

# 动植物转录组 ( Transcriptome ) 产品说明书

( 内部资料，请勿外传 )



科技服务体系    动植物研究方向

科技服务体系    动植物研究方向



版本信息：

版本编号	名称	完成日期	作者
v 1.0	动植物转录组 ( Transcriptome ) 产品说明书 v1.0	2011-06-26	明瑶
v1.1	动植物转录组 ( Transcriptome ) 产品说明书 v1.1	2011-07-08	明瑶

2011 年 07 月 08 日

# 目 录

1	产品概述 .....	1
1.1	什么是转录组测序 .....	1
1.2	转录组测序的产品功能 .....	1
1.3	转录组测序产品优势 .....	1
1.4	转录组测序产品发展史 .....	1
1.5	项目执行时间 .....	3
1.6	产品交付结果 .....	3
2	转录组测序研究方法 .....	4
2.1	产品策略 .....	4
2.2	样品准备 .....	5
2.2.1	RNA 样品要求 .....	5
2.2.2	RNA 样品送样标准 .....	6
2.2.3	RNA 提取的组织用量建议 .....	6
2.3	样品运输要求 .....	7
2.3.1	样品包装 .....	7
2.3.2	样品标识 .....	8
2.3.3	样品运输条件 .....	8
2.4	文库的构建及测序 .....	9
2.4.1	实验流程 .....	9
2.4.2	测序及数据处理 .....	10
2.5	转录组生物信息学分析 .....	10
2.5.1	没有参考序列的转录组 De novo.....	10
2.5.2	有参考序列的转录组 Re-sequencing .....	18
2.5.3	参考文献 .....	24
3	成功案例 .....	25

3.1 华大成功案例 .....25.....

3.2 相关文献解读 .....26.....

# 1 产品概述

## 1.1 什么是转录组测序？

转录组测序的研究对象为特定细胞在某一功能状态下所能转录出来的所有 RNA 的总和，包括 mRNA 和非编码 RNA。转录组测序是指用新一代高通量测序技术对物种或者组织的转录本进行测序并得到相关的转录本信息。

## 1.2 转录组测序的产品功能

1. 获得物种或者组织的转录本信息；
2. 得到转录本上基因的相关信息，如：基因结构，功能等；
3. 发现新的基因；
4. 基因结构优化；
5. 发现可变剪切；
6. 发现基因融合；
7. 基因表达差异分析。

## 1.3 转录组测序产品优势

覆盖度高：检测信号是数字信号，几乎覆盖所有转录本；

检测精度高：几十到数十万个拷贝精确计数；

分辨率高：可以检测到单碱基差异，基因家族中相似基因及可变剪切造成的不同转录本的表达；

完成速度快：整个项目周期只需要 50 个工作日时间；

成本低：基本上每个实验室可以承担相关研究经费。

## 1.4 转录组测序产品发展史

转录组的研究手段大体包括：EST 序列构建及研究，芯片研究，运用第二代测序技术研究等。EST 是从一个随机选择的 cDNA 克隆进行 5 端和 3 端单次 sanger 测序获得的短的 cDNA 部分序列，代表一个完整基因的一小部分，在

数据库中其长度一般从 20 到 7000 bp 不等 ,平均长度为 360 ± 120 bp。EST 来源于一定环境下一个组织总 mRNA 所构建的 cDNA 文库 ,因此 EST 也能说明该组织中各基因的表达水平。基因芯片研究 ( microarray ) 是将大量探针分子固定于支持物上 , 然后与标记的样品进行杂交 , 通过检测杂交信号的强度及分布来进行分析基因表达差异。 高通量测序技术研究转录本则是利用第二代测序技术 , 直接对全部转录本进行研究 , 无需繁琐的建库流程 , 就可以得到高覆盖度高精度的转录本信息。 尤其是基于 Illumina 高通量测序平台的转录组测序技术使能够在单核苷酸水平对任意物种的整体转录活动进行检测。 在分析转录本的结构和表达水平的同时 , 还能发现未知转录本和稀有转录本 , 精确地识别可变剪切位点以及 cSNP( 编码序列单核苷酸多态性 ) , 提供最全面的转录组信息。

表 1-1 转录组研究技术比较

Technology	Tiling Array	cDNA or EST sequencing	Transcriptome sequencing
Principle	Hybridization	Sanger sequencing	Next-Gen sequencing
Resolution	From several to 100	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simutaneously map transcribed regions and gene expression	Yes	Limited for expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundred fold	Not practical	>8,000-fold
Ability to distinguish different isoforms and allelic expression	Limited	Yes	Yes


高通量测序技术研究转录本以低成本为主要特征 , 目前世面上的转录组测序产品主要有 Roche 454 和 Illumina HiSeq 2000 为平台的产品。Roche 454 测序仪读长较长 , 但是在判断连续单碱基重复区时准确度不高 ; Illumina HiSeq 兼有高通量、高准确度、低成本的优点 , 美中不足的就是读长低于 454。为此 , 华大基因专门针对 Solexa , 开发了专门的基因组组装软件 SOAPdenovo。随着第一篇完全利用 Solexa 技术完成的熊猫基因组文章的发表 , 华大 SOAPdenovo 软件的组

装效果也同时获得了科学界的一致肯定。至此，华大基因组测序与组装走在了世界的前列。同时，华大的基因组测序也走向了产业化，大量的转录组项目为华大的信息分析人员积累了更多的测序和组装经验，将为客户提供世界顶级的转录组测序组装服务。

## 1.5 项目执行时间


从样品检测合格开始，不包括由于样品问题停滞的时间，动植物转录组测序、及生物信息分析整体的完成周期为 50 个工作日。

## 1.6 产品交付结果

 交付指标：

数据量

1. 基因组极大的物种，如：小麦，玉米等，建议 8 Gb；
2. 普通基因组物种，建议 4 Gb。

 交付数据：

 标准信息分析（无参考序列）

1. 对原始数据进行去除接头、污染序列及低质量 reads 的处理
2. 数据产出统计及测序数据的成分和质量评估
3. 组装结果分析（Contig 长度分布、Scaffold 长度分布、Unigene 长度分布）
4. Unigene 功能注释
5. Unigene 的 GO 分类
6. Unigene 的 COG 分类
7. Unigene 代谢通路分析
8. 预测编码蛋白框（CDS）
9. Unigene 表达差异分析（两个或两个以上样品）
10. Unigene 在样品间的差异 GO 分类（需两个或两个以上样品）和 Pathway 富集性分析

### 定制化信息分析

1. 多个样品做 de novo 分析时 , 分析并提供每个样品的 Unigene 的 GQ pathway 等结果
2. 将 Hiseq 数据与其他数据如 EST 等联合组装 ( 需要客户提供其他数据 )

### 标准信息分析 ( 需提供参考基因序列、参考基因组序列及基因注释结果 )

1. 对原始数据进行去除接头、污染序列及低质量 reads 的处理
2. 测序评估 ( 比对统计、测序随机性评估、 Reads 在基因组上的分布 )

### 高级信息分析 ( 基于 1-2 标准分析 )

3. 基因表达注释 ( 基因覆盖度、覆盖深度分布等 )
4. 基因差异表达分析 ( 两个或两个以上样品 )
5. 对基因结构进行优化 ( 仅针对真核生物 )
6. 鉴定基因的可变剪接 ( 仅针对真核生物 )
7. 预测新转录本
8. SNP 分析 ( 仅针对真核生物 )

### 定制化信息分析

1. 将两个样品之间的可变剪切进行比较分析 , 统计几个样品间的可变剪切和新转录本的异同
2. 基因融合分析
3. 组与组之间的差异分析
4. 重复间的数据相关性分析

## 2 转录组测序研究方法

### 2.1 产品策略

转录组 de novo 产品策略 :



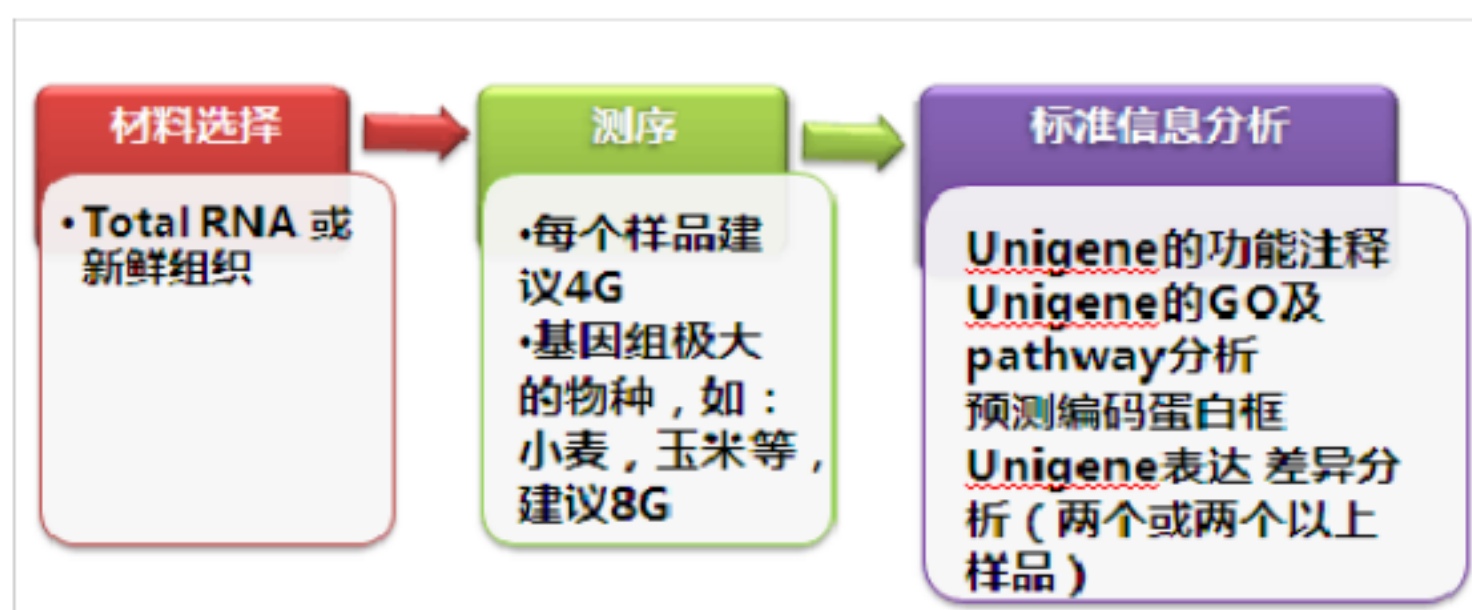


图 2-1 ：转录组 de novo 产品策略

有 ref 的转录组 re-sequencing 产品策略：



图 2-2 ：有 ref 的转录组 re-sequencing 产品策略

## 2.2 样品准备

### 2.2.1 RNA 样品要求

1. 样品类型：去蛋白并进行 DNase 处理后的完整总 RNA；
2. 样品需求量（单次）：植物样品：20 g  $\mu$ 人、大鼠、小鼠样品：5 g  $\mu$ 其他类型动物：10 g  $\mu$
3. 样品浓度：植物样品：400 ng/l； $\mu$ 人、大鼠、小鼠样品：80 ng/l；其它类型动物样品：200 ng/l  $\mu$ 原核生物样品：500 ng/l  $\mu$
4. 样品纯度：OD<sub>260/280</sub> = 1.8~2.2；OD<sub>260/230</sub> 2.0；动物植物样品：RIN 7.0，28S:18S 1,原核生物样品：RIN 6,023S:16S= 1.2~2.2。

2.2.2 RNA 样品送样标准

合作伙伴需要提供 Nanodrop、Gel-Electrophotometric 或者 Aglient 中一种或多种形式的样品分析结果；

应仔细纯化样品，尽量避免多 糖、蛋白质、和外切酶的残留；样品必须注明溶剂成分。

表 2-1：转录组送样标准

Analysis	Total RNA Amount	concentration (ng/ μ l)	OD <sub>260/280</sub>	OD <sub>260/230</sub>	RIN	28S: 18S	23S: 16S
转录组	20ug (plant), 5ug (human, rat, mouse), 10ug (other animals)	80 ng/ul ( human, rat, mouse); 200ng/ul (other animals); 400ng/ul (plants); 500ng/ul (for prokaryotes)	1.8 ~ 2.2	2.0	7.0 (animal s, plants) 6.0 (prokar yotes)	1.0	1.2~2.2 (for prokaryotes)

