转录组测序分析

作者QQ：894064647

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，主要包括 mRNA和非编码RNA 。

名词解释

Fastq：Fastq是测序技术中一种反映测序序列的碱基质量的文件格式。第一行以“@”符号开头，后面紧跟一个序列的描述信息；第二行是该序列的内容；第三行以“+”符号开头，后面可以是该序列的描述信息，也可省略；而第四行是第二行中的序列内容每个碱基所对应的测序质量值。

```
@ERR329500.1 HWUSI-EAS697:8:87:3296:18042#CGATGT/1
CACAAATTAAGCAGCCATAGATGGGTCATTTTTACTGTAAAGGCTGATCAAGGAAGATACCCTG
+
IIIIIIIIIIIIIIIIHIIGIIIIIEGDGGIIIFIIIIHFIIIIIIIIIIIIHBIIHHIHHIHI
```

名词解释

RPKM：Reads Per Kilobase per Million mapped reads，是指每 1 百万个map 上 的reads 中 map 到外显子的每1K 个碱基上的reads 个数。计算公式四$RPKM=10^6C/NL/10^3$，其中C为唯一比对到目的基因的reads数；N为唯一比对到参考基因的总reads数，L是目的基因编码区的碱基数。RPKM法可以消除基因长度、数据量之间的差异进行计算基因表达量。

FPKM计算的是片段（fragments），而RPKM计算的是数据（reads）。Fragment比read的含义更广，因此FPKM包含的意义也更广，可以是pair-end的一个fragment，也可以是一个read。

数据库准备(hg19)

UCSC(fasta、bed)

http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz

bowtie2-build hg19.fa

下载fastq数据

http://www.ebi.ac.uk/

"lung cancer rna-seq" --> Nucleotide sequences (74)

http://www.ncbi.nlm.nih.gov/

SRA --> "lung cancer rna-seq"

http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.4.5-2/sratoolkit.2.4.5-2-ubuntu64.tar.gz

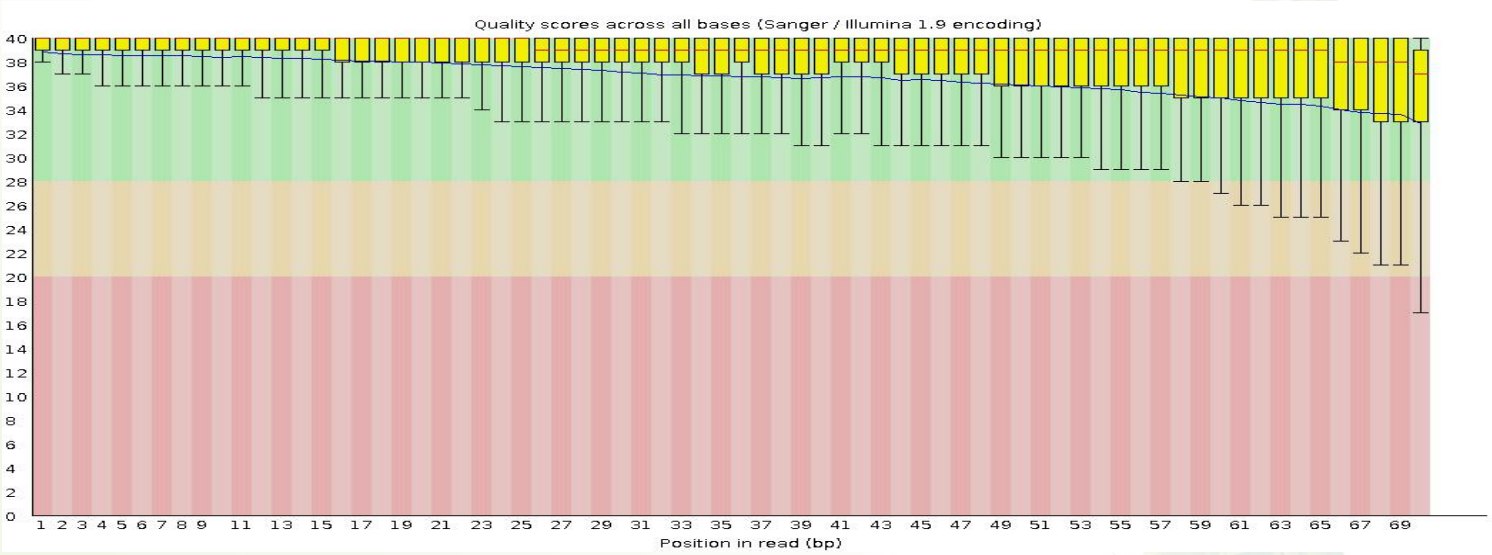
质控

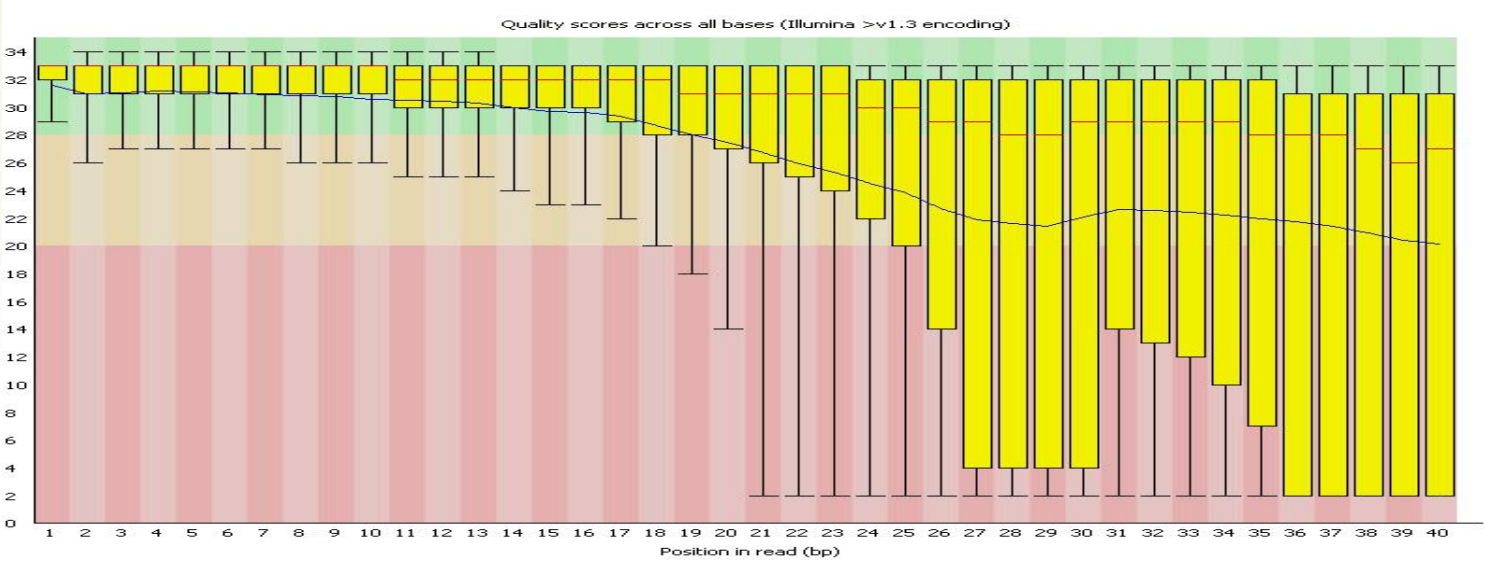
fastqc ERR499.read1.fq ERR499.read2.fq -t 2 -o qc_report_dir

