Table of Contents

Introduction

第一部分 项目概况

1.1 项目概况

项目编号	**xxxxx**
开题单号	xxxxx
项目类型	有参基因转录组测序
物种名称	
样本形式	
测序平台	
数据量	
分析项目	
分析者工号	
完成日期	
产品经理	
技术支持	xxx

1.2 实验流程

我们从样本提取到上机测序，有严格的样本检测、质控流程，通过各环节对样本质量控制， 确保数据的真实可信。操作流程如下:

graph TD subgraph 数据质控和过滤 A[测序下记数据] -->B[数据过滤] B[数据过滤]-->C[数据质量评估] end subgraph 参考基因组比对 C[数据质量评估] --> D[比对到基因组] D[比对到基因组] --> E[比对结果评估] end subgraph 分析 E[比对结果评估] --> F[表达量计算] F[表达量计算] --> G[差异表达分析] G[差异表达分析] -->H[GO/KEGG富集分析] G[差异表达分析] --> I[聚类分析] end

本项目通过 Oligo(dT) 磁珠富集总 RNA 中带有 polyA 结构的 mRNA，采用离子打断的方式，将 RNA 打断到长度300bp左右的片段。选择长度为300bp的片段，这是因为接头长度是固定的，如被打断的片段长度较短，将导致接头序列的比例偏高，从而降低了有效数据的比例；如被打断的片段长度较长，则不利于上机测序过程中簇的生成。以 RNA 为模板，用6碱基随机引物和逆转录酶合成 cDNA 第一链，并以第一链 cDNA 为模板进行第二链 cDNA 的合成。

文库构建完成后，采用PCR扩增进行文库片段富集，之后根据片段大小进行文库选择，文库大小在450bp。接着，通过 Agilent 2100 Bioanalyzer 对文库进行质检，再对文库总浓度及文库有效浓度进行检测。然后根据文库的有效浓度以及文库所需数据量，将含有不同 Index 序列（各样本加上不同的 Index，最后根据 Index 区分各样本的下机数据）的文库按比例进行混合。混合文库统一稀释到2nM，通过碱变性，形成单链文库。

样品经过 RNA 抽提、纯化、建库之后，采用第二代测序技术（Next-Generation Sequencing，NGS），基于 Illumina HiSeq 测序平台，对这些文库进行双末端 （Paired-end，PE）测序。

1.3 参考基因组

参考基因组信息

Genome	Vitis_vinifera.12X.dna.toplevel.fa
Genebuild by	Ensembl
Database version	96.3
Base Pairs	486,175,922

参考基因组注释信息

Database	Number	Percentage
xxx	111	111
Xxx	111	111
Xxx	111	111
Xxx	111	111
xxx	111	111

1.4 有参转录组分析流程

首先对原始下机数据（Raw Data）进行过滤，将过滤后得到的高质量序列(Clean Data)比对到该物种的参考基因组上。

根据比对结果，计算每个基因的表达量。在此基础上，进一步对样品进行表达差异分析、富集分析和聚类分析。对比对上的 Reads 进行拼接，还原出转录本序列。

graph TD subgraph 数据质控和过滤 A[测序下记数据] -->B[数据过滤] B[数据过滤]-->C[数据质量评估] end subgraph 参考基因组比对 C[数据质量评估] --> D[比对到基因组] D[比对到基因组] --> E[比对结果评估] end subgraph 分析 E[比对结果评估] --> F[表达量计算] F[表达量计算] --> G[差异表达分析] G[差异表达分析] -->H[GO/KEGG富集分析] G[差异表达分析] --> I[聚类分析] end

第二部分 原始数据整理

2.1 原始数据整理

2.1.1 文库基本情况

Sample	Lib Name	Lib Insert Size	Sequencing Platform	Sequencing Mode
xxx	xxx	xxx	xxx	xxx

注:

- Sample，样品名称
- Lib Name，文库名
- Lib Insert Siz，文库插入片段长度
- Sequencing Platform，测序平台
- Sequencing Mode，测序方式

