拟南芥 *clf-29* 突变体的转录组分析

完成人

卫子源

指导小组成员

董爱武　教授

目　录

摘　要

组蛋白 H3 第 27 位赖氨酸三甲基化（H3K27me3）是一种重要的表观遗传标记，与基因沉默相关。真核生物中，PRC2（Polycomb Repressive Complex 2）复合物负责 H3K27me3 的建立，其亚基 CLF（CURLY LEAF）是具有催化功能的组蛋白甲基化转移酶之一。我们希望利用高通量转录组数据（RNA-sequencing，RNA-seq），分析 CLF 蛋白的缺失在不同条件下，对植物全局基因表达的影响，挖掘 CLF 参与调控的生物学途径。以野生型拟南芥 Col-0 为对照，对拟南芥 *clf-29* 突变体进行了转录组测序分析，我们发现有 834 个基因表达上调，684 个基因表达下调。我们同时还对上调与下调的基因进行了聚类分析，发现部分功能与拟南芥 *clf-29* 突变体被报道的表型变化相对应。同时，我们也发现了 CLF 参与的其他生物学过程，为进一步拓展 CLF 的功能研究奠定了基础。

关键词： 拟南芥，H3K27me3，CLF，RNA-seq

Abstract

Trimethylation of lysine 27 of histone H3 (H3K27me3) is an important epigenetic marker and is thought to be involved in gene silencing. PRC2 (Polycomb Repressive Complex 2) is thought to play a significant role in establishing H3K27me3 in eukaryotes. Its subunit CLF (CURLY LEAF) is one of the histone methylation transferases with catalytic function. We hope to use the data of RNA-sequencing (RNA-seq) to analysis the effects of loss of CLF protein on global gene expression in plants under different conditions, and explore the biological pathways CLF regulated. Using wild-type Arabidopsis col-0 as a control, the result of RNA-sequencing analysis of clf-29 mutant in Arabidopsis thaliana showed that 834 genes were up-regulated and 684 genes were down-regulated. We also performed cluster analysis of up-regulated and down-regulated genes, and found that some changes of gene expression quantity corresponded to the reported phenotypic changes in *clf-29* Arabidopsis. At the same time, we also found some new biological processes, which laid a foundation for the further functional research of CLF.

Keywords： Arabidopsis，H3K27me3，CLF，RNA-seq

一、前言

1.1 表观遗传学

生物到底是由什么决定的？长久以来，这一问题一直被人们所关注。生物是从何而来的？为什么每个生命之间既有相同点又有不同点？到底是什么决定了生命？古往今来众说纷纭。随着科技的进步和人们的不断探索，虎克首次观察到了细胞。之后，施莱登和施旺提出了细胞学说，标志着人们对于生命的认识有了重要的进步。人们开始意识到生物是由细胞构成的，细胞的命运也决定了生物的命运。早在几个世纪以前，孟德尔通过豌豆的实验，在缺乏对细胞组成结构的认识的情况下，跨时代的提出了遗传因子成对存在的概念。之后，摩尔根与白眼果蝇的故事也被大家所知晓，这一果蝇的杂交实验，证明了基因在染色体上[1]。随着人们对微观世界探索不断深入，我们知道了 DNA 是遗传物质，进一步的，沃森和克里克又帮助我们了解了 DNA 的结构[2]。四种碱基的编列组合记载了遗传信息，DNA 转录为 RNA 进而翻译成为蛋白质行使功能。人们曾经认为，只要解析了全部的人类基因组数据，就可以完全掌握人类生命的奥义。然而，直到众多生物的基因组数据均已被解析的今天，我们对生命的认识依旧不完善，还有很多的谜团没有被解开。我们的基因数量并不多，但是仍然可以使整个生命系统正常运行[3]，这无疑警示我们 DNA 序列并不是影响生物性状的唯一方式。例如有实验证明，在老鼠怀孕期间，暴露在双酚环境下，会导致其后代产生肥胖，而且这种表型的变化在 F6 代之后才会恢复[4]。很多的证据证明，生物体会利用一些特殊的方式，使细胞“记忆”外界的变化或者响应环境的改变，进而调控基因的表达。在这些过程中，生物的 DNA 序列并没有发生改变，但是他们的基因表达水平却会产生变化，这就是表观遗传现象。表观遗传学就是研究在基因的核苷酸序列不发生改变的情况下，基因表达的可遗传变化的一门遗传学[5]。

表观遗传调控会伴随生物一生，真核生物会通过对 DNA 以及染色体进行修饰来调节基因的表达，而这些修饰被称为表观遗传标记。表观遗传标记包括 DNA 甲基化、组蛋白变体以及组蛋白修饰在内的多种形式（图 1）。DNA 不是无修饰的，我们现在的研究发现，虽然 DNA 可以简化为 ATCG 四种碱基的排列组合，但是除此之外还有很多其他的信息。在生物体中存在多种酶可以修饰 DNA，例

如替换掉 DNA 上的常规碱基，插入特殊碱基，或者对碱基进行修饰，例如对其甲基化等[6]。这些额外的修饰也会影响基因表达。除此之外，DNA 还会和特异的蛋白质结合，这些蛋白质会帮助 DNA 压缩并螺旋化形成染色质。生物的染色质的紧密程度会影响 DNA 序列的可接近程度，由此我们可以将其分为常染色质和异染色质，常染色质相对松散，RNA 聚合酶等蛋白易于接近，这些区域的基因也就更容易表达，而异染色质比较紧密，这一区域的基因表达困难。染色质的基本单位是核小体，核小体由 DNA 和缠绕其上的组蛋白构成。一般来说，有五种常规组蛋白：H1，H2A，H2B，H3，H4，其中 H2A，H2B，H3，H4 组成核心组蛋白八聚体，与 DNA 一起构成核小体，H1 位于核小体之间。组蛋白的 N 端残基会伸出核小体之外，这些残基之上往往存在多种共价修饰，如甲基化、乙酰化、磷酸化、泛素化等，这些修饰影响染色质的紧密程度，进而影响 DNA 的复制、修复和转录等过程[7]。生物体还会通过某种方式，用组蛋白变体替换掉常规组蛋白。目前包括 H2A，H3 等组蛋白均被发现存在多种组蛋白变体。这些组蛋白变体并非是原有组蛋白基因序列发生了突变，而是原本就存在的变体蛋白，生物体会利用他们调控基因的表达[8]。总而言之，生物体的表观遗传标记多种多样，功能不尽相同，其中组蛋白甲基化是目前研究相对较多的一种表观遗传标记，H3K27me3 就是其中一种。

