上海伯豪生物技术有限公司

地 址：上海市李冰路151号 (201203)
电 话：021-51320288
传 真：021-51320266
网 址：www.shanghaibiotech.com
邮 箱：market@shbiochip.com

SHANGHAI BIOTECHNOLOGY CORPORATION

Add: No.151, Libing Rd., Zhangjiang Hi-tech Park,
Pudong, Shanghai 201203
Tel: +86-21-51320288
Fax: +86-21-51320266
Website: www.ebioservice.com/eng/index.asp
E-mail: market@shbiochip.com

上海伯豪生物技术有限公司
SHANGHAI BIOTECHNOLOGY CORPORATION

新一代测序服务解决方案

RNA 水平研究

用 我 们 的 平 台 · 加 速 您 的 研 究 !

/ 02 上海伯豪新一代测序服务平台
/ 04 转录组测序服务
/ 13 miRNA 测序服务
/ 17 RIP-SEQ服务

技术服务网站 **WWW.ebioservice.com**
技术服务热线 **800-820-5086 / 400-880-5086**

上海伯豪生物技术有限公司
Shanghai Biotechnology Corporation

上海伯豪生物技术有限公司是上海生物芯片有限公司/生物芯片上海国家工程研究中心根据国内外研发外包发展的需要，整合旗下系统技术平台、商业化服务体系、高素质服务团队等资源成立的致力于研发外包服务公司。

上海伯豪生物技术有限公司拥有五大服务平台：样品处理平台、微阵列芯片平台、高通量测序平台、生物标志物平台、生物信息平台，凭借高标准的技术平台和多样化的服务等竞争优势，公司向国内外企业和相关单位提供系统的生物学研究全面解决方案。目前正在为多达18家跨国制药企业（包括排名前10位的跨国制药企业）和超过1100家的国内科研机构、医院等提供基因表达谱、基因分型、比较基因组学、DNA甲基化、miRNA、生物标志物筛选及确认、生物信息等技术服务。

公司拥有一支以上海为基础，辐射全国的强大市场营销队伍和销售网络，已设立华北、华南、华东三个大区八个办事处，推广公司的主导技术服务、代理产品，快速提升公司品牌的知名度，扩大影响力。过去10年承接项目数超过3000个，用户单位1100家以上，客户发表论文337篇（IF1004），全面推动了中国基因组学服务产业的发展。

公司的系统化技术平台和高质量的服务体系得到了广大客户青睐和合作伙伴的度评价和认可。公司不仅是Affymetrix公司在中国第一家认证的服务提供商，是Agilent公司和ABI公司在国内唯一认证的服务提供商，先后被授予“Agilent公司亚太区最佳服务供应商”和“Affymetrix公司优秀服务商”称号。上海伯豪生物技术有限公司利用高端的现代化仪器设备、多项成熟完善的技术平台、高素质的技术团队、国际水平的商业化服务流程、积极进取的经营理念，为广大客户提供持续、高效和稳定的技术支持和服务。

SBC国内外客户分布图

上海伯豪/SBC
新一代测序服务平台

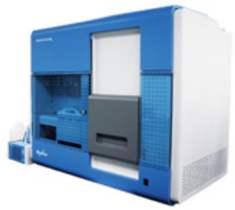
新一代测序技术作为对传统测序的变革，近几年迅速发展。目前，新一代测序技术已广泛应用于生物学研究的各个领域，很多生物学问题都可以借助新一代测序技术予以解决。上海伯豪生物技术有限公司自2009年起先后建立了ABI SOLiD新一代测序服务平台和Illumina新一代测序服务平台，拥有 Illumina HISeq2500(1台)，Illumina *GAIIx*(1台)，ABI SOLiD5500 （2台）和ABI SOLiD4 （2台）。

2010年，上海伯豪生物技术有限公司正式对外推出“SBC新一代测序系统解决方案”，将新一代测序技术与生物芯片技术相结合，协助广大科研工作者以更全面、更深入和更低廉的价格获得基因组、转录组及修饰组的各项数据。力争尽快将新一代测序技术及生物芯片技术相结合，更加系统地分析基因组的结构和表达的变化。

Illumina *GAIIx* /HiSeq2500 新一代测序服务平台

Illumina Solexa *GAIIx* /HiSeq测序系统是一种基于荧光标记的可逆终止化学反应原理实现的单分子簇边合成边测序技术。由于全部四种可逆终止dNTP都会在同一测序循环中出现，其中的自然竞争过程可使合成偏差降至最低，所以对于同聚物的检测也不会有太大问题，这种可逆终止的化学方法严格确保了逐个碱基的测序。目前，illumina solexa *GAIIx* 测序系统双端读长可达300nt，HiSeq2500测序系统高通量模块双端读长可达到200nt，单次运行可产生600G的原始数据。快速模块单次运行仅需27小时，可产生120G的原始数据（以2x100计算）。此外，快速模块读长最长可达到双端300nt，仅需40小时，即可产生180G的原始数据。

Illumina Hiseq2500系统



Illumina solexa *GAIIx* 系统

cBot Cluster Generation

ABI SOLiD 新一代测序服务平台

SOLiD系统包含测序组件、化学组件、计算集群和数据存储组件。这个平台基于通过寡核苷酸连接和检测进行测序。与聚合酶测序方法不同的是，SOLiD系统利用专利的逐步连接（stepwise ligation）技术来产生高质量的数据。目前，SOLiD系统单次运行最大可产生200G的原始数据（以2x60nt计算）。

ABI SOLiD平台一览

上海伯豪新一代测序服务项目

- 基因组测序
  de novo测序、全基因组重测序、DNA序列捕获重测序、RAD-SEQ、单细胞全基因组测序及单细胞外显子捕获测序
- 转录组测序
  转录组测序（无参考基因组）、转录组测序（有参考基因组）、small RNA测序、微量样本转录组测序，定向转录本测序
- 表观基因组测序
  染色质免疫沉淀测序、甲基化DNA免疫共沉淀测序、全基因组Bisulfite甲基化测序

典型应用文章

1. Song Y, Ma K, Ci D, Zhang Z, Zhang D. Sexual dimorphism floral microRNA profiling and target gene expression in andromonoecious poplar (Populus tomentosa). PLoS One. 2013, 8(5):e62681.（IF4.092）（客户发表文章）**使用服务种类：miRNA测序服务**
2. Zou W, Chen D, Xiong M, Zhu J, Lin X, Wang L, Zhang J, Chen L, Zhang H, Chen H, Chen M, Jin M. Insights into the increasing virulence of the swine-origin pandemic H1N1/2009 influenza virus. Sci Rep. 2013, 3:1601.（客户发表文章）**使用服务种类：RNA测序服务**
3. Wang Q, Du HB, Li M, LI Y, Liu SN, Gao P, Zhang XL, Cheng J. MAPK Signal Transduction Pathway Regulation: A Novel Mechanism of Rat HSC-T6 Cell Apoptosis Induced by FUZHENGHUAYU Tablet. Evid Based Complement Alternat Med. 2013.（IF4.774）（客户发表文章）**使用服务种类：miRNA测序服务**
4. Wang T, Cui Y, Jin J, Guo J, Wang G, Yin X, He QY, Zhang G. Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. Nucleic Acids Res. 2013 Mar 21.(IF8.026)（客户发表文章）**使用服务种类：RNA测序服务**
5. Yang J, Duan S, Zhong R, Yin J, Pu J, Ke J, Lu X, Zou L, Zhang H, Zhu Z, Wang D, Xiao H, Guo A, Xia J, Miao X, Tang S, Wang G. Exome Sequencing Identified NRG3 as a Novel Susceptible Gene of Hirschsprung's Disease in a Chinese Population. Mol Neurobiol. 2013 Jan 12.（IF5.735）**使用服务种类：外显子捕获测序数据分析服务**
6. Li J, Li X, Chen Y, Yang Z, Guo S. Solexa sequencing based transcriptome analysis of Helicoverpa armigera larvae. Mol Biol Rep. 2012, 39(12):11051-9.（IF2.929）（客户发表文章）**使用服务种类：mRNA测序服务**
7. Ma L, Yang S, Zhao W, Tang Z, Zhang T, Li K. Identification and analysis of pig chimeric mRNAs using RNA sequencing data. BMC Genomics. 2012, 13(1):429.（IF4.073）（客户发表文章）**使用服务种类：mRNA测序服务**
8. Ye L, Su X, Wu Z, Zheng X, Wang J, Zi C, Zhu G, Wu S, Bao W. Analysis of Differential miRNA Expression in the Duodenum of Escherichia coli F18-Sensitive and-Resistant Weaned Piglets. PLoS One. 2012, 7(8):e43741.(IF4.092)（客户发表文章）**使用服务种类:miRNA测序服务**
9. Huang J, Deng Q, Wang Q, Li KY, Dai JH, Li N, Zhu ZD, Zhou B, Liu XY, Liu RF, Fei QL, Chen H, Cai B, Zhou B, Xiao HS, Qin LX, Han ZG. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet. 2012 Aug 26.（IF35.532）**使用服务种类：全外显子组测序服务**
10. Wu H, Qu S, Lu C, Zheng H, Zhou X, Bai L, Deng Z. Genomic and transcriptomic insights into the thermo-regulated biosynthesis of validamycin in Streptomyces hygroscopicus 5008. BMC Genomics. 2012, 13:337.（IF4.073）（客户发表文章）**使用服务种类：Roche 454测序服务/Agilent earray定制芯片服务**
11. Tao XY, Xue XY, Huang YP, Chen XY, Mao YB. Gossypol-enhanced P450 gene pool contributes to cotton bollworm tolerance to a pyrethroid insecticide. Mol Ecol. 2012 Apr 20.（IF6.457）（客户发表文章）**使用服务种类：Roche 454测序/NimbleGen芯片服务**
12. Gai S, Zhang Y, Mu P, Liu C, Liu S, Dong L, Zheng G. Transcriptome analysis of tree peony during chilling requirement fulfillment: Assembling, annotation and markers discovering. Gene. 2011 Dec 16.(IF2.266)（客户发表文章）**使用服务种类：Roche 454测序服务**
13. Wei X, Ju X, Yi X, Zhu Q, Qu N, Liu T, Chen Y, Jiang H, Yang G, Zhen R, Lan Z, Qi M, Wang J, Yang Y, Chu Y, Li X, Guang Y, Huang J. Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. PLoS One. 2011;6(12):e29500.（IF4.411）**使用服务种类：序列捕获测序服务**
14. Wu DQ, Ye J, Ou HY, Wei X, Huang X, He YW, Xu Y. Genomic analysis and temperature-dependent transcriptome profiles of the rhizosphere originating strain Pseudomonas aeruginosa M18. BMC Genomics. 2011, 12:438.（IF4.206）（客户发表文章）**使用服务种类：Roche 454测序服务 / Agilent earray定制芯片（假单胞菌M18）服务 / SBC Analysis System分析服务**
15. The Schistosoma japonicum Genome Sequencing and Functional Analysis Consortium. The Schistosoma japonicum genome reveals features of host-parasite interplay. Nature, 2009,460(7253):345-51.（IF34.480）**使用服务种类：Roche 454 高通量测序**

