达的转录后调控功能，并经常与癌症相关的，可能是一个重要的致病因素2.提供了基因转录起始位点的信息（TSS）和microRNA的TSS |
| methCancer | 人；  肿瘤；  肿瘤组织。 | DNA甲基化相关基因、突变和癌症信息，和CpG岛（CGI）。图形界面methyview显示在基因组学和遗传学数据融合的研究中DNA甲基化 | 访问所有的数据和数据连接。CGI克隆和全球的CGI的预测，DNA甲基化数据，癌基因和突变的信息，DNA甲基化，基因表达和肿瘤。可以在特定组织中某些癌基因的表达状况中查询三个数据类型和研究甲基化之间的联系。可以定义特定的表达与启动子区域的甲基化数据的组织和基因，返回一个基因和甲基化数据的列表 | 20 000的CGI克隆序列的大规模测序，与UHN人类CpG岛微阵列数据库的17 132个CGI克隆序列。共有34738的CGI克隆。485个注释的癌基因，323个是由实验验证的甲基化数据，114个与 CGI的预测相匹配，6615候选的癌基因，3698个甲基化数据和1900个CGI甲基化预测数据，24020个其他基因。 | / |
| DiseaseMeth | 人；  72种人类疾病 | 基因组甲基化浏览器和定制的视图显示基因中心疾病甲基化信息结合基因组信息；通过它可以同时可视化多个基因组和表观基因组资源；还可以链接到其他表观基因组数据库。 | 通过基因ID和疾病名称；搜索工具检索启动子的甲基化状态当提供起始和末端染色体坐标时，可以检索给定染色体区域的所有甲基化状态。还可以检索所选疾病，组织，细胞系，技术和基因ID的数据。 | 支持（下载页面列出了有关数据集的详细信息，包括名称/ ID，疾病，数据分析，发布链接，实验平台，样本大小和下载链接） | / |
| DbDAD | 人，小鼠；  人类老年疾病 | ChainMap可视化工具，它可以显示甲基化途径的地图；浏览列表基本信息，如相关疾病，基因ID，甲基化风格（高甲基化或低甲基化）和原始文章名称。 | 通过基因ID和疾病名称；除基本信息更详细信息可以通过点击详细链接访问。详细信息包括致病部位，诱导因子，致病结果和实验室实验样品。 | / | / |
| MethDB | / | 详细的DNA甲基化数据序列信息和实验结果。 | 搜索DNA甲基化和环境表观遗传防治效果 | / | 在线提交功能 |
| MethBase | 人，老鼠 | MethBase作为UCSC基因组浏览器中的轨道中心，可作为科学社区公开获得。如果使用UCSC Genome Browser的主要网站，默认情况下内置MethBase跟踪中心 | 加载MethBase轨道中心设置文件并转到基因组浏览器页面后，它显示一组预选甲基化的甲基化水平轨迹和HMR轨迹；点击任何轨道转到该轨道的说明页面，其中提供有关该甲基酯的详细元和概要统计 | 可提供下载 | 对于每种甲基酯，Methbase在单个位点，等位基因特异性甲基化区域，低甲基化或高甲基化区域，部分甲基化区域以及详细的元数据和简要统计提供甲基化水平。 |
| Pubmeth | 癌症 | 选择基因或搜索癌症类型或亚型 | 1.指定感兴趣的基因，得到被选择的基因的总结概述；2.搜索癌症类型或亚型，查看是不是在PubMeth并用选择的癌症类型查询PubMeth |  | 主要面向各种文献而非数据。 |
| MethylomeDB | 人，小鼠；  精神分裂症和抑郁症甲状腺；  脑部； | 使用摆动和微阵列格式选项以单CpG分辨率查看甲基化数据。 | / | Methylome浏览器提供了表功能，方便下载。用户可以根据所选样品的位置下载全基因组或甲基化数据。高级用户可能希望下载原始数据用于更复杂的生物信息学分析。 “下载”页面提供了当前构建数据库中所有30个样品的甲基化数据的链接。 | / |

参考文献

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

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

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.