2.2.3 RNA 提取的组织用量建议

表 2-2：转录组 RNA 提取组织用量建议

组织类型	送样量
新鲜动物组织干重	1-2 g
新鲜植物组织干重	3-4 g
新鲜培养细胞数	8*10 <sup>6</sup> ~9*10 <sup>7</sup> 个
血清	N/A

注：不同类型的样品 RNA 产量差别较大，像人或哺乳动物的全血中红细胞没有细胞核，每毫升血液中实用细胞数少，RNA 得率低，送样量需要加大；鸟类或鱼类的血液中红细胞含有细胞核，可适当减少送样量；含肌纤维和脂肪一类物质以及含多糖多酚较高的复杂植物，RNA 得率一般较低，送样量需要增加；代谢活跃的肝脏组织细胞量旺盛，每 50 mg 组织可达 20~30 ug RNA，可适当降低送样量。

## 2.3 样品运输要求

### 2.3.1 样品包装

对于 RNA 样品，我们建议合作伙伴尽量用 1.5 ml Eppendorf 管装载样品，为了防止 Eppendorf 管在运输过程中受到挤压破裂，导致样品损失，最好将 Eppendorf 管装在 50 ml 离心管（或其他支撑物）中，里面还可以添加棉花、吸水纸等固定。如是大批量样品，请用冻存盒之类的存放盒装好样品，防止样品受损。（注意：切勿在 50 ml 管内或其他支撑物内加入液氮等危险品）。

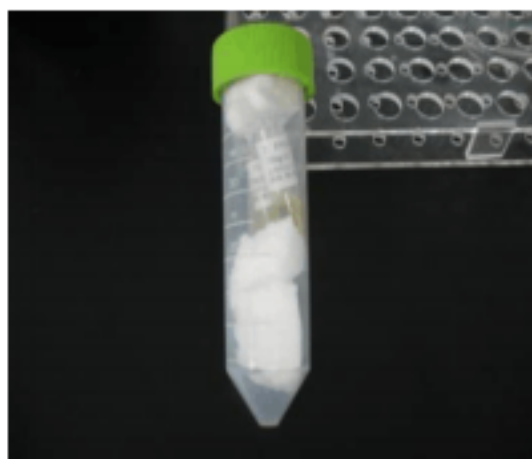


图 2-3：在 15 ml EP 管中装棉花或包膜固定样品管

对于组织样品，一般建议用 1.5 ml 的 Eppendorf 管，或 2 ml 的螺旋管装载。

在样品运输过程中请用 parafilm 膜将管口密封好。不建议样品溶于无水乙醇、异丙醇等有机溶剂邮寄，因为有机试剂比较容易泄露、泄露后容易使管壁字迹模糊，甚至造成样品交叉污染。如果一定需溶于有机溶剂运输，那么 Eppendorf 管的管口至少要用 parafilm 膜封 5 圈以上。

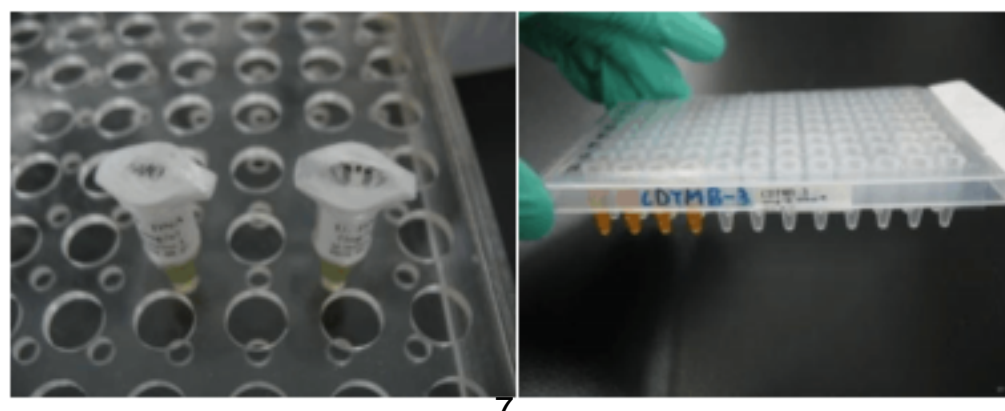


图 2-4：用封口膜封好

对于血液样品，可用 5~10 ml 抗凝管装载，但为了防止抗凝管在运输过程中受到碰撞而破裂，需要将抗凝管放在泡沫或棉花中固定，并彼此隔开。

### 2.3.2 样品标识

不建议用油性笔直接在管壁或管盖上写样品名称等信息，最好将样品名称等各种信息写在标签纸上，贴在管壁，外面再用透明胶带缠绕一圈（一方面防止样品名称被泄露的有机溶剂溶掉，另一方面也可以防止标签纸没有粘牢脱落，导致样品无法应用）。

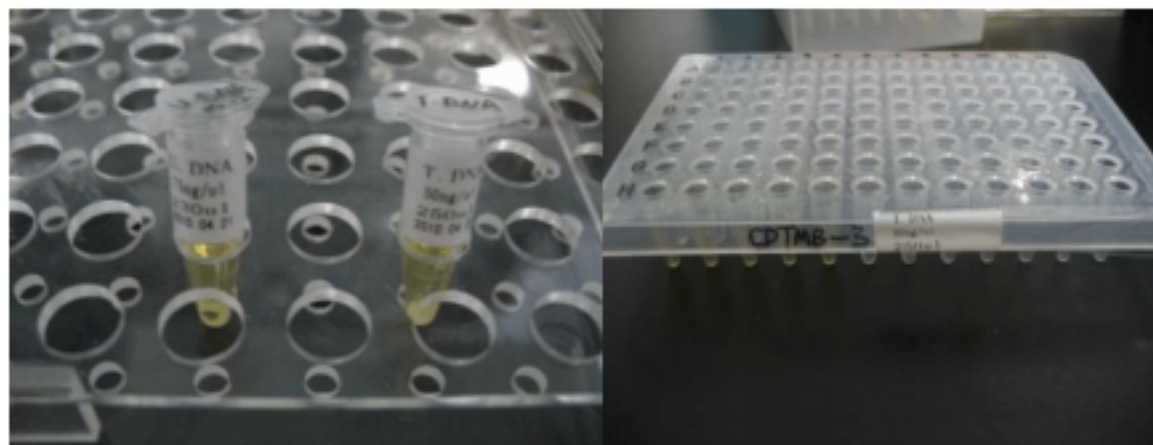


图 2-5：管壁上先用纸条写好，再用胶片缠绕；PCR 板请在侧面标记，再用胶布贴好

邮寄样品时，必须附有我们华大提供的标准格式的样品信息单（电子版、文字版），请合作伙伴仔细检查，务必保证信息单中填写的样品名称、数量需要与实际邮寄的样品名称标识、样品数量完全一致。

### 2.3.3 样品运输条件

- a) DNA 样品如果用乙醇沉淀，则可以常温运输，否则在运输过程中，应放于干冰中，时间不要超过 72 小时；或利用冰袋运输，时间最好不要超过 24 小时；
- b) RNA、组织样品无论溶于什么溶剂，都需放于干冰中运输，时间不要超过 72 小时；

- c) 血浆要保存在干冰中运输，确保样品送达接收地点时有足量的干冰剩余，并及时存放血浆于 -80℃ 冰箱中，禁止将样品在室温状态下放置；全血要在生物冰袋条件下运输，且在 12 小时内送达；
- d) 运输过程中需要添加的干冰和冰袋的量与季节、运输时间长短、泡沫盒的薄厚有关（为更有利于保温，尽量选用大块的干冰，如果条件允许，建议可在邮寄的泡沫盒的上下填充一些棉花等，以隔绝热量的传递）。

## 2.4 文库的构建及测序

### 2.4.1 实验流程

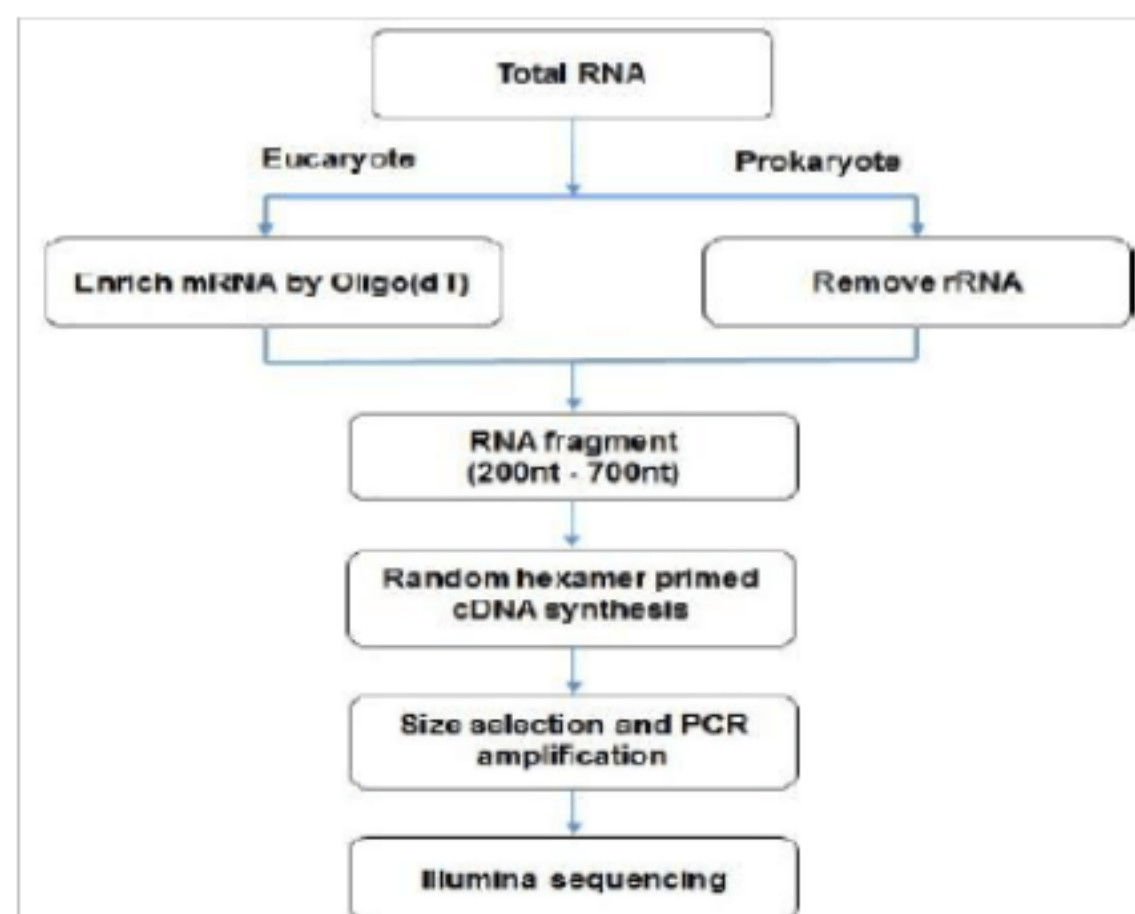


图 2-6：转录组实验流程

提取样品总 RNA 后，用带有 Oligo (dT) 的磁珠富集真核生物 mRNA（若为原核生物，则用试剂盒去除 rRNA 后进入下一步）。加入 fragmentation buffer 将 mRNA 打断成短片段，以 mRNA 为模板，用六碱基随机引物（random hexamers）合成第一条 cDNA 链，然后加入缓冲液、dNTPs、RNase H 和 DNA polymerase I 合成第二条 cDNA 链，在经过 QiaQuick PCR 试剂盒纯化并加 EB 缓冲液洗脱之后做末端修复、加 poly (A) 并连接测序接头，然后用琼脂糖凝胶电泳进行片段大小选择，最后进行 PCR 扩增，建好的测序文库（200 bp）用 Illumina HiSeq 2000 进行测序。



