美吉生物结题报告

真核转录组纯测序项目

客户：董阁

2025 年 11 月 12 日

目　录

一、项目信息

项目名称			
真核转录组纯测序项目			
合同编号			
MJ20250904414			
项目样本信息			
样本来源	肺		
样本形式	组织		
备注	32个样品		
客户信息			
单位名称	天津医科大学		
单位地址			
课题组	蔡志刚	电话	18322398279
		邮箱	dongge@tmu.edu.cn
项目联系人	董阁	电话	18322398279
		邮箱	dongge@tmu.edu.cn
公司联系人信息			
销售员	夏应升	电话	15137601880
		邮箱	yingsheng.xia@majorbio.com
技术支持	张锟	电话	8185
		邮箱	rna@majorbio.com
项目审批人			
签名：_____ _____年___月___日			

二、测序实验流程

真核mRNA测序是基于Illumina平台，对真核生物特定组织或细胞在某个时期转录出来的所有mRNA进行测序，测序实验采用Illumina TruseqTM RNA sample prep Kit方法进行文库构建，其操作流程图及仪器试剂如下图和下表所示：

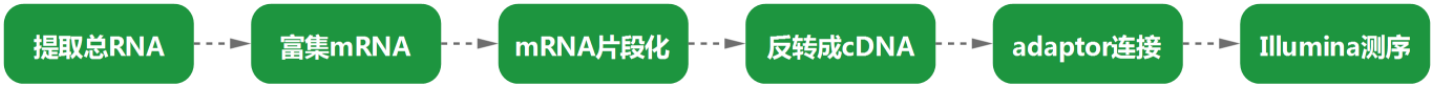

图2-1 真核转录组实验流程

表2-1 试剂仪器表

实验步骤	试剂仪器名称	厂商
mRNA分离	磁力架	Invitrogen
建库试剂	TruseqTM RNA sample prep Kit	Illumina
定量	Quantus™ Fluorometer QuantiFluor® dsDNA System	Promega
文库回收	Agencourt AMPure XP	Beckman
上机测序	HiSeq X Reagent Kits NovaSeq Reagent Kits	Illumina

2.1 提取total RNA

从组织样品中提取total RNA，利用Nanodrop2000对所提RNA的浓度和纯度进行检测，琼脂糖凝胶电泳检测RNA完整性，Agilent2100测定RIN值。单次建库要求RNA总量1ug，浓度≥50ng/μL，OD260/280介于1.8~2.2之间。

2.2 Oligo dT富集mRNA

真核生物mRNA 3'末端具有ployA尾的结构，利用带有Oligo(dT)的磁珠与ployA进行A-T碱基配对，可以从总RNA中分离出mRNA，用于分析转录组信息。

2.3 片段化RNA

Illumina平台是针对短序列片段进行测序，富集得到的mRNA是完整的RNA序列，平均长度达几

kb，因此需要对其进行随机打断。加入fragmentation buffer，选择合适条件，可以将mRNA随机断裂成300bp左右的小片段。

2.4 反转合成cDNA

在逆转录酶的作用下，利用随机引物，以mRNA为模板反转合成一链cDNA，随后进行二链合成，形成稳定的双链结构。

2.5 连接adaptor

双链的cDNA结构为粘性末端，加入End Repair Mix将其补成平末端，随后在3’末端加上一个A碱基，用于连接Y字形的接头。

2.6 片段筛选和文库富集

对连接adapter后的产物进行纯化和片段分选，用分选产物进行PCR扩增，纯化得到最终的文库。

2.7 Illumina平台上机测序

1) QuantiFluor® dsDNA System定量，按数据比例混合上机；

2) cBot上进行桥式PCR扩增，生成clusters；

3) Illumina平台测序。

三、项目结果报告

本项目采用Illumina测序平台完成转录组测序，构建Illumina PE文库进行2×150 bp测序，对获得的测序数据进行质量控制（质控），之后利用生物信息学手段对转录组数据进行分析。

3.1 原始测序数据说明

为方便测序数据的分析、发布和共享，Illumina平台测序得到的原始图像数据经过Base Calling转化为序列数据，得到最原始的测序数据文件。原始数据一般存储为FASTQ格式。FASTQ格式文件可记录所测读段（Reads）的碱基及其质量分数。如图3-1所示，FASTQ格式以测序读段为单位进行存储，每条Reads在FASTQ格式文件中占四行，其中第一行和第三行由文件识别标志（Sequence Identifiers）