转录组测序服务

转录组是特定物种、组织或细胞类型转录的所有RNA（转录本）的集合，包括mRNA和非编码RNA(Non-coding RNA，非编码RNA又包括：tRNA，rRNA，snoRNA，microRNA，piRNA，lncRNA等。通过比较转录组或基因表达谱的研究以揭示生物学现象或疾病发生的分子机制是高通量组学研究的一个常用策略。利用高通量测序技术研究转录组在全面快速得到基因表达谱变化的同时，还可以通过测定的序列信息精确地分析转录本的cSNP（编码序列单核苷酸多态性）、可变剪接等序列及结构变异，另外对于检测低丰度转录本和发现新转录本具有其独特的优势。

转录组测序技术优势

- **直接得到核酸序列信息，除了得到基因表达量的差异，更可以检测RNA的结构和结构变异。**
- **开放性的转录组分析：无需参考基因组信息，无需设计探针，不但能检测已知基因还能够发现新的转录本。**
- **在测序覆盖率足够大时能够检测到细胞中的低丰度转录本。**
- **随着测序深度的增加可以获得更广的动态检测范围，能够同时鉴定和定量高丰度转录本和低丰度转录本。**

转录组测序研究内容

Illumina Solexa *GAIIx* /HiSeq测序系统是一种基于荧光标记的可逆终止化学反应原理实现的单分子簇边合成边测序技术。由于全部四种可逆终止dNTP都会在同一测序循环中出现，其中的自然竞争过程可使合成偏差降至最低，所以对于同聚物的检测也不会有太大问题，这种可逆终止的化学方法严格确保了逐个碱基的测序。目前，illumina solexa *GAIIx* 测序系统双端读长可达300nt，HiSeq2500测序系统高通量模块双端读长可达到200nt，单次运行可产生600G的原始数据。快速模块单次运行仅需27小时，可产生120G的原始数据（以2x100计算）。此外，快速模块读长最长可达到双端300nt，仅需40小时，即可产生180G的原始数据。

基因表达水平研究

基因表达是将基因中蕴含的遗传信息通过转录、剪接、翻译等转变成功能产物的所有加工过程，是生物生命活动的基础和关键。基因从DNA 转录成RNA 是基因表达的首要步骤。基因转录形成的RNA 丰度不同很大程度地影响到基因最终功能产物的分子浓度。因此，检测基因在RNA水平的表达量即基因转录得到的RNA 的丰度成为人们研究基因表达的关键方法。

研究目的	样本类型	实验流程	信息分析	后期验证
基因表达 水平研究	组织、细胞 RNA、全血	抽提、质检 建库、测序	差异表达分析 功能注释	大样本验证 生物学功能验证

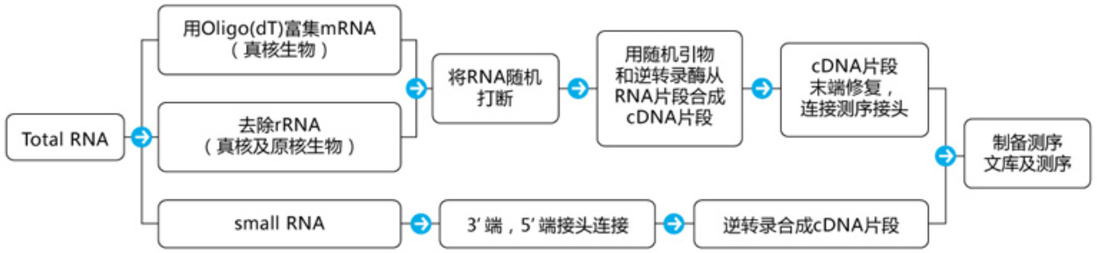
转录本的结构及结构变异研究

利用单碱基分辨率的RNA-Seq　技术可极大地丰富基因注释的很多方面，　包括5′/3′边界鉴定、UTRs区域鉴定以及新的转录区域鉴定等。在发现序列差异(如融合基因鉴定、编码序列多态性研究)方面, RNA-Seq 也展示了其很大的潜力。

研究目的	样本类型	实验流程	信息分析	后期验证
转录本结构及 结构变异研究	组织、细胞 RNA、全血	抽提、质检 建库、测序	转录本结构研究 转录本结构变异研究	

转录组测序服务

转录组测序实验技术路线

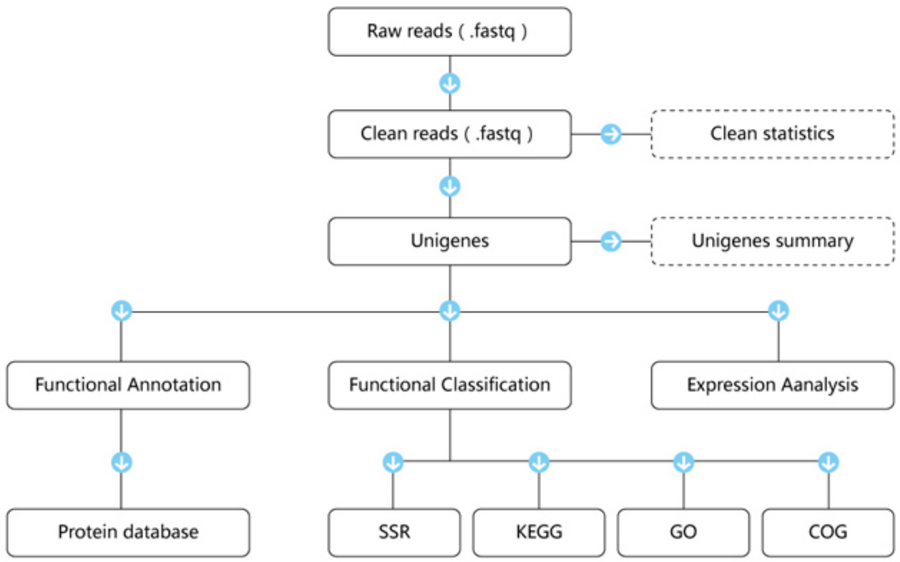


RNA-Seq实验流程

转录组测序样品要求（RNA）

- **样品纯度：**OD 260/280值应在1.9～2.2 之间，RNA 28S:18S≥1.5推荐 RIN≥8；DNA 应该去除干净。
- **样品浓度：**最低浓度不低于100ng/μl。
- **样品总量：**每个样品总量不少于15μg。
- **样品溶剂：**要溶解在H20或TE (pH 8.0)中。
- **样品运输：** RNA用冻存管保存，并用干冰或液氮运输。