2.4.2 测序及数据处理

数据处理的步骤：（ de novo 和有 ref 的转录组 re-sequencing 是相同的 ）

- 1 去除含 adaptor 的 reads
- 2 去除 N 的比例大的 reads
- 3 去除低质量 reads （质量值  $Q \leq 10$  的碱基数占整个 read 的 50% 以上 ）
- 4 去重复 （ duplication ）

2.5 转录组生物信息学分析

2.5.1 没有参考序列的转录组 De novo

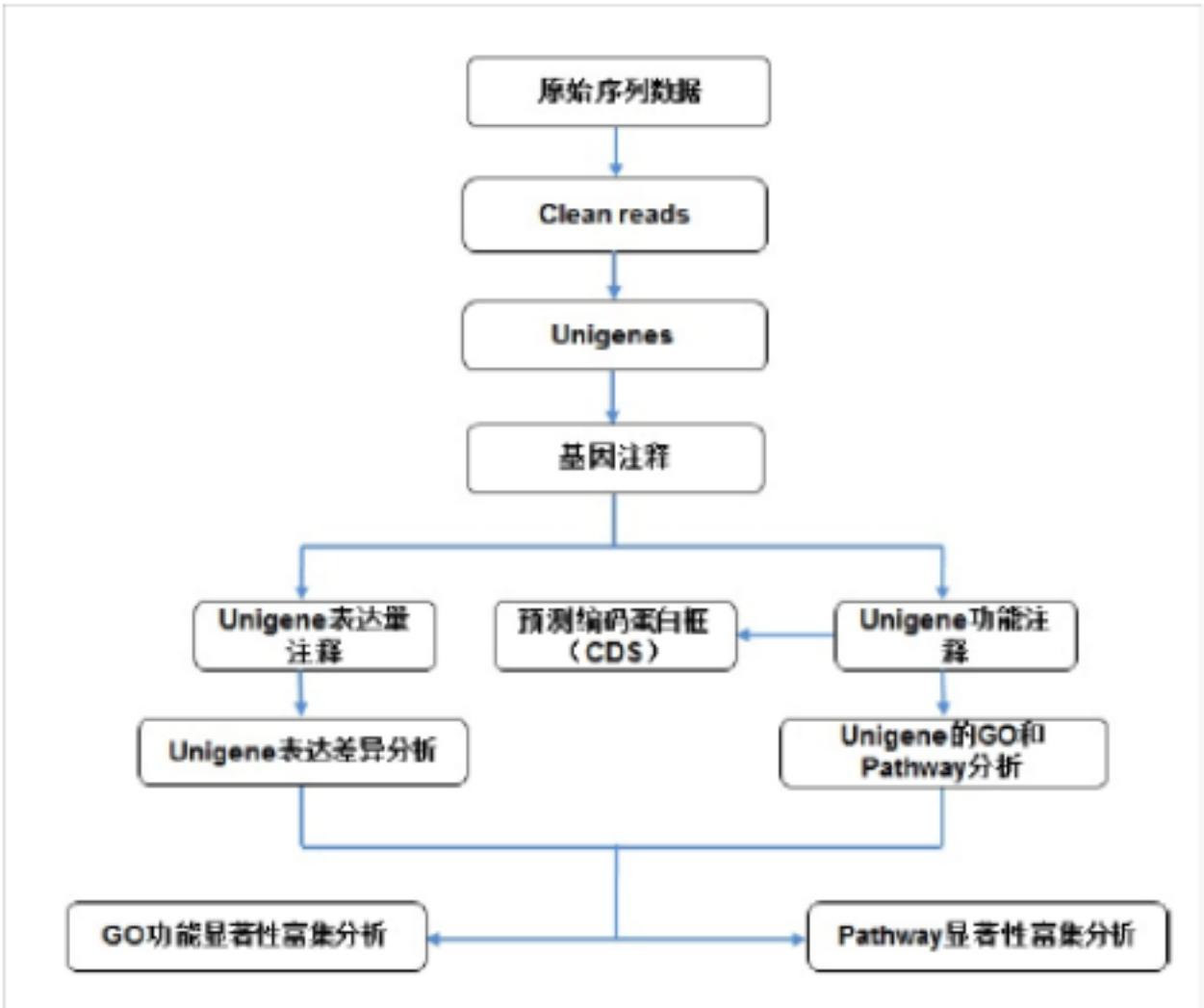


图 2-7 ：转录组 de novo 信息分析流程及详情

产量统计

测序的数据产量是合同的重要指标，一个样品由 clean reads 得到的测序碱基数应不少于合同的规定，该项工作的完成情况见下：

表 2-3 ： 测序产量统计

Samples	Total Reads	Total Nucleotides (nt)	Q20 percentage	N percentage	GC percentage *
ChuLi	26,359,524	2,372,357,160	94.67%	0.00%	44.07%
DuiZhao	27,051,222	2,434,609,980	95.28%	0.01%	42.15%

\* Total Nucleotides = Total Reads1 x Read1 size + Total Reads2 x Read2 size

测序得到的原始图像数据经 base calling 转化为序列数据，我们称之为 raw data 或 raw reads，结果以 fastq 文件格式存储，fastq 文件为用户得到的最原始文件，里面存储 reads 的序列以及 reads 的测序质量。在 fastq 格式文件中每个 read 由四行描述：

```
@FC61FL8AAXX:1:17:1012:19200#GCCAAT/1
CCACTGTCATGTGAACATCACAGAGACATTTCTTGA
+
bbbbbbbbbbbbbbbbbbbbbaaaaaaaaaa_\
```

图 2-8 ： 测序数据描述

每个序列共有 4 行，第 1 行和第 3 行是序列名称（有的 fq 文件为了节省存储空间会省略第三行“+”后面的序列名称），由测序仪产生；第 2 行是序列；第 4 行是序列的测序质量，每个字符对应第 2 行每个碱基，第 4 行每个字符对应的 ASCII 值减去 64，即为该碱基的测序质量值，比如 c 对应的 ASCII 值为 99，那么其对应的碱基质量值是 35。从 Illumina GA Pipeline v1.3 开始（目前为 v1.6），碱基质量值范围为 2 到 35。表 1 为测序错误率与测序质量值简明对应关系。具体地，如果测序错误率用 E 表示，碱基质量值用 sQ 表示，则有下列关系：

$sQ = -10\log_{10}E$

表 2-4 ： 测序错误率与测序质量值简明对应关系

测序错误率	测序质量值	对应字符
5%	13	M
1%	20	T
0.1%	30	^

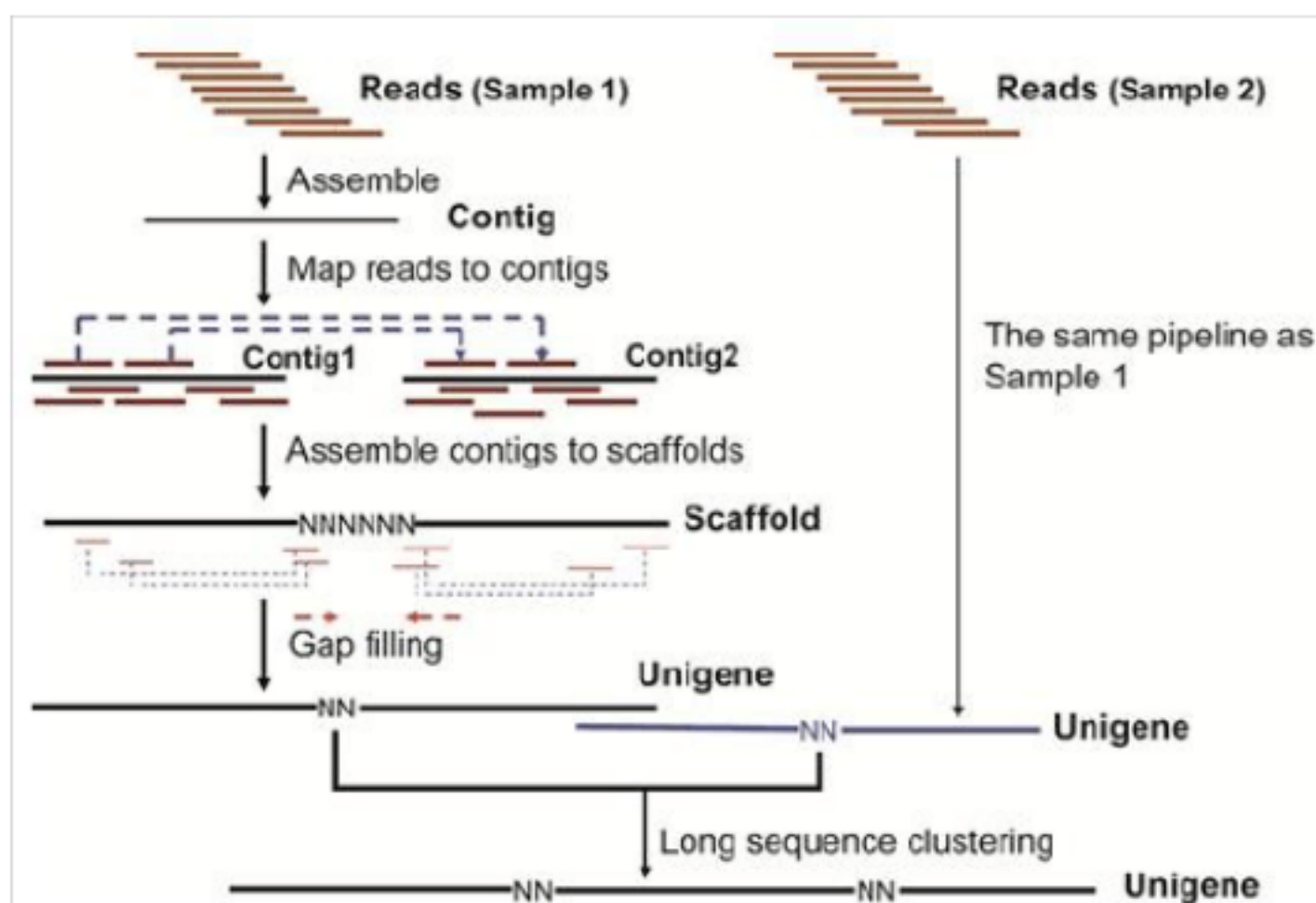


图 2-9：转录组 de novo 组装示意图

组装方法：

我们使用短 reads 组装软件 SOAPdenovo (Li et al. 2009) 做转录组从头组装。SOAPdenovo 首先将具有一定长度 overlap 的 reads 连成更长的片段，这些通过 reads overlap 关系得到的不含 N 的组装片段我们称之为 Contig。然后，我们将 reads 比对回 Contig，通过 paired-end reads 能确定来自同一转录本的不同 Contig 以及这些 Contig 之间的距离，SOAPdenovo 将这些 Contig 连在一起，中间未知序列用 N 表示，这样就得到 Scaffold。进一步利用 paired-end reads 对 Scaffold 做补洞处理，最后得到含 N 最少，两端不能再延长的序列，我们称之为 Unigene。如果同一物种做了多个样品测序，则不同样品组装得到的 Unigene 可通过序列聚类软件做进一步序列拼接和去冗余处理，得到尽可能长的非冗余 Unigene。最后，将 Unigene 序列与蛋白数据库 nr、Swiss-Prot、KEGG 和 COG 做 blastx 比对 ( $e\text{-value} < 0.00001$ )，取比对结果最好的蛋白确定 Unigene 的序列方向。如果不同库之间的比对结果有矛盾，则按 nr、Swiss-Prot、KEGG 和 COG 的优先级确定 Unigene 的序列方向，跟以上四个库皆比不上的 Unigene 我们用软件 ESTScan (Iseli et al. 1999) 预测其编码区并确定序列的方向。对于能确定序列方向的 Unigene 我们给出其从 5' 到 3' 方向的序列，对于无法确定序列方向的 Unigene 我们给出组装软件得到的序列。