2.1.2 数据整理

样品经过上机测序，得到图像文件，由 测序平台自 带软件进行转化，生成 FASTQ 的原始数据（Raw Data），即 下机数据。我们对每个样品的下机数据（Raw Data）分别进行统计，包括样品名、 Q30 、模糊碱基所占 百分比、以及 Q20(%) 和 Q30(%) 。

统计结果:

Sample	Read No	Base(bp)	Q20	Q30
G42H-1	49019750	7352962500	7108881025	6751905726
G42H-2	48002056	7200308400	6957803614	6613248663
G42H-3	49674216	7451132400	7212856698	6867699302
G42L-1	38428812	5764321800	5592675905	5322828533
G42L-2	46586046	6987906900	6798398700	6489313303
G42L-3	40635504	6095325600	5913343486	5636312799
G63H-1	56699692	8504953800	8245831114	7898876218
G63H-2	38238760	5735814000	5560556709	5320309608
G63H-3	52899018	7934852700	7584530425	7243930176
G63L-1	45463306	6819495900	6635563488	6365612382

注:

- Sample，样品名称
- Read No，Reads总数
- Base(bp)，碱基总数
- Q20，碱基识别准确率在99%以上的碱基总数
- Q30，碱基识别准确率在99.9%以上的碱基总数
- Q20，碱基识别准确率在99%以上的碱基所占的百分比
- Q30，碱基识别准确率在99.9%以上的碱基所占的百分比

2.1.3 数据过滤

测序数据包含一些带接头、低质量的 Reads，这些序列会对后续的信息分析造成很大的干扰，因此需要对测序数据进行进一步过滤。

主要处理过程包括:

- 过滤掉低质量，太短和太多N的"bad reads"
- 从头部(5')或尾部(3')计算滑动窗内的碱基质量均值，并切除低质量碱基
- 头尾剪裁（上下机序列可能不稳定）
- 去除接头
- 不匹配碱基对矫正
- 修剪尾部(3')的polyG（修剪ployX尾，可以去掉不想要的ploy。如：mRNA-Seq 中的ployA）
- 处理使用了唯一分子标识符（UMI）的数据，并将UMI转换为序列名称

Sample	Read No	Base(bp)	Q20	Q30
G42H-1	47791424	7157479504	6974419410	6637845191
G42H-2	46686136	6992210290	6815114985	6492335743
G42H-3	48464034	7257044527	7079814582	6754855716
G42L-1	37723104	5654419595	5517038300	5258614726
G42L-2	45884794	6877245244	6723337529	6426057958
G42L-3	39754954	5958889600	5819237005	5556269902
G63H-1	54920952	8140924783	7965881594	7653112302
G63H-2	37257718	5541201516	5416821502	5195961828
G63H-3	47884648	6808921603	6674441178	6427487681
G63L-1	44591548	6650727919	6511153568	6258595421

注:

- Sample，样品名称
- Read No，Reads总数
- Base(bp)，碱基总数
- Q20，碱基识别准确率在99%以上的碱基总数
- Q30，碱基识别准确率在99.9%以上的碱基总数
- Q20，碱基识别准确率在99%以上的碱基所占的百分比
- Q30，碱基识别准确率在99.9%以上的碱基所占的百分比

2.1.4 碱基质量分布

测序错误率受测序仪本身、测序试剂、样品等多个因素共同影响。 对于RNASeq 技术，测序错误率分布具有两个特点:

- 测序错误率会随着测序序列的长度增加而升高，这是由测序过程中化学试剂的消耗导致的，是 Illumina 高通量测序平台都具有的特征
- 前6个碱基的位置（即建库过程中反转录所需要的随机引物的长度）也会发生较高的测序错误率，这种错误是由随机引物和 RNA 模板的不完全结合引起的

我们用测序数据的单碱基质量分布图评价单个位置的碱基质量。一般而言，Reads 的 5’端和 3’端的碱基质量较低，中间部分的碱基质量较高。大部分序列的碱基质量在20以上，代表测序质量较好。

