Transcriptome Denovo Project

BGI Co., Ltd.

October 23, 2018

Project: **RNA deneovo**

Customer: **Yue Yao**

Company/Institute: **华大基因科技有限公司**

Project Code: **RNA-denovo-Rubber**

Oraganism: **Rubber**

Contents

1 分析结果

1.1 摘要

本项目使用 Illumina Hiseq 平台一共测了 4.58Gb 数据。组装并去冗余后得到 44,842 个 Unigene，总长度，平均长度，N50 以及 GC 含量分别为 39,619,114 bp，883bp，1,618bp 和 37.35 %。然后将 Unigene 比对到七大功能数据库进行注释，最终分别有 24,679(NR:55.04%)，12,614(NT: 28.13%)，20,152(Swissprot: 44.94%)，8,544(COG:19.05%)，19,226(KEGG: 42.87%)，6,207(GO:13.84%) 以及 19,339(Interpro:43.13%) 个 Unigene 获得功能注释。根据注释结果共检测出 24,730 个 CDS，未注释上的 Unigene 使用 ESTScan 预测后获得 3,199 个 CDS。同时还检测出 3,768 个 SSR 分布于 3,088 个 Unigene 中。

1.2 测序数据过滤

测序的原始数据包含低质量、接头污染以及未知碱基 N 含量过高的 reads，数据分析之前需要去除这些 reads 以保证结果的可靠性。过滤后 reads 的质量指标见表 1，碱基含量分布以及质量分布见图 1 和图 2。

Table 1: 过滤后的 reads 质量统计

Sample	Total Raw Reads(Mb)	Total Clean Reads(Gb)	Total Clean Bases(Gb)	Clean Reads Q20(%)	Clean Reads Ratio(%)
SA	35.41	30.52	4.58	99.17	86.18

[1] Q20: 质量值大于 20 的碱基数目占总碱基数目的比例.

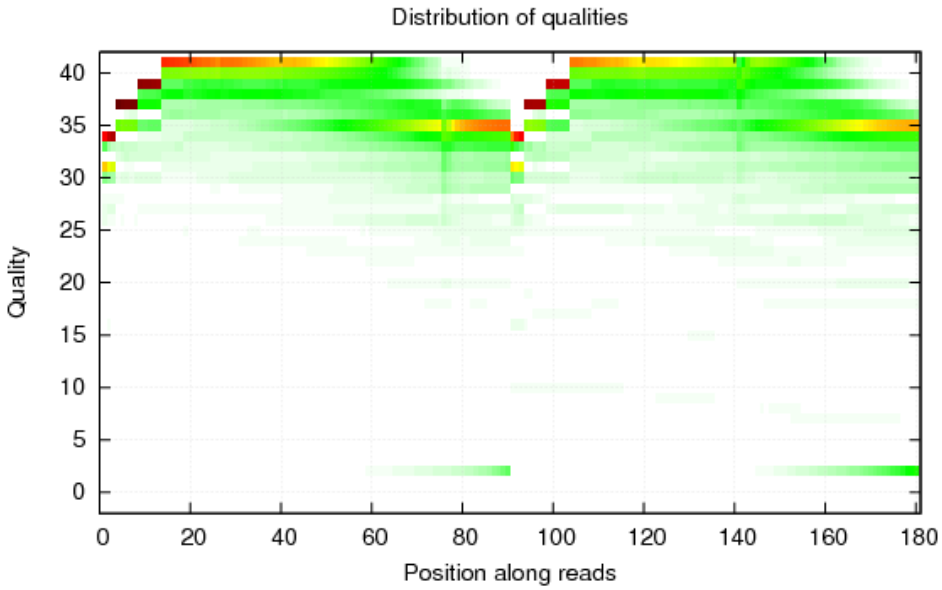

Figure 1: Clean reads 的碱基含量分布图。X 轴代表碱基在 read 中的位置，Y 轴代表此类碱基的含量比例。正常情况下，reads 每个位置的碱基含量分布稳定，无 AT 或 GC 分离现象。由于 Illumina 平台在 RNA-Seq 测序中，反转录成 cDNA 时所用的 6bp 随机引物会引起前 6 个位置的 GC 含量组成存在偏好性，故图中前 6bp 的波动为正常现象。

1.3 Denovo 组装

数据过滤后, 我们使用 Trinity [15] 对 cleanreads 进行组装, 组装质量指标见表 2 , 组装的转录本长度分布见图 3 . 接下来我们使用 Tgicl [16] 对转录本进行聚类去冗余得到 Unigene，聚类后的 Unigene 质量指标见表 3 , Unigene 的长度分布见图 4 。(对于多个样品的研究，我们使用 Tgicl 对每个样品的 Unigene 进行再一次的聚类去冗余得到最终的 Unigene 用于后续分析，命名为"All-Unigene")

Table 2: 转录本的质量指标

Sample	Total Number	Total Length	Mean Length	N50	N70	N90	GC(%)
OB	71,815	51,583,421	718	1,359	655	261	37.34