转录组测序数据分析技术路线及数据分析内容（无参考基因组）



转录组测序数据分析（无参考基因组）

数据预处理

目的： 对原始测序数据进行一定程度的过滤。

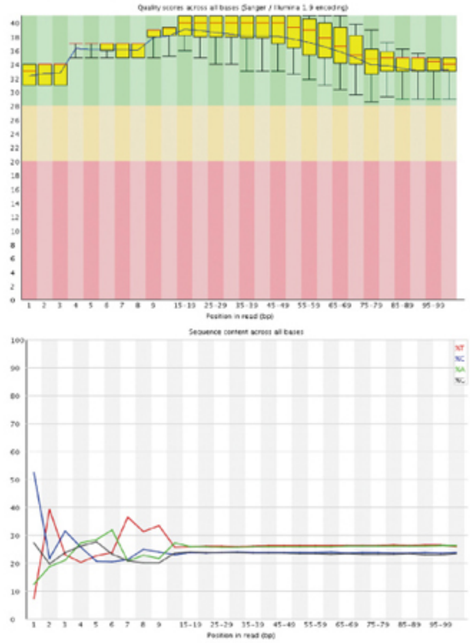
原理： 根据测序接头以及测序质量对原始的测序数据进行预处理，其中，测序质量Q与测序错误E之间的关系如下：

$$Q=-10Log_2E$$

质量与错误率对照表

测序错误率(E)	测序质量值(Q)
5%	13
1%	20
0.1%	30

结果： 对预处理后质量以及碱基分布统计进行统计。

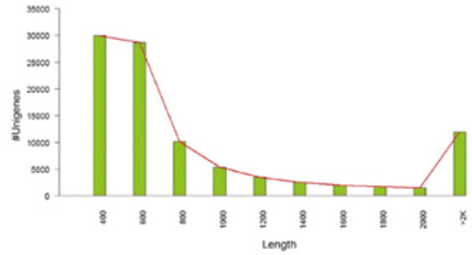


UniGene 拼接

目的： 将预处理后reads进行拼接，得到拼接结果。

原理： 应用 de Bruijn graph path 算法对reads进行denovo拼接；对上一步的拼接结果，再用Hamilton Path算法拼接。

结果： UniGene序列，UniGene统计信息，序列长度分布图。

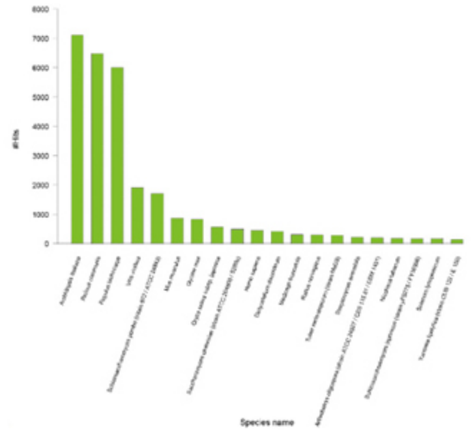


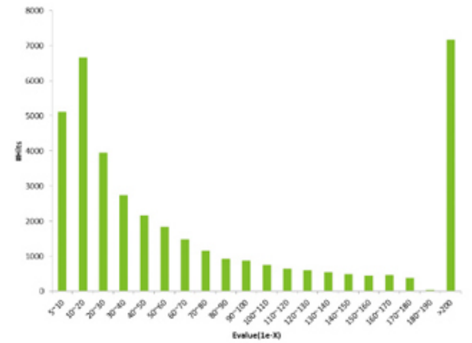
数据库注释

目的： 对拼接得到的UniGene进行功能注释。

原理： 通过blast+算法将拼接得到的UniGene序列与数据库进行比对。

结果： 比对结果表格，物种分布统计和Evalue分布统计。





UniGene 表达分析

目的： UniGene定量分析。

原理： 以UniGene为reference，分别将每个样本的reads进行reference mapping ,从而得到每个样本在每个UniGenes中的一个reads覆盖度，然后应用RPKM/FPKM标准化公式对富集片段的数量进行归一化。

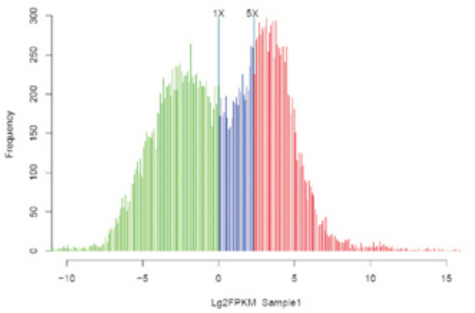
RPKM： Reads Per Kilobase of exon model per Million mapped reads，公式下:

$$RPKM = \frac{\text{total exon Reads}}{\text{mapped reads (Millions)} \times \text{exon length (KB)}}$$

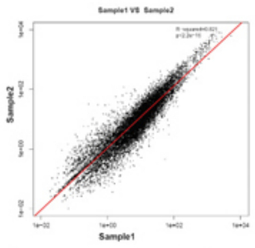
FPKM： Fragments Per Kilobase of exon model per Million mapped reads，公式下:

$$FPKM = \frac{\text{total exon Fragments}}{\text{mapped reads (Millions)} \times \text{exon length (KB)}}$$

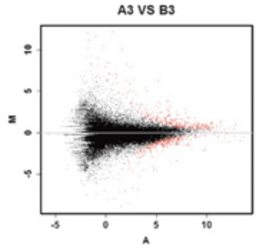
转录组测序服务



UniGene表达分布图，1X，5X分别为FPKM=1，FPKM=5分界点，可以大体观察到低表达，中表达以及高表达的比例关系



UniGene样本间表达相关性散点图



样本间表达差异程度的MA图，可以体现差异表达总体偏差

UniGene表达差异分析

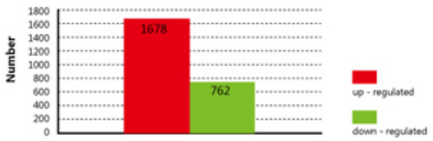
目的： 对定量结果进行统计检验分析，找出差异表达UniGene。

原理： 双层过滤筛选差异基因。

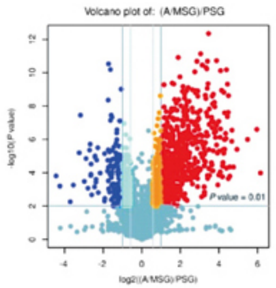
FC值筛选： 采用Fold-change(FC)，表达差异倍数进行第一层此的差异基因筛选。

FDR检验： 一般采用卡方检验中的fisher精确检验进行p值检验，采用Benjamini FDR(False discovery ratio)校验方法对p值进行假阳性检验，即，通过FDR显著性参数进行第二层次的差异基因筛选。

结果展示：



组间差异基因上调与下调个数统计，可以通过此图观察上调与下调的一个总体趋势



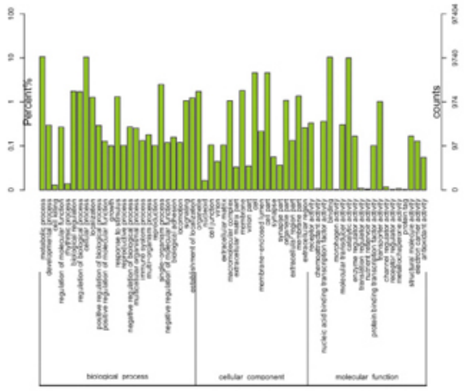
差异基因火山图，可以观察到差异基因总体分布

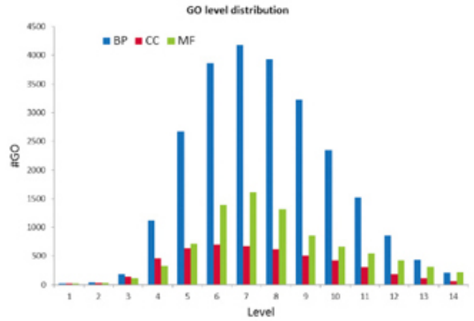
GO 功能分类

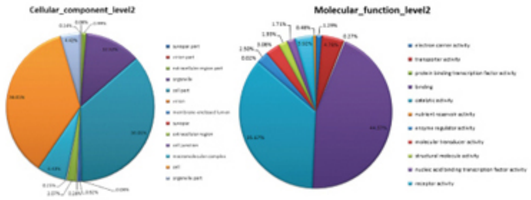
目的： 利用数据库注释信息将 UniGene进行 GO 功能分类。

原理： 利用数据库的注释结果，应用blast2GO算法进行GO功能分类，得到所有序列在 Gene Ontology 的三大类：molecular function, cellular component, biological process 的各个层次所占数目，一般取到14层。