组装质量统计：

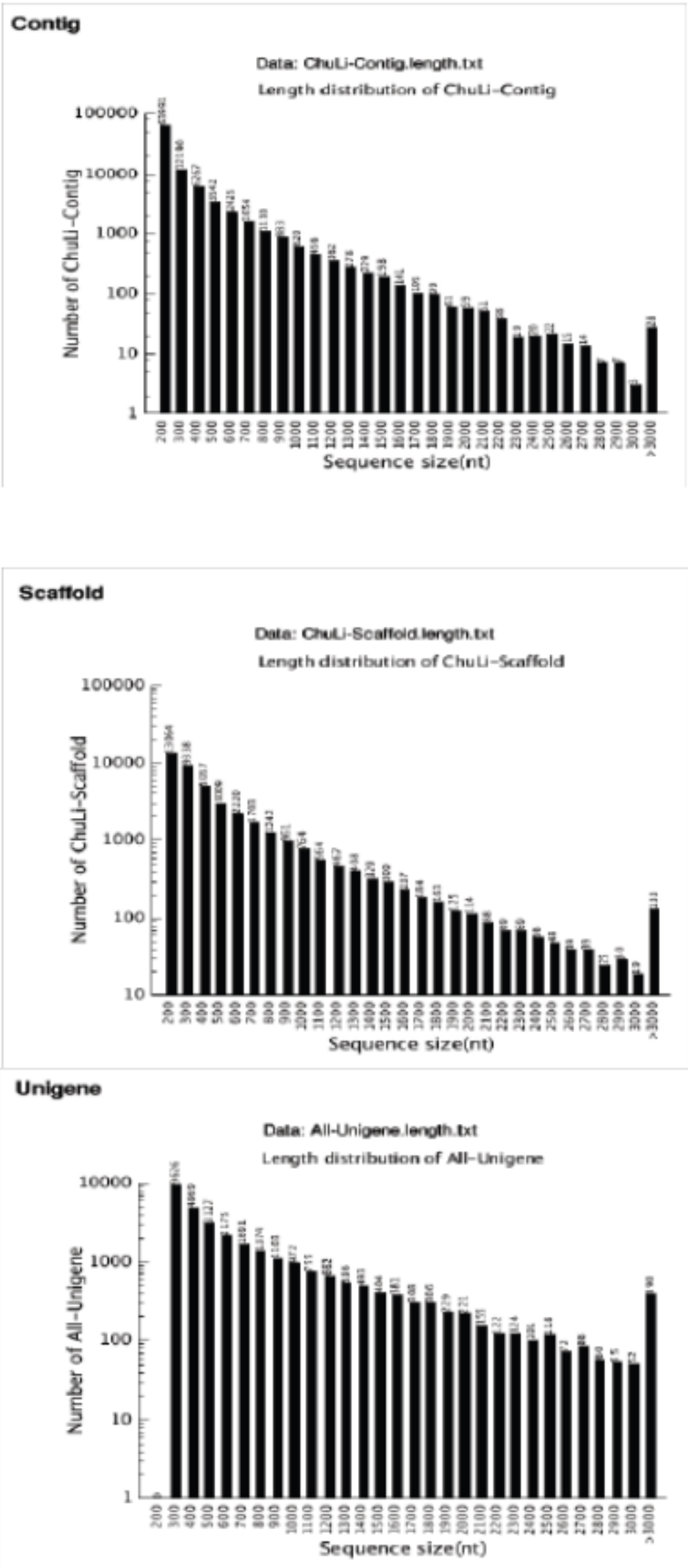


图 2-10 ：组装质量统计图

基因注释

功能注释信息给出 Unigene 的蛋白功能注释、 COG功能注释。

首先，通过 blastx 将 Unigene 序列比对到蛋白数据库 nr、Swiss-Prot、KEGG和 COG( e-value<0.00001 )，得到跟给定 Unigene 具有最高序列相似性的蛋白，从而得到该 Unigene 的蛋白功能注释信息。

COG是对基因产物进行直系同源分类的数据库，每个 COG蛋白都被假定来自祖先蛋白，COG数据库是基于细菌、藻类、真核生物具有完整基因组的编码蛋白、系统进化关系进行构建的，我们将 Unigene 和 COG数据库进行比对，预测 Unigene 可能的功能并对其做功能分类统计，从宏观上认识该物种的基因功能分布特征。

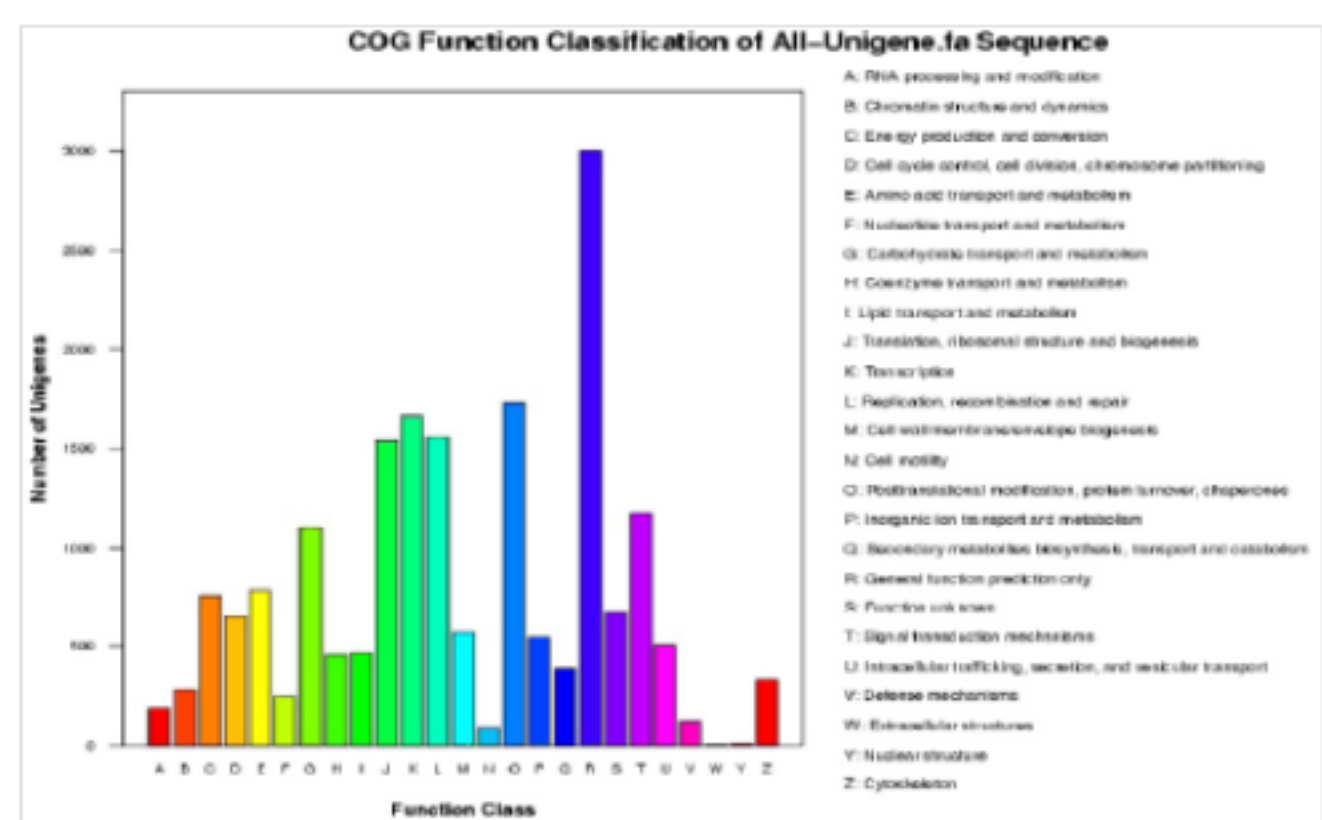


图 2-11 : COG功能聚类

### Unigene 的 GO分类

根据 nr 注释信息我们能得到 GO功能注释。Gene Ontology (简称 GO) 是一个国际化的基因功能分类体系，提供了一套动态更新的标准词汇表 (controlled vocabulary) 来全面描述生物体中基因和基因产物的属性。GO 总共有三个 ontology，分别描述基因的分子功能 (molecular function)、所处的细胞位置 (cellular component)、参与的生物过程 (biological process)。我们根据 nr 注释信息，使用 Blast2GO 软件 (Conesa et al. 2005) 得到 Unigene 的 GO 注释信息。Blast2GO 已被其它文献引用超过 150 次，是同行广泛认可的 GO 注释软件。得到每个 Unigene 的 GO 注释后，我们用 WEGO 软件 (Ye et al. 2006) 对所有 Unigene 做 GO 功能分类统计，从宏观上认识该物种的基因功能分布特征。

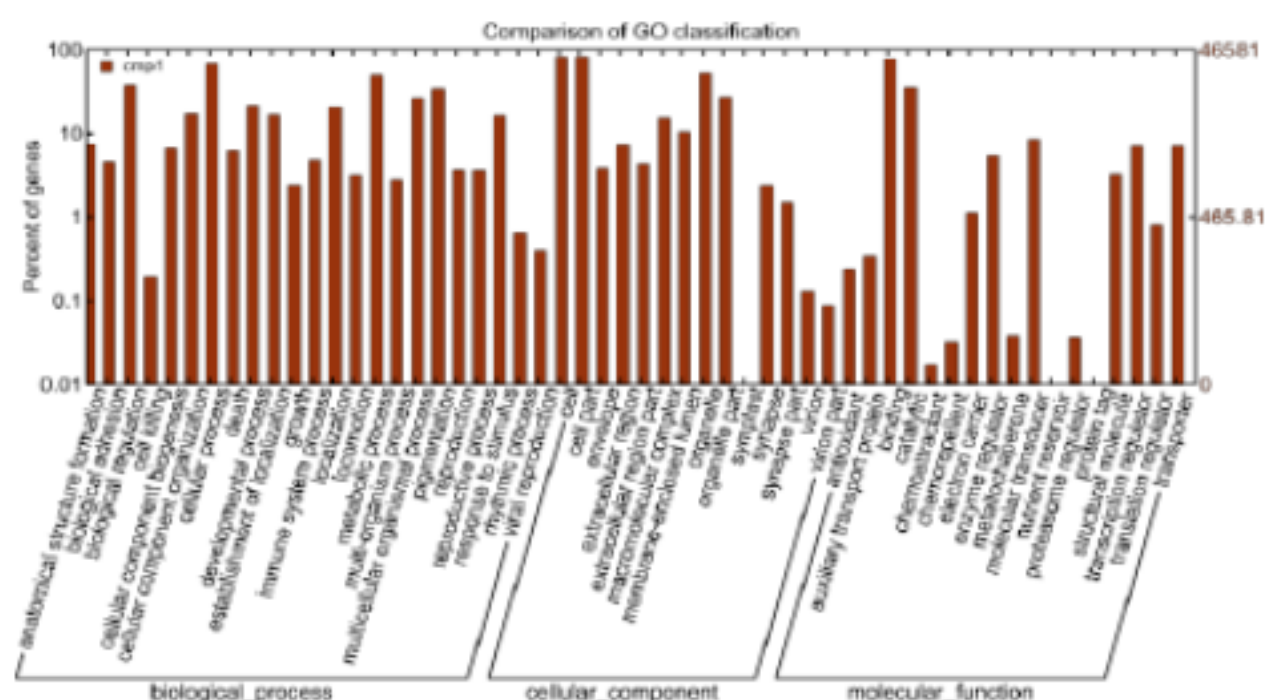


图 2-12 : Unigene 的 GO 分类

## Unigene 代谢通路分析

KEGG是系统分析基因产物在细胞中的代谢途径以及这些基因产物的功能的数据库，利用 KEGG可以进一步研究基因在生物学上的复杂行为。根据 KEGG注释信息我们能进一步得到 Unigene 的 Pathway 注释。

预测编码蛋白框 ( CDS)