[1] N50: 用于衡量组装的连续性，数值越大说明组装效果越好，计算方法为：按转录本长度从大到小排序后逐个累加至所有转录本总长度的 50% 时，最后一个累加的数值大小即为 N50。

[2] GC(%): 碱基 G 和 C 的比例。

Table 3: Unigene 的质量指标

Sample	Total Number	Total Length	Mean Length	N50	N70	N90	GC(%)
OB	71,815	51,583,421	718	1,359	655	261	37.34

[1] N50: 用于衡量组装的连续性，数值越大说明组装效果越好，计算方法为：按转录本长度从大到小排序后逐个累加至所有转录本总长度的 50% 时，最后一个累加的数值大小即为 N50。

[2] GC(%): 碱基 G 和 C 的比例。

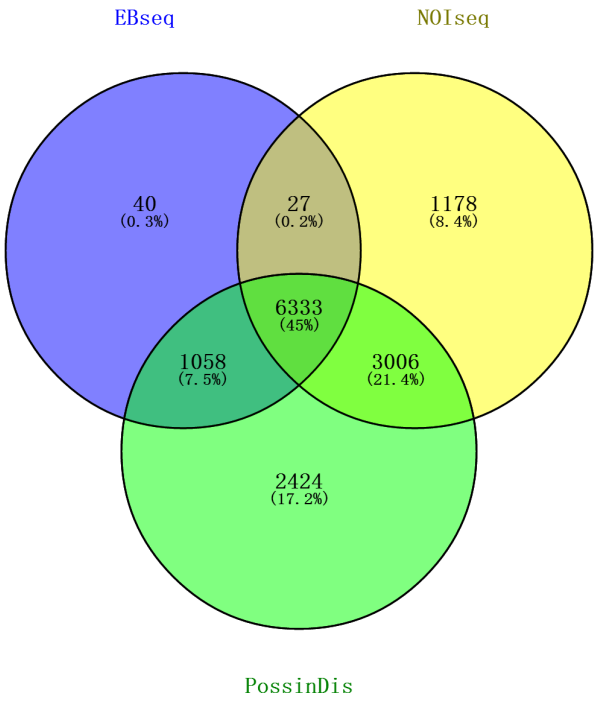

Figure 2: Unigene 的长度分布图。X 轴代表 Unigene 长度，Y 轴代表相应 Unigene 的数目。

1.4 Unigene 功能注释

组装完毕后，我们将对组装得到的 Unigene 进行七大功能数据库注释 (NR, NT, GO, COG, KEGG, Swissprot and Interpro)，注释结果见表 4。根据 NR 注释结果，我们统计了注释结果的物种分布，见图 5。根据 COG、GO 和 KEGG 注释结果，我们统计了各自的功能分类，见图 6，图 7 和图 8。同时，维恩图来展示 NR、COG、KEGG、Swissprot 以及 Interpro 的注释结果，见图 9。

Table 4: 功能注释的统计结果

Values	**Total**	**Nr**	**Nt**	**Swissprot**	**KEGG**	**COG**	**Interpro**	**GO**	**Overall**
Number	51,582	25,417	15,359	21,425	20,249	9,744	20,199	5,802	29,407
Percentage	100%	49.27%	29.78%	41.54%	39.26%	18.89%	39.16%	11.25%	57.01%