结果： MF，BP，CC三大分类结果文件以及 UniGene2GO 关系列表，三大类别中第二层次上的柱状分布图和饼图，GO功能的层次分布图。





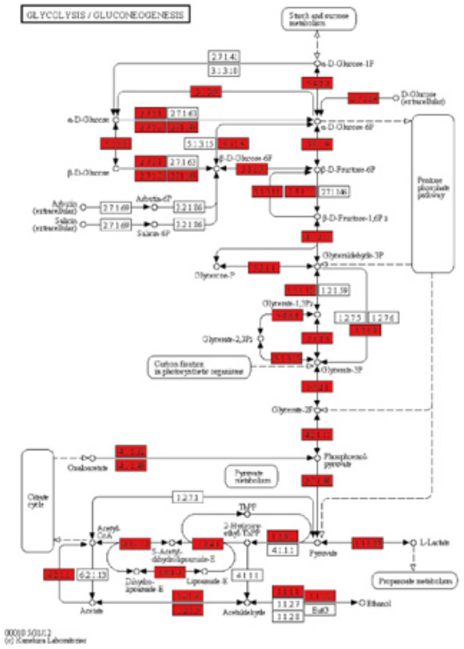


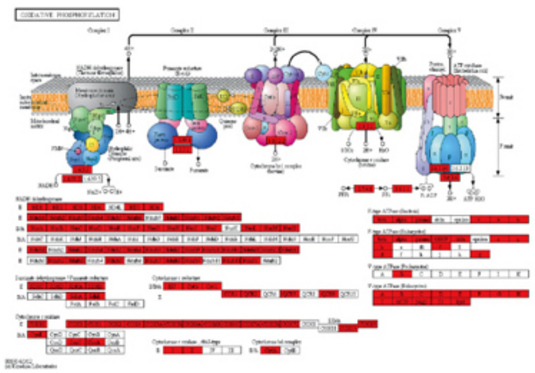
KEGG代谢通路分析

目的： 对拼接得到 UniGene 进行 KEGG pathway 映射。

原理： 应用KEGG KAAS在线 pathway比对分析工具对拼接得到的UniGene进行KEGG映射分析。

结果： 标记的Pathway通路图。



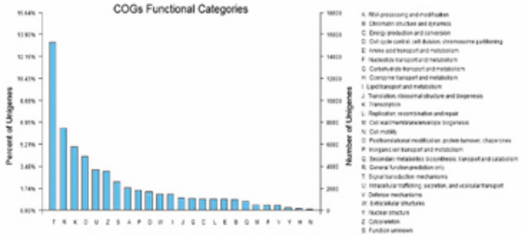


COG 注释

目的： 对拼接得到 UniGene 进行 COG功能分类。

原理： 利用blast+算法将拼接得到的UniGene与CDD库中的COG/KOG库进行比对，进行COG功能分类预测，将其映射到COG分类中。

结果： COG分类分布情况图。



SSR 重复序列注释

目的： 对拼接得到 UniGene进行 SSR 简单重复序列的查找。

原理： 筛选标准：单核苷酸重复的次数在10次或10次以上，二核苷酸重复的次数在 6次或6次以上，三至六核苷酸重复的次数在 5次或 5次以上。同时，也筛选中间被少数碱基 (间隔小于100或等于100)打断的不完全重复的SSR。

结果： 重复序列的信息文件以及统计文件。

LncRNA 预测

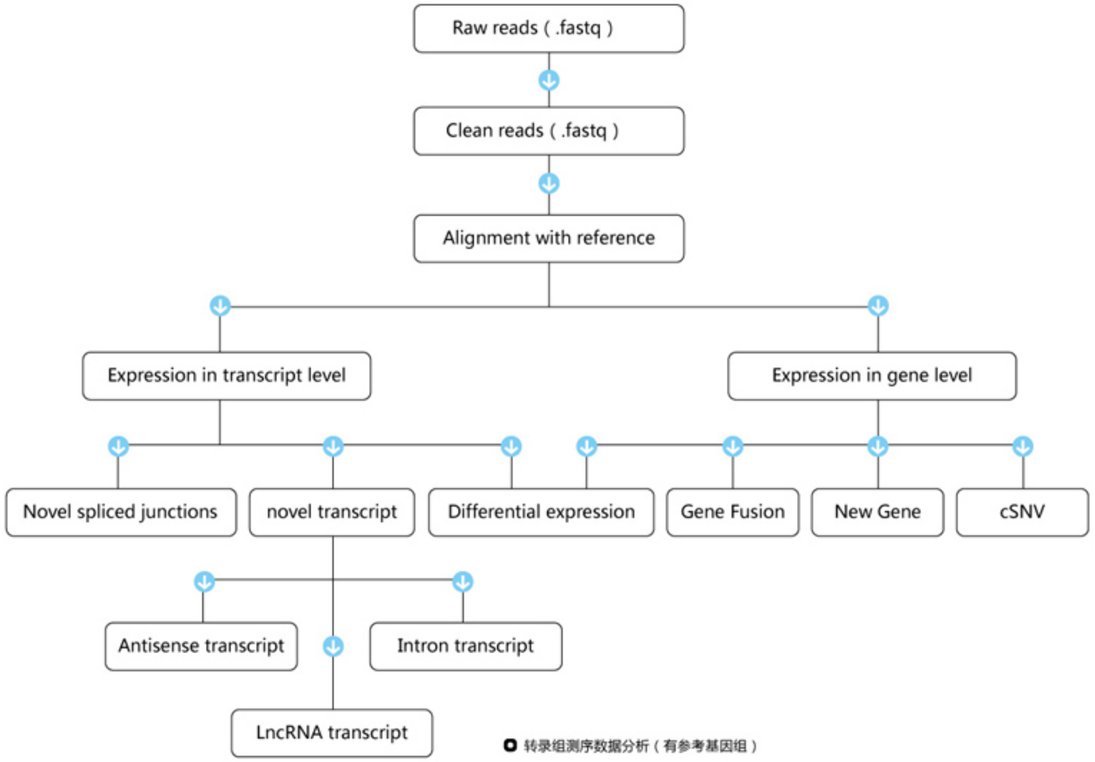
目的： 对拼接得到的UniGene进行LncRNA(Long noncoding RNA)预测。

原理： 通过以下过程对UniGene进行过滤，最终得到候选LncRNA序列。

- Unigene length > 200bp；
- Unigene ORF(Open Reading Frame) length < 300；
- 将满足长度条件的UniGene与多个近源物种进行进化分析，得到序列的保守性和进化特性；
- 根据上述的特性和已知数据库中coding、noncoding区域的特性建立编码筛选模型；
- 将符合noncoding模型的UniGene与Pfam等蛋白域数据库进行同源性比对，进一步去除可能的编码特性，最终得出LncRNA预测结果。

转录组测序服务

转录组测序数据分析技术路线及数据分析内容（无参考基因组）



转录组测序数据分析（有参考基因组）

数据预处理

目的： 对原始测序数据进行一定程度的过滤。

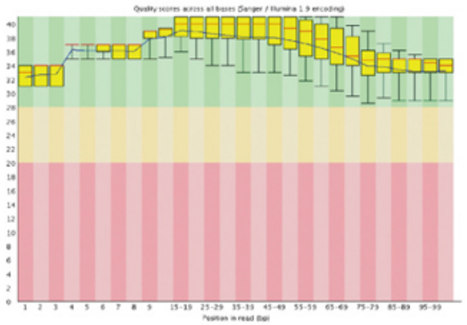
原理： 根据测序接头以及测序质量对原始的测序数据进行预处理，其中，测序质量Q与测序错误E之间的关系如下：