首先，我们按 nr、Swiss-Prot、KEGG和 COG的优先级顺序将 Unigene 序列与以上蛋白库做 blastx 比对（ $e\text{-value} < 0.00001$ ），如果某个 Unigene 序列对上高优先级数据库中的蛋白，则不进入下一轮比对，否则自动跟下一个库做比对，如此循环直到跟所有蛋白库比对完。我们取 blast 比对结果中 rank 最高的蛋白确定该 Unigene 的编码区序列，然后根据标准密码子表将编码区序列翻译成氨基酸序列，从而得到该 Unigene 编码区的核酸序列（序列方向 5'→3'）和氨基酸序列。最后，跟以上蛋白库皆比对不上的 Unigene 我们用软件 ESTScan (Iseli et al. 1999) 预测其编码区，得到其编码区的核酸序列（序列方向 5'→3'）和氨基酸序列。

## 基因表达量的计算

Unigene 表达量的计算使用 RPKM 法 ( Reads Per kb per Million reads ) ( Mortazavi et al. 2008 ) , 其计算公式为 :

$$RPKM = \frac{10^6 C}{NL / 10^3}$$

设 RPKM(A)为 Unigene A 的表达量，则 C 为唯一比对到 Unigene A 的 reads 数，N 为唯一比对到所有 Unigene 的总 reads 数，L 为 Unigene A 的碱基数。

差异 Unigene 的 GO 和 Pathway 分析

## GO 功能分析

Gene Ontology (简称 GO) 是一个国际化的基因功能分类体系，提供了一套动态更新的标准词汇表 (controlled vocabulary) 来全面描述生物体中基因和基因产物的属性。GO 总共有三个 ontology (本体)，分别描述基因的分子功能 (molecular function)、所处的细胞位置 (cellular component)、参与的生物过程 (biological process)。GO 的基本单位是 term (词条、节点)，每个 term 都对应一个属性。GO 功能分析一方面给出差异表达基因的 GO 功能分类注释；另一方面给出差异表达基因的 GO 功能显著性富集分析。

GO 功能分类注释给出具有某个 GO 功能的基因列表及基因数目统计。

GO 功能显著性富集分析给出与基因组背景相比，在差异表达基因中显著富集的 GO 功能条目，从而给出差异表达基因与哪些生物学功能显著相关。该分析首先把所有差异表达基因向 Gene Ontology 数据库的各个 term 映射，计算每个 term 的基因数目，然后应用超几何检验，找出与整个基因组背景相比，在差异表达基因中显著富集的 GO 条目，其计算公式为

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

其中，N 为所有 Unigene 中具有 GO 注释的基因数目；n 为 N 中差异表达基因的数目；M 为所有 Unigene 中注释为某特定 GO term 的基因数目；m 为注释为某特定 GO term 的差异表达基因数目。计算得到的 pvalue 通过 Bonferroni 校正之后，以 corrected-p value 0.05 为阈值，满足此条件的 GO term 定义为在差异表达基因中显著富集的 GO term。通过 GO 功能显著性富集分析能确定差异表达基因行使的主要生物学功能。

Gene Ontology 数据库为 <http://www.geneontology.org/>。

我们的 GO 功能分析同时整合了表达模式聚类分析，研究人员能方便地看到具有某一功能的所有差异基因的表达模式。例，immune response 为在差异表达



基因中最显著富集的一个 GO term（表 2-5）。图 2-13 显示了参与 immune response 的差异基因的表达模式。

表 2-5：差异表达基因中显著富集的 GO-term

Terms from the Process Ontology with p-value as good or better than 0.05				
Gene Ontology term	Cluster frequency	Genome frequency of use	Corrected P-value	Expression Profile
<a href="#">immune response view genes</a>	82 out of 807 genes, 10.2%	863 out of 13525 genes, 4.9%	2.74e-07	<a href="#">View Result</a>
<a href="#">immune system process view genes</a>	100 out of 807 genes, 12.4%	921 out of 13525 genes, 6.8%	3.77e-06	<a href="#">View Result</a>
<a href="#">response to virus view genes</a>	21 out of 807 genes, 2.6%	105 out of 13525 genes, 0.8%	0.00138	<a href="#">View Result</a>
<a href="#">regulation of apoptosis view genes</a>	63 out of 807 genes, 7.8%	583 out of 13525 genes, 4.3%	0.00508	<a href="#">View Result</a>
<a href="#">regulation of programmed cell death view genes</a>	63 out of 807 genes, 7.8%	592 out of 13525 genes, 4.4%	0.00840	<a href="#">View Result</a>
<a href="#">regulation of cell death view genes</a>	63 out of 807 genes, 7.8%	593 out of 13525 genes, 4.4%	0.00888	<a href="#">View Result</a>
<a href="#">regulation of cell death view genes</a>	63 out of 807 genes, 7.8%	593 out of 13525 genes, 4.4%	0.00888	<a href="#">View Result</a>

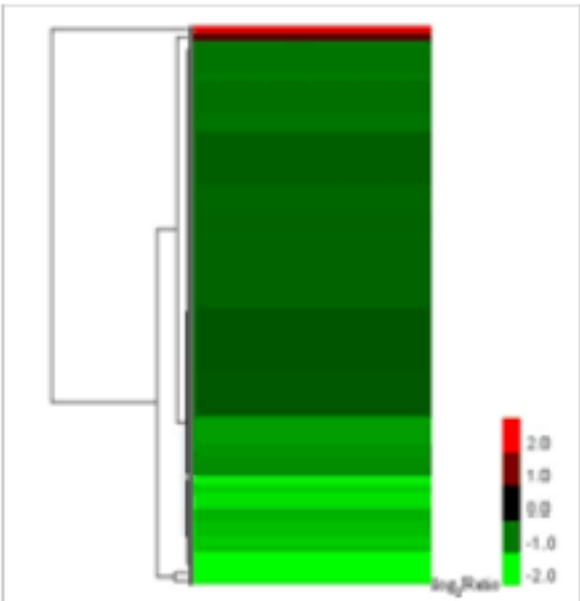


图 2-13：参与 immune response 的差异基因的表达模式

KEGG Pathway 分析

在生物体内，不同基因相互协调行使其生物学，基于 Pathway 的分析有助于更进一步了解基因的生物学功能。KEGG 是有关 Pathway 的主要公共数据库（Kanehisa et al. 2008），Pathway 显著性富集分析以 KEGG Pathway 为单位，应用超几何检验，找出与整个基因组背景相比，在差异表达基因中显著性富集的 Pathway。该分析的计算公式同 GO 功能显著性富集分析，在这里 N 为所有 Unigene 中具有 Pathway 注释的基因数目；n 为 N 中差异表达基因的数目；M 为所有 Unigene 中注释为某特定 Pathway 的基因数目；m 为注释为某特定 Pathway 的差异表达基因数目。FDR 0.05 的 Pathway 定义为在差异表达基因中显著富集

的 Pathway。通过 Pathway 显著性富集能确定差异表达基因参与的最主要生化代谢途径和信号转导途径。结果如表 2-6 所示。

表 2-6：显著性富集分析列表

#	Pathway	DEGs with pathway annotation (2085)	All genes with pathway annotation (8986)	Pvalue	Qvalue	Pathway ID
1	Metabolic pathways	307 (14.72%)	1081 (12.03%)	1.354119e-05	0.002911356	ko01100
2	Proteasome	23 (1.1%)	48 (0.53%)	0.0001482570	0.015937627	ko03050
3	B cell receptor signaling pathway	29 (1.39%)	70 (0.78%)	0.0005085341	0.036444944	ko04662
4	Apoptosis	34 (1.63%)	89 (0.99%)	0.001018471	0.045737882	ko04210
5	Hematopoietic cell lineage	31 (1.49%)	80 (0.89%)	0.001271905	0.045737882	ko04640
6	Primary immunodeficiency	16 (0.77%)	33 (0.37%)	0.001276406	0.045737882	ko05340
7	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	13 (0.62%)	25 (0.28%)	0.001618825	0.049721054	ko00563
8	N-Glycan biosynthesis	18 (0.86%)	40 (0.45%)	0.001901140	0.051093137	ko00510
9	Huntington's disease	60 (2.88%)	187 (2.08%)	0.003132988	0.074843602	ko05016
10	Other glycan degradation	9 (0.43%)	16 (0.18%)	0.004361382	0.093769713	ko00511
11	Alzheimer's disease	56 (2.69%)	176 (1.96%)	0.005133783	0.100342122	ko05010
12	Biosynthesis of steroids	13 (0.62%)	28 (0.31%)	0.005747707	0.102979750	ko00100
13	Chronic myeloid leukemia	27 (1.29%)	74 (0.82%)	0.006681019	0.110493776	ko05220
14	Epithelial cell signaling in Helicobacter pylori infection	25 (1.2%)	68 (0.76%)	0.007943713	0.121992735	ko05120

#	序号
Pathway	通路名
DEGs with pathway annotation (2085)	注释到该通路的差异表达基因的数目
All genes with pathway annotation (8986)	注释到该通路的所有基因的数目
Pvalue	超几何检验的P值
Qvalue	Q值（Q < 0.05为在差异表达基因中显著富集的Pathway）
Pathway ID	KEGG数据库中的Pathway ID

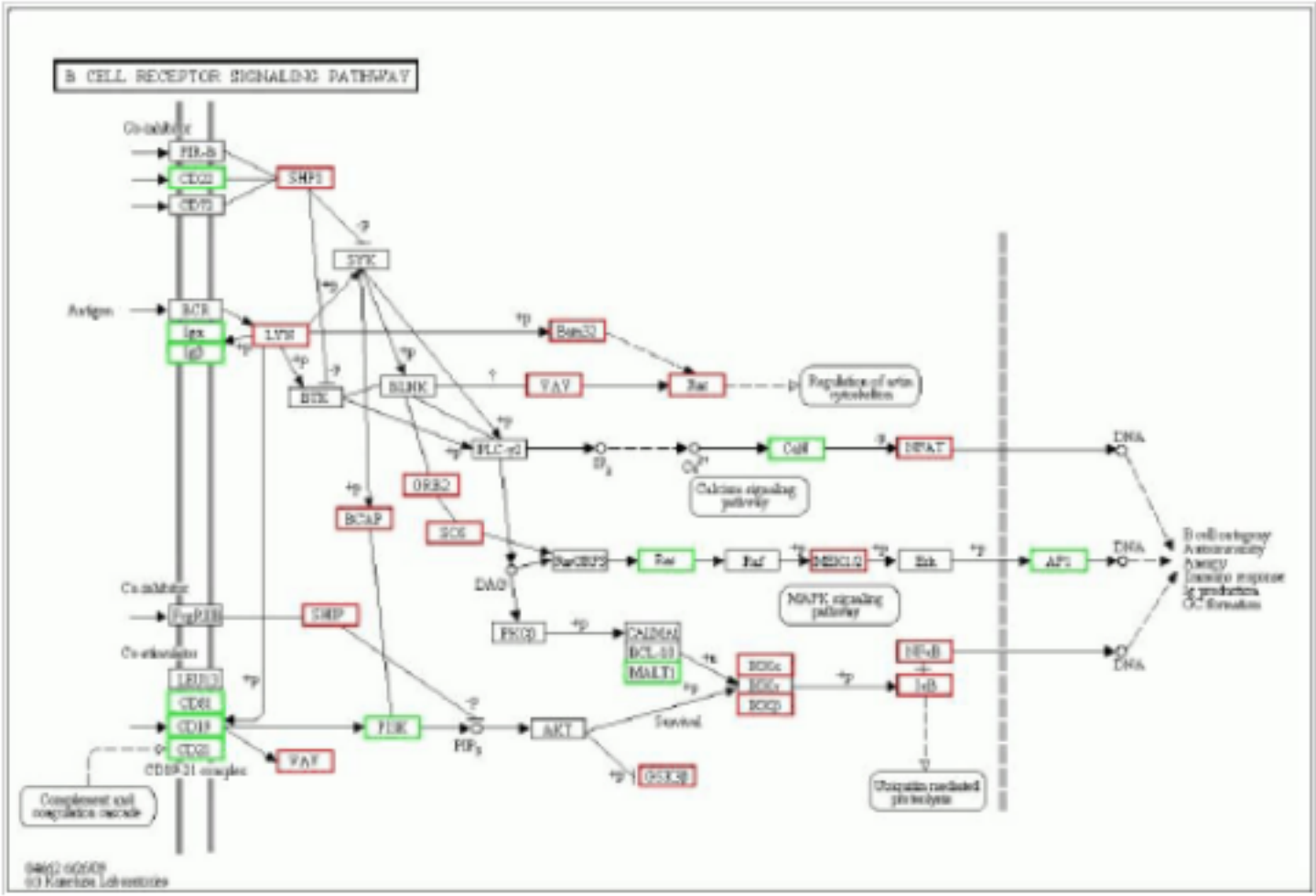


图 2-14：KEGG 数据库中 B cell receptor signaling pathway 的详细信息

2.5.2 有参考序列的转录组 Re-sequencing



图 2-15 ：有 ref 的转录组 re-sequencing 信息分析流程及详情

比对统计

我们使用短 reads 比对软件 SOAPaligner/soap2（Li et al., 2009）将 clean reads 分别比对到参考基因组和参考基因序列，然后统计出比对结果。