[1] Overall: 被七大数据库中任意一个数据库注释上的 Unigene 总数。

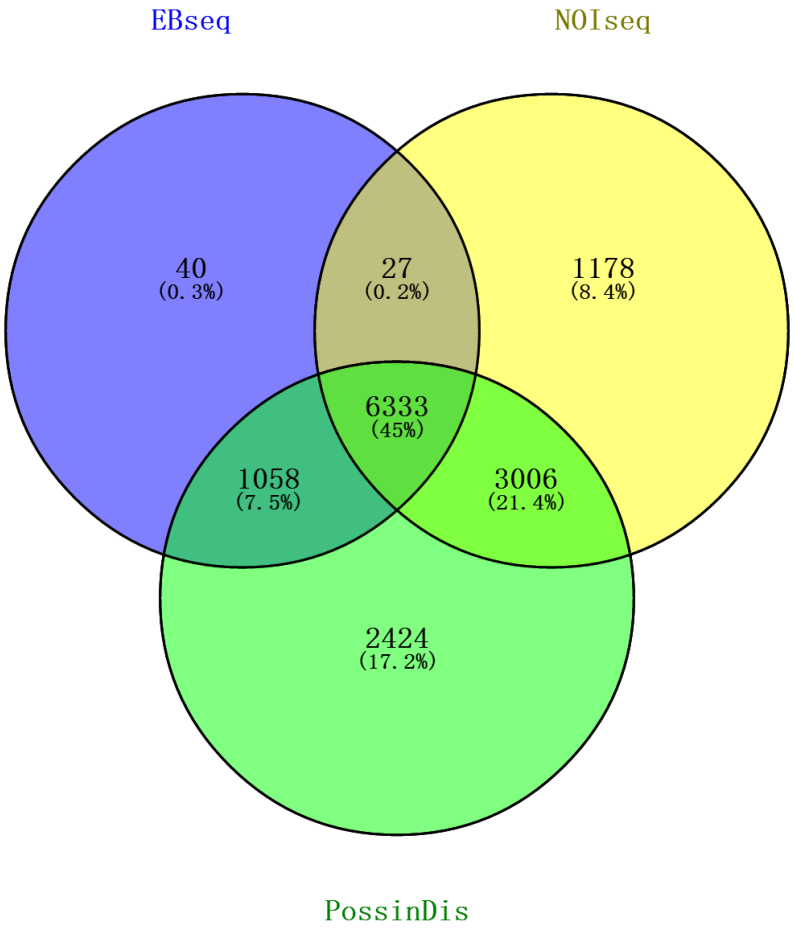

Figure 3: 注释物种分布

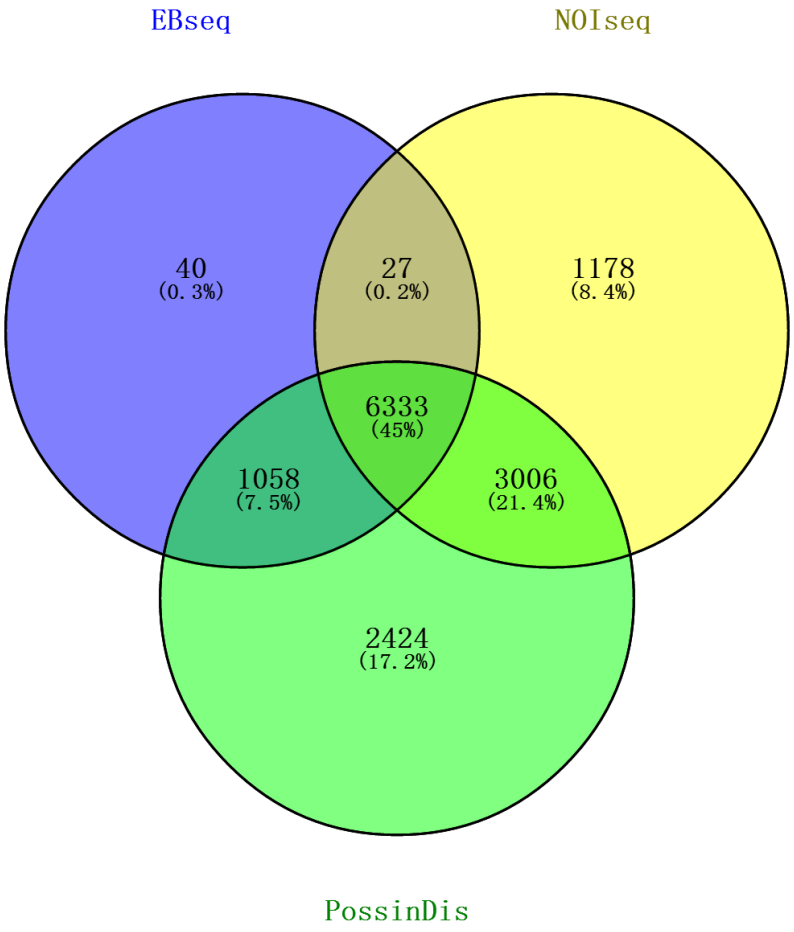

Figure 4: COG 功能分布统计图。X 轴表示相应的 Unigene 基因数，Y 轴表示相应的 COG 分类

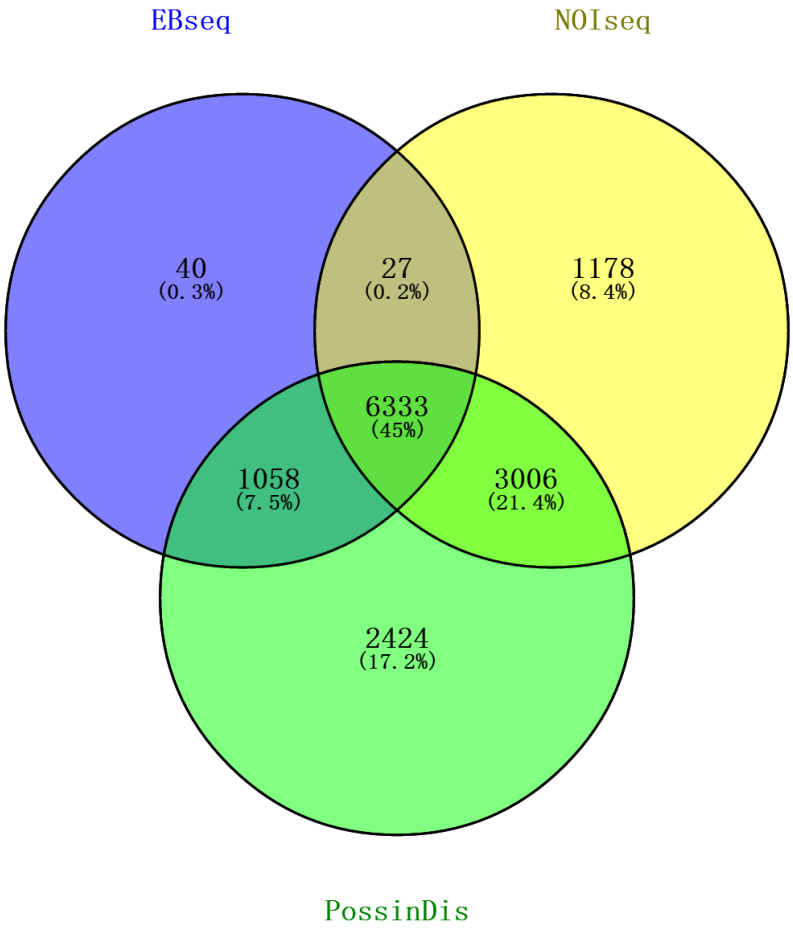

Figure 5: GO 功能分布统计图。X 轴表示相应的 Unigene 基因数，Y 轴表示相应的 COG 分类