质控和过滤前质量分布图:

▶ G42H-1
▶ G42H-2
▶ G42H-3
▶ G42L-1
▶ G42L-2
▶ G42L-3
▶ G63H-1
▶ G63H-2
▶ G63H-3
▶ G63L-1
▶ G63L-2

质控和过滤后质量分布图:

▶ G42H-1
▶ G42H-2
▶ G42H-3
▶ G42L-1
▶ G42L-2
▶ G42L-3
▶ G63H-1
▶ G63H-2
▶ G63H-3
▶ G63L-1
▶ G63L-2

2.1.5 碱基含量分布

碱基含量分布一般用于检测有无AT、GC分离现象。对于RNASeq来说，鉴于序列打断的随机性和G/C、A/T含量分别相等的原则，理论上每个测序循环中的GC含量相等、AT含量相等（如果是链特异性建库，可能会出现AT分离和/或GC分离），且在整个测序过程基本稳定不变，呈水平线。但在现有的高通量测序技术中，反转录合成 cDNA 时所用的6bp的随机引物会引起前几个位置的核苷酸组成存在一定的偏好性，这种波动属于正常情况。碱基含量分布结果见图

质控和过滤前碱基含量分布图:

- G42H-1
- G42H-2
- G42H-3
- G42L-1
- G42L-2
- G42L-3
- G63H-1
- G63H-2
- G63H-3
- G63L-1

质控和过滤后碱基含量分布图:

- G42H-1
- G42H-2
- G42H-3
- G42L-1
- G42L-2
- G42L-3
- G63H-1
- G63H-2
- G63H-3
- G63L-1
- G63L-2

2.2 比对分析

2.2.1 比对结果基本统计

使用 TopHat2的升级版HISAT2（http://ccb.jhu.edu/software/hisat2/index.shtml) 软件将过滤后的 Reads 比对到参考基因组上。HISAT2使用改进的BWT算法（Sirén et al. 2014）具有更快的速度并且资源占 用较少。 HISAT2比对时，对于非链特异性文库使用默认参数，链特异性文库需要指定文库类型（即first 使用 --rna-strandness RF，second 使用--rna-strandness FR）。若参考基因组选择合适，且相关实验不存在污染，测序序列的 Mapping 比例一般会高于70%。

Mapping 比例较低时的原因可能是:

- 参考基因组组装不好，或者所测物种与参考基因组的亲缘关系较远；
- 样品的特殊前处理或者相对于参考基因组此样品本身的变异太大，导致 Mapping Rate 相对较低。

比对结果为Bam文件。序列比对的基本信息统计:

Sample	Clean Reads	Total Mapped	Multiple Mapped	Uniquely Mapped
xxxxx	12121212	12121212	2222222	22222222