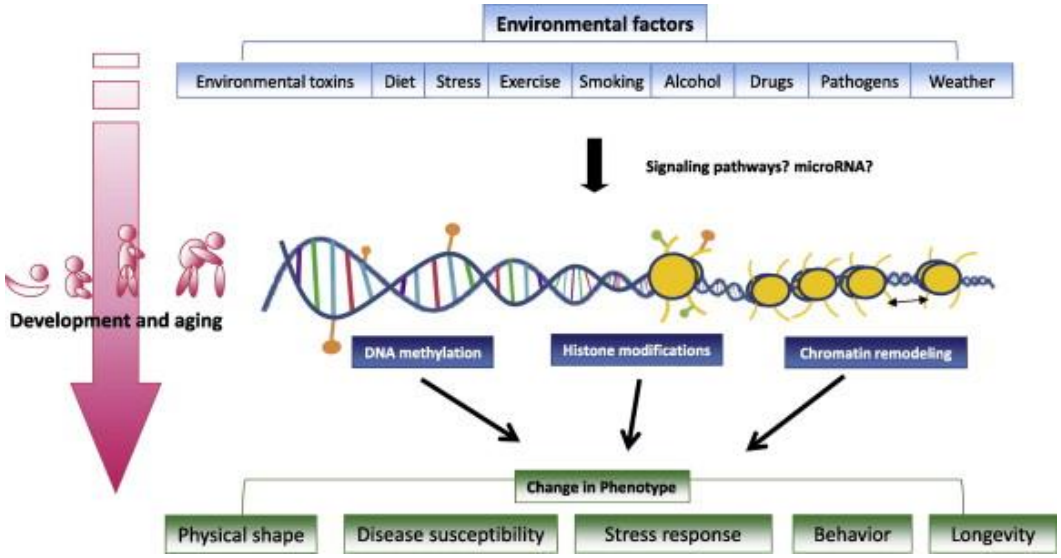

图 1：表观遗传与表观遗传标记[9]

表观遗传与外界的环境有关，会随着生物的发育和成长不断变化，影响表型。具体的影响方式包括 DNA 甲基化、组蛋白修饰甚至是染色质重塑。

1.2 H3K27me3

H3K27me3 代表 H3 组蛋白残基第 27 位赖氨酸的三甲基化修饰，这种表观遗传修饰被认为与基因表达的抑制有关。其在维持植物正常发育和生存上有重要作用，这种表观遗传标记的正常建立，被认为与植物的种子萌发、叶的生长以及正常开花等众多表型有关[10]。

在真核生物中，H3K27me3 由 PRC2 复合物负责建立。PRC2 全称 Polycomb Repressive Complex 2，是一种 Polycomb group (PcG) 蛋白复合物[11]，其在真核生物中保守存在。CURLY LEAF (CLF) 、SWINGER (SWN)和 MEDEA（MEA）是拟南芥的三种 PRC2 催化亚基，其中 MEA 仅在胚乳中发挥作用，而 CLF 和 SWN 在一般的植物发育过程中发挥作用，受特殊 DNA 序列等因素的招募，在特点位点建立 H3K27me3 修饰（图 2）。作为 PRC2 复合物的一部分，CLF 除了催化 H3K27me3 的建立，还被认为与体细胞同源重组有关[12]。也有相关的研究认为 CLF 蛋白也可能通过调节 AGAMOUS（AG）和 SHOOTMERISTEMLESS（STM）蛋白的表达来影响叶和花的形态[13][14]。而 CLF 也受其他因子调节，例如有研究发现 COOLAIR 可能与 PRC2 的募集有关[15]，染色质重塑蛋白 PICKLE 等会拮抗 CLF 蛋白的功能等[16]。综上，CLF 对表观遗传标记 H3K27me3 的正常建立对植物的正常发育有重要的作用，深入探究 CLF 的功能有重要的科学意义。

研究 CLF 蛋白，不仅仅是将其作为一种特殊的酶来探究其功能，由于它与表观遗传标记 H3K27me3 联系密切，对 CLF 的研究更偏向与对 H3K27me3 正常建立的作用的研究。CLF 蛋白的缺失，无疑会导致 H3K27me3 在拟南芥基因组上的分布和含量产生较大的变化，这种变化对植物来说意味着什么是很重要的。之前有很多的研究探究了 CLF 缺失会导致的植物表现变化，而我们更希望能从整体水平探究 CLF 缺失，对拟南芥的全局基因表达会有什么影响。

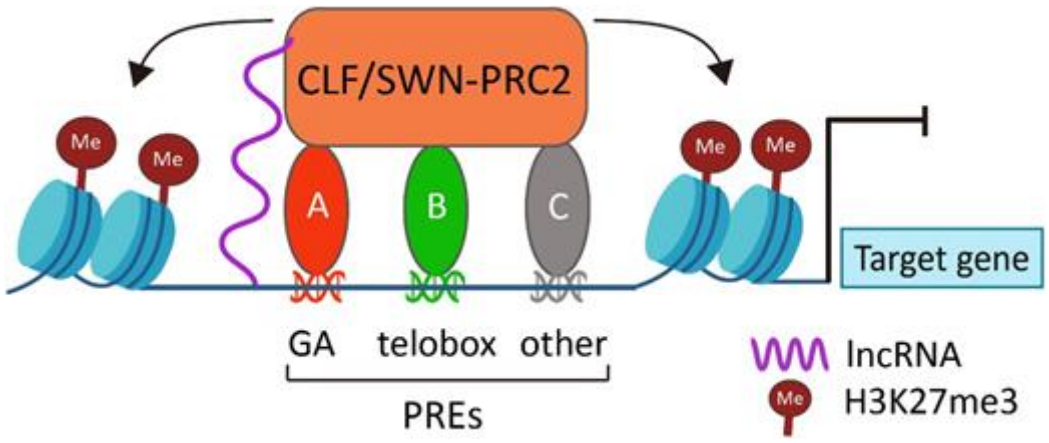

图 2：PRC2 建立 H3K27me3[17]

PRC2 复合物会建立 H3K27me3，主要由亚基 CLF 和 SWN 负责催化，其定位目前认为与特殊序列和 lncRNA 有关

1.3 高通量测序技术

除了常规的突变体构建与表型观察外，高通量测序技术的快速发展给我们的研究提供了新的思路。从很早以前，人们就对探明生物的基因组序列有着极高的热情，虽然现在已经证明仅仅知道人类的全基因组序列和人类有哪些基因是远远不够的。但是由此发展出的各种 DNA 测序技术以及现在的高通量测序技术却为我们现在进一步的研究提供了便利。