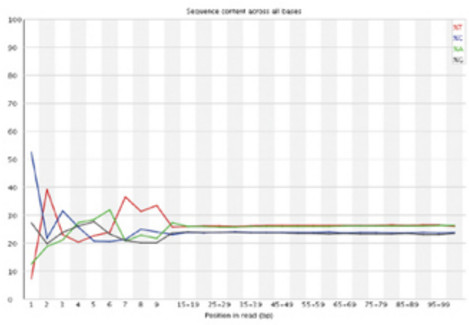
$$Q=-10Log_2E$$

结果： 对预处理后质量以及碱基分布统计进行统计。

质量与错误率对照表

测序错误率(E)	测序质量值(Q)
5%	13
1%	20
0.1%	30





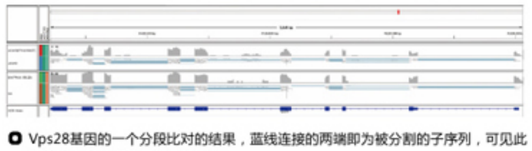
比对基因组

目的： 将经过预处理的测序数据与参考基因组进行相似性比对。

原理： Burrower-Wheeler转换算法与splicing比对算法。

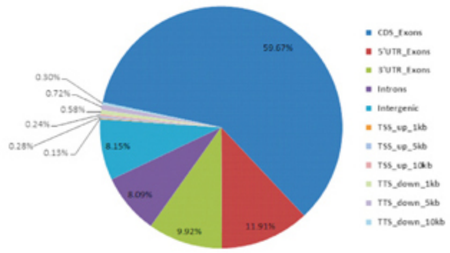
结果： 对预处理后质量以及碱基分布统计进行统计。

- **Burrower-Wheeler转换算法：**由于测序数据量非常大，与整条基因组比对所需资源与时间是较为巨大的。目前，我们采用Burrower-Wheeler(BWT)算法对基因进行建立索引、碱基压缩等过程，这样可以很大程度上加快比对速度，减少比对过程中所需资源。
- **splicing比对算法：**即分段比对算法，当某条测序序列位于转录本剪切位点时，也就是这条序列同时属于两个外显子，如果将它与参考基因组进行比对，由于基因组两个外显子之间含有intron区，那么它将无法找到它合适的位置；但是应用分段比对算法就可以将这条测序序列分割变成多段子序列，然后应用这些段子序列与基因组进行比对，这样就可以找到它们真正的位置。

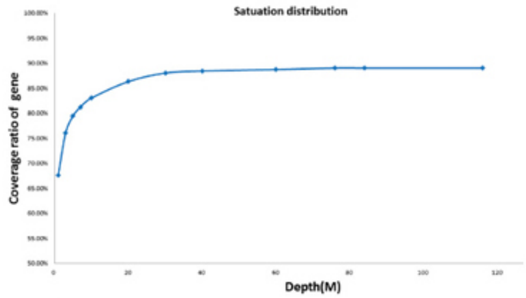


Vps28基因的一个分段比对的结果，蓝线连接的两端即为被分割的子序列，可见此种算法非常的适用于转录组测序。

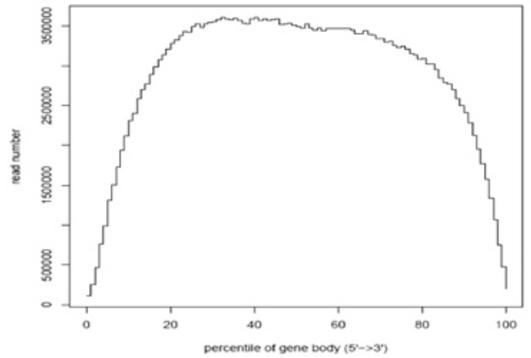
结果展示：应用比对结果进行一些相关mapping统计，测序饱和度及测序5', 3'bias统计。



Multi mapping，Unique mapping及Unique gene-body mapping 统计.



饱和度分析，当reads达到一定测序量后，基因覆盖率基本达到饱和。



测序3', 5'偏好性统计，测序主要集中于基因bady区，两端偏向性较轻。

基因表达水平研究

目的： 应用基因组比对结果进行基因定量。

原理： 从指定物种基因模型(基因结构)中得到gene、exon、intron以及UTR等位置信息，通过基因组比对结果计算出在不用区域富集片段数目，然后应用RPKM/FPKM标准化公式对富集片段的数量进行归一化。

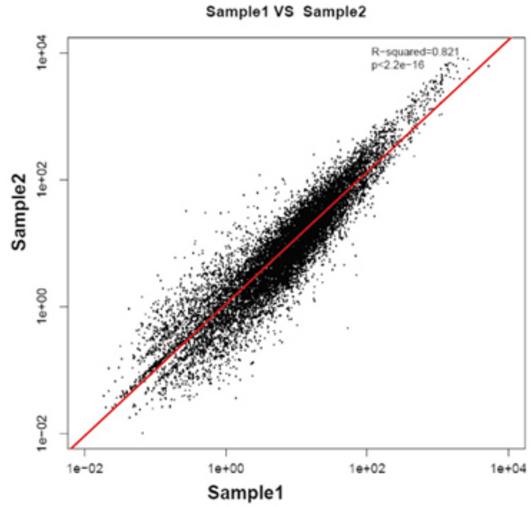
RPKM： Reads Per Kilobase of exon model per Million mapped reads，公式如下：

$$RPKM = \frac{\text{total exon Reads}}{\text{mapped reads (Millions)} \times \text{exon length (KB)}}$$

FPKM： Fragments Per Kilobase of exon model per Million mapped reads，公式如下：

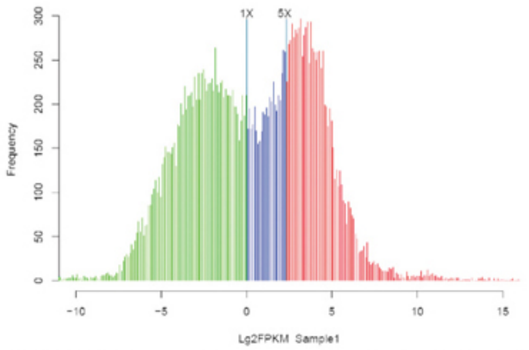
$$FPKM = \frac{\text{total exon Fragments}}{\text{mapped reads (Millions)} \times \text{exon length (KB)}}$$

结果展示：表达量相关性图以及表达量分布图。

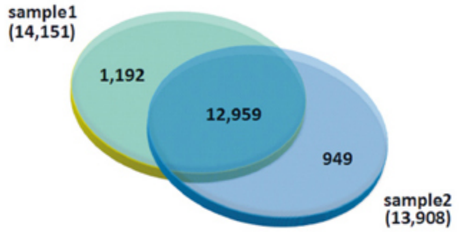


样本表达相关性，通过相关系数R2与显著性p值进行相关性检验。

转录组测序服务



基因表达分布图，1X，5X分别为FPKM=1，FPKM=5分界点，可以大体观察到低表达，中表达以及高表达的比例关系。



样本间表达基因关系饼图，可以看出共同表达以及独有表达的概况。

差异表达分析

目的： 应用统计学方法对基因在样本间的表达差异进行分析。

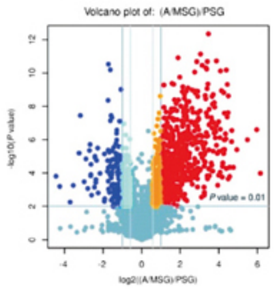
原理： 双层过滤筛选差异基因。

FC值筛选： 采用Fold-change(FC)，表达差异倍数进行第一层此的差异基因筛选。

FDR检验： 一般采用卡方检验中的fisher精确检验进行p值检验，采用Benjamini FDR(False discovery ratio)校验方法对p值进行假阳性检验，即，通过FDR显著性参数进行第二层次的差异基因筛选。

结果展示：

组间差异基因上调与下调个数统计，可以通过此图观察上调与下调的一个总体趋势



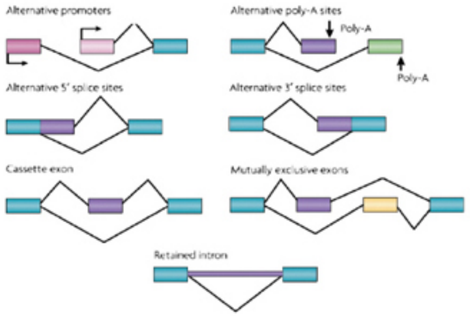
差异基因火山图，可以观察到差异基因总体分布

转录本结构分析

目的： 侦测不同类型的可变剪切事件。

原理： 通过测序序列的splicing事件来侦测可能发生剪切连接的候选exon，通过已有可变剪切方式进行验证，最终得出真实的可变剪切事件。

结果： 对常见的可变剪切方式进行统计分析。



新转录本预测

目的： 预测antisense transcript以及intron transcript。

原理： 通过测序序列在基因组上富集的方向性进行反义转录本预测，如果有富集区域方向与基因转录本方向相反且达到一定的富集阈值，即可认为其为antisense transcript。将完全位于intron区的一段富集片段作为intron transcript。