质量值图
GC含量图
ATGC比例图

质控

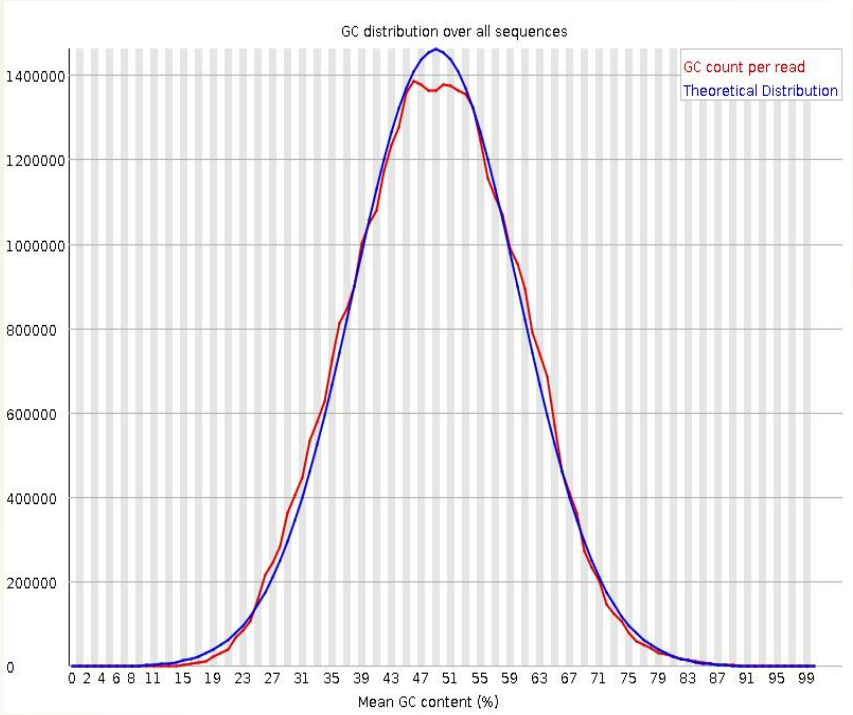
- 质量值

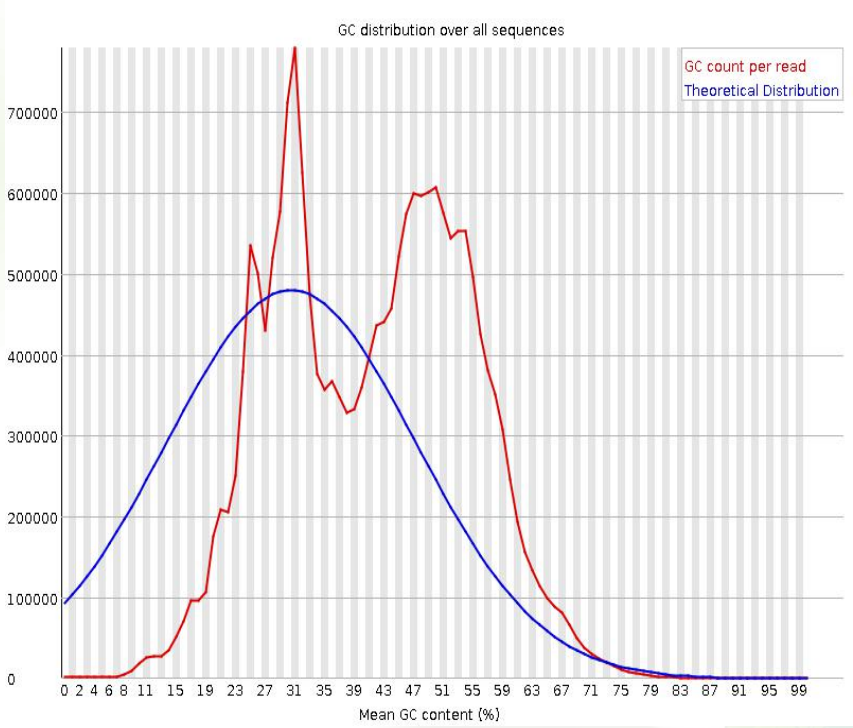

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


Quality scores across all bases (Illumina >v1.3 encoding)
Position in read (bp)

质控

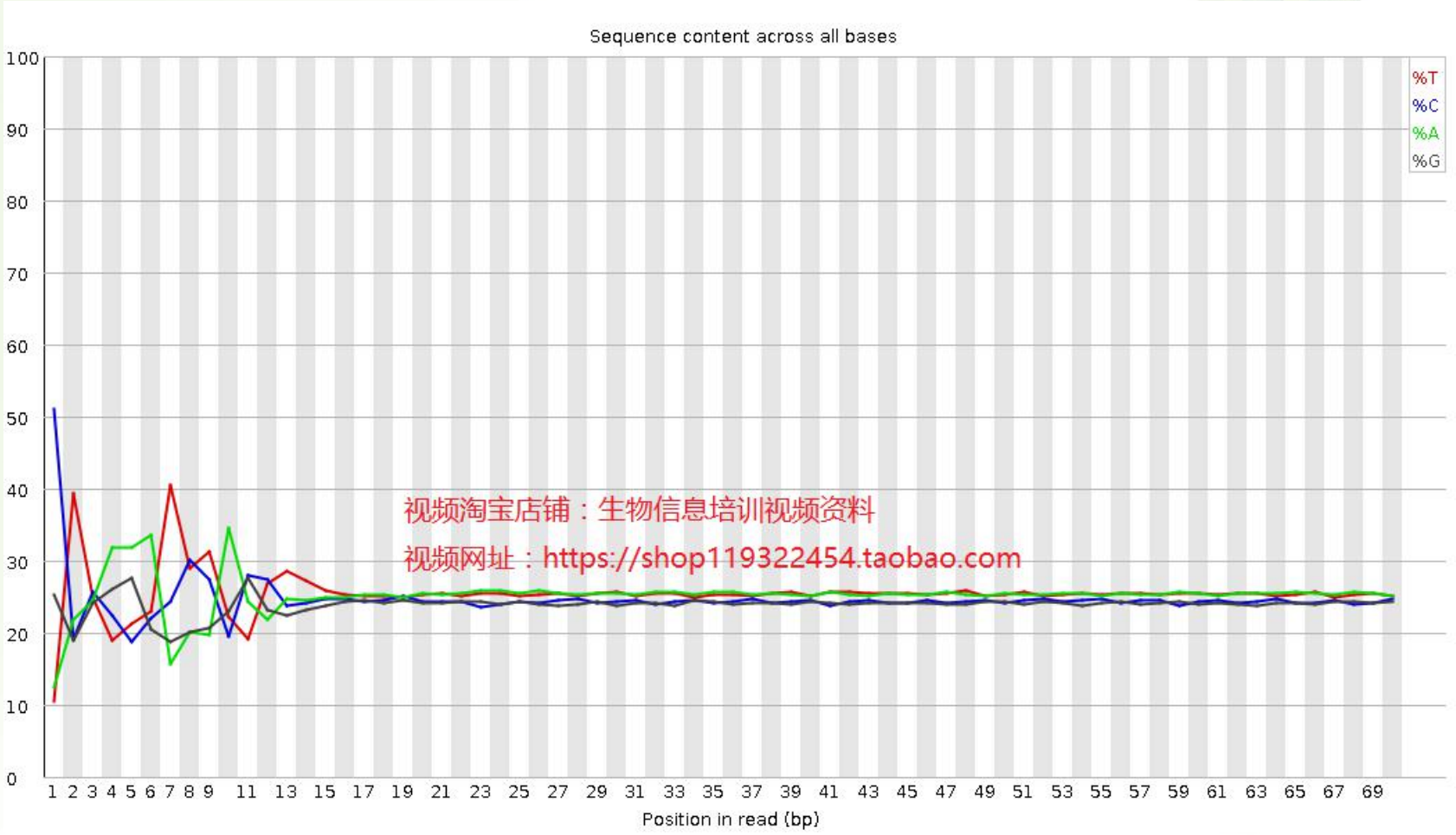
- GC含量


GC distribution over all sequences
GC count per read
Theoretical Distribution
1400000
1200000
1000000
800000
600000
400000
200000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)