和读段名（ID）组成（第一行以“@”开头而第三行以“+”开头；第三行中ID可以省略，但“+”不能省略），第二行为碱基序列，第四行为对应位置碱基的测序质量分数。

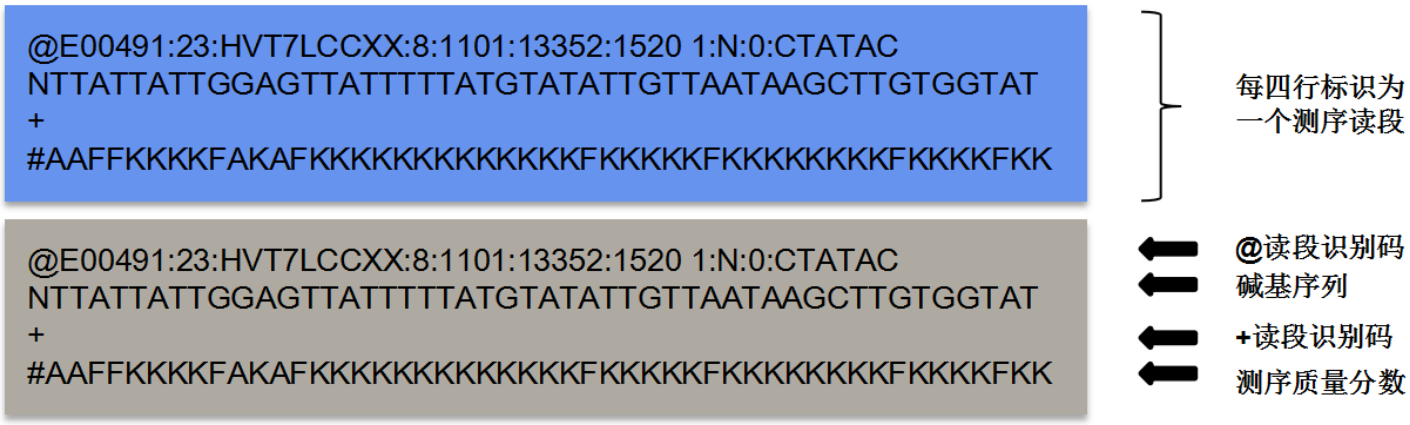

图3-1 读段FASTQ数据格式示例

Illumina测序仪一个Run有2个Flowcell，一个Flowcell中包含8个Lane，其中一个Lane包含2列，每一列又包含60个Tile，每一个Tile又会种下不同的Cluster，其产生的测序文件识别标志（Sequence Identifiers）的详细说明如表3-1所示：

表3-1 测序文件读段识别码说明

标识	英文描述
E00491	Unique instrument name
23	Run ID
HVT7LCCXX	Flowcell ID
8	Flowcell lane
1101	Tile number within the flowcell lane
13352	'x'-coordinate of the cluster within the tile
1520	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise

0	0 when none of the control bits are on, otherwise it is an even number
CTATAC	Index sequence

Reads的质量分数以不同的字符来表示，在Illumina平台中，将每个字符对应的ASCII码减去33，即为对应的测序质量值。一般地，碱基质量从0-40，即对应的ASCII码为从“!”（0+33）到“I”(40+33），碱基质量越大，可信度越高。用e表示测序错误率，用Q表示Illumina HiSeqTM的碱基质量值，则有下列关系：

$$Q = -10 \times \lg e$$

表3-2 测序错误率与测序质量值对应关系简明表

测序错误率（e）	测序质量值（Q）	对应ASCII码
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

Illumina测序属于第二代测序技术，单次运行能产生数百万级的Reads，如此海量的数据无法逐个展示每条Reads的质量情况；运用统计学的方法，对所有测序Reads的每个Cycle进行碱基分布和质量波动的统计，可以从宏观上直观地反映出样本的测序质量和文库构建质量。我们针对每一个样本的原始测序数据进行测序相关质量评估，包括A/T/G/C碱基含量分布统计和碱基错误率分布统计。