新基因预测

目的： 预测 intergenic 区可能存在的新基因并对新基因进行功能注释。

原理： 首先，得到在基因间区有测序序列富集的一些段区域；然后，排除那些已经有注释的那些段区域作为候选的新基因。

结果展示：

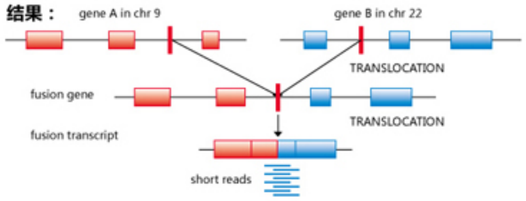
novel_gene_ID	chr_ID	start	end	strand	S1_RPKM	S2_RPKM
XLOC_009269	chr3L	23712607	23712829	+	141.962095	24.358539
XLOC_006159	chr2RHet	2773549	2774009	+	123.868536	4.645177
XLOC_006062	chr2R	15728925	15729265	+	113.252178	10.787401
XLOC_016026	chrX	8235861	8235980	-	110.055983	129.838445
XLOC_002947	chr2L	22888366	22888609	+	108.881655	27.201717
XLOC_006112	chr2R	20626356	20626726	-	105.763197	8.669353
XLOC_002884	chr2L	15107516	15107665	-	103.177186	165.740292
XLOC_009103	chr3L	9060330	9061299	+	98.347576	163.966484
XLOC_009160	chr3L	12701929	12702611	+	93.382837	103.650211
XLOC_009231	chr3L	20756764	20757020	+	91.713319	10.731791

基因融合分析

目的： 寻找可能发生融合功能的基因。

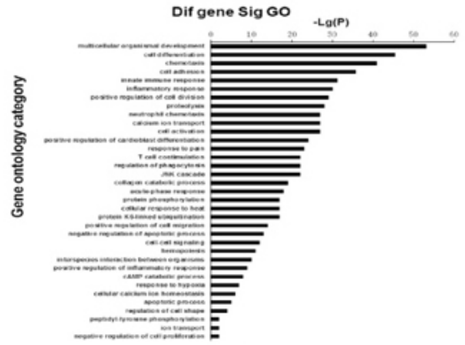
原理： 通过测序片段的splicing事件以及pair-end测序的距离信息进行基因融合位点的定位，如果一个测序片段的一个子片段与geneA匹配，另一子片段与geneB匹配，那么geneA与geneB有可能为一个融合基因，而当pair-end双向测序时，一对测序片段中一个与geneA匹配，另一个与geneB匹配，那么geneA与geneB有可能为一个融合基因。如果同时满足两个条件，那么融合发生的可能性就较大。

结果：



GO 富集分析

目的： 对差异基因相关GO功能进行富集分析。

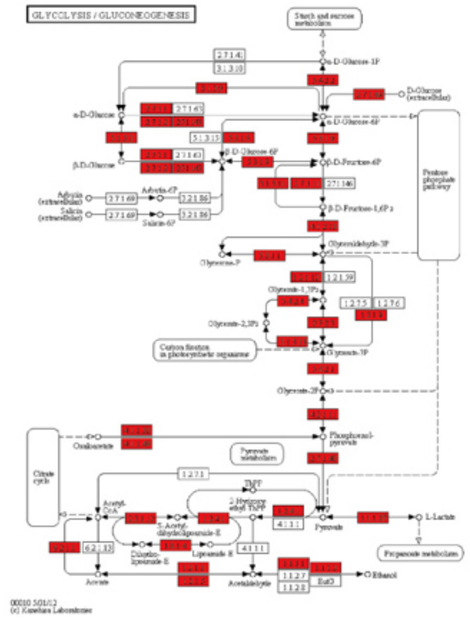


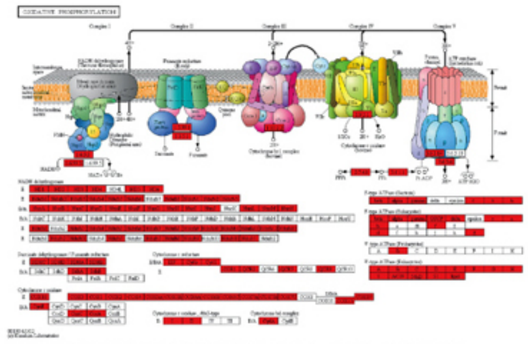
显著富集GO功能图形统计

KEGG 富集分析

目的： 对差异基因进行KEGG通路富集分析。

原理： 应用物种自己的KEGG pathway进行富集分析，富集结果更加贴近物种现实功能实现的通路，尤其对目前功能注释尚不完全的物种，如：大豆、玉米、葡萄、杨树、白菜、牛、羊等物种的KEGG通路分析。





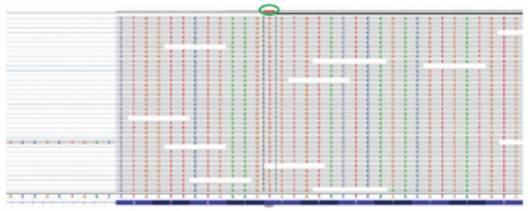
有显著富集功能KEGG通路图，其中，红色标记为差异基因

cSNV 查找

目的： 在转录水平找出变异位点或者片段。

原理： 通过测序数据得到基因组每个位点的碱基富集情况；然后，统计每个点富集富集的碱基种类，得出可能存在的变异(即，与参考基因组碱基不同且富集程度较高的碱基类别)。

结果：



LncRNA 预测

目的： 对新转录本进行LncRNA(Long noncoding RNA) 预测。

原理： 通过以下过程对新转录本进行过滤，最终得到候选LncRNA序列：

- 通过基因组比对得到4类新转录本：Intergenic transcript、Full intron transcript、Antisense transcript、Overlapped with known transcript，将这些新转录本用于LncRNA预测；
- New Transcript length > 200bp；
- New Transcript ORF(Open Reading Frame) length < 300；
- 将满足长度条件的New Transcript与多个近源物种进行进化分析，得到序列的保守性和进化关系；
- 根据上述的特性以及已知数据库中coding、noncoding区域的特性建立编码筛选模型；
- 将符合noncoding模型的New Transcript与Pfam等蛋白域数据库进行同源性比对，进一步去除可能的编码特性，最终得出LncRNA预测结果。

miRNA 测序服务

microRNA(miRNA)是一种大小约21—23个碱基的单链小分子RNA，是由具有发夹结构的约70—90个碱基大小的单链RNA前体经过Dicer酶加工后生成，不同于siRNA（双链）但是和siRNA密切相关。microRNA通过和靶基因mRNA碱基配对引导沉默复合体（RISC）降解mRNA或抑制mRNA的翻译，从而在转录后水平调控蛋白表达（最新发现microRNA也能在转录水平调控基因表达）。microRNA在物种进化中相当保守，在动物、植物和真菌等中发现的microRNA表达均有严格的组织特异性和时序性。microRNA在细胞生长和发育过程中起多种作用，包括调控发育、分化、凋亡和增殖等。

目前研究microRNA的方法主要是realtime-PCR、生物芯片技术以及第二代测序技术。基于第二代测序技术的microRNA测序，可以一次获得数百万条microRNA序列，能够快速鉴定出不同组织、不同发育阶段、不同疾病状态下已知和未知的micoRNA及其表达差异，为研究microRNA对细胞进程的作用及其生物学影响提供了有力工具。

miRNA 测序技术优势

- **高灵敏度：理论上可以检测单个细胞中一个拷贝的microRNA；**
- **高精度：可以检测microRNA单个碱基的差异；**
- **不受先验信息的干扰，既能鉴定已知microRNA，又有能力发现新的microRNA；**
- **保留定向信息，用于链特异的表达分析；**
- **利用Barcode在单次运行中经济地分析多个样品。**

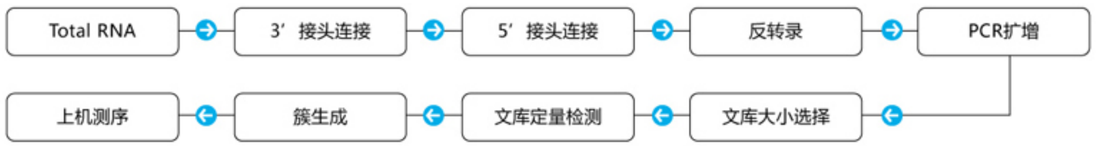
miRNA 测序研究内容

- miRNA表达谱水平研究及新miRNA预测研究：

目前研究microRNA表达量的方法主要是realtime -PCR和生物芯片技术，这些方法主要关注microRNA的表达与定量，局限于那些序列信息已知的microRNA，无法发现和研究未知的microRNA分子。随着第二代测序技术的发展，这一问题得到了解决。基于第二代测序技术的microRNA测序，可以一次性获得数百万条microRNA序列，从而帮助广大科研工作者进行miRNA表达谱水平研究，并预测未知的microRNA。