表 2-7 ：样品与参考基因比对的统计图

样品BG_TR_0040_001和参考基因比对的统计结果		
Map to Gene	reads number	percentage
Total Reads	138896384	100.00%
Total BasePairs	13889638400	100.00%
Total Mapped Reads	83894807	60.40%
perfect match	58198874	41.90%
<=5bp mismatch	25695933	18.50%
unique match	53945047	38.84%
multi-position match	29949760	21.56%
Total Unmapped Reads	55001577	39.60%

随机性评估

在转录组实验过程中，首先要通过物理或化学方法将转录本打断成短片段，然后上机测序。如果打断随机性差，reads 偏向于来自基因特定区域，将会直接影响转录组的各项分析结果。我们利用 reads 在基因上的分布来评价打断随机性。由于不同参考基因有不同长度，我们把 reads 在基因上的位置标准化到相对位置（reads 在基因上的位置与基因长度的比值），然后统计基因的不同位置比对上的 reads 数。如果打断随机性好，reads 在基因各部位应分布得比较均匀。



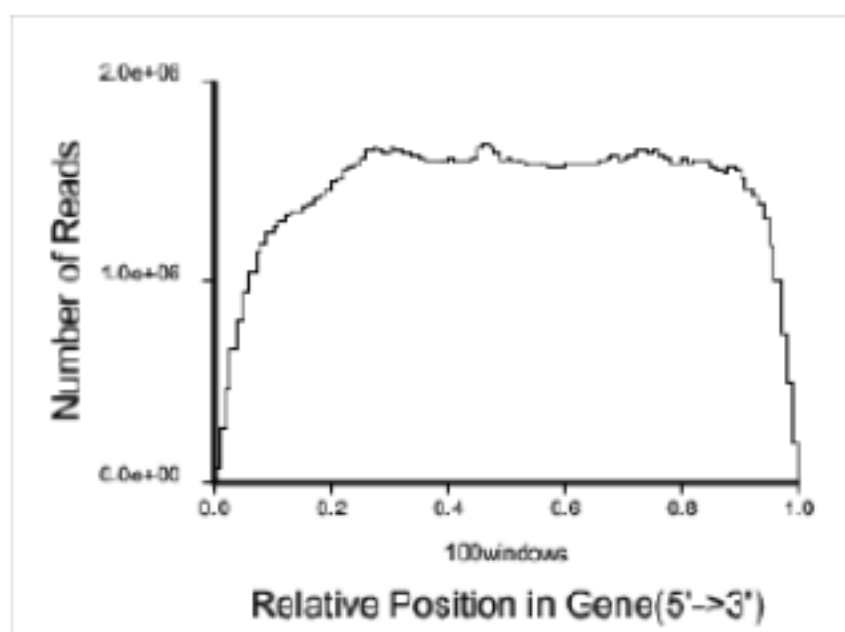


图 2-16 ：随机性评估示意图

## 基因表达注释

本分析给出基因覆盖度、表达量和功能注释信息。

## 基因覆盖度（ Coverage ）统计

基因测序覆盖度指每个基因被 reads 覆盖的百分比，其值等于基因中 unique mapping reads 覆盖的碱基数跟基因编码区所有碱基数的比值。

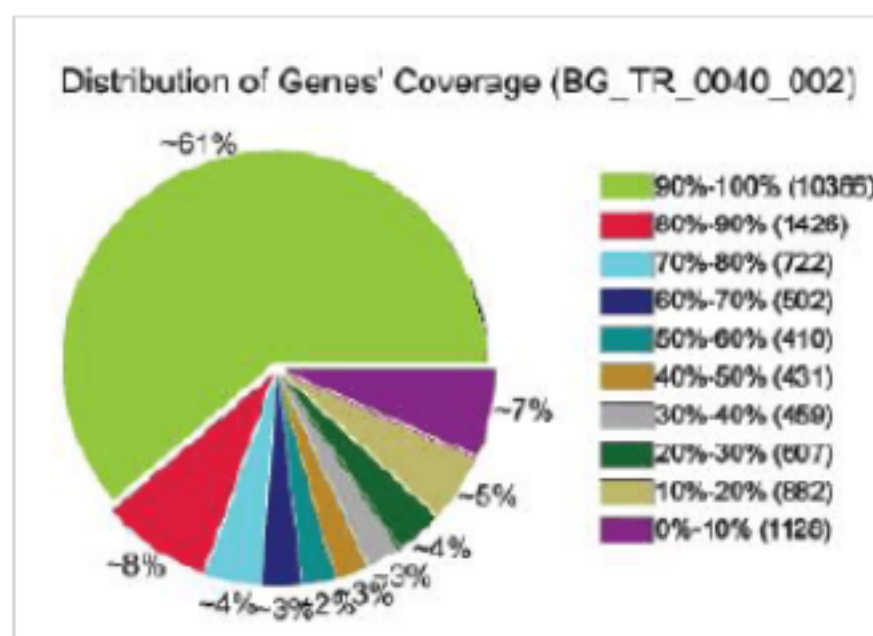


图 2-17 ：基因覆盖度统计

## 基因表达量

基因表达量的计算使用 RPKM 法( Reads Per Kb per Million reads )( Mortazavi et al., 2008 )，其计算公式为：

$$RPKM = \frac{10^6 C}{NL / 10^3}$$



设 RPKM(A)为基因 A 的表达量，则 C 为唯一比对到基因 A 的 reads 数，N 为唯一比对到参考基因的总 reads 数，L 为基因 A 编码区的碱基数。RPKM 法能消除基因长度和测序量差异对计算基因表达的影响，计算得到的基因表达量可直接用于比较不同样品间的基因表达差异。

如果一个基因存在多个转录本，则用该基因的最长转录本计算其测序覆盖度和表达量。

### 基因差异表达分析

差异表达分析找出在不同样本间存在差异表达的基因，并对差异表达基因做 GO 功能分析和 KEGG Pathway 分析。

参照 Audic S 等人发表在 Genome Research 上的数字化基因表达谱差异基因检测方法 (Audic et al., 1997) ,我们开发了严格的算法筛选两样本间的差异表达基因。

假设观测到基因 A 对应的 clean tag 数为 x，已知在一个大文库中，每个基因的表达量只占有所有基因表达量的一小部分，在这种情况下，p(x) 的分布服从泊松分布：

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (\lambda \text{ 为基因 A 的真实转录数})$$

已知，样本一总 clean tag 数为 N1，样本二总 clean tag 数为 N2，基因 A 在样本一中对应的 clean 数为 x，在样本二中对应的 clean 数为 y，则基因 A 在两样本中表达量相等的概率可由以下公式计算：

$$2 \sum_{i=0}^{i=y} p(y|x)$$

或  $2 \times (1 - \sum_{i=0}^{i=y} p(y|x))$  (如果  $\sum_{i=0}^{i=y} p(y|x) > 0.5$ )

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x! y! \left(1 + \frac{N_2}{N_1}\right)^{(x+y+1)}}$$

然后，我们对差异检验的 p value 作多重假设检验校正，通过控制 FDR ( False Discovery Rate ) 来决定 P Value 的域值。

假设挑选了 R 个差异表达基因，其中 S 个是真正有差异表达的基因，另外 V 个是其实没有差异表达的基因，为假阳性结果。希望错误比例  $Q = V/R$  平均而言

不能超过某个可以容忍的值（比如 0.001 %），则在统计时预先设定 FDR 不能超过 0.001。在我们的分析中，差异表达基因定义为  $FDR \leq 0.001$  且倍数差异在 2 倍和 2 倍以上的基因。

基因结构优化：

本分析通过比较测序结果和现有基因注释结果，对基因的 5'端或 3'端进行延长。如下图所示，首先，将 reads 比对到基因组，提取基因组中被 unique mapping reads 覆盖的次数大于或等于某阈值（默认为 2）且位置连续的区域作为转录活性区 (Transcription Active Region，TAR，图中蓝色方块区域)；然后通过 paired-end reads（图中紫色线条）将不同的 TAR 连接形成潜在的 gene model；最后，通过比较潜在 gene model 与现有基因注释的差别，对基因的 5'端和 3'端进行延长（图 2-18 中表现的仅是基因 3'端发生延长的情况）。

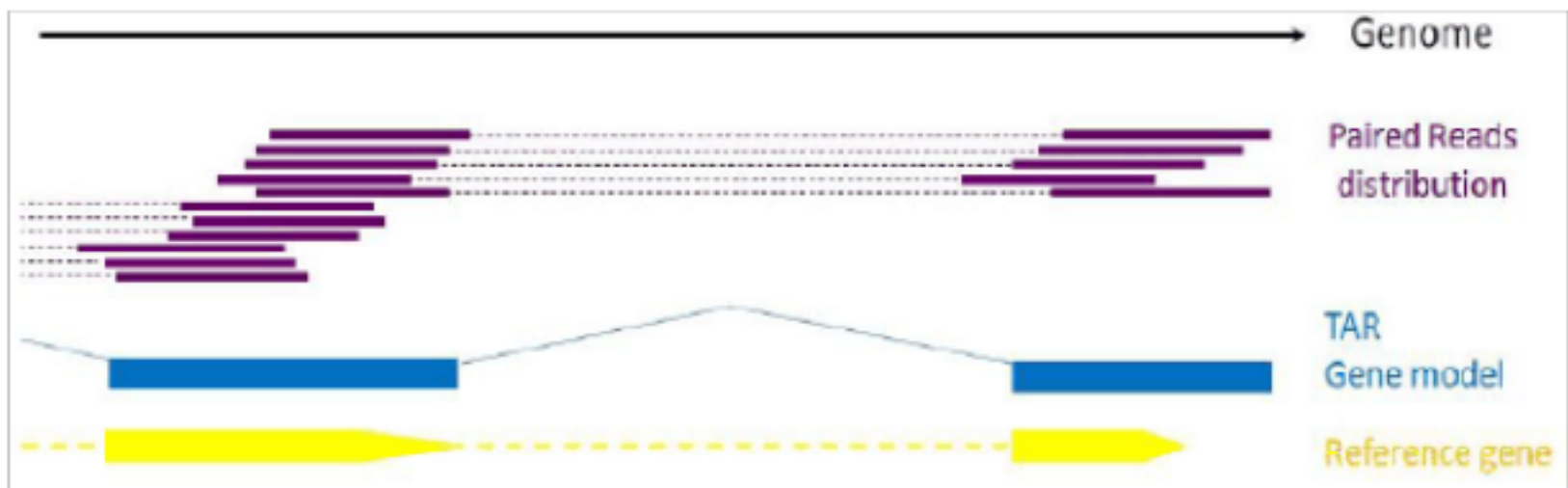


图 2-18：基因结构优化示意图

表 2-8：样品基因 5 和 3 端延长统计

样品BG_TR_0040_001基因5'端和3'端的延长					
gene	5' or 3' end	Chromosome	Strand	original region	extended region
NM_033543.42082530-42093196	5	chr19	+	42082531-42093196	42082494-42082531
NM_001018136.49230919-49249104	5	chr17	+	49230920-49249104	49230859-49230920
NM_0001495842898-5851485	5	chr19	-	5842899-5851485	5851485-5851531
NM_03259335812958-35815042	5	chr9	-	35812959-35815042	35815042-35815127
NM_016816113344738-113357711	5	chr12	+	113344739-113357711	113344670-113344739
NM_01410635270541-35280497	5	chr15	-	35270542-35280497	35280540-35280639
NM_024618119039439-119054723	5	chr11	+	119039440-119054723	119038903-119039163
NR_029481.96941115-96941202	5	chr9	+	96941116-96941202	96940993-96941116
NM_00244640697650-40721481	5	chr19	+	40697651-40721481	40697603-40697651
NM_01803541937223-41945843	5	chr19	-	41937224-41945843	41945682-41945742