GC distribution over all sequences
GC count per read
Theoretical Distribution
700000
600000
500000
400000
300000
200000
100000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

质控

- ATGC比例


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69
Position in read (bp)
视频淘宝店铺：生物信息培训视频资料
视频网址：https://shop119322454.taobao.com

过滤

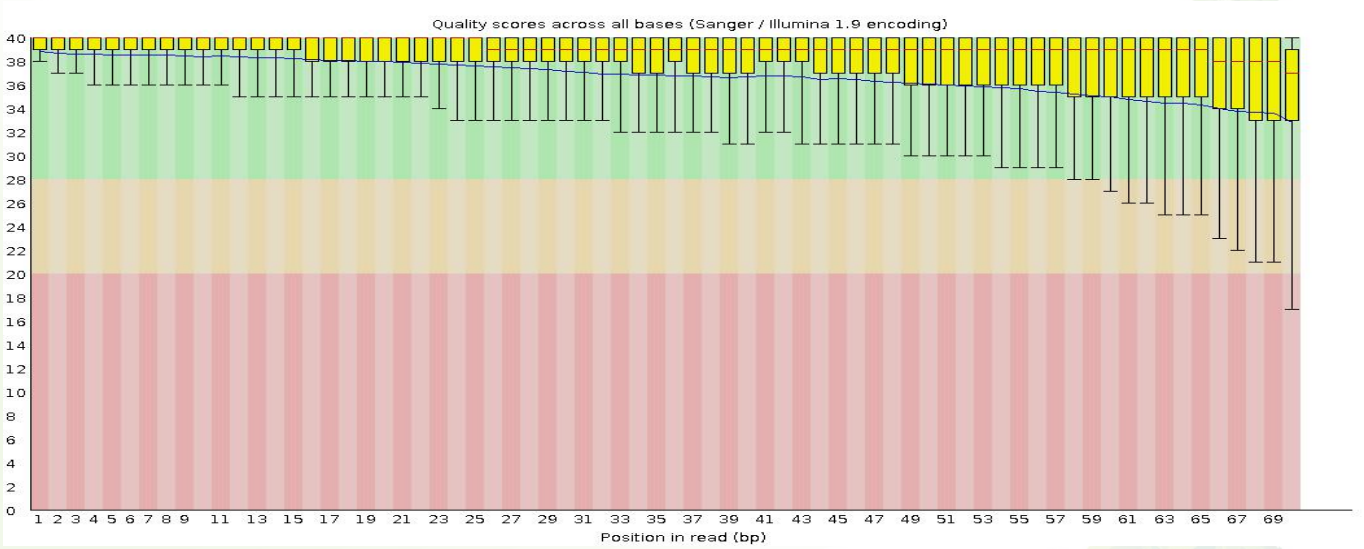
1. 过滤接头。对含接头的reads去除接头序列。
2. 一条reads上N（未能确定出具体的碱基类型）的比例大于5%，则过滤掉该reads。
3. 过滤低质量reads，过滤掉Q20<80% reads。

过滤统计

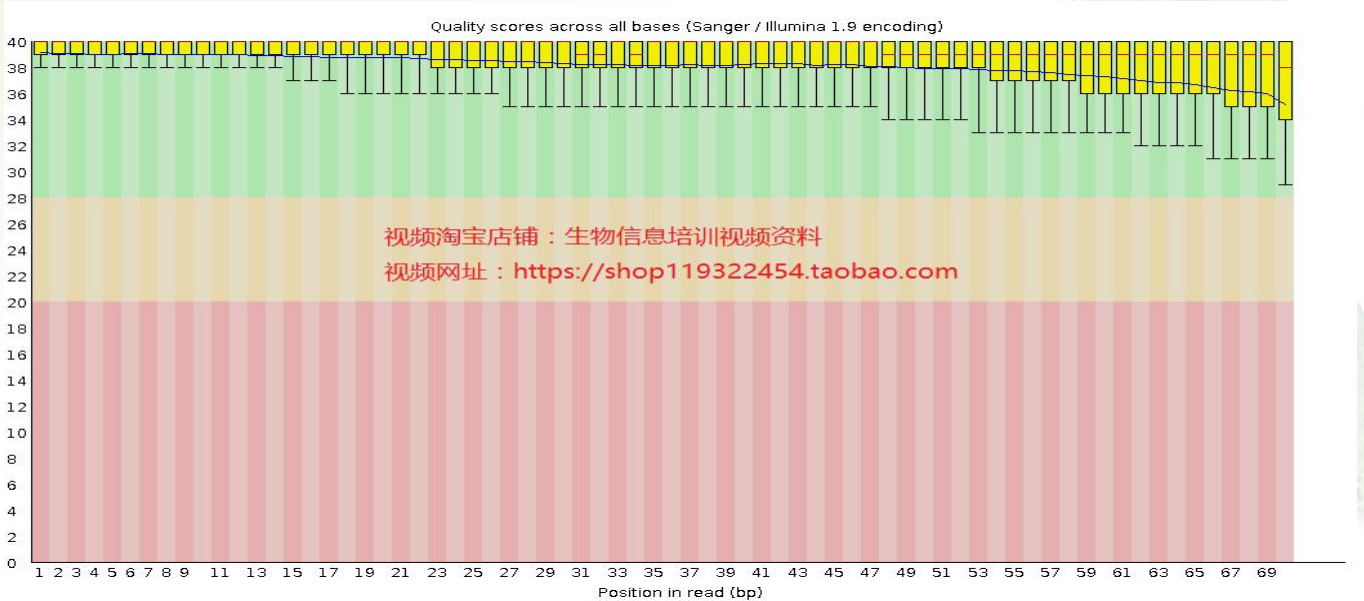
read	raw	adapter	N	Low qual	clean
read1	28701483 (100%)	53761(0.19%)	29(0.00%)	1849520 (6.44%)	25779623 (89.82%)
read2	28701483 (100%)	22886(0.08%)	58357(0.20%)	1798947 (6.27%)	25779623 (89.82%)