1975 年，Frederick Sanger 发明了 Sanger 双脱氧链终止法，这个方法的创立为第一代测序技术奠定了基础。其利用带荧光信号的 ddNTP 取代了 dNTP，使得被测序序列在扩增时反应终止，产生众多不同长度的核苷酸，再利用电泳分离，检测荧光信号，进而获取 DNA 序列[18]。一代测序准确率高，是一种金标准，但是其效率较低，测序通量低，如果想要测得一套基因组的数据需要花费很长的时间，这显然远远不能满足人们研究的要求。之后，以高通量为特点的新一代测序出现了[19]。二代测序利用桥式 PCR，为大量待测序的 DNA 片段加上接头，使其能固定在固体支撑物上，同时保障每个小区域仅有一条 DNA 片段，再利用带荧光的 dNTP 进行多轮的聚合-清洗-读取拍照来获取序列。二代测序技术能够对十几万到几百万条 DNA 分子进行测序。对于强调多样品，大数据分析的各种组学研究来说，二代测序无疑为科研提供了极大的方便。人们可以同时对好几个样品，一个家系，甚至是一类人群进行大规模的测序，从中利用统计学的方法获取想要的信息，这对于研究遗传病、肿瘤以及表观遗传都有很重要的作用。然而，二代

测序一个很严重的问题就是测序长度短，以至于需要研发许多的配套方法来尽可能的延长其能测序的 DNA 片段长度。随着人们对于测序长度的要求不断提升，三代测序也应运而生。三代测序又被称为单分子测序，利用 DNA 分子穿过纳米蛋白孔进行测序，其可以测序很长的序列，也适用于临床样品，但现在其测序成本较高，错误率也较高，还有待继续发展。

与此同时，配套的算法也在不断的精进，从提高 reads 片段的 mapping 效率，到提高差异基因分析的准确度，各样的分析软件都在同步的更迭，不断地提高计算速度，提高计算精度，提供个性化分析的方法。测序技术的发展，让我们能够从更多的角度，更高的层次去分析生物相关的问题。从大数据分析寻找基因突变位点，到转录组，表观遗传组的分析。现在，随着我们对生物的精细化研究，诸如单细胞测序，Hi-C 技术，空间转录组技术，甚至空间表观组学技术等也在不断发展。高通量测序技术已经能让我们从更加宏观，更加系统的层面去分析某个基因的作用，某种表观遗传标记的作用。

大规模测序技术的发展，允许我们对拟南芥的转录组进行分析[20]。这使得我们可以从比较宏观的方面去了解 CLF 缺失所导致的表观遗传变化究竟会从哪些角度影响植物正常生理功能。我们利用构建好的*clf-29*突变体拟南芥作为实验组，与作为对照组的野生型拟南芥共同进行 RNA 提取，得到两组共六份 RNA 样品送交公司进行测序，我们利用获得的高通量测序的结果，进行 RNA-seq 分析，对 CLF 蛋白所调控的基因进行较为全面地探究。

二、材料与方法

2.1 植物材料

本论文实验所涉及的拟南芥材料其生态型为 Col-0, 突变体 *clf-29* (SALK_N521003)订购于 ABRC（Arabidopsis Biological Reserch Center）种子库。文库构建所用材料为长日照条件（22℃，白光 16h，黑暗 8h）生长 14 天的拟南芥幼苗。

2.2 RNA-seq

在新一代高通量测序技术发展的背景下诞生的一项技术。通过提取样本的全 RNA，构建 cDNA 文库，利用高通量平台进行测序的方式，解读出含有大量 reads 的原始数据。再通过映射读数、汇总每个基因的读数计数、标准化和检测差异表达的基因来获取我们想要的信息[21]。该项技术已经被运用于包括临床研究[22]、动植物研究等多个方面。在植物领域，RNA-seq 已经被用于包括水稻[23]、玉米[24]、拟南芥等植物的研究，对于一些非模式生物的研究也有进展[25]。除了研究基因的表达量差异之外，可变剪切[26][27]，非编码 RNA[28]，环状 RNA[29]等的研究也有长足的进展。在这里，我们主要应用其分析突变体较野生型在基因表达上的变化。

我们使用天根生化科技有限公司开发的植物 RNA 提取试剂盒对 14 天的野生型 Col 以及突变体 *clf-29* 进行 RNA 抽提。得到高质量的 RNA 后，使用 KAPA 的 mRNA capture 磁珠和链特异性 mRNA-seq 建库试剂盒进行 RNA-seq 文库构建。文库扩增纯化后送上海晶能有限公司进行双端 150 bp 读长的高通量测序，测序深度为 15-30X。每个样本三个独立的生物学重复。

2.3 序列片段的质量检测与去接头

高通量测序并不是百分之百准确的，质量控制和预数据处理是必要的。以 Illumina 的测序为例，其能读数的碱基长度为 30-300bp。序列信息是由测序仪器内的可逆终止子循环反应产生，这些信号会以不同的比色信号形式呈现，不同颜色的荧光信号代表不同的碱基。除了碱基信息外，该测序数据还会附带一个质量数据，即该处碱基的置信度，错误率的范围从 7.94e-5 到 1。无论是杂质还是机器本身故障造成的误差，其都会提供 Q 值给我们作为参考，供我们判断此处的碱基测序结果是否可信[30]。因此，无论是单端测序还是双端测序，都需要先对原

始数据的测序质量进行控制。我们使用 fastqc 工具来帮助我们进行该步骤，该工具是一个基于 java 的软件，能够快速的帮助我们了解原始数据是否存在问题。

循环的过程中，通常会在被测序列的 3’段添加一个通用的引物。我们在之后会将测序得到的 reads 序列对应到基因组上，而该引物的序列无疑会影响该过程。我们利用 cutadapt 进行引物序列的切除。Cutadapt 的开发是在多特蒙德大学 Sven Rahmann 教授的团队中开始的，它可以通过容错的方式查找接头或引物序列来帮助完成这些序列去除，还可以通过各种方式修改和过滤单端和双端 reads[31]。在双端测序时，还需要去除第二条 reads 的 3’ 接头。我们所进行的是双端测序，去除的 3’接头为 AGATCGGAAGAGCACACGTCTGAACTCCAGTCA 与 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT。

2.4 reads 在基因组上的匹配

我们进行全转录组的高通量测序，是为了确定样品基因表达水平的高低。为此，我们需要将测序所得到的 reads 对应到基因组的基因上，再借由表达量和每个基因所对应的 reads 数的关系来判断每个基因的表达量高低。实现这一目的的第一步就是完成 reads 到拟南芥基因组上的对应。

目前有很多的对比用软件，例如 bwa，bowtie 等。这里我们选用的是 hisat2，这是一种快速而灵敏的软件，其基于 BWT 和 ferragina-manzini index 两种索引框架。在存在参考基因组时，hisat2 是最快速的 mapping 用工具[32]。我们所得是双端数据，设置“—rna-strandness RF”参数把干净的 reads 比对到拟南芥的 TAIR10 基因组上。

2.5 RPKM

我们所关注的是拟南芥基因的表达量，那么如何通过 reads 数来判断基因的表达量高低呢？一般来讲，如果一个基因的表达量高，那么在样本中提取到的相应的 mRNA 数量也会很多，测序测得的 reads 也会多，在利用 hisat2 进行匹配的时候 mapping 上的 reads 数也会多。但是与此同时，如果有一个基因特别的长，其转录产生的 mRNA 也很长，那么测序所得的 reads 数也会较多，而这一点与基因的表达量无关。为了避免这种影响，我们需要对数据进行标准化，RPKM 就是一种标准化的方式。RPKM 全称为 Reads Per Kilobase per Million mapped reads，其代表每百万 reads 中来自于某基因每千碱基长度的 reads 数。其公式为：

$$RPKM = \frac{\text{Total exon reads}}{\text{Mapped reads(Millons)} * \text{Exon length(Kb)}}$$