3.2 原始测序数据质控

由于原始测序数据中会包含测序接头序列、低质量读段、N（N表示不确定碱基信息）率较高序列及长度过短序列，这将严重影响后续分析的质量。所以，在分析之前会先对原始测序数据进行质控，从而得到高质量的质控数据（clean data）以保证后续分析结果的准确性。

使用软件：fastp

具体步骤及顺序如下：

1) 去除reads中的接头序列，去除由于接头自连等原因导致没有插入片段的reads；

2) 将序列末端（3’端）低质量（质量值小于20）的碱基修剪掉，如剩余序列中仍然有质量值小于10的碱基，则将整条序列剔除，否则保留；

3) 去除含N（模块碱基）的reads；

4) 舍弃去adapter及质量修剪后长度小于30bp的序列。

数据质控完成后，对质控后的数据再次进行统计以及质量评估，同样包括：

① 碱基错误率分布统计；

② A/T/G/C碱基含量分布统计。

接头序列为：

5’：AGATCGGAAGAGCACACGTC

3’：AGATCGGAAGAGCGTCGTGT

3.3 测序数据统计

利用Illumina的建库测序平台，构建插入片段大小为400 bp左右的测序文库，按照项目合同要求进行测序。Illumina测序单次运行能产生数十亿级的reads，如此海量的数据无法逐个展示每条read的质量情况；因此我们运用统计学的方法，对所测序列进行统计，可以从宏观上直观地反映出样本的文库构建质量和测序质量。详细结果见表3-3：

表3-3 测序数据统计表

Sample ID	Raw Reads	Raw Base	Q20(%)	Q30(%)	GC(%)
n20_2	21527854	6501411908	99.1034	96.3463	50.7135
n20_1	21151346	6387706492	99.0644	96.1832	50.9273
w20_2	24578593	7422735086	98.9597	95.9424	50.6465
w20_1	21303955	6433794410	98.9685	95.9677	50.5326
n18_2	23405120	7068346240	99.0237	96.216	49.9035

n18_1	19944193	6023146286	99.0086	96.1142	49.8807
w15_1	20972169	6333595038	99.0032	96.0937	48.5477
n12_2	20879747	6305683594	98.9977	96.1086	49.9473
w9_2	21223866	6409607532	99.0028	96.0751	50.1303
w6_2	18208161	5498864622	99.0065	96.1742	48.7299
w6_1	22627416	6833479632	98.9771	96.0696	48.341
n4_2	20846321	6295588942	98.9773	96.1019	48.3455
n4_1	24961933	7538503766	99.0133	96.1902	49.0778
w4_2	24994587	7548365274	98.9808	96.0737	49.2694
w4_1	20629425	6230086350	98.9814	96.1342	49.0562
left_w12_1	23479819	7090905338	99.0367	96.151	48.629
left_n9_2	23146310	6990185620	99.0572	96.2249	49.77
left_n9_1	23260677	7024724454	99.0299	96.106	49.5389
left_w9_2	24411420	7372248840	98.9979	96.1147	48.7213
left_w9_1	22423207	6771808514	99.034	96.131	49.4096
left_n6_2	24207136	7310555072	99.0565	96.2507	48.9169
left_w6_2	21898610	6613380220	99.0384	96.1191	49.2239
left_w6_1	21322180	6439298360	99.0257	96.0849	48.5574
left_n5_2	19493590	5887064180	99.0089	96.0867	47.8397
left_n5_1	23470015	7087944530	99.0676	96.3161	48.8303
left_w5_2	21364864	6452188928	99.0427	96.1159	49.6553
left_w5_1	23788379	7184090458	99.0498	96.2174	49.6269
left_n15_2	18219593	5502317086	98.7683	95.4115	48.5249

left_n15_1	20734226	6261736252	98.7915	95.3457	48.8804
left_n12_2	19746645	5963486790	98.7902	95.3758	50.4034
left_n12_1	19146060	5782110120	98.8824	95.6293	49.3778
left_w12_2	18392876	5554648552	98.7615	95.4315	48.6175