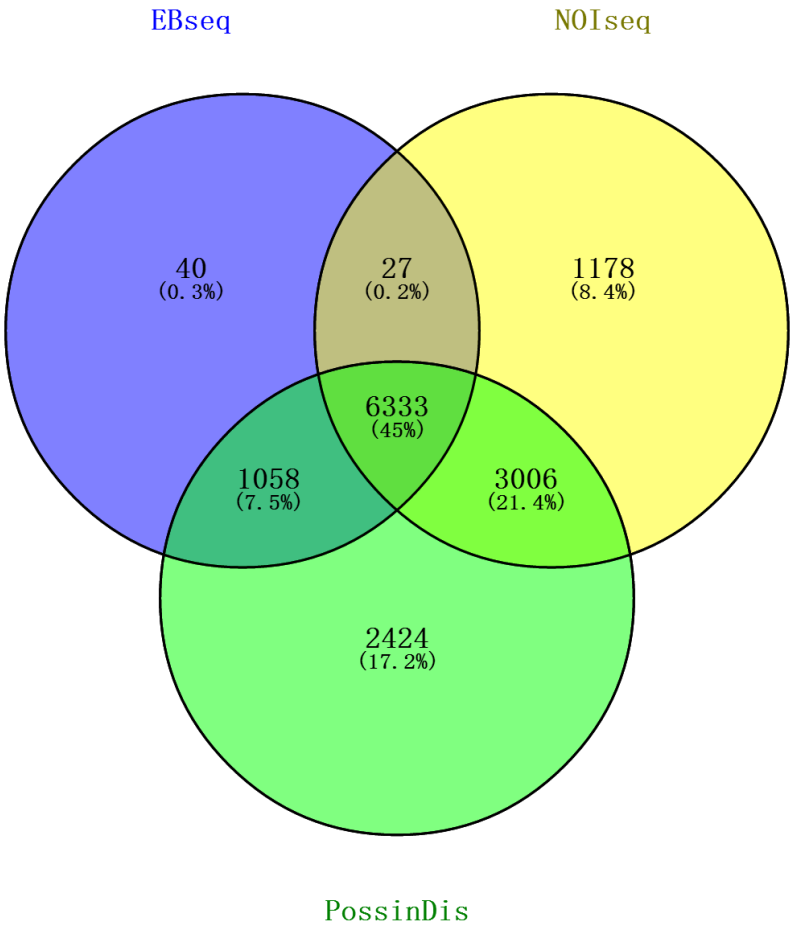

Figure 6: KEGG 功能分布统计图。X 轴表示相应的 Unigene 基因数，Y 轴表示相应的 COG 分类

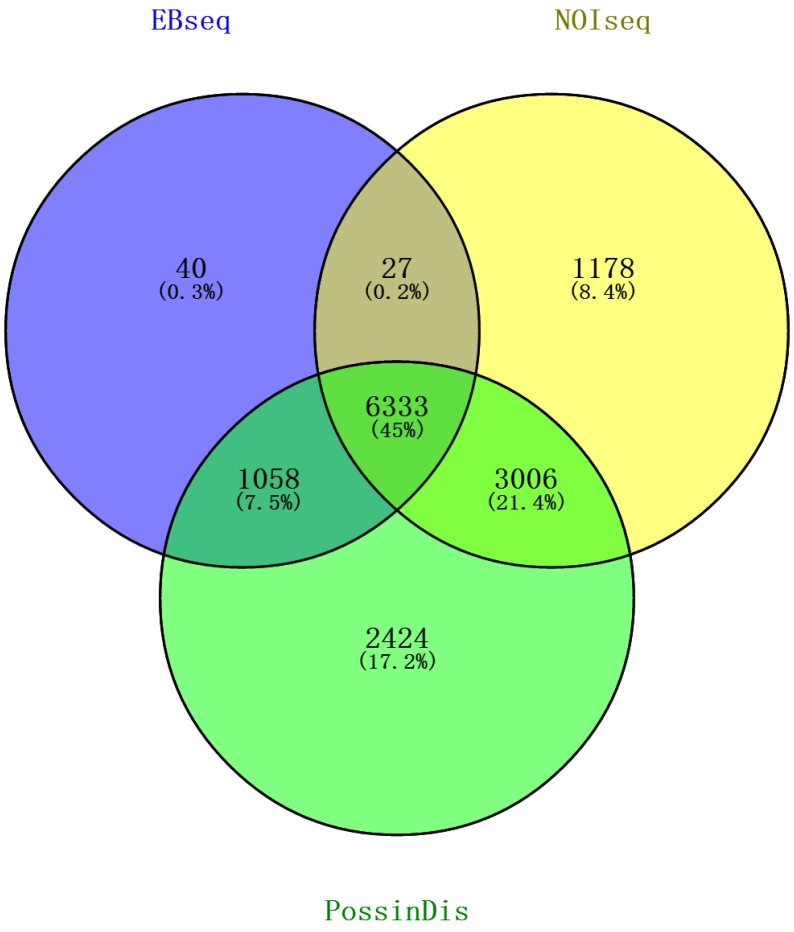

Figure 7: NR、COG、KEGG、Swissprot 以及 Interpro 功能注释维恩图。

1.5 Unigene 的 CDS 预测

根据功能注释结果，我们挑选 Unigene 的最佳比对片段作为该 Unigene 的 CDS。对于未能注释上的 Unigene，我们使用 ESTScan[5] 进行 CDS 预测。预测结果见表 13 ，预测的 CDS 长度分布见图 10。

Table 5: CDS 的质量指标

Software	Total Number	Total length	Mean length	N50	N70	N90	GC(%)
Blast	25,437	19,268,922	757	1,167	720	324	41.74
ESTScan 3,866	1,418,124	366	384	282	222	42.25	
Overall 29,303	20,687,046	705	1,086	648	294	41.77	

[1] N50: 用于衡量序列的连续性，数值越大说明连续性越好，计算方法为：按 CDS 长度从大到小排序后逐个累加至所有 CDS 总长度的 50(%): 碱基 G 和 C 的比例。