过滤统计

过滤前：



过滤后：



比对

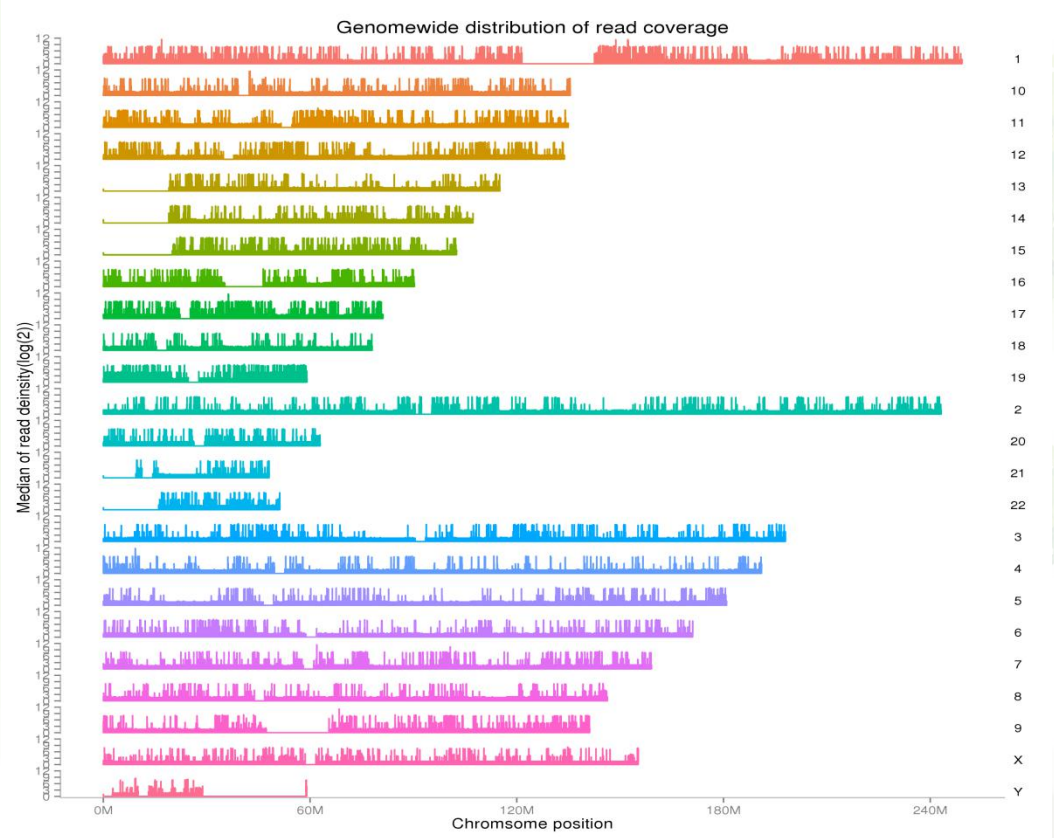
1. 建索引　bowtie2-build hg19.fa
2. 比对

tophat2 -o tophat_ERR500 -p 8 --read-mismatches 2 --read-edit-dist 2 -r 50 --library-type fr-unstranded /data/hg19/hg19 ERR500.clean 1.fq ERR500.clean2.fq

比对统计

statistics	Input Reads	mapped	Multiple	Unique	R1 mapped	R2 mapped
Percentage	48948102(100%)	48212358(98.50%)	561991(1.15%)	47650367(97.35%)	24118187(49.27%)	24094171(49.22%)

覆盖度


Genomewide distribution of read coverage
Median of read deinsity(log(2))
1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
3
4
5
6
7
8
9
X
Y
0M
60M
120M
180M
240M
Chromsome position

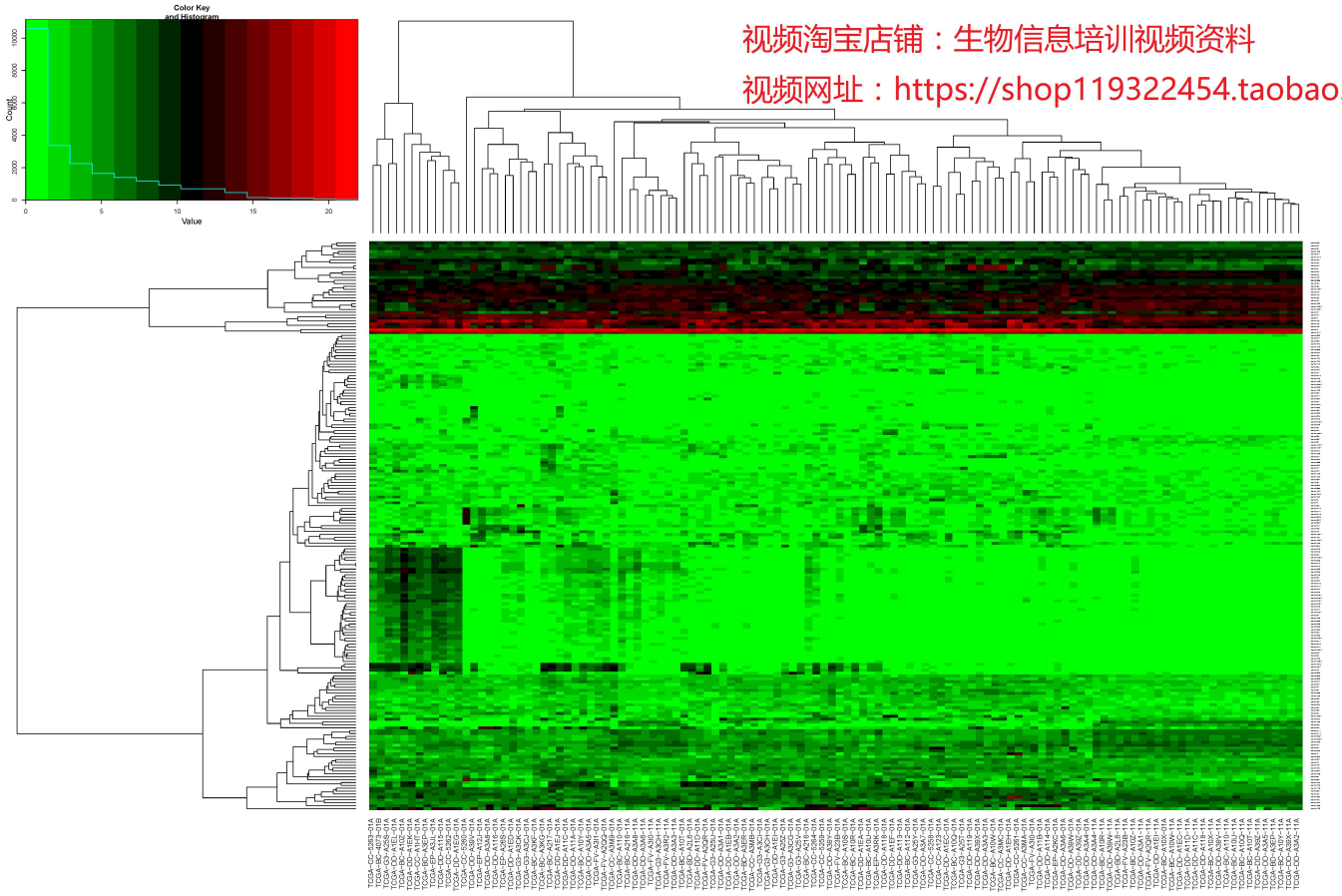
表达量计算

- ✓ cufflinks -o outDir -p 10 --library-type fr-unstranded -G gtfFile unique.bam
- ✓ RSEM
- ✓ HTSeq