注：

Sample ID：样本编号；

Raw reads：原始测序数据的总条目数（reads，代表测序读段，一个reads即为一条）；

Raw bases：原始测序总数据量（即Raw reads数目乘以reads读长）；

Q20（%）、Q30（%）：Q20、Q30分别指测序质量在99%和99.9%以上的碱基占总碱基的百分比，一般Q20在85%以上，Q30在80%以上；

GC(%)：G和C碱基总和占总碱基的百分比。

3.5 测序碱基含量分布统计

碱基含量分布检查一般用于检测有无AT、GC分离现象。鉴于序列的随机性和碱基互补配对的原则，理论上每个测序循环上的GC含量相等、AT含量相等，且在整个测序过程基本稳定不变，呈水平线。N为测序仪无法判断的碱基类型。本项目中样品的碱基含量分布图如图3-2所示，反映出该样品的文库构建质量和测序质量均可满足后续分析。

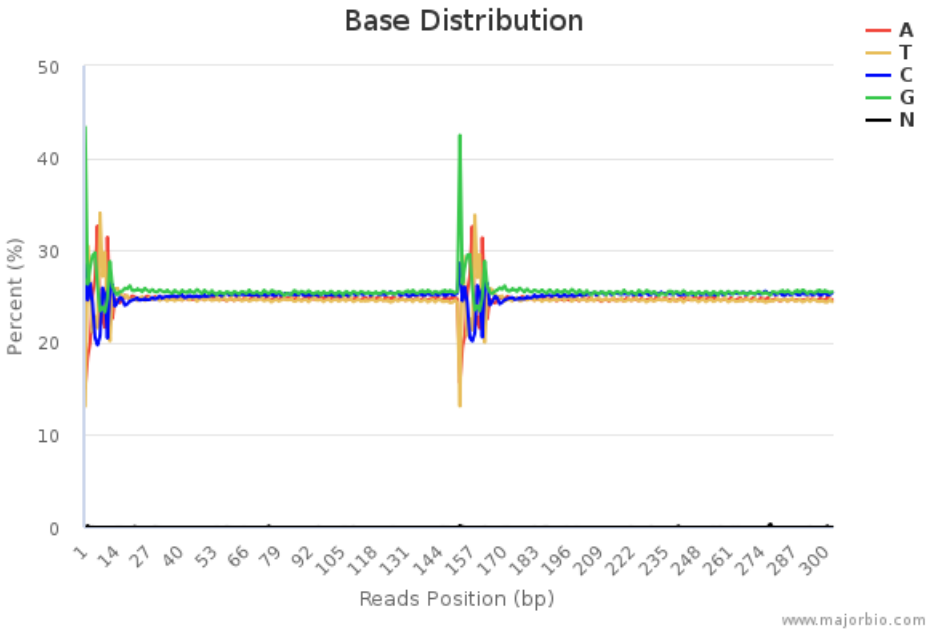

图3-2 样品的碱基含量分布图

注：横坐标是Reads碱基坐标，坐标表示Reads上从5'到3'端依次碱基的排列；纵坐标是所有Reads在该测序位置A、C、G、T、N碱基分别占的百分比，不同碱基用不同的颜色表示。序列的起始位置与测序的引物接头相连，因此A、C、G、T在起始端会有所波动，后面会趋于稳定。模糊碱基N所占比例越低，说明未知碱基数越少，测序样本受系统AT偏好影响越小。虚线左侧为Read1的统计，虚线右侧为Read2的统计结果。

3.6 测序碱基错误率分布统计

测序错误率会随着测序序列长度的增加而升高，这是由于测序过程中化学试剂的消耗导致的，另外，由于Illumina HiseqTM测序的技术特点，测序片段前端几个Cycles和末端的错误率会偏高。本项目中样品的测序错误率分布如图3-3所示：