我们利用 deepTools 软件中的 bamCoverage 工具将比对基因组后的 bam 文件转换成 bigwig 文件。通过这种标准化的方式，可以反映出真实的基因表达量，但是这里还需要注意一些问题，测序的 RNA 库在不同的实验条件和/或测序协议之间可能存在显着差异，因此完全不同的两组数据并不能直接对比[33]。

2.6 数据可视化

Integrative Genomics Viewer (IGV)是一种高性能、易于使用的交互式工具，它可以将基因组数据进行可视化，便于使用者分析基因组数据的结果[34]。通过这种工具，我们可以容易的看出那些基因没有表达。

2.7 数据的统计学意义

在对数据进行基本处理后，我们需要确定实验组和对照组之间的比较是否是有意义的。我们选用了 plotCorrelation 来完成这件事。我们可以先利用 multiBamSummary 或 multiBigwigSummary 输出一个样本数据的集合矩阵。然后在 plotCorrelation 中利用 Spearman 方法计算相关系数，进而从统计学意义上看两组数据是否满足“组间数据差距较为明显，组内数据差距不明显”的特点。只有在满足这一条件后，我们才能认为对突变体和野生型这两组数据的对比是有意义的。我们选用的是 Spearman 方法，绘制了 heatmap。

2.8 差异基因的分析

在后期的分析中我们借助了 R 语言的部分工具，来帮助我们进行诸如火山图绘制、差异基因聚类分析等工作。先使用 FeatureCounts 软件[35]计算转录本上（所有外显子）的 reads 数，再利用 R 程辑包分析表达量差异。我们设定 |log 2 [Foldchange]| ≥ log 2 (1.5) 和 P 值 ≤ 0.05 作为筛选条件提取差异基因。在这里介绍主要的两个程辑包。

2.8.1 DESeq2

RNA-seq 需要对大量的数据进行处理和比较，我们需要对比实验组和对照组之间的基因表达水平的差异，以此来判断哪些基因在突变体中表达上调，哪些基因在突变体中表达下调，以此来明确 CLF 蛋白的作用。我们还需要对比组内三个重复样本之间的数据，以此来判断某个基因的表达量变化结果是否可行，排除偶然性的偏差的影响。DEseq2 是一种可以帮助我们处理 RNA-seq 数据的 R 语言

程辑包。他是原开发人员开发的第二代的 DEseq，具有很高的灵敏度和精度，以及较低的误报率，为 RNA-seq 数据的基因水平分析提供了一个全面而通用的解决方案[36]。

2.8.2 TopGO

在对突变体拟南芥的整体基因表达水平变化进行分析后，我们会希望能够进一步了解这些表达量上调和下调的基因分别有什么功能，借此来解释 CLF 蛋白缺失后突变体拟南芥的表型变化。我们更希望能够与拟南芥的 H3K27me3 的分布变化联系起来，进一步明确拟南芥 H3K27me3 的正常建立对植物表型正常的影响。因此，我们需要对表达量变化的基因进行聚类分析。TopGO 是一个 R 语言的程辑包，由 Adrian Alexa，Jorg Rahnenfuhrer 创建，可以帮助我们完成基因的聚类分析。我们选用 org.At.tair.db 数据库，BP 分类，将基因根据他们参与的生物过程进行聚类。

三、实验结果

3.1 数据质量控制

在对数据进行分析之前，我们首先需要确保获得的测序数据的质量。在高通量测序时，由于技术限制存在极限测序长度，一般来讲越往后出现测序错误的概率越大，如果该测序结果较差，那么最后将其对应到基因组上并进行分析，得到的结果就剩不可信的[30]。我们利用 fastqc 对原始数据进行质量检测（附图 1）。六份样品均为双端测序，原始测序数据的 DNA 片段平均长度约为 150bp，质量均处于较高水平，即碱基测序错误概率较小（图 3 A、B）。此外，考虑到测序时 DNA 片段的接头也会对后续的分析造成影响，我们也利用 cutadapt 对测序结果进行了接头的去除（附图 2）。相比较于原始数据，去掉接头的数据质量更高，平均每个碱基的测序可信度也有提升（图 3 C、D）。

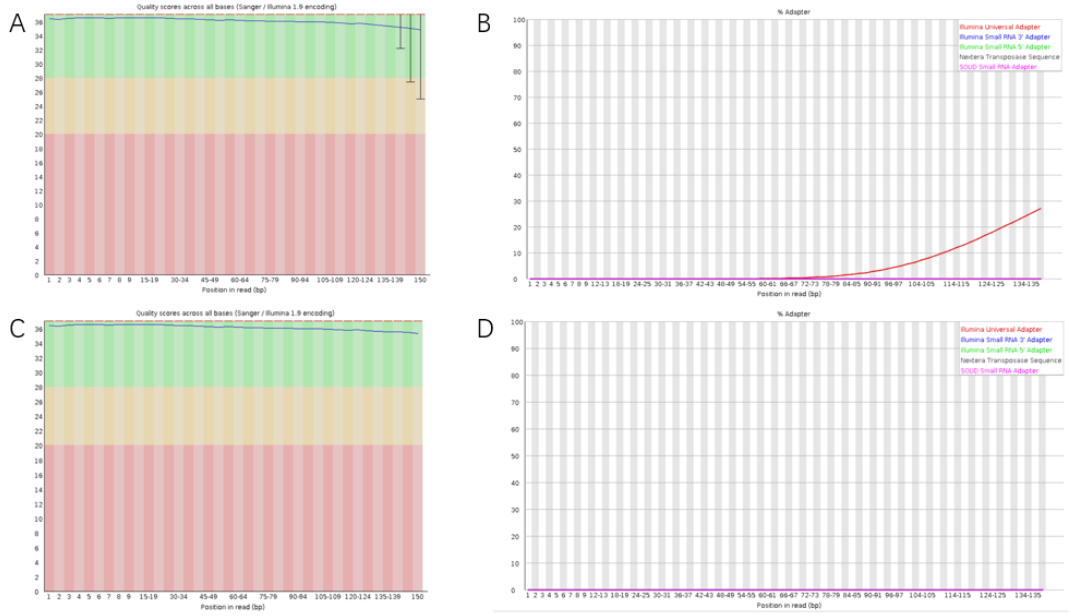

图 3：对原始测序数据的质量控制

（A 和 B）代表了数据在去接头前的整体数据质量；（C 和 D）为去接头处理之后的整体数据质量；（A 和 C）所示的图表示了测序在整体可信度，曲线在绿色部分表示可信度高，在完成去接头处理后，整体可信度均较高；（B 和 D）所示的图表示了接头对于原始数据的影响，由（D）图可见去接头处理效果较好