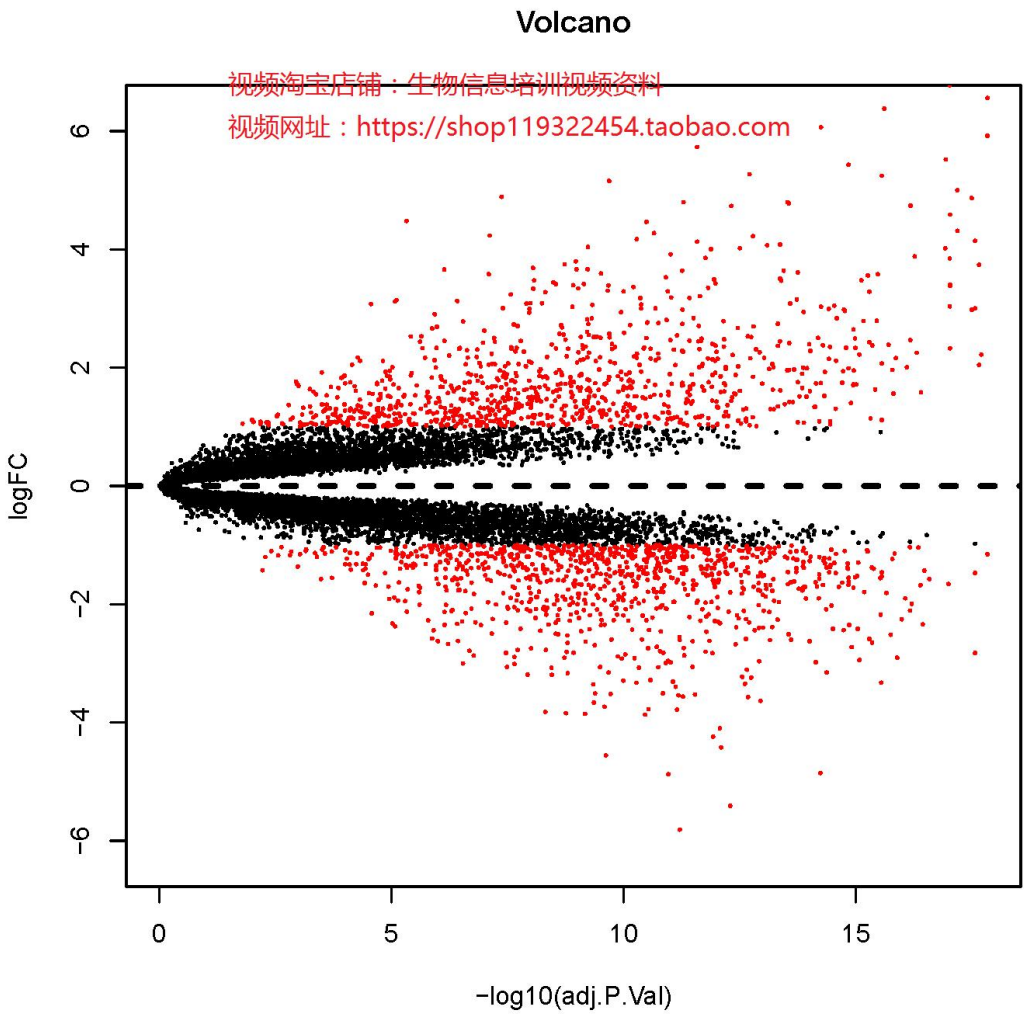
差异表达

Name	normalAve	tumorAve	logFC	pValue	qValue
ADAMTS13|11093	1913.115	209.3295	-3.19175	1.68E-62	3.45E-58
CSRNP1|64651	6868.116	1335.735	-2.36223	1.01E-46	1.04E-42
GABRD|2563	6.405442	170.6827	4.703357	7.26E-45	4.96E-41
DBH|1621	2031.552	106.1415	-4.25681	8.30E-44	3.70E-40
CDCA5|113130	39.37348	561.616	3.830401	9.02E-44	3.70E-40
PLVAP|83483	640.8951	4969.907	2.954596	1.95E-43	6.68E-40
COL15A1|1306	47.58915	1023.457	4.419871	2.66E-43	7.78E-40
CDKN3|1033	13.03448	335.4387	4.670194	1.59E-42	4.06E-39
OIT3|170392	4593.752	360.2504	-3.67222	2.45E-42	5.59E-39
ECM1|1893	3110.492	384.9053	-3.01411	1.82E-41	3.74E-38