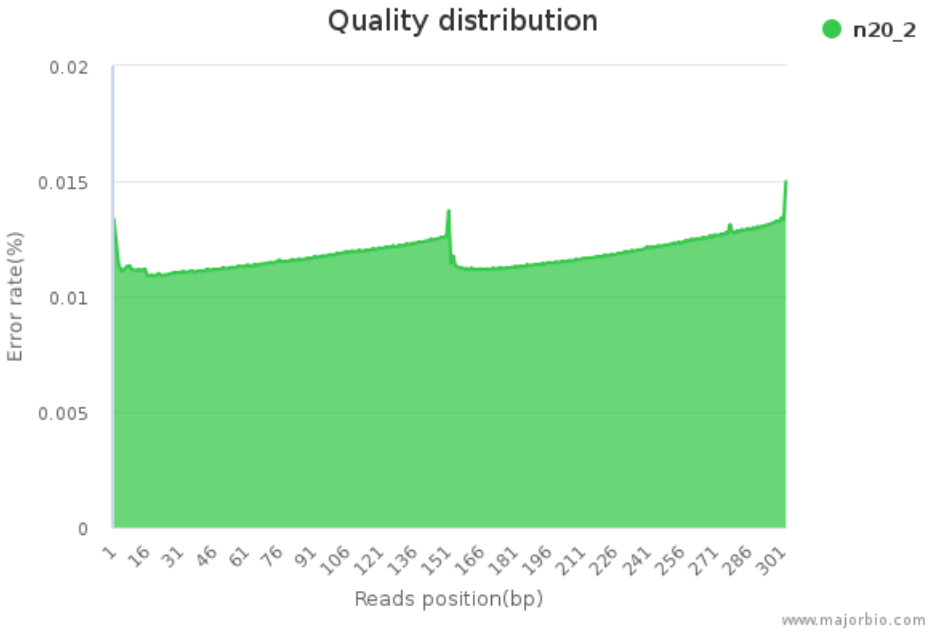

图3-3 样品的碱基错误率分布图

注：横坐标是Reads碱基坐标位置，表示Reads上从5'到3'端依次碱基的排列；纵坐标是所有Reads在该位点处碱基的平均错误率（%）。虚线左侧为双端测序的Read1的错误率分布情况，虚线右侧为Read2的错误率分布情况。

四、附录

4.1 结果目录文件说明

| |--rawdata/ 【原始下机数据】
| | |--L1EGC010491--L_0_1*R1.raw.fastq.gz 【read1的原始序列，*代表样本名称】
| | |--L1EGC010491--L_0_1*R2.raw.fastq.gz 【read2的原始序列，*代表样本名称】
| | |--rawdata.stat.xls 【各样本原始数据统计表】
| | |--md5.txt 【各样本测序数据MD5值】
| | |--report.pdf 【项目结题报告】
| | |--qc_pdf.tar.gz/ 【各样本原始测序数据统计图】
| | |--L_0_1*raw_qc_base.line.pdf 【碱基含量分布图，*代表样本名称】
| | |--L_0_1*raw_qc_error.line.pdf 【碱基错误率分布图，*代表样本名称】

4.2 文件打开或浏览方法

所有提供的文件均为Linux系统下的文件，压缩包使用“tar -zcvf ”命令压缩，以下为不同系统用户解压缩的方法：

Unix/Linux/Mac用户: 使用tar -zxvf *.tar.gz命令

Windows用户：使用WinRAR软件解压缩

如果在本附录中无特殊说明，所有提供的文件均为Linux系统下文本文件，Unix/Linux用户可以使用more或less命令查看文本文件内容。对于Windows用户，一般文本文件可以使用写字板或excel打开。推荐使用开源文本编辑器gedit for win32版本(http://projects.gnome.org/gedit/)或者商业文本编辑器UltraEdit。当文件比较大时，打开文件可能导致Windows系统死机，建议使用性能较好的计算机或者使用更适合处理大量数据的Unix/Linux系统打开。

数据中可能包含部分图像文件，一般图像文件后缀名为.png、.jpg、.gif等，对于图像文件，Windows用户可以使用图片浏览器打开，Linux/Unix用户使用display命令打开。

后缀名为svg的文件为文本格式描述的图像文件，Windows用户需要安装Adobe Illustrator软件打开。Linux/Unix用户可以使用rsvg-view命令查看。 公司默认提供为“pdf“格式的矢量图，可利用"Adobe Illustrator"软件对该格式图片进行编辑。

Linux下的表格均为制表符(Tab)分割的文本，为了便于阅读，建议使用excel或openoffice等办公软件用表格形式打开，打开时请用“Tab分割”方式。

4.3 联系方式

	上海总部
地址:	上海市浦东新区康新公路3399号时代医创园3号楼
电话:	400-660-1216
E-mail:	rna@majorbio.com
	广州分公司
地址:	广州市海珠区荔福路68号广州市微生物所六楼西
电话:	020-61130189
E-mail:	seqgz@majorbio.com
	北京子公司
地址:	北京市海淀区安宁庄东路18号光华创新园新科研楼6楼

电话：	010-51293026,010-51293126
E-mail：	seqbj@majorbio.com