研究目的	样本类型	实验流程	信息分析	后期验证
miRNA表达谱研究 新miRNA预测研究	组织、细胞 全血、miRNA	抽提、质检 建库、测序	数据过滤 数据分析	大样本验证 qPCR、QGP

miRNA 测序实验技术路线



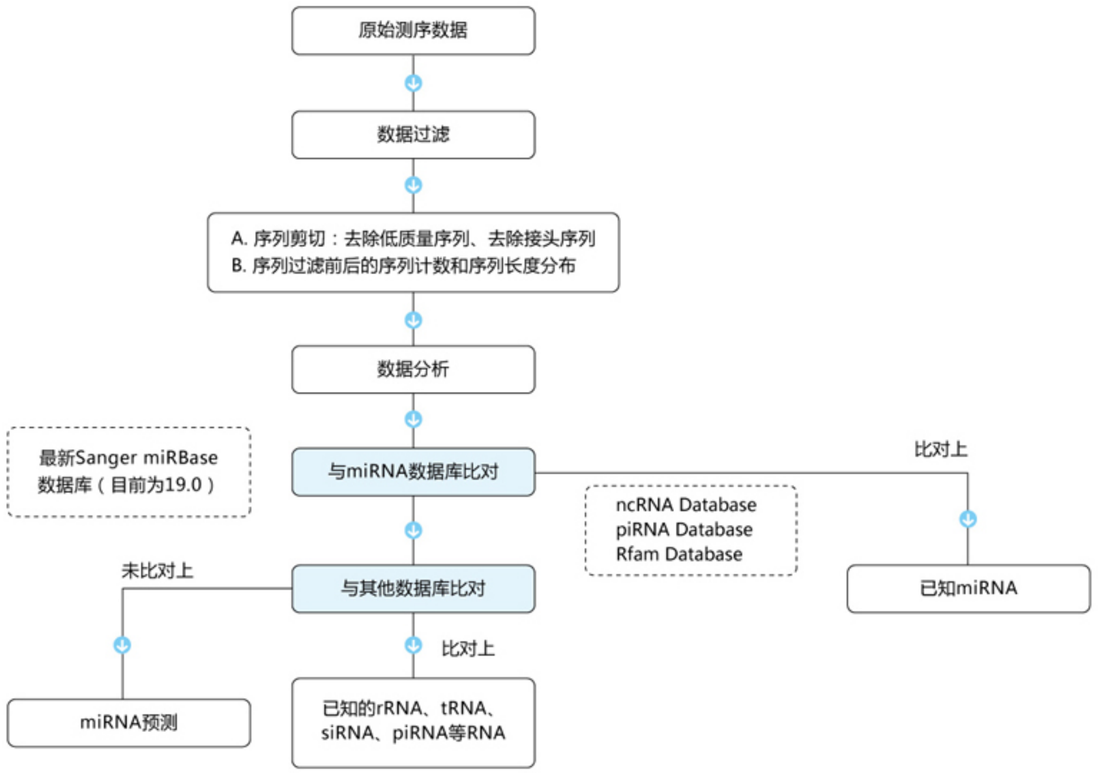
样品要求（含有small RNA的Total RNA）

- **样品纯度：**OD 260/280值应在1.9～2.2 之间，RNA 28S:18S≥1.5推荐 RIN≥8；DNA 应该去除干净；
- **样品浓度：**最低浓度不低于100ng/ul；
- **样品总量：**每个样品总量不少于6ug；
- **样品溶剂：**溶解在H20中；
- **样品运输：** RNA用冻存管保存，并用干冰或液氮运输。

附：样本准备应该遵循的基本原则

- **代表性原则：**取样的代表性关系到实验结果是否具有科学意义，因此您应该根据实验目的慎重选择您的取样方案。病变组织样本中应不夹带正常组织，正常组织样本中不能含有病变组织。有条件时，应做到实验组与对照组的样本在取材时间、部位、处理条件等方面尽可能保持一致，否则可能会影响实验结果的可信度。
- **准确性原则：**代表性样本的各种特征数据必须被准确记录，并按要求（低温、迅速）采集、制备、贮存、运输，最终正确地按实验设计进行实验和数据处理。
- **迅速性原则：**样本质量是实验中影响实验结果的最关键因素，因此用于实验的样本，在采集、制备、贮存、运输过程中应尽可能地做到迅速，最大限度的缩短从样本采集到实验的时间。
- **低温原则：**所取样本离体后，应尽快置于液氮、干冰或-80℃冰箱中，并保证在实验前始终处于-70℃以下，以避免RNA的降解。

miRNA 测序实验技术路线

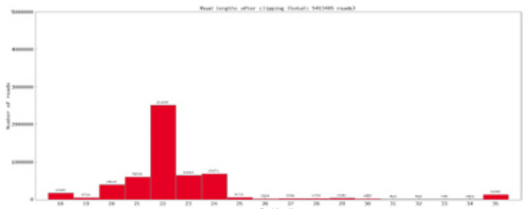


miRNA 测序服务

miRNA测序数据分析内容

microRNA长度分布统计以验证试验可靠性

应用fastx(fastx_toolkit-0.0.13.2)对测序原始reads进行预处理，去除接头序列以及低质量序列。



经过长度统计，各长度上序列的分布情况

比对注释

将测序得到的序列与miRBBase以及其他非编码数据库ncRNA，pirna，Rfam数据库里的序列进行比对，对已知microRNA进行注释：

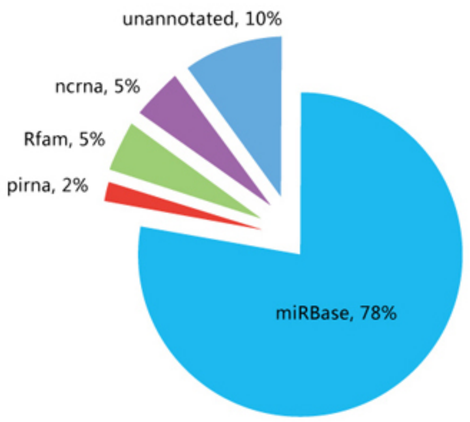
下图为经过注释的结果，其中分别列出和miRBase数据库，pirna数据库，Rfam数据库以及ncRNA数据库的比对情况。

Resource	Sequences in resource	Sequences found	Percentage found
miRBase (Homo sapiens)	1,600	705	44.1%
pirna	171,551	655	0.4%
Rfam v10	444,417	25,940	5.8%
Homo_sapiens.GRCh37.68.ncrna	20,677	2,409	11.7%

下图为针对miRBase种Sus scrofa物种进行的比对注释统计：

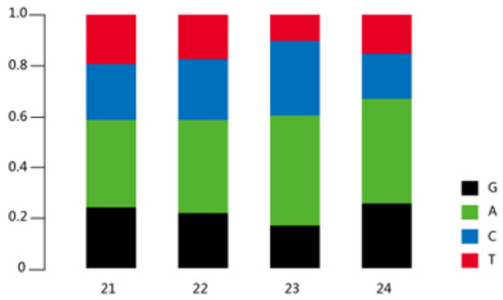
Organism	Total	Mature 5' total	Mature 5' exact matches	Mature 5' length variants	Mature 5' mutant variants	Non-mature total	Mature 3'	Precursor
Sus scrofa	9,323,383	764,675	394,687	308,794	61,794	8,558,708	8,409,720	148,988

由之前所得的注释结果，可以作图来更进一步展示其结果：

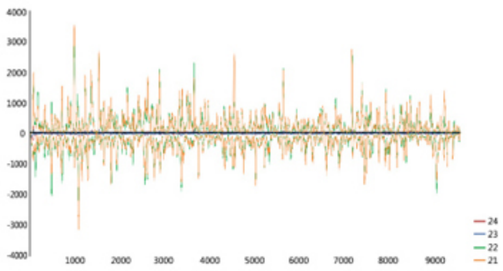


对整体的注释结果，还可以采取进一步的分析，例如：

1. 统计碱基偏好性，下图就是测序所得序列分别在21, 22, 23, 24长度上的5’碱基分布情况。



2. 对于测序所得序列，可以统计出其正负链分布情况，以找寻生物学上的特征。



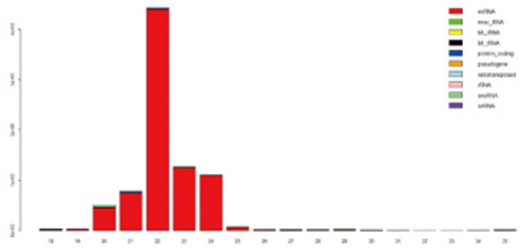
针对某单一microRNA，也可以对其进行更深度的分析。
例如：对其序列的匹配情况进行分别统计：