差异基因热图

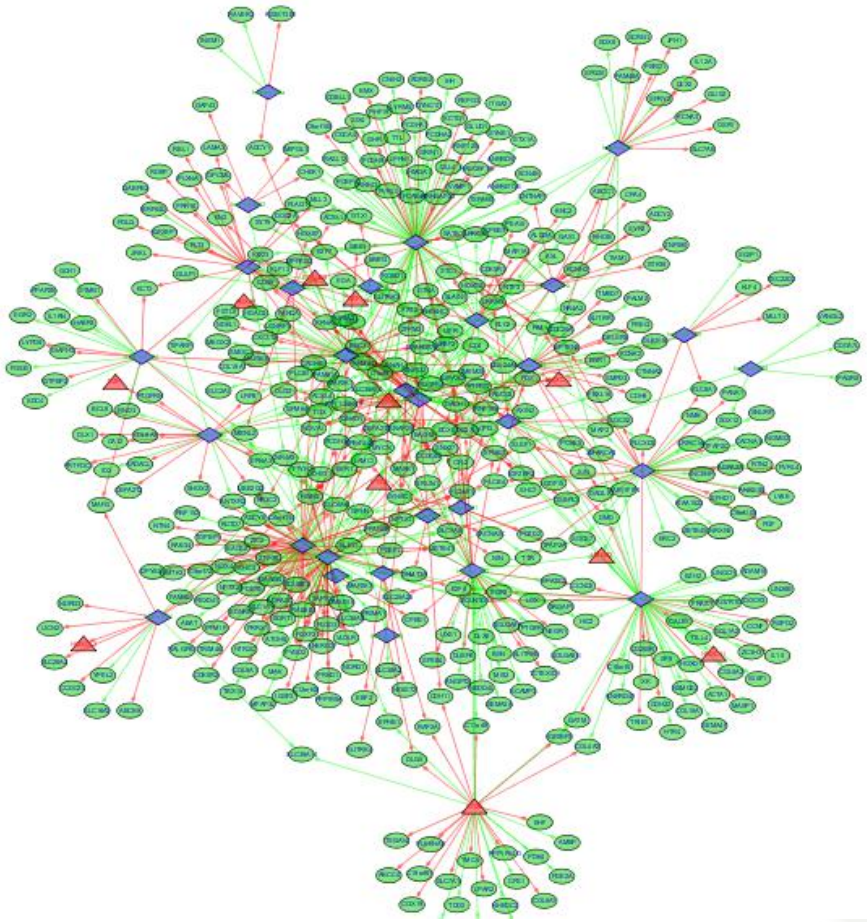


火山图

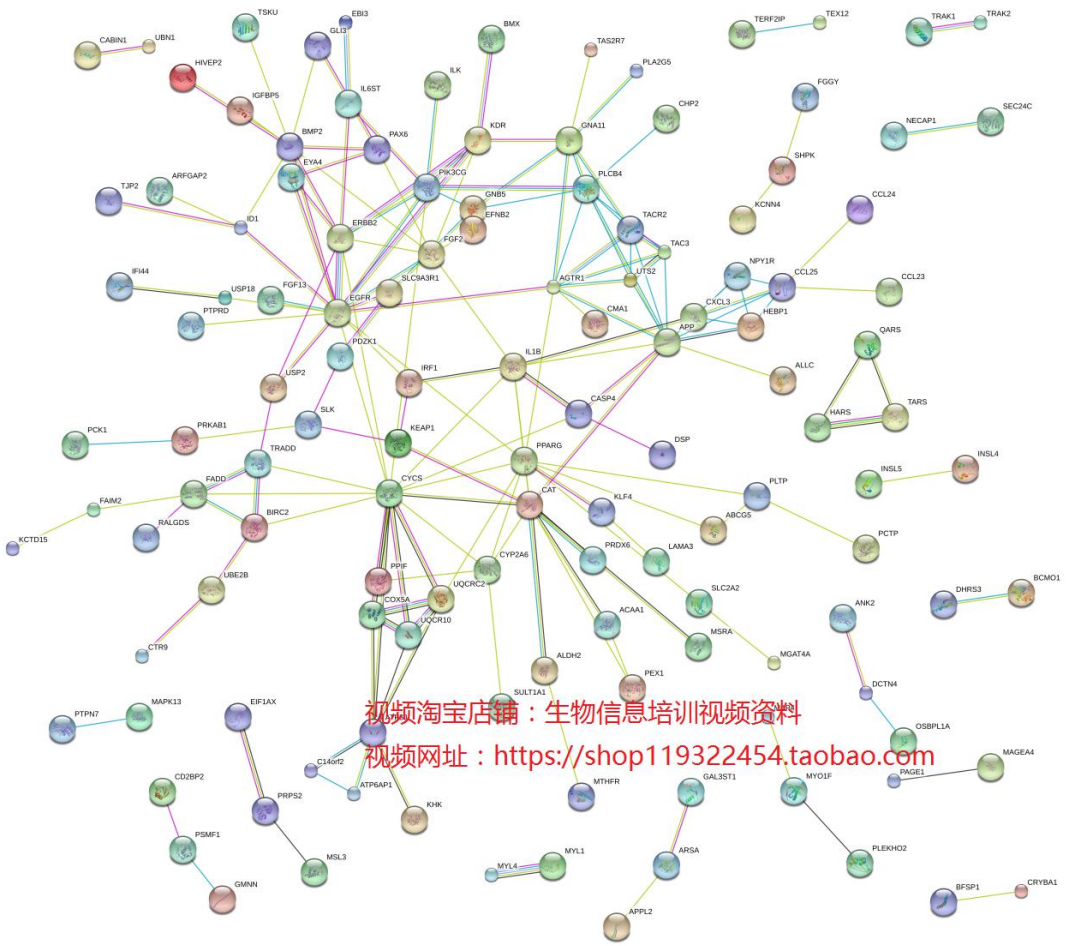


横坐标为-log10(adj.P.Val)，纵坐标为logFC，红色得点代表的差异基因，黑色代表非差异表达基因。

共表达网络

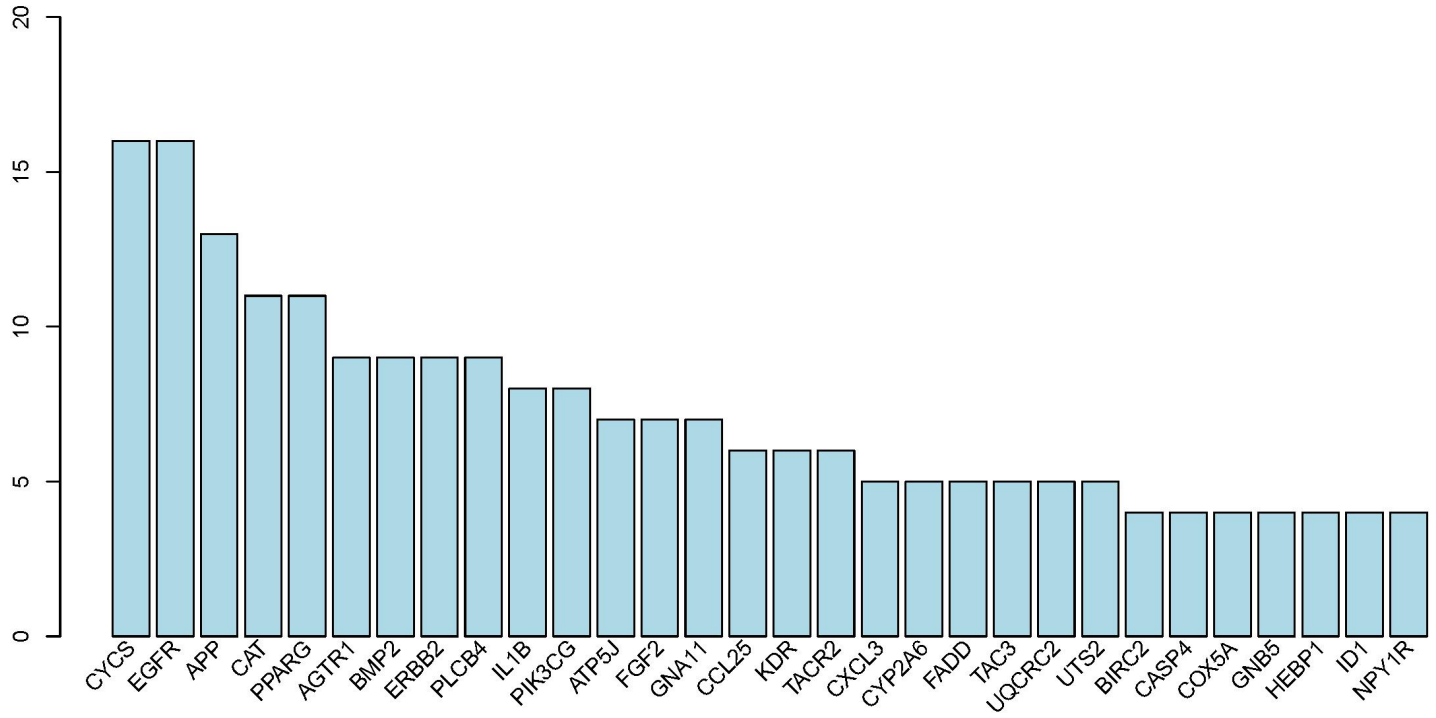


蛋白互作网络



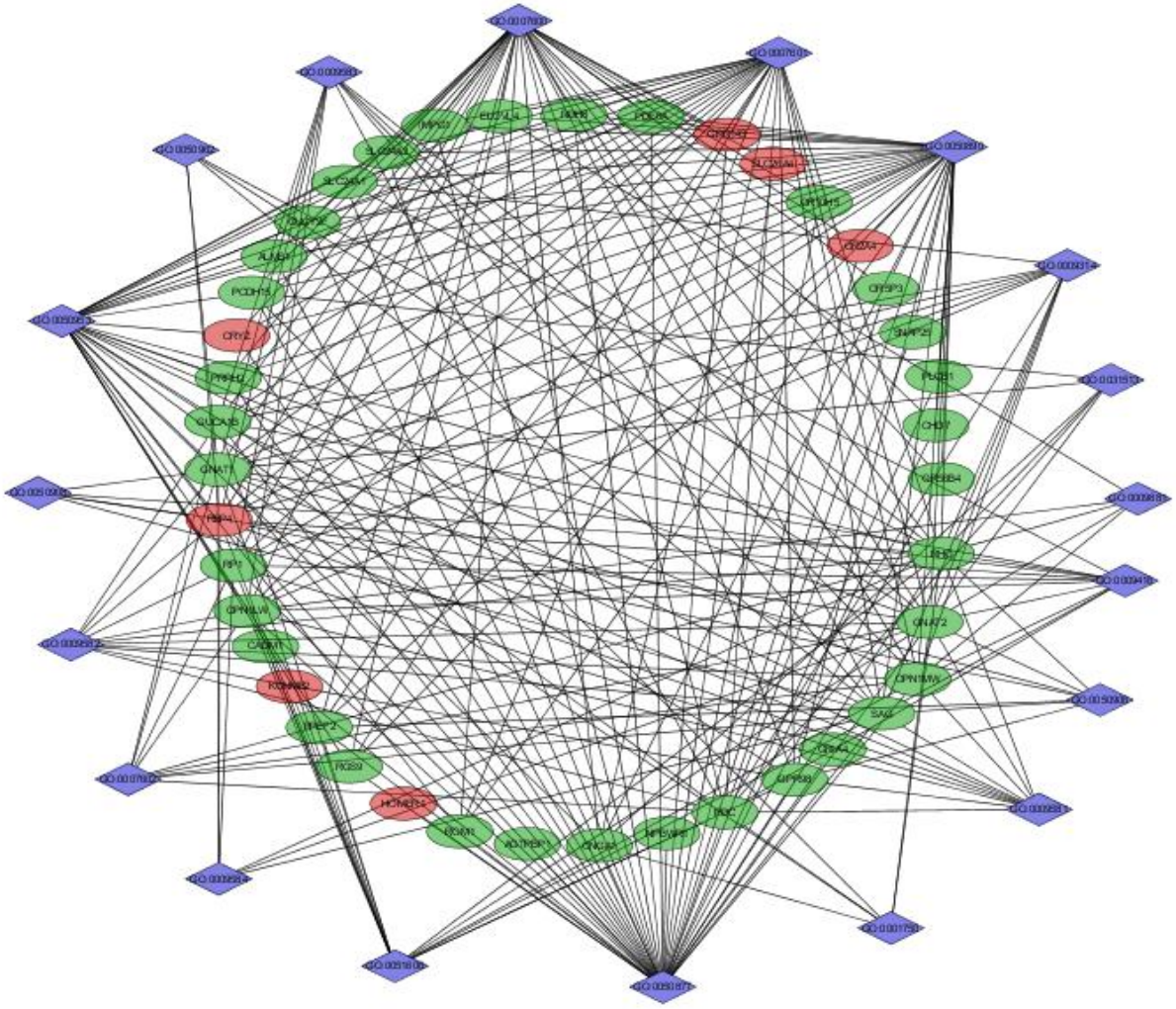
圆圈代表基因，线条代表基因间存在相互作用，圆圈内部的结果代表蛋白的结构。线头颜色代表证明蛋白之间存在相互作用的不同证据。(A red line indicates the presence of fusion evidence; a green line - neighborhood evidence; a blue line - coocurrence evidence; a purple line - experimental evidence; a yellow line – text mining evidence; a light blue line - database evidence; a black line - coexpression evidence.)