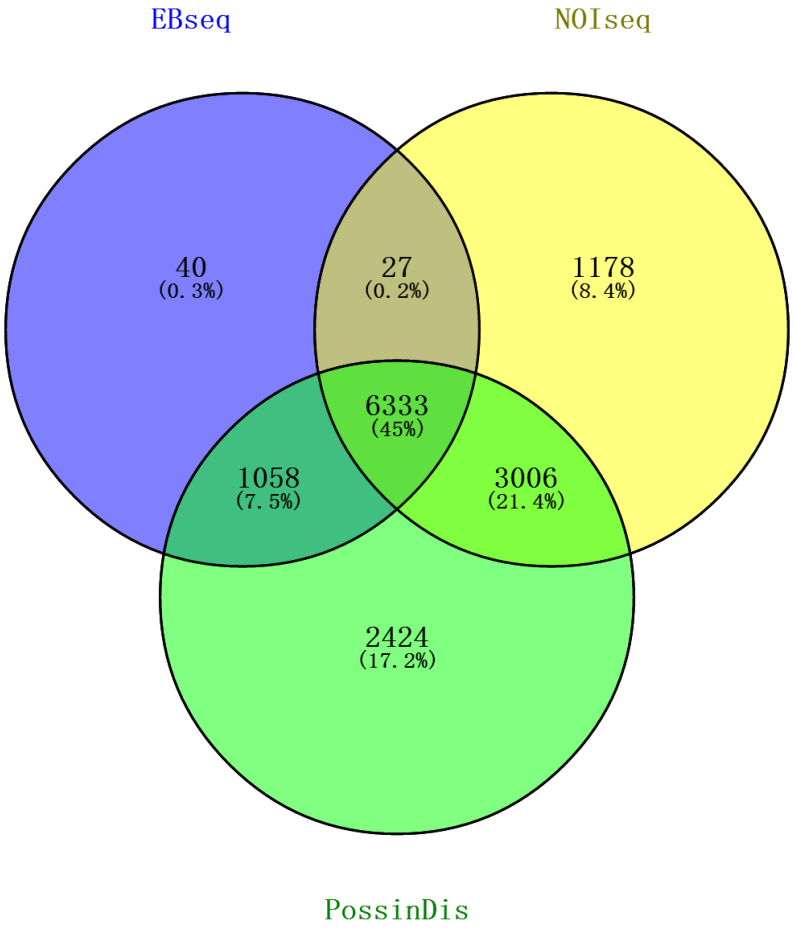

Figure 8: CDS 长度分布图。X 轴代表 CDS 的长度，Y 轴代表相应的 CDS 数目。

1.6 Unigene 的 SSR 预测

根据组装结果，我们对 Unigene 的 SSR 进行检测，同时为每个 SSR 设计引物。SSR 长度特征见表 14 和图 11 。引物设计结果见表 15 。

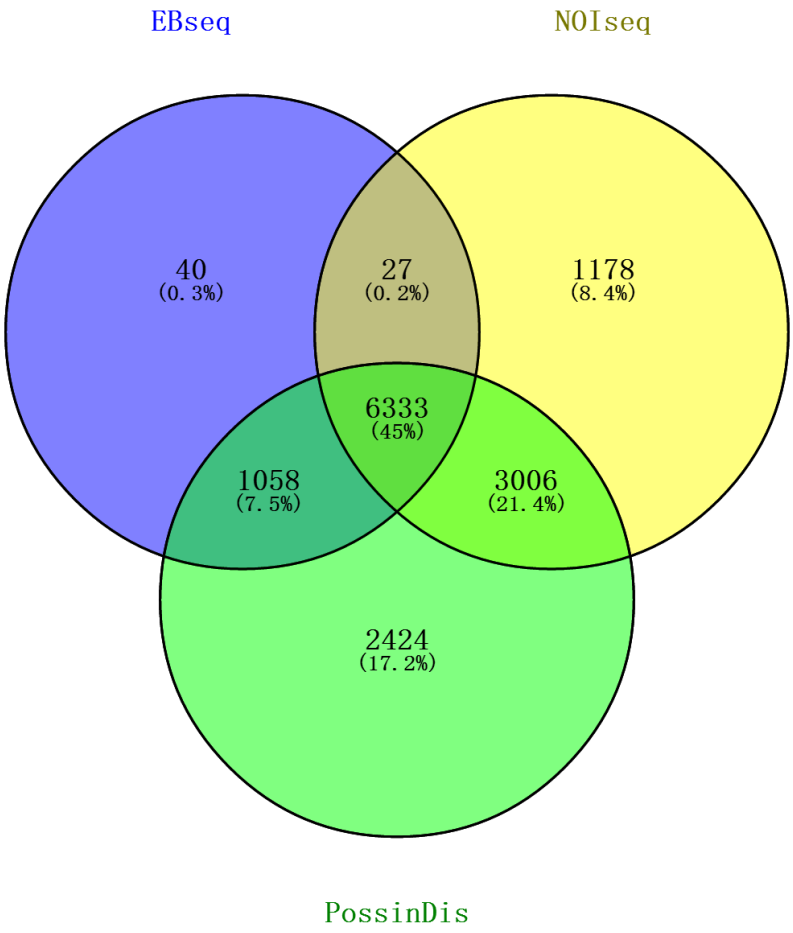

Figure 9: SSR 长度分布图。X 轴代表 SSR 的类型，Y 轴代表相应的 SSR 的数目。

1.7 Unigene 的 SNP 检测

根据组装结果，我们使用 GATK [9] 对每个样品进行 SNP 检测，结果存储为以 VCF 格式。SNP 检测结果见表 16 和图 12 。

Table 6: SNP 类型统计

Sample	A-G	C-T	Transition	A-C	A-T	C-G	G-T	Trancversion	Total
OB	6,461	6,182	12,643	2,141	2,697	1,033	1,996	7,867	20,510