2.2.2 比对区域分布统计

2.2.3 基因覆盖均一度

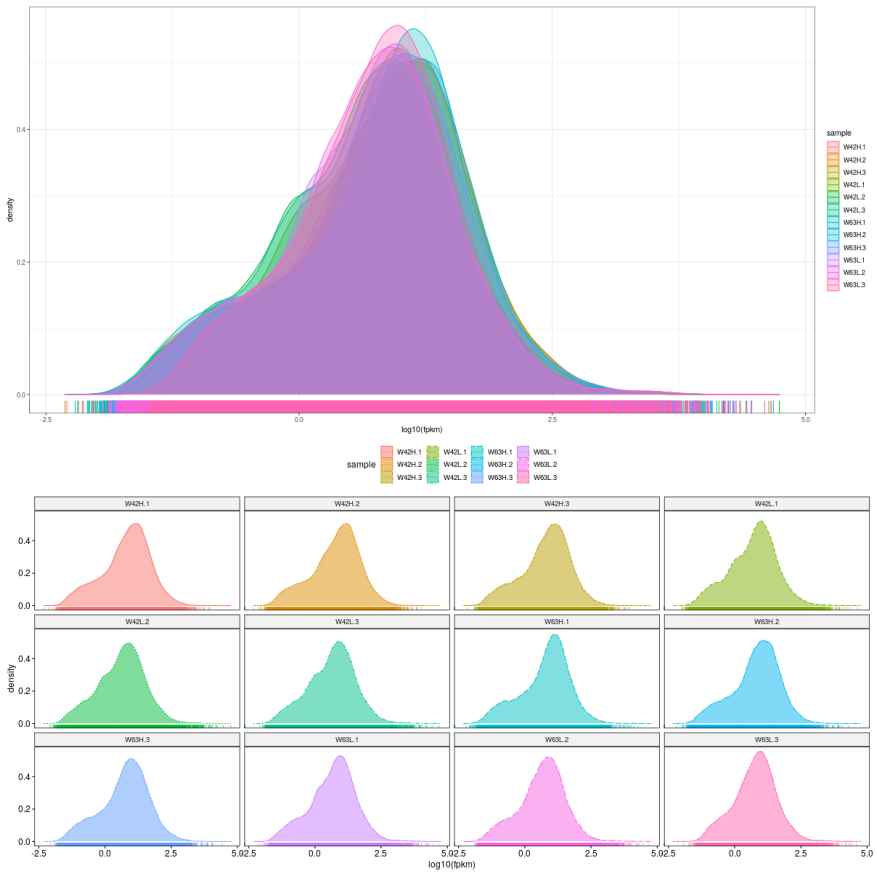
第三部分 差异分析

3.1 表达量分析

3.1.1 表达量分析

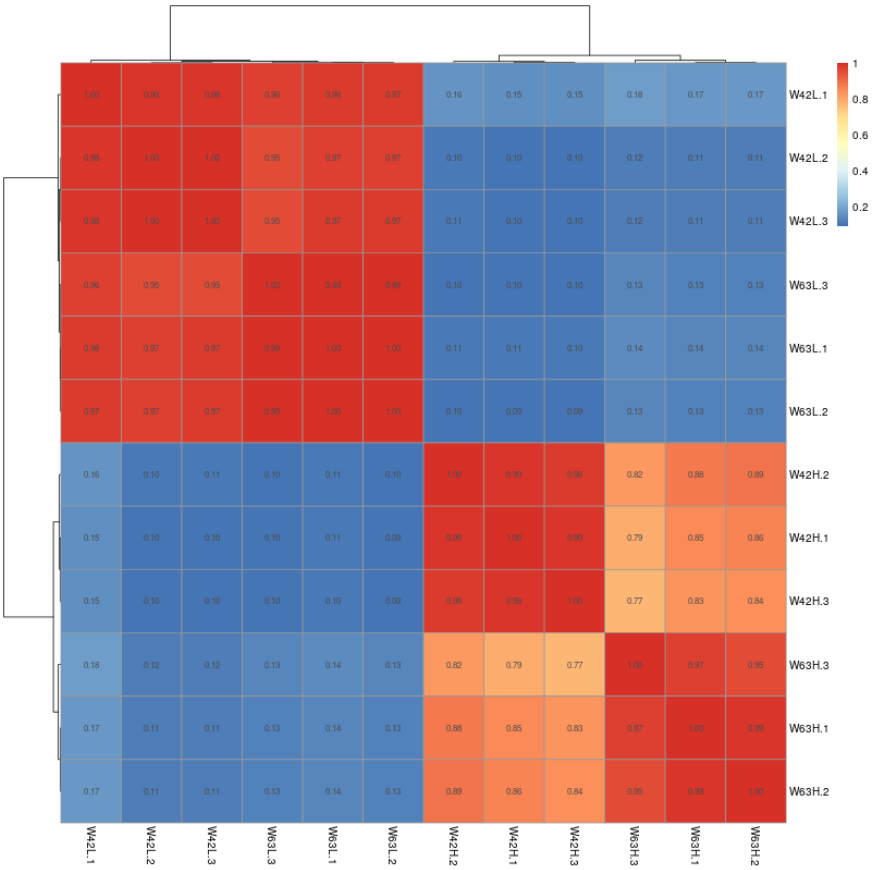
3.1.2 FPKM密度分布

一般认为转录组的表达分析包括三个水平层面，即基因表达水平，转录本表达水平，外显子表达水平。同一个基因不同剪切形式的表达，可能引起截然不同的生物学效应。FPKM密度分布能整体地考察样品所有基因的表达量模式，一般来说中等表达的基因占绝大多数，低表达和高表达的基因占一小部分。基因在各个样品中的表达量特征统计结果见图