互作网络柱状图



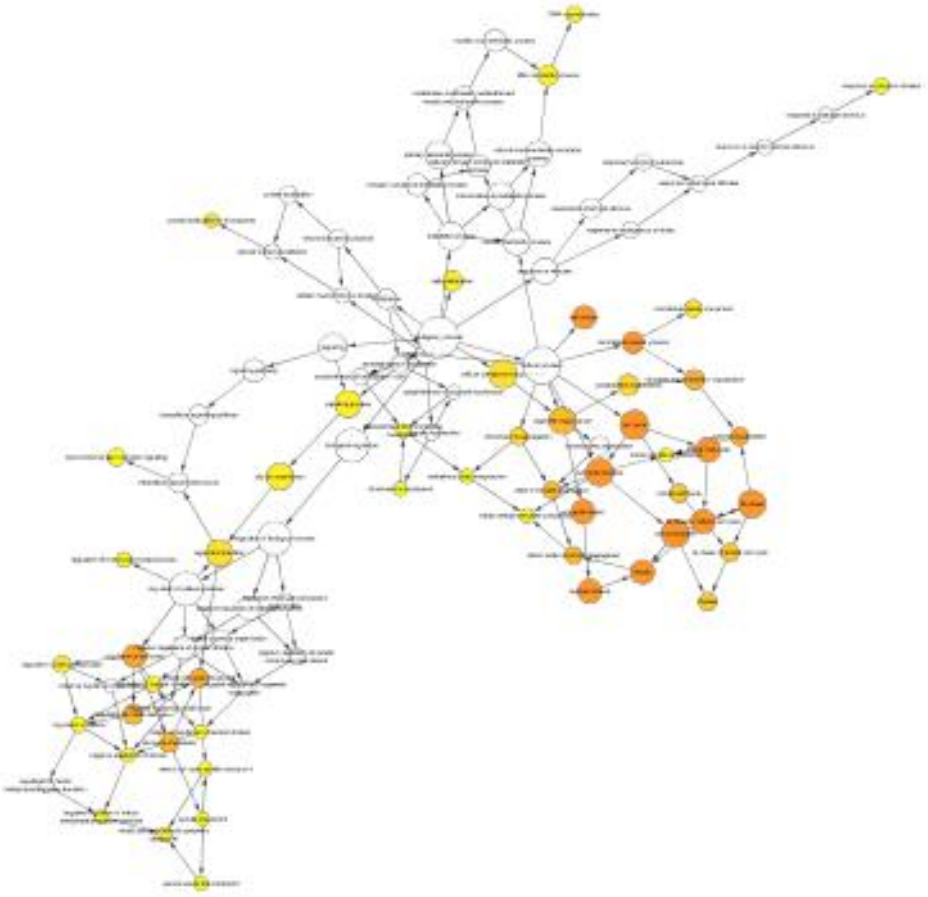
互作网络邻接基因数目柱状图

GO分析

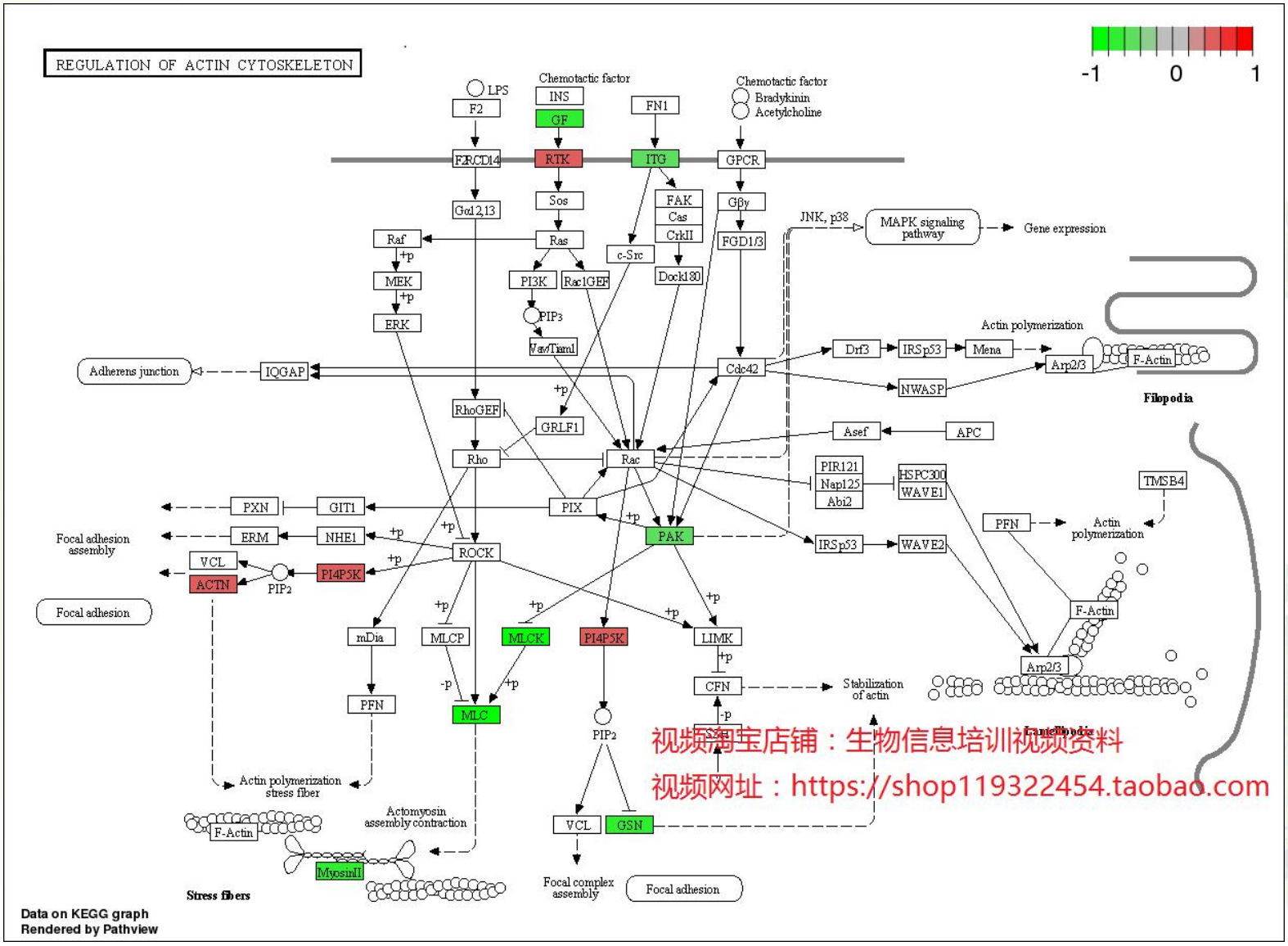


GO有向无环图

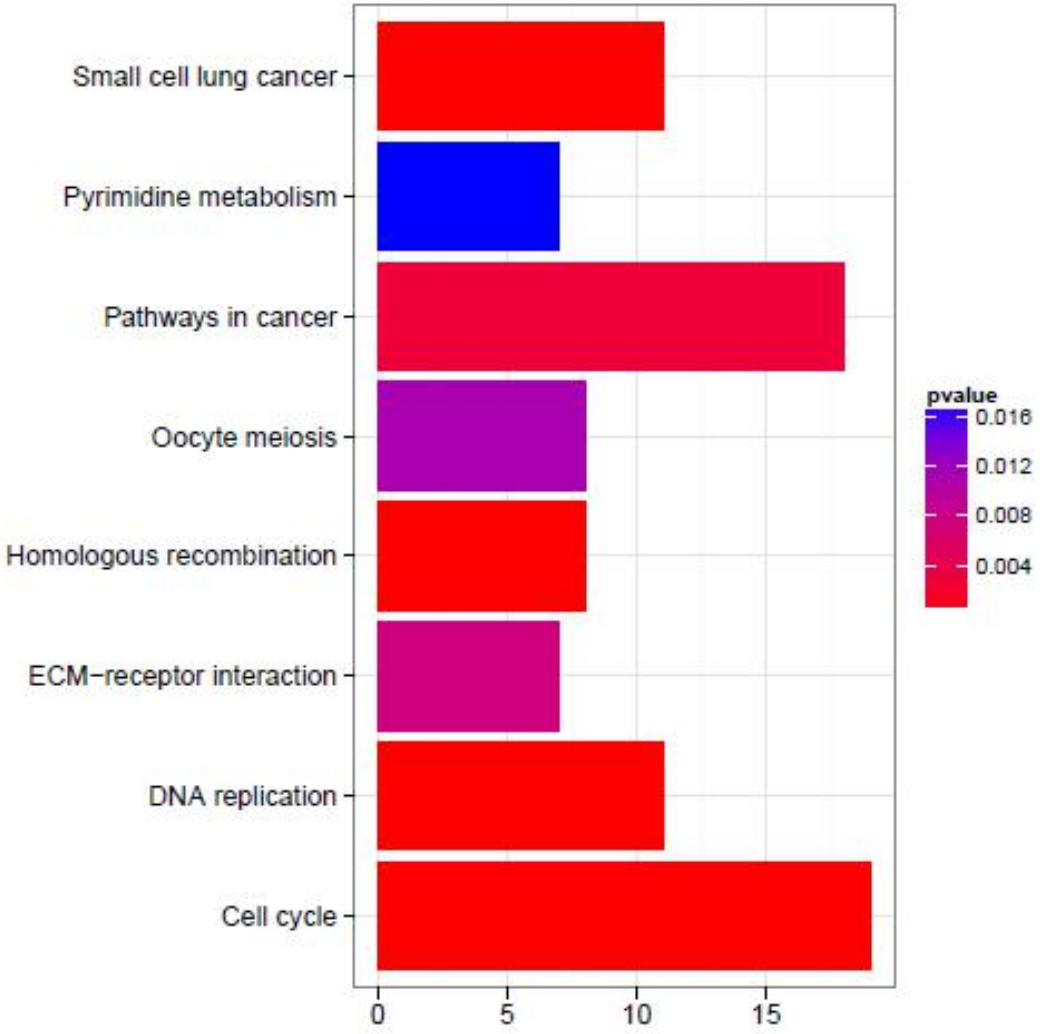
该有向无环图为差异基因GO富集分析的结果图形化展示方式，分支代表包含关系，箭头方向从上之下所定义的功能范围越来越小，并通过包含关系，将相关的GO Term一起展示，颜色深浅代表富集程度，越深富集水平越高，反之，则越低。



KEGG分析



KEGG柱状图


Small cell lung cancer
Pyrimidine metabolism
Pathways in cancer
Oocyte meiosis
Homologous recombination
ECM−receptor interaction
DNA replication
Cell cycle
0
5
10
15
pvalue
0.016
0.012
0.008
0.004

新转录本

Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)

新lncRNA

长链非编码RNA(Long non-coding RNA, lncRNA)是长度大于 200 个核苷酸的非编码 RNA。研究表明, lncRNA 在剂量补偿效应、表观遗传调控、细胞周期调控和细胞分化调控等众多生命活动中发挥重要作用，成为遗传学研究热点。

characteristic	threshold
length	200-100000nt
orf length	0-100aa
protein functional domain	no
Coding Potential	no

tp://cpc.cbi.pku.edu.cn/view/cpc_view.jsp?ID=65BBE6C0-E4EE-11E4-9691-CF7AFB039FF

其他分析

- 基因融合
- RNA编辑
- 新lncRNA预测与功能分析

谢谢大家！

作者QQ： 894064647