[1] Transition: 嘌呤和嘌呤之间的替换，或嘧啶和嘧啶之间的替换。Transversion: 嘌呤和嘧啶之间的替换。

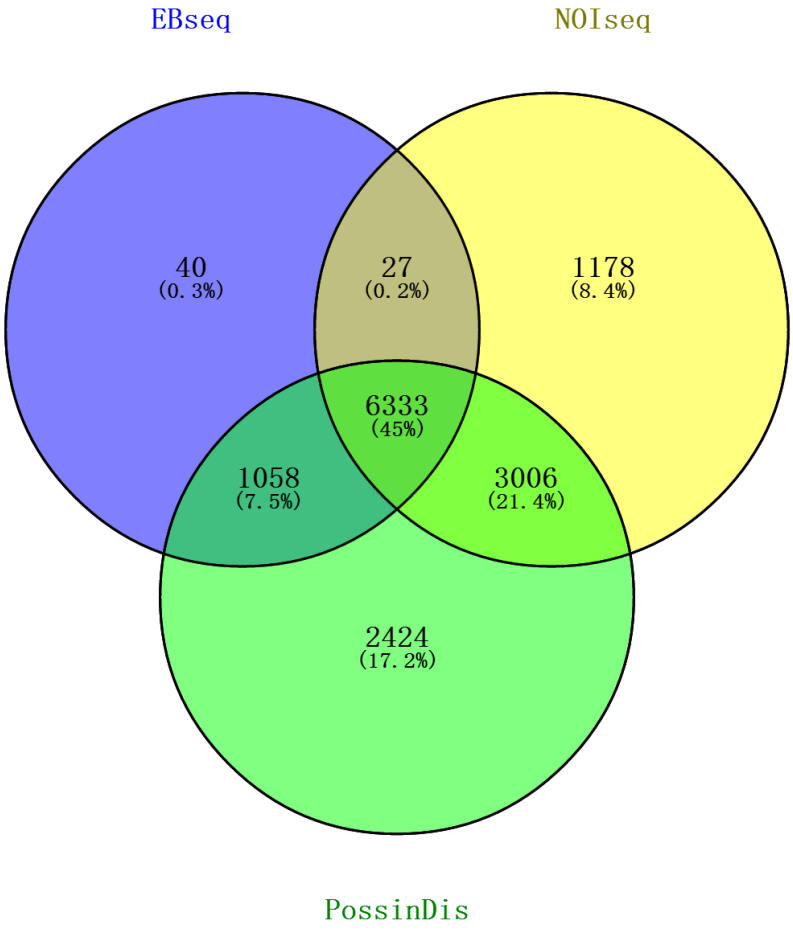

Figure 10: SNP 变异分布图。X 轴代表变异类型，Y 轴代表相应的 SNP 数目。

1.8 Unigene 的表达量的计算

2 分析方法

2.1 转录组 De novo 研究流程

提取样品总 RNA 并使用 DNase I 消化 DNA 后，用带有 Oligo（dT）的磁珠富集真核生物 mRNA；加入打断试剂在 Thermomixer 中适温将 mRNA 打断成短片段，以打断后的 mRNA 为模板合成一链 cDNA ，然后配制二链合成反应体系合成二链 cDNA ，并使用试剂盒纯化回收、粘性末端修复、cDNA 的 3’末端加上碱基“A”并连接接头，然后进行片段大小选择，最后进行 PCR 扩增；构建好的文库用 Agilent 2100 Bioanalyzer 和 ABI StepOnePlus Real-Time PCR System 质检，合格后使用 IlluminaHiSeq4000 或其他平台进行测序

测序所得数据称为 raw reads。首先，我们过滤掉低质量、接头污染以及未知碱基 N 含量过高的 reads，过滤后的数据称为 clean reads。然后对 clean reads 进行组装得到 Unigene，之后对 Unigene 进行 SSR 检测、功能注释，对每个样品计算 Unigene 表达水平以及检测 SNP 。最后，对于多个样品根据需求检测不同样品之间的差异表达基因，并对差异表达基因做深入的聚类分析和功能富集分析。完整的分析流程图见图 1 。

2.2 测序数据过滤

测序的原始数据包含低质量、接头污染以及未知碱基 N 含量过高的 reads，数据分析之前需要去除这些 reads 以保证结果的可靠性。我们使用内部软件进行过滤，具体步骤如下：

1) 去除包含接头的 reads(接头污染)；

2) 去除未知碱基 N 含量大于 5% 的 reads；

3) 去除低质量的 reads (我们定义质量值低于 15 的碱基占该 reads 总碱基数的比例大于 20% 的 reads 为低质量的 reads)。

过滤后的 reads 称为”CleanReads” 并保存为 FASTQ [1] 格式 (格式说明请查阅帮助页面)。

2.3 De novo 组装

我们使用 Trinity 对 clean reads(去除 PCR 重复以提高组装效率) 进行 de novo 组装, 然后使用 Tgicl 将组装的转录本进行聚类去冗余，得到 Unigene。Trinity 包含三个独立模块: Inchworm，Chrysalis 以及 Butterfly，依次顺序处理大量 reads. Trinity 首先把 reads 构建成大量单独的 deBruijn 图，然后对每个图分别提取全长的转录本剪切亚型。简要处理过程如下：

Inchworm: 构建 k-mer 库，K=25。过滤低频 k-mer 选择最高频度的 k-mer 作为种子（不包括复杂度和单一的 k-mers, 一次用完即从 k-mer 库中剔除），用来 Contig 组装。以 k-mer 间 overlap 长度等于 k-1 对种子进行延伸，直到不能再延伸，形成线性 Contig。