3.1.3 样品相关性检验

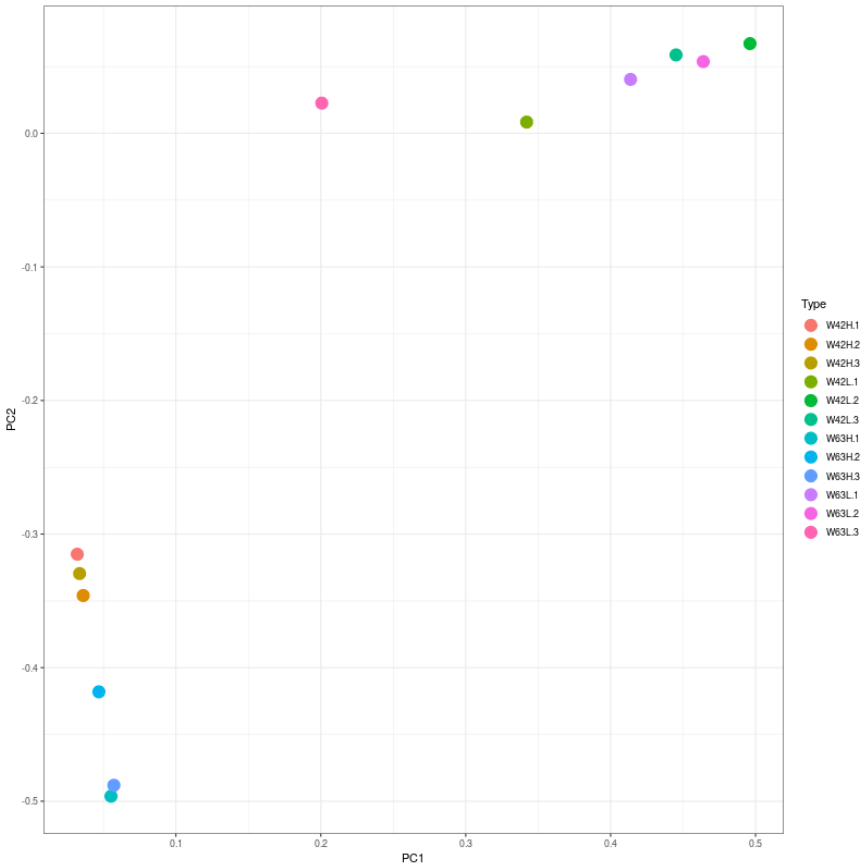
样品间基因表达水平相关性是检验实验可靠性和样本选择是否合理的重要指标，在做差异表达分析之前，应先检查样品间基因的表达水平相关性。我们用皮尔逊相关系数表示样品间基因的表达水平相关性，相关系数越接近1，表明样品之间表达模式的相似度越高。一般来说，相关系数在0.8-1之间属于极强相关，如果生物学重复的样本之间相关系数低于0.8，表示样品之间的相关性较低，见图



3.1.4 PCA分析

PCA 主成分分析（Principal Components Analysis），通过线性变换，将高维数据降低至二维或三维，同时保持各方差贡献最大的特征，即降低数据复杂度。当有多个样品时，我们使用R语言的 DESeq软件包，根据表达量对各样品进行 PCA 主成分分析。PCA 分析可以把相似的样本聚到一起，距离越近表明样本间相似性越高。

结果见图



3.2 表达差异分析

3.2.1 样品差异表达检测

3.2.2 聚类分析

3.3 差异表达基因功能富集分析

3.3.1 GO富集分析

3.3.2 KEGG富集分析

第四部分 附录