还可以对其对应的microRNA前体二级结构进行观察。

分类注释

将测序得到的序列与物种所对应的基因组数据库比对，对有注释的reads的来源进行分类统计，鉴定并统计出已知的microRNA以及各种不同种类的RNA分子。

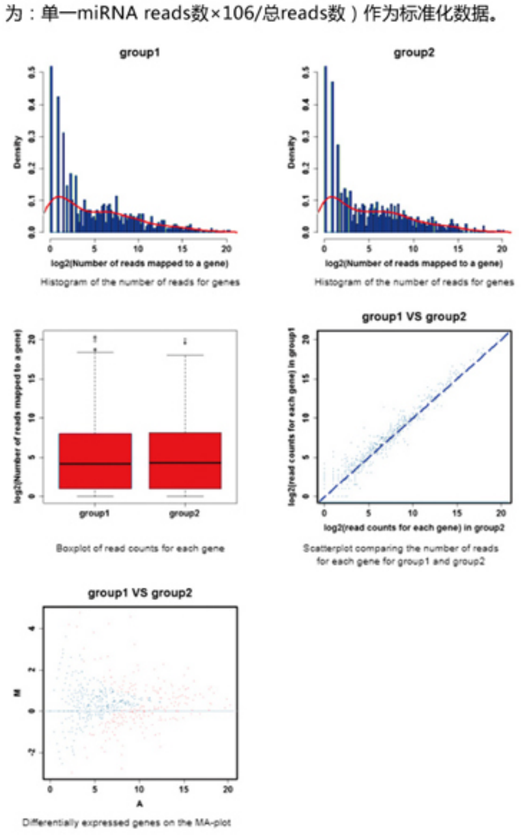
如图，经过与数据库进行分别比对，可以鉴定并统计出包括tRNA，rRNA,snoRNA,snRNA的数量及分布。



差异分析

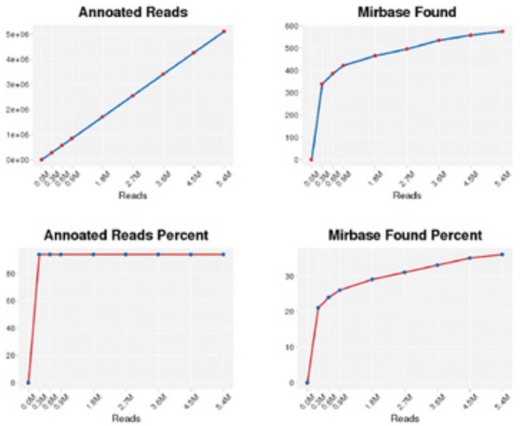
我们采取用DEGseq R语言包结合perl脚本将样品按照客户的分组情况，进行表达量的比较分析。

在差异分析中，我们会采用TPM（Transcripts per million,公式为：单一miRNA reads数×106/总reads数）作为标准化数据。



饱和度分析

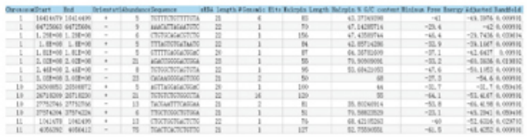
将注释结果按比例划分作图，以观察样品注释的趋势，发现其在生物学上的合理性。



新 microRNA 预测

对于未注释上的序列，我们将其与该物种全基因组序列进行比对分析，通过折叠模型预测新的microRNA,通过折叠模型分析,若有序列位于茎环结构上,则初步判定该序列为一个候选的新microRNA。

对于预测出的新microRNA,我们会统计并列出其所位于的染色体,起始位置,终止位置,正负链,以及数目,长度,GC含量,最小自由能等数值。



对于新microRNA，我们还会计算并绘制出其前体的二级结构，以及其与成熟microRNA之间的位置关系。

mircoRNA 作用靶基因预测

采用miranda软件，对microRNA序列以及对应物种的基因组cDNA序列进行可能的靶位点预测

Miranda软件比对结果示意图如下：

```
miRNA   3’  GUGUUG--GUCG--AUUGGUGUGACGGU  5’
                 |||:|   ||||   ||:| |||||||||
Target  5’  AACAGCAGCAGCAATAGCAACGCTGCCA  3’
```

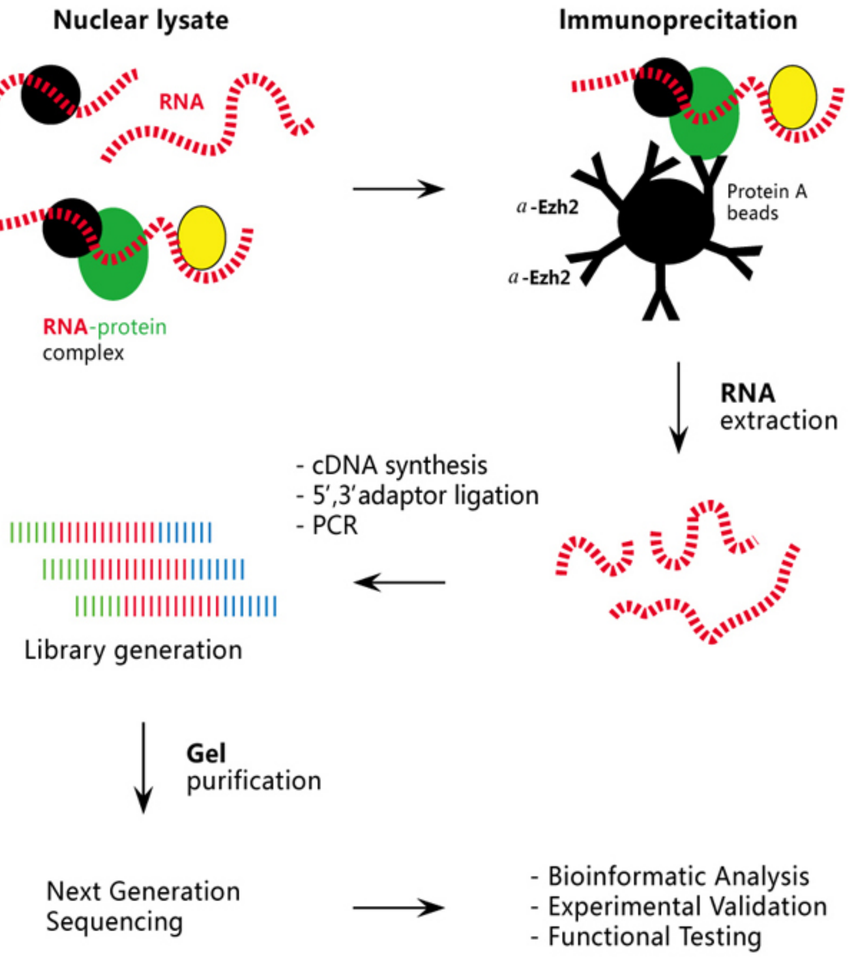
Match Score: 115

RIP-SEQ 服务

RNA Immunoprecipitation（RIP）是研究细胞内RNA与蛋白结合情况的技术，是了解转录后调控网络动态过程的有力工具，能更有效地发现miRNA的调节靶点。RIP技术针对目标蛋白的抗体把相应的RNA-蛋白复合物沉淀下来。之后，经过分离纯化就可以对结合在复合物上的RNA进行测序、芯片等高通量分析。

上海伯豪将RIP技术结合新一代测序技术，推出了RIP-SEQ服务，将有助于广大科研工作者更高通量地了解癌症以及其它疾病整体水平的RNA变化。

RIP -SEQ 实验基本原理



- 用抗体或表位标记物捕获细胞核内或细胞质中内源性的RNA结合蛋白。
- 防止非特异性的RNA的结合。
- 免疫沉淀把RNA结合蛋白及其结合的RNA一起分离出来。
- 结合的RNA序列通过高通量测序（RIP-Seq）方法来鉴定。

上海伯豪RIP-SEQ数据分析服务内容

- **原始数据的QC、数据预处理**
- **RIP-seq 分析**

- **将reads比对到基因组：**
  将预处理reads与reference genome进行mapping，最后得到mapping的sam结果文件，给出mapping结果统计。
- **peak detect：**
  应用sam文件进行peak富集区查找。
- **motif detect：**
  查找结合位点区结构特性，寻找转录因子结合区域的motif结构。
- **基因关联：**
  应用结合位点区位置，确定其周围所涉及的基因。
- **关联基因GO差异分析：**
  以参考基因组为背景集对结合位点关联基因进行GO富集分析。

技术服务网站
WWW.ebioservice.com