3.2 数据的初步处理

在确保原始测序数据的可信度较高后，我们利用获得的高通量测序数据进行拟南芥基因组的匹配。原始的数据是众多的碱基序列，此时我们并不能对其进行

分析，我们需要找到这些片段所对应的基因是什么。利用 histat2，我们成功获得了初步的 mapping 数据，结果显示这些测序结果的 mapping 率较高，均有超过96%的片段找到了其对应的在拟南芥基因组是的位置（表 1）。但是同时，许多片段的虽然能够成功的对映到基因组上，但是其并不能完全的匹配，可能存在错配的情况。我们进一步对 mapping 片段的质量进行检测，并过滤掉对比率较低的数据。在去除错误率大于 1%的数据后，数据的整体保留率依旧较高，大多数实验组的数据保留率都达到了 80%以上（表 1）。此外，测序深度也是评估一组数据是否可信的标准。我们需要保证测序所得的序列长度能够较好的覆盖拟南芥的基因组，因此我们利用测序序列的总长度与拟南芥基因组大小的比值作为测序深度。一般来讲，测序深度越高，证明所测得的序列对基因组的覆盖率越好[37]。如表 1，所有六组测序数据的测序深度均良好，表明该数据的覆盖度较好，进一步加强了该原始数据的可信性。

表 1：RNA-seq 数据处理汇总

	sequencing depth	raw	trimmed	trimmd%	mapping	mapping%	Q>=20	(Q>=20)%
clf29-1	43	34956528	34950686	99.98%	34048958	97.42%	31056109	88.86%
clf29-2	40	32170570	32163718	99.98%	31124830	96.77%	27521976	85.57%
clf29-3	41	33342874	33336138	99.98%	32489400	97.46%	29393087	88.17%
col-1	46	37244778	37242310	99.99%	36542155	98.12%	32494386	87.25%
col-2	28	22235624	22230806	99.98%	21583890	97.09%	19355789	87.07%
col-3	41	33187892	33184346	99.99%	32474201	97.86%	23442498	70.64%

注：该表为六组数据的测序深度，原始 reads 数，去接头后，基因组对应后，质量筛选后剩余的 reads 数以及和原有 reads 相比所占的比率。可以借此进一步分析数据的质量。

3.3 测序数据的可靠性分析

在保证了我们所获得的六组的数据的测序质量之后，我们需要进一步确定每一组的数据是否都达到了我们的实验要求，如特定基因是否没有表达，实验组和对照组之间的组间差异是否明显等。首先我们分析了六组数据在拟南芥基因组上的分布情况，利用 IGV 对 mapping 后的数据进行可视化，发现在 *CLF* 基因处，实验组的该基因表达量为 0，而对照组的有明显的峰值（图 4）。证明对照组的 *CLF* 基因正常表达，而实验组的 *CLF* 基因已经成功沉默，没有转录出对应 CLF

蛋白的 mRNA。

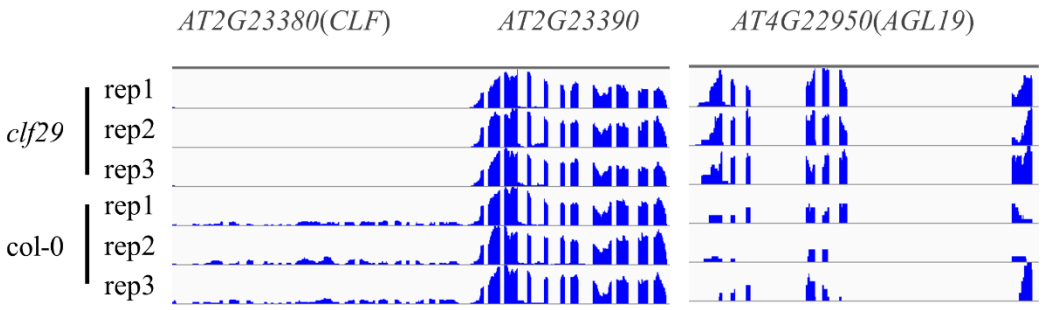

图 4: *CLF* 在突变体 *clf-29* 及野生型 Col-0 中的表达情况

我们利用 IGV 可视化 reads 在 mapping 后的结果，IGV 可以将每个基因每处的 reads 数通过峰值图的形式展示出来，图中 *clf-29* 组缺失的位置，即为 *CLF* 基因处，可知表达量为 0。*AT2G23390* 为 *CLF* 附近基因，表达量基本不受影响。*AGL19* 为被报道的与 *PRC2* 复合体有关的影响早花表型的基因之一[38]，突变体中表达上调

在确保了目标基因成功沉默之后，我们又分析了实验组与对照组之间的差异性。我们利用 Spearman Correlation 进行分析，发现实验组和对照组组内的差异相对较小，重复性好，而实验组和对照组之间的差异相对较大（图 5），说明我们可以通过对比两个组之间的差异，来分析 CLF 基因缺失对拟南芥转录组的影响。

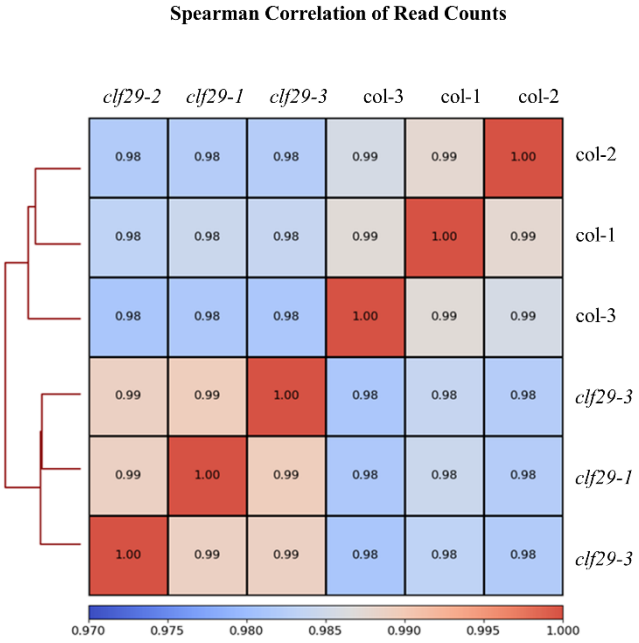

图 5:组间组内相关性分析

如图，我们利用 Spearman 相关性进行分析，发现实验组与对照组之间的组内相关性较

高，组间相关性相对较低。可知野生型与突变体之间存在较为明显的差异，可以进行进一步的对比进而得出结论

3.4 表达量差异分析

我们将此前 mapping 所得的数据进行 RPKM 的标准化。由此可以反应出拟南芥不同基因的表达量。我们将实验组的数据与对照组的数据进行分析。将实验组的各个基因表达量与对照组野生型的各个基因表达量进行对比，由此得到 *CLF* 突变后拟南芥基因表达的变化情况。我们观察到有 834 个基因表达上调，684 个基因表达下调（图 6）。