可变剪接分析

可变剪接使一个基因产生多个 mRNA 转录本，不同 mRNA 可能翻译成不同蛋白。因此，通过可变剪接一个基因可能产生多个蛋白，极大地增加了蛋白多样

性。虽然已知可变剪接在真核生物中普遍存在，但我们可能仍低估了可变剪接的比例，最近，基于高通量测序的可变剪接研究在人、小鼠、拟南芥中发现了很多新的可变剪接事件。

在生物体内，主要存在 7 种可变剪接类型：A) Exon skipping; B) Intron retention; C) Alternative 5' splice site; D) Alternative 3' splice site; E) Alternative first exon; F) Alternative last exon; G) Mutually exclusive exon。图 2-19 是我们利用高通量测序数据鉴别出来的 4 种可变剪接，图中每个位置的 Exp. Level 等于  $\log_2(\text{Reads 数})$ 。

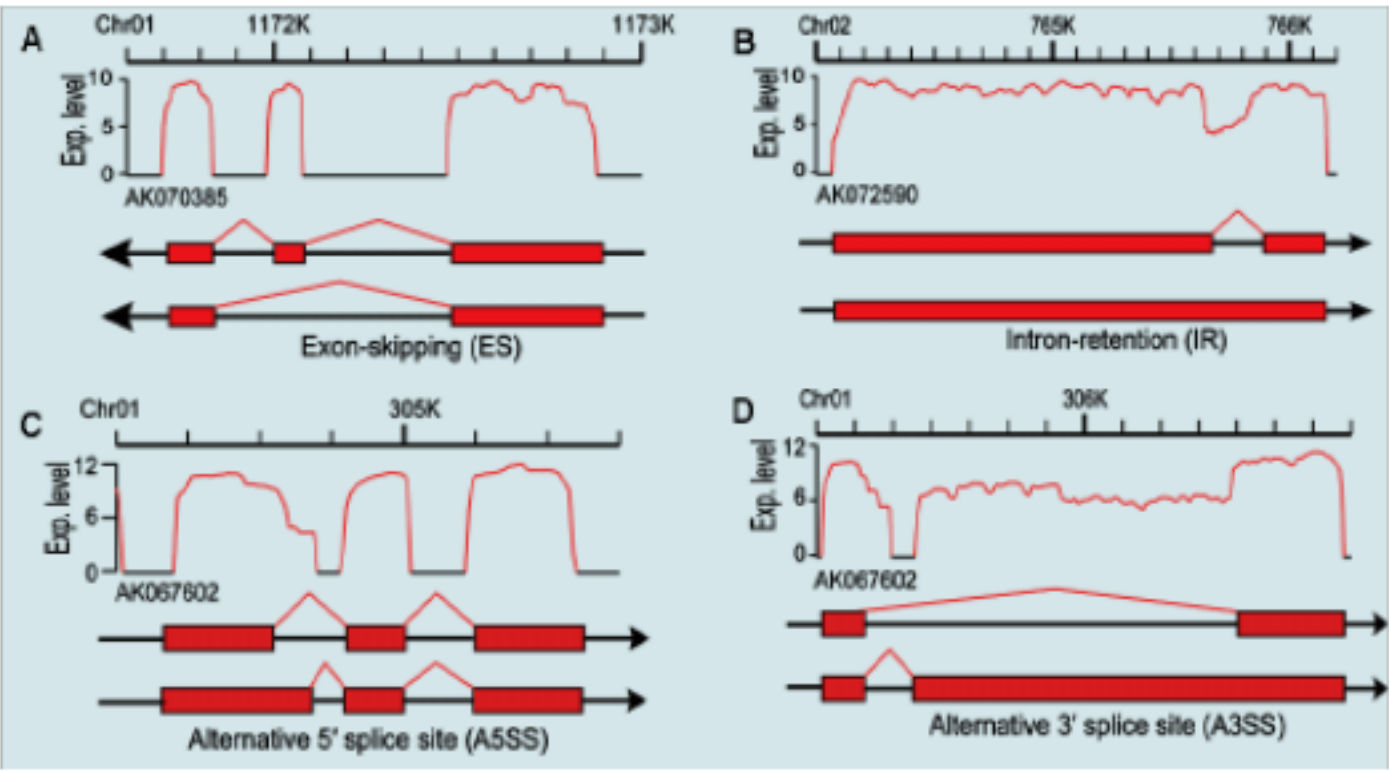


图 2-19: 4 种可变剪接示意图

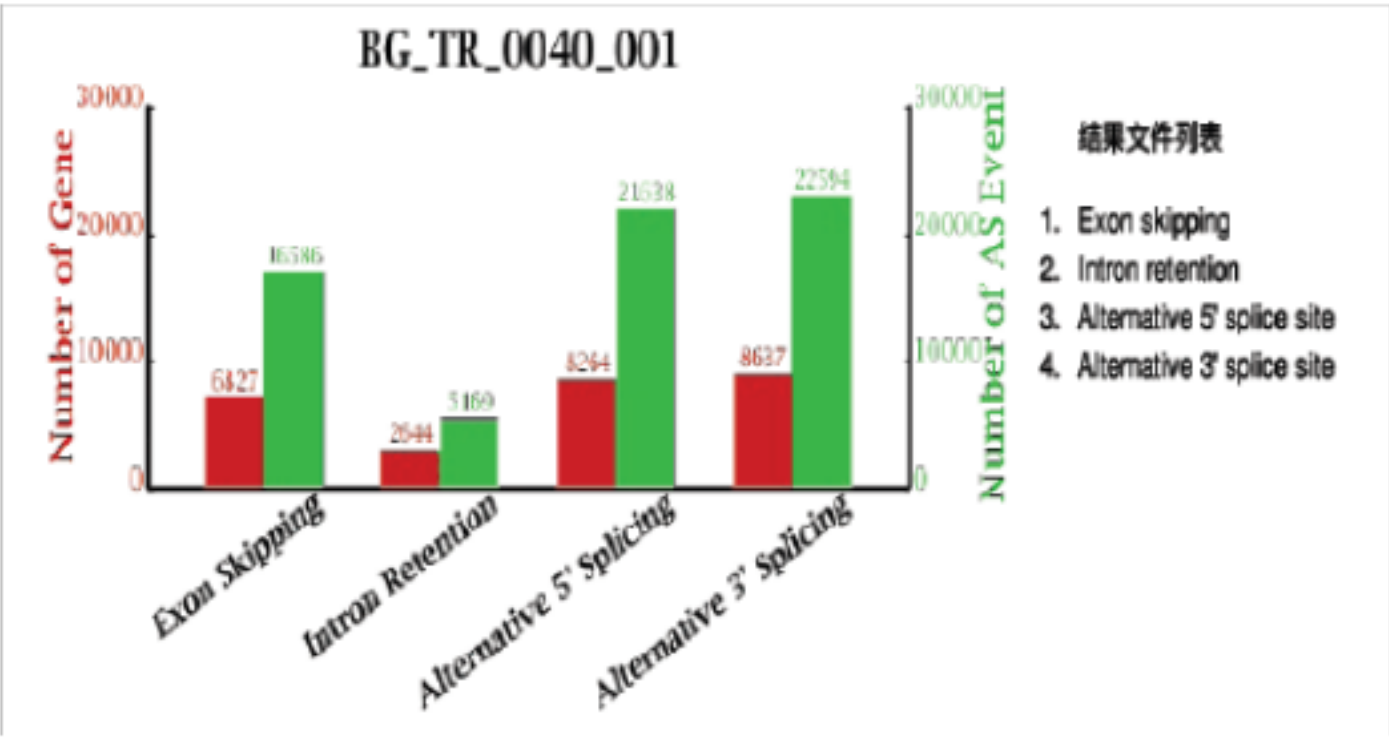


图 2-20 ：结果统计示意图

预测新转录本

现有数据库中对转录本的注释可能还不全面，通过高通量测序我们能检测到新的转录本（Mortazavi et al., 2008）。我们首先从潜在 gene model 中挑选出长度大于 150 bp 且平均覆盖度大于 2 的 gene model，再从中找出位于基因间区域（一个基因 3'端下游 200 bp 到下一个基因 5'端上游 200 bp 之间的区域）的潜在 gene model 作为候选的新转录本。

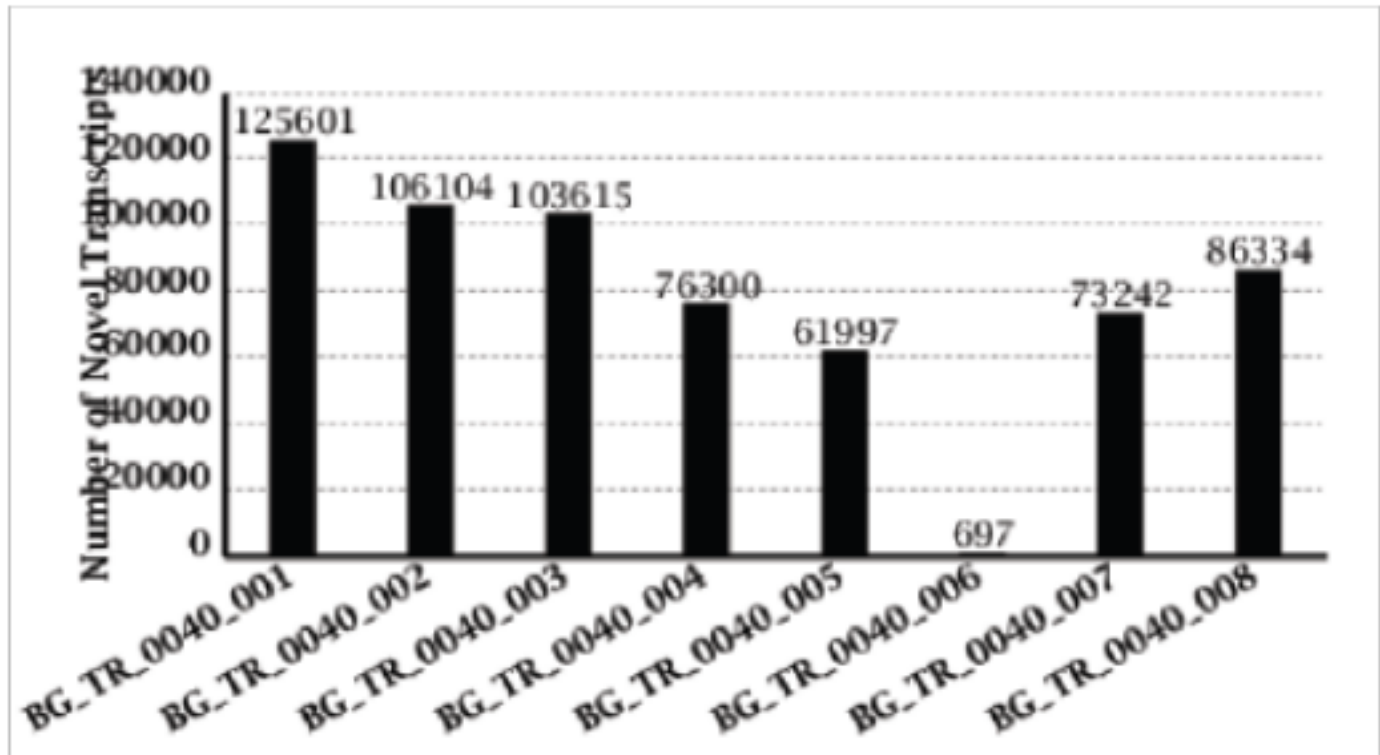


图 2-21：各样品新转录组数量统计图

### SNP 分析

对于多样品的项目，我们使用 SOAPsnp( Li et al., 2009 ) 检测样品间的单核苷酸多态性（SNP）。

### InDel 分析

我们使用华大自主开发的软件 SOAPindel 来检测样本中的短插入或短缺失的序列片段（10 bp）。SOAPindel 是针对重测序技术而专门开发的检测插入和缺失片段的工具（尚未发布）。

### 2.5.3 参考文献

Audic, S. and Claverie, J. M. The significance of digital gene expression profiles. Genome Research (1997).

