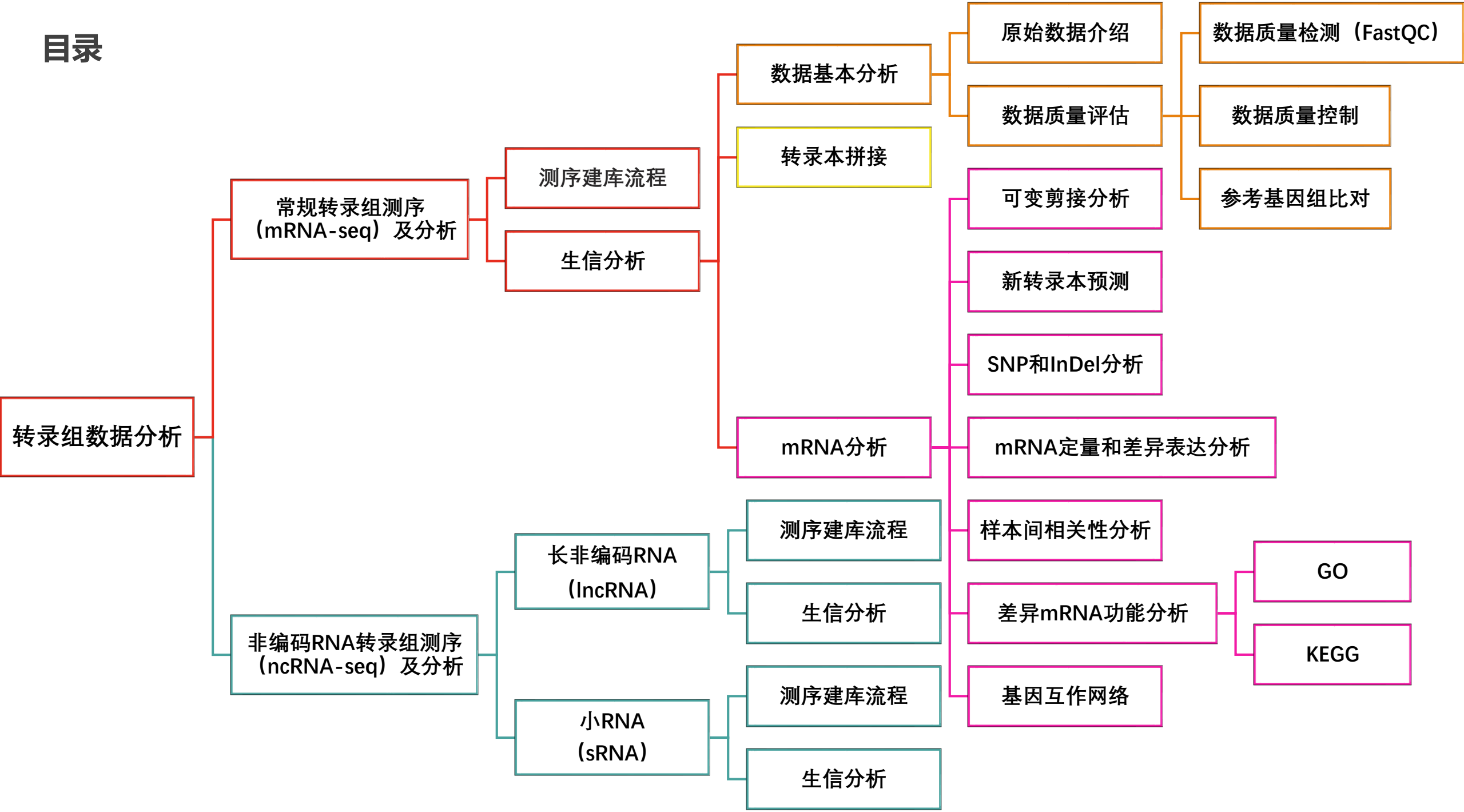
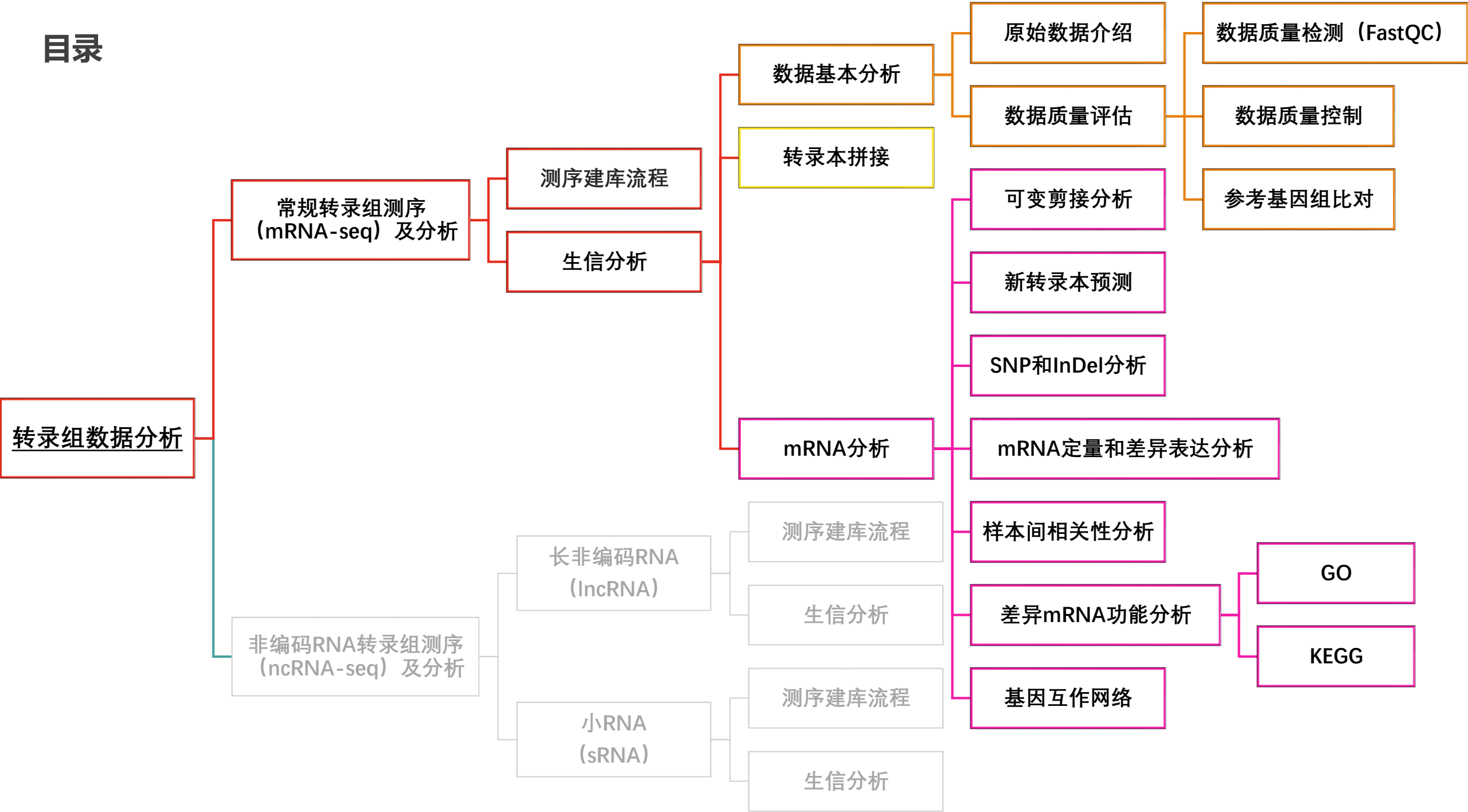
转录组数据分析

（mRNA）

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