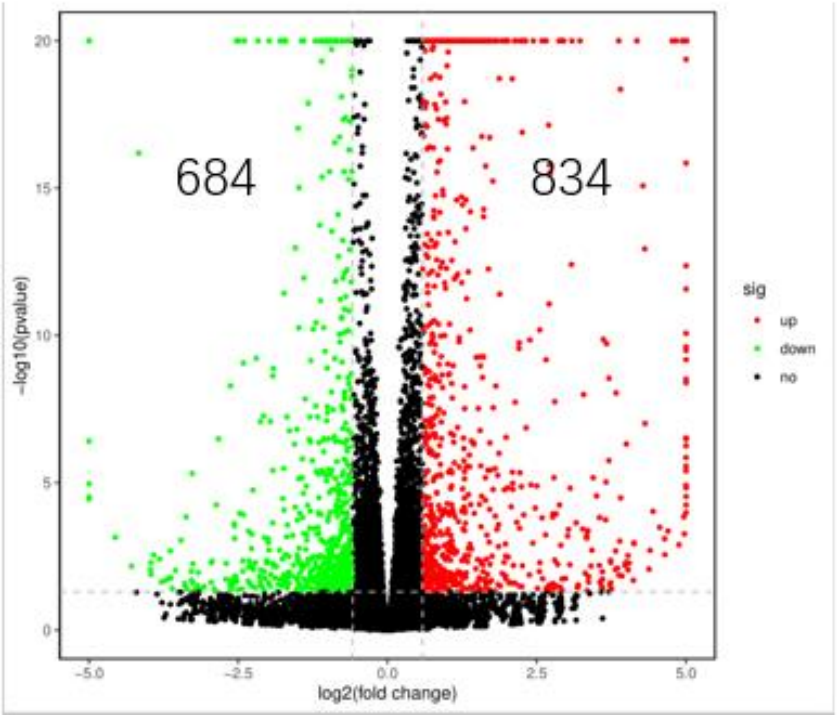

图 6：差异基因火山图

图中横轴代表突变体基因表达量相对与野生型的倍数，在原点右侧的代表表达量上调，在原点左侧的代表表达量上调。以变化倍数大于 1.5 倍为标准，即突变体基因表达量高于野生型一点五倍的为表达量上调，野生型表达量高于突变体 1.5 倍的为下调；纵轴代表可信度，越接近横轴的代表可信度越低，以 0.05 为界，置信度大于 0.95 的数据予以采纳

3.5 差异基因的聚类分析

我们进一步对于这些产生明显变化的基因进行 topGO 分析，发现在上调的基因中，与花的发育有关的基因数量是最多的（图 7 A）。这也与之前多项研究中发现的 H3K27me3 缺失，以及 PRC2 复合物缺失会导致拟南芥出现早花表现是

相吻合的[38]。除了与花的发育有关的基因之外，其他上调的基因大多是与生理生化反应，以及 DNA 和 RNA 正常行使功能有关的。这也可能表明 H3K27me3 不仅仅是通过自身的建立来抑制基因表达，也有可能是集中抑制和 RNA 合成有关基因，由此来使全局水平的基因表达受抑制。

除了上调表达的基因外，我们还探究了下调基因的特点（图 7 B）。我们发现下调基因多数是与植物应对生物胁迫或者非生物胁迫相关的。很多的基因与对抗外来生物的侵犯，如细菌、真菌等是有关的。这说明 CLF 与 H3K27me3 的建立与植物在发育较为成熟时的正常生理功能，抵抗外界环境都是有关系的。目前，对于 CLF 蛋白在应对胁迫时的功能研究较少，未来可以对 CLF 蛋白在这方面的功能进行更加深入的研究。同时，也有较多的下调基因可能是与植物对植物激素的反应有关的，包括生长激素，茉莉酸在内。这些基因与拟南芥的开花，生长，发育有关，他们的缺失也会导致拟南芥抗逆能力的下降。这也说明 CLF 蛋白有可能是通过影响激素发挥作用，进而影响植物表型的。