Conesa, A., Gotz, S. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics (2005).



Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol (1999).

Kanehisa M, Araki M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Research (2008).

Li R, Zhu H, Ruan J, et al .De novo assembly of human genomes with massively parallel short read sequencing. Genome Research (2009).

Mortazavi A, Williams BA, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods (2008).

Pertea G, Huang X, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics (2003).

Ye J, Fang L,et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Research (2006).

### 3 成功案例

#### 3.1 华大成功案例

表 3-1 ： 华大转录组成功案例

物种名	拉丁名	文章
红豆杉	Taxus mairei	Hao DC, Ge GB, Xiao PG, Zhang YY, Yang L. The First Insight into the Tissue Specific Taxus Transcriptome via Illumina Second Generation Sequencing. PIOS ONE2011
水稻	Indic acv.9311	Zhang GJ, Guo GW, Hu XD,et al . Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Research2010
甘薯	Ipomoea batatas	Wang ZY, Chen JY, Zhang XJ et al . De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato ( Ipomoea batatas ). BMC Genomics. 2010
东亚飞	Locusta	Chen S., Yang PC., Jiang F, Wei YY, Ma ZY, Kang LDe Novo

蝗	migratoria	Analysis of Transcriptome Dynamics in the Migratory Locust during the Development of Phase Traits. PLOS ONE2010
米曲霉	Aspergillus oryzae	Wang B, Guo GW, Wang C, Lin Y, Wang XN, Zhao MM, Guo Y, He MH, Zhang Y, and Pan L. Survey of the transcriptome of Aspergillus oryzae via massively parallel mRNA sequencing. Nucleic Acids Research.2010
烟粉虱	Bemisia tabaci	Wang XW, Luan JB, Li JM, Bao YY, Zhang CX and Liu SS.De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics. 2010

3.2 相关文献解读

案例 1 没有参考基因组的转录组研究。

基因组序列未知的物种，通过转录组 de novo 测序，研究转录本序列和基因表达情况，深入研究目标性状形成的分子机制，并可开发 cDNA SSRs( cSSRs) 标记。

案例： Wang Z, Fang B, et al. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato ( Ipomoea batatas ). BMC Genomics2010.

研究背景：

甘薯的块状根是重要农业经济生物器官，可以作为人类主食、动物饲料、工业原材料或生产乙醇原料。 因此，对甘薯这种重要经济物种的深入研究十分必要。然而，到目前为止，因为缺少有用的甘薯基因数据库，加上甘薯是六倍体，本身在基因组组装上有一定难度，这些制约了甘薯分子生物学研究的发展。

研究目的：

利用 Illumina 的 HiSeq 2000 第二代测序平台，建立了第一个甘薯转录组数据库，开发了 cSSRs标记，用于甘薯育种。

案例流程：

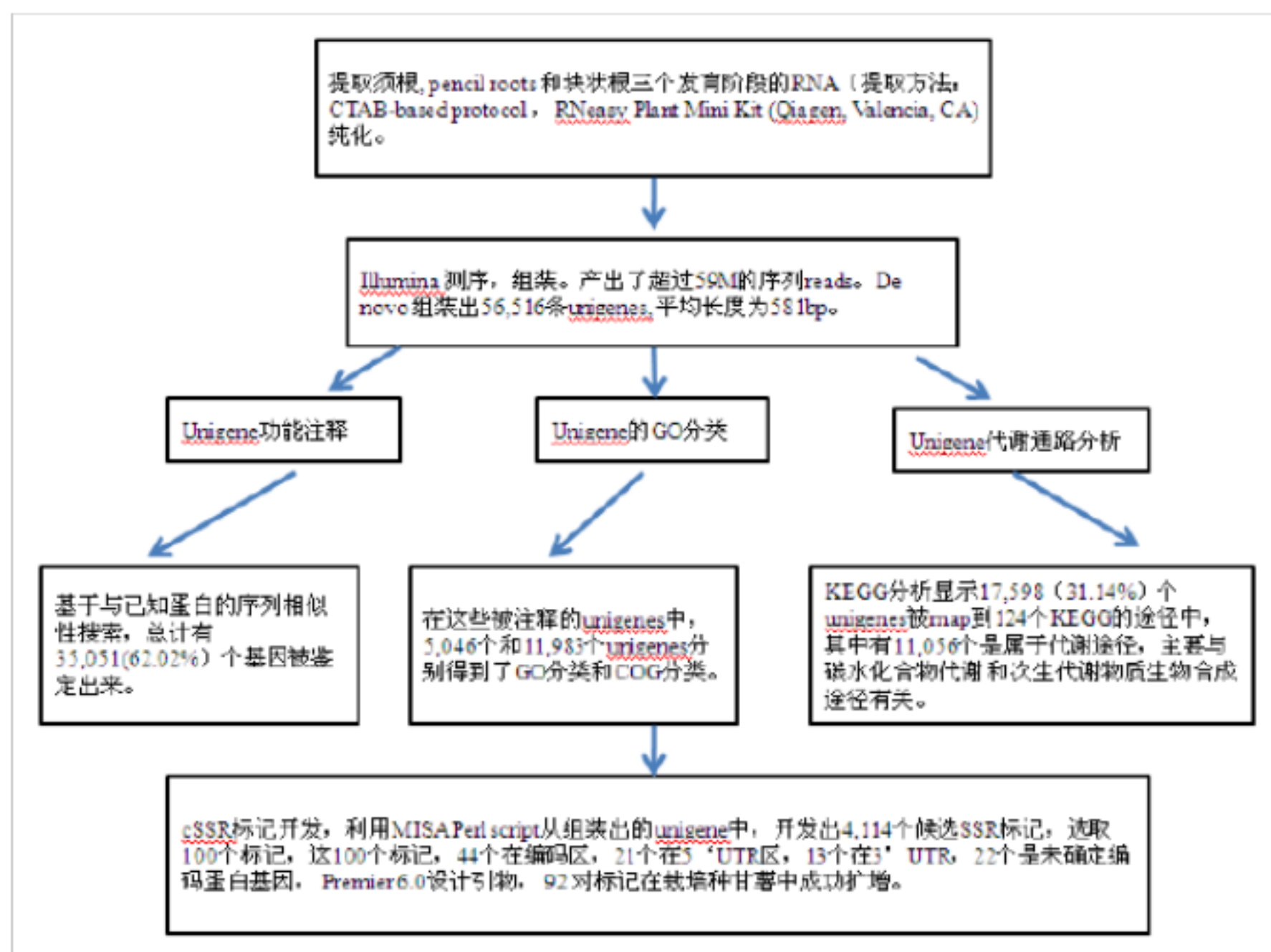


图 3-1：案例 1 流程图

结果：

研究结果表明基于 Illumina HiSeq 2000 的双末端测序的转录组分析可在非模式物种，特别是在基因组大且复杂的物种中，能有效地用于新基因发现和新的分子标记的开发。

## 案例 2 有参考序列的转录组

转录组测序能够全面快速地获得某一物种特定器官或组织在某一状态下的几乎所有转录本，对已知基因组序列的物种进行转录组测序可以研究不同组织不同时期的基因表达变化，鉴定更多的新转录本、融合基因、新 miRNA 以及可变剪切的方式。

案例． Zhang GJ, Guo GW, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Research 2010.

研究背景：

在单碱基水平上对水稻转录组进行系统的研究尚属空白。

研究目的：  
利用转录组测序来研究水稻不同组织的基因表达， 建立较全面的水稻转录组数据库。

案例流程：

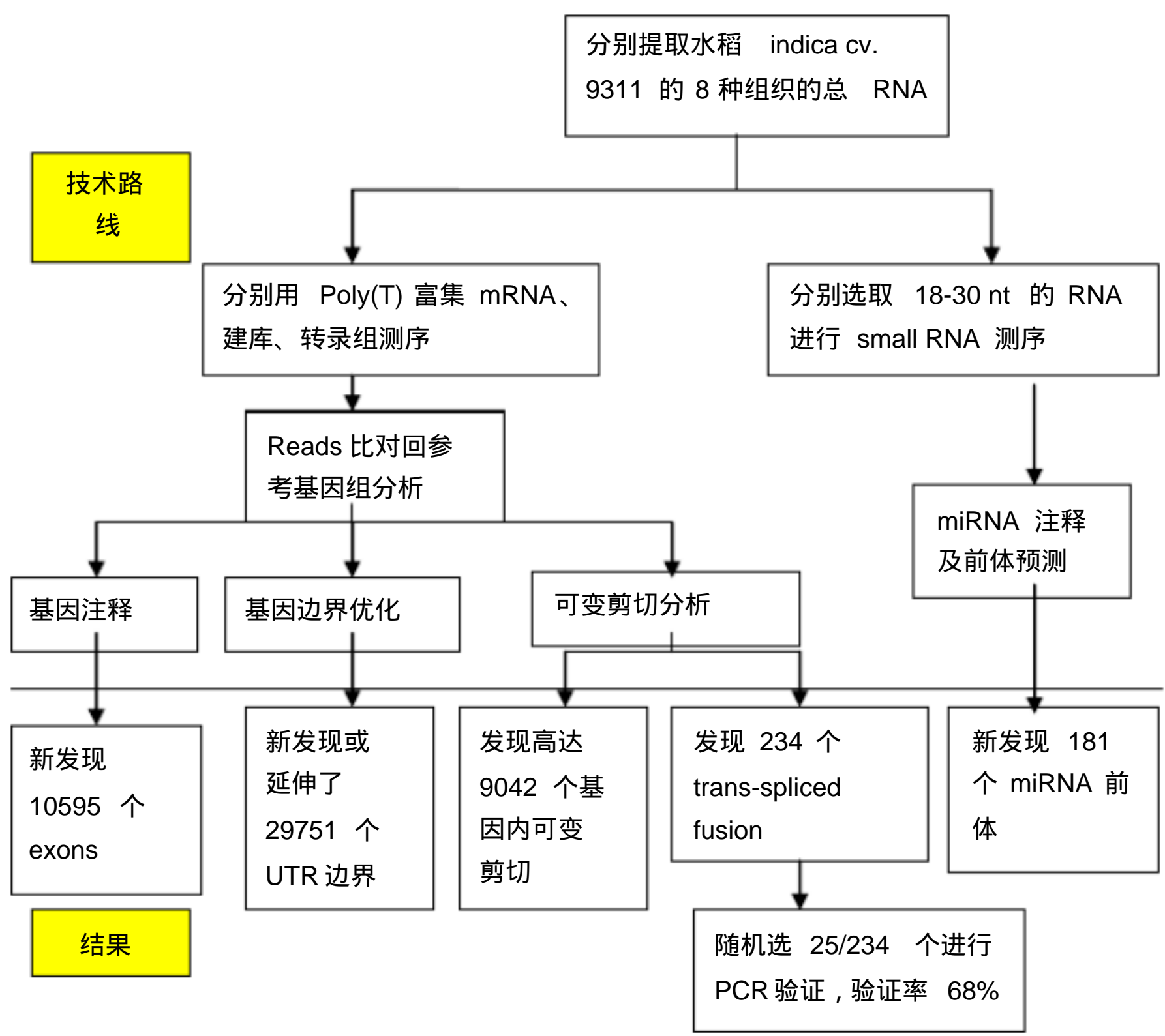


图 3-2：案例 2 流程图