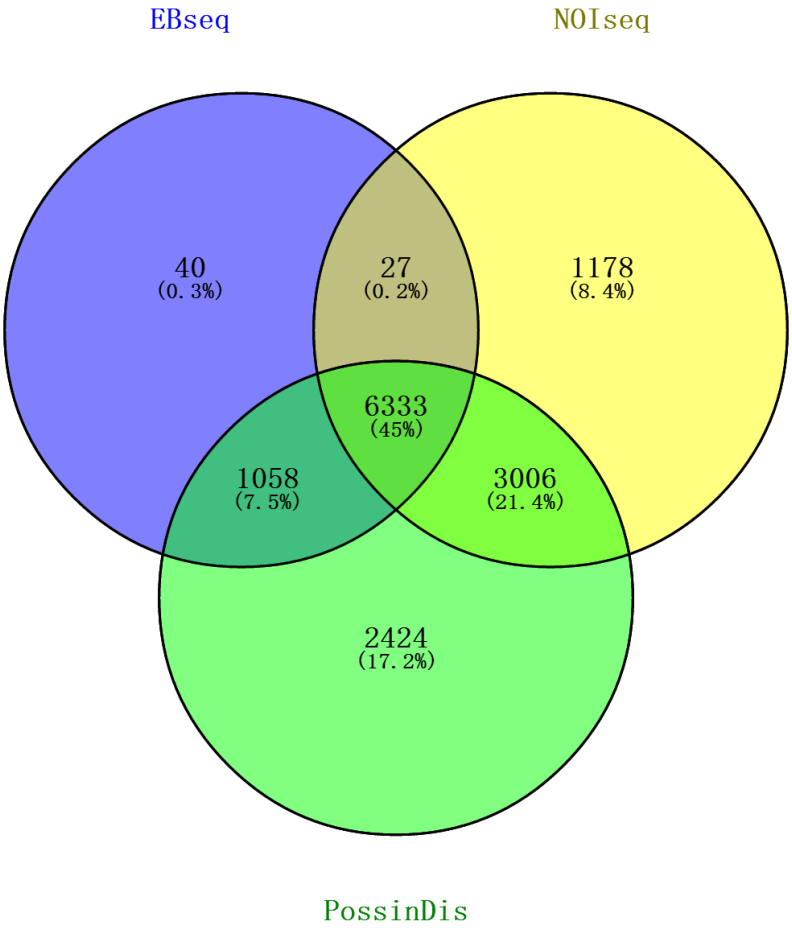

Figure 11: 转录组 de novo 研究流程。

Chrysalis: 把可能存在可变剪切及其他平行基因的 Contigs 聚类。每个 Contig 集定义成一个 Component，对每个 Component 构建 de Bruijn graphs。拿 reads 验证，看每个 Component 的 reads 支持情况。

Butterfly: 合并在 deBruijn 图中有连续节点的线性路径，以形成更长的序列。剔除可能由于测序错误（只有极少 reads 支持）的分叉，使边均匀。用动态规划算法打分，鉴定被 reads 和 readpairs 支持的路径，剔除 reads 支持少的路径。

Trinity 的组装结果我们称为转录本，然后使用 Tgicl 进行聚类去冗余得到 Unigene(对于多个样品，将再次使用 Tgicl 对每个样品的 Unigene 进行聚类去冗余得到最终的 Unigene 用于后续分析)。Unigene 分为两部分，一部分是 clusters，同一个 cluster 里面有若干条相似度高（大于 70 基因家族的编号）。其余的是 singletons（以 Unigene 开头），代表单独的 Unigene。

3 参考文献

[1] Cock P., et al.(2010). The SangerFASTQ file format forsequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6): 1767-1771.

[2] Altschul SF, et al.(1990).Basic local alignment search tool.J Mol Biol. 1990 Oct 5;215(3):403-10.

[3] Conesa A, et al.(2005).Blast2GO: a universal tool forannotation, visualization and analysis in functional genomics research.Bioinformatics. 2005 Sep 15;21(18):3674-6.

[4] Quevillon E, et al.(2005).InterProScan: protein domains identifier.Nucleic Acids Res. 2005 Jul 1;33(Web Serverissue):W116-20.

[5] Iseli C, et al.(1999).ESTScan: a program fordetecting, evaluating, and reconstructing potential coding regions in EST sequences.Proc Int Conf Intell Syst Mol Biol. 1999:138-48.

[6] Thiel T, et al.(2003).Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.).Theor Appl Genet. 2003 Feb;106(3):411-22.

[7] UntergrasserA, et al.(2012).Primer3 -newcapabilities and interfaces.Nucl. Acids Res. (2012)40 (15): e115.

[8] Kim D, et al.(2015).HISAT: a fast spliced alignerwith lowmemory requirements. Nature Methods 2015.

[9] McKenna A, et al.(2010).The Genome Analysis Toolkit: a MapReduce framework foranalyzing next-generation DNA sequencing data.Genome Res. 2010 Sep;20(9):1297-303.

[10] Langmead B, et al.(2012).Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

[11] Li B, et al.(2011).RSEM: accurate transcript quantification from RNA-Seq data with orwith-

out a reference genome.BMCBioinformatics. 2011 Aug 4;12:323.

[12] Eisen, M. B., et al. (2001). Clusteranalysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, (1998)95(25): 14863-8. 2001.29: 1165-1188.

[13] M. J. L. de Hoon, et al. (2004). Open Source Clustering Software.Bioinformatics, 20(9): 1453-1454.

[14] Saldanha, A. J. (2004). Java Treeview–extensible visualization of microarray data. Bioinformatics, 20(17): 3246-8.

[15] GrabherrMG, et al.(2011).Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52.

[16] Pertea G, et al.(2002).TIGRGene Indices clustering tools (TGICL): a software system forfast clustering of large EST datasets.Bioinformatics (2003)19 (5): 651-652.