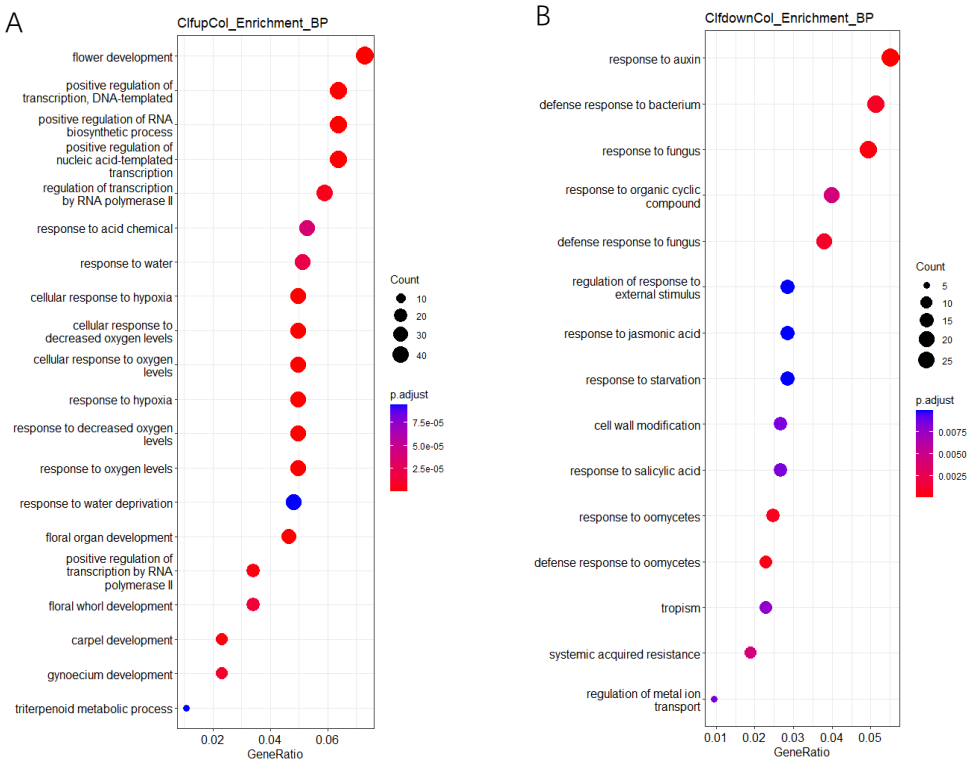

图 7：基因聚类分析

（A）为突变体中表达上调的基因聚类分析，（B）中为基因表达下调的聚类分析；从高到低，每种类型的基因数量依次减少

四、讨论

在本研究工作中，我们利用 RNA-seq 技术分析了在 CLF 蛋白缺失的条件下，拟南芥转录组的变化。CLF 蛋白被认为是与 H3K27me3 的建立有密切关系的，是 PRC2 复合物的组成部分之一，CLF 蛋白的缺失也就代表着正常的 H3K27me3 建立会部分受阻，而 H3K27me3 被认为是与基因转录抑制有关的。我们发现较多的基因在 CLF 蛋白缺失后表达量上调，这点与 H3K27me3 的功能以及 CLF 蛋白的功能是相吻合的。此外，我们对这些表达的基因进行了聚类分析，也发现上调和下调的基因有着明显的特点。而且 H3K27me3 的变化也会影响植物的抗逆能力，大部分此类基因在正常情况下是被抑制的，只有外界胁迫来临的时候才会被激活。在 CLF 蛋白缺失后，包括响应氧气水平，水以及化学酸的基因都有上调。与此同时，还有许多表达上调的基因是直接与 RNA 的合成以及转录的调控有关的。这也许表明 H3K27me3 在基因组上的分布有着更加明确的目的性，可能会偏向于与修饰某些基因来影响 RNA 的合成，或者影响抗逆相关通路，进而影响全局基因表达水平。另外需要注意一件事，虽然有 834 个基因表达下调，但是这个数量并不是非常多，因为除 CLF 之外，诸如 SWN 等也能参与 H3K27me3 的建立，在此次并未干预这些蛋白的正常表达。

与此同时，我们还注意到一件事情，虽然在基因表达量分析时观察到较多的基因是上调的，但是仍然有 684 个基因是下调的。例如，和茉莉酸以及生长素有关的基因下调。在过去的一些研究中，曾经发现在过表达 H3K27me3 去甲基化酶 ELF6 和 JMJ13 蛋白后，生长素和茉莉酸相关途径激活，这会影响植物的自花授粉比例[39]。但是在我们上述的研究中，发现有一些该功能相关的基因在 CLF 缺失后出现下调。包括一些响应胁迫的基因，而在 CLF 蛋白缺失后，也有很多该类型的基因表达下调。

这些下调基因的出现与预期结果是有一定出入的，那么他们下调的原因是什么呢？原因之一可能是，在缺失某些表观遗传标记构建蛋白时，突变体的整体表观遗传水平会下降，并不代表所有基因的表观遗传水平都会下降。在一些情况下，缺失关键性蛋白后，表观遗传标记的分布会产生较大的变化，甚至可能与正常情况下植物表观遗传标记的分布几乎完全不同。也就是说，CLF 蛋白以及 PRC2 复

合物，他们所起的作用不仅仅是建立 H3K27me3 这一表观遗传标记，而且还会影响 H3K27me3 的分布。即使全局水平的 H3K27me3 水平下降，依然会有一小部分基因的表观遗传修饰水平上升，进而影响一系列的基因表达和植物功能。

除此之外，与基因表达抑制有关的表观遗传修饰不止有 H3K27me3，包括 H2A.Z[40]，DNA 甲基化[6]等。有许多的表观遗传修饰的功能是类似的，在 H3K27me3 缺失时，有可能其他抑制性表观修饰会进行补充，甚至有可能对部分基因的抑制会更甚，反而导致这部分基因表达受抑制。除了表观遗传标记之外，许多基因有可能是属于同一通路相互调控的，部分基因的上调也有可能进而会导致另一部分基因的下调。即部分基因表达下调是另一部分基因表达上调的结果。在原定的实验与分析计划中，我们还会利用 ChIP-seq 技术对突变体的 H3K27me3 分布进行探究，我们可以借此详细了解下调基因和突变体 H3K27me3 富集基因之间的关系，但是由于疫情原因，这一部分数据尚不完善，有待进一步的探究。

对于 CLF 蛋白本身来说，还有很多值得探究的问题，例如 CLF 蛋白在染色质上的分布，之前已经有工作对 CLF 蛋白在基因组上的分布进行过研究，但是可能是受限于所选用的标签，其研究所获取的分布较为局限，位点相对较少[41]。此外，CLF 的结合蛋白也是值得发掘的，这也可以帮我们明确 H3K27me3 在特定位点上的精确建立机制，也可以与 GO 分析中找到的途径关联起来，帮助我们理解 CLF 与植物生理途径之间的关系。但是由于疫情原因，这些工作的样品收集工作无法进行，导致缺乏高通量测序的原始数据，无法进行分析，少了很多的结果且无法进一步的进行分析，也是十分遗憾。在未来，包括 CLF 的分布，定位的原理，什么样的外界条件会影响 H3K27me3 的建立等问题，都需要进一步的研究。

附图：

附图 1：去接头前全部测序数据的质量检测

（A-F）分别代表实验组三组与对照组三组去接头之前双端测序的结果，左侧为 R1 测序质量，右侧为 R2 测序质量

附图 2：去接头后全部测序数据的质量检测

（A-F）分别代表实验组三组与对照组三组去接头之后双端测序的结果，左侧为 R1 测序质量，右侧为 R2 测序质量。分布展示了整体测序质量与去接头质量

参考文献

1.Morgan TH. Sex limited inheritance in Drosophila[J]. Science, 1910, 32(812):120-122.

2.J.D. Watson, F.H.C. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid[J]. Nature, 1953, 171: 737

3.Pennisi E. Why do humans have so few genes[J]. Science, 2005, 309(5731):80.

4.Corces V, Yoon Hee Jung, Brianna J. Bixler, Daniel Ruiz, Hsiao-Lin V. Wang, Hannah Linsenbaum, et al. Transgenerational inheritance of BPA-induced obesity correlates with transmission of new CTCF sites in the *Fto* gene[J]. bioRxiv 2020.

5.Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape[J]. Cell, 2007, 128(4):635-638

6.Moore LD, Le T, Fan G. DNA methylation and its basic function[J]. Neuropsychopharmacology, 2013, 38(1):23-38.

7.Lawrence M, Daujat S, Schneider R. Lateral thinking: How histone modifications regulate gene expression[J]. Trends Genet, 2016, 32(1):42-56.

8.Redon C, Pilch D, Rogakou E, Sedelnikova O, Newrock K, Bonner W. Histone H2A variants H2AX and H2AZ[J]. Curr Opin Genet Dev, 2002, 12(2):162-169.

9.Tammen S, Friso S, Choi S. Epigenetics: The link between nature and nurture[J]. Mol Aspects Med, 2013, 34(4):753-764.

10.Molitor A, Shen WH. The polycomb complex PRC1: composition and function in plants[J]. J Genet Genomics, 2013, 40(5):231-238.

11.Jiao H, Xie Y, Li Z. Current understanding of plant Polycomb group proteins and the repressive histone H3 Lysine 27 trimethylation. Biochem Soc Trans[J]. 2020，8;48(4):1697-1706.

12.Chen N, Zhou WB, Wang YX, Dong AW, Yu Y. Polycomb-group histone methyltransferase CLF is required for proper somatic recombination in Arabidopsis[J]. J Integr Plant Biol, 2014, 56(6):550-558.

13. Goodrich J, Puangsomlee P, Martin M, Long D, Meyerowitz EM, Coupland G. A

Polycomb-group gene regulates homeotic gene expression in Arabidopsis[J]. Nature, 1997, 386, 44–51.

14. Schubert D, Primavesi L, Bishopp A, Roberts G, Doonan J, Jenuwein T, et al. Silencing by plant Polycomb-group genes requires dispersed trimethylation of histone H3 at lysine 27[J]. EMBO J, 2006, 25, 4638–4649.

15.Tian Y, Zheng H, Zhang F, Wang S, Ji X, Xu C, et al. PRC2 recruitment and H3K27me3 deposition at FLC require FCA binding of COOLAIR[J]. Sci Adv, 2019, 5(4):eaau7246.

16. Aichinger E, Villar CB, Di Mambro R, Sabatini S, Köhler C. The CHD3 chromatin remodeler PICKLE and Polycomb group proteins antagonistically regulate meristem activity in the Arabidopsis Root[J]. Plant Cell, 2011, 23(3):1047-1060.

17.Jie Shu, Chen Chen, Chenlong Li, Yuhai Cui; The complexity of PRC2 catalysts CLF and SWN in plants[J]. Biochem Soc Trans, 2020, 48 (6): 2779–2789.

18.F. Sanger, S. Nicklen, A.R. Coulson. DNA sequencing with chain-terminating inhibitors[J]. PNAS, 1977, 74: 5463

19. E.R. Mardis. Next-generation sequencing platforms[J]. Annu Rev Anal Chem (Palo Alto Calif), 2013, 6: 287

20.Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview[J]. Hum Immunol, 2021, 82(11):801-811.

21.Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data[J]. Am J Bot, 2012, 99(2):248-256.

22.Yamada S, Nomura S. Review of Single-Cell RNA Sequencing in the Heart[J]. Int J Mol Sci, 2020, 21(21):8345.

23. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq[J]. Genome Res, 2010, 20: 1238–1249.

24. Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, et al. The developmental dynamics of the maize leaf transcriptome[J]. Nature Genetics, 2010, 42: 1060–1067.

25.Strickler SR, Bombarely A, Mueller LA. Designing a transcriptome next-generation

sequencing project for a nonmodel plant species[J]. Am J Bot, 2012, 99(2):257-266.

26. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Mockler. Genome-wide mapping of alternative splicing in Arabidopsis thaliana[J]. Genome Res, 2010, 20: 45–58.

27.Wang W, Qin Z, Feng Z, Wang X, Zhang X. Identifying differentially spliced genes from two groups of RNA-seq samples[J]. Gene, 2013, 518(1):164-170.

28.Yamada M. Functions of long intergenic non-coding (linc) RNAs in plants[J]. J Plant Res, 2017, 130(1):67-73.

29.Luo J, Wang XL, Sun ZC, Wu D, Zhang W, Wang ZJ. Progress in circular RNAs of plants[J]. Yi Chuan, 2018, 40(6):467-477.

30. He B, Zhu R, Yang H, Lu Q, Wang W, Song L, et al. Assessing the impact of data preprocessing on analyzing next generation sequencing data[J]. Front Bioeng Biotechnol, 2020, 8:817.

31.Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads[J]. EMBnet J, 2011, 17:10–12.

32.Lachmann A, Clarke DJB, Torre D, Xie Z, Ma'ayan A. Interoperable RNA-Seq analysis in the cloud[J]. Biochim Biophys Acta Gene Regul Mech, 2020, 1863(6):194521.

33.Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols[J]. RNA, 2020, 26(8):903-909.

34.Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration[J]. Briefings in Bioinformatics, 2013, 14, (2):178–192

35.Liao Y, Smyth G K, Shi W. FeatureCounts: an efficient general-purpose program for assigning sequence reads to genomic features[J]. Bioinformatics, 2014, 30(7): 923-930.

36.Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. Genome Biol, 2014, 15(12):550.

37. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses[J]. Nat Rev Genet, 2014, 15(2):121-132.

38. Schönrock N, Bouveret R, Leroy O, Borghi L, Köhler C, Gruissem W, et al. Polycomb-group proteins repress the floral activator AGL19 in the FLC-independent vernalization pathway[J]. Genes Dev, 2006, 20(12):1667-1678.

39.Keyzor C, Mermaz B, Trigazis E, Jo S, Song J. Histone Demethylases ELF6 and JMJ13 Antagonistically Regulate Self-Fertility in Arabidopsis [J]. Front Plant Sci, 2021, 12:640135.

40.Long J, Carter B, Johnson ET, Ogas J. Contribution of the histone variant H2A.Z to expression of responsive genes in plants [J]. Semin Cell Dev Biol. 2022, S1084-9521(22)00132-X.

41. Shu J, Chen C, Thapa RK, Bian S, Nguyen V, Yu K, et al. Genome-wide occupancy of histone H3K27 methyltransferases CURLY LEAF and SWINGER in Arabidopsis seedlings[J]. Plant Direct, 2019, 3(1).

致谢

时光飞逝，日月如梭。大学本科四年转眼就要结束了，在这里有许多的经历。学到了许多，体验了许多，有收获，也有遗憾。繁花春水，具自飘落。在这里，我要向一些人表示真挚的感谢。

首先，我要感谢董爱武老师。在我真正开始科研相关的学习时给我提供了很大的帮助。她向我阐释了科学研究的精神，让我更深入的明白了什么是科学研究，科学研究的意义是什么。董老师对科学的态度和理解，也对我有很大的帮助，也对我未来的道路指出了方向。在我做毕业设计时，董老师从基础知识的学习到课题选取，以及具体的工作，都给我提供了无私的帮助。在这里我首先要向董老师表示真挚的感谢。

其次，我要感谢实验室里的师兄师姐。我首先要感谢谢文浩师兄，感谢他让我这个生信分析的小白，在不断地学习和试错中掌握了一些数据分析的方法。然后我要感谢李成章师兄和黄亚雪师姐，感谢李成章师兄提供的基础数据，让我能够完成毕业设计。同时也要感谢他们在实验上给予我的帮助，让我在湿实验上也学到了许多东西。我也要感谢杜康兮师兄，感谢他在论文修改上给我的建议。此外，我还要感谢其他的师兄师姐，感谢他们的无私帮助。

同时，我也要感谢我的同学和朋友，感谢他们在平时学习时给予我的帮助，感谢他们在我烦恼或者失落时给予我的支持，我们共同学习，共同前进，鲜衣怒马，不负韶华。

最后我尤其要感谢我们家人，我的父母。没有他们的支持和帮助，我也不可能到今天。希望我未来也能够不失所望，实现抱负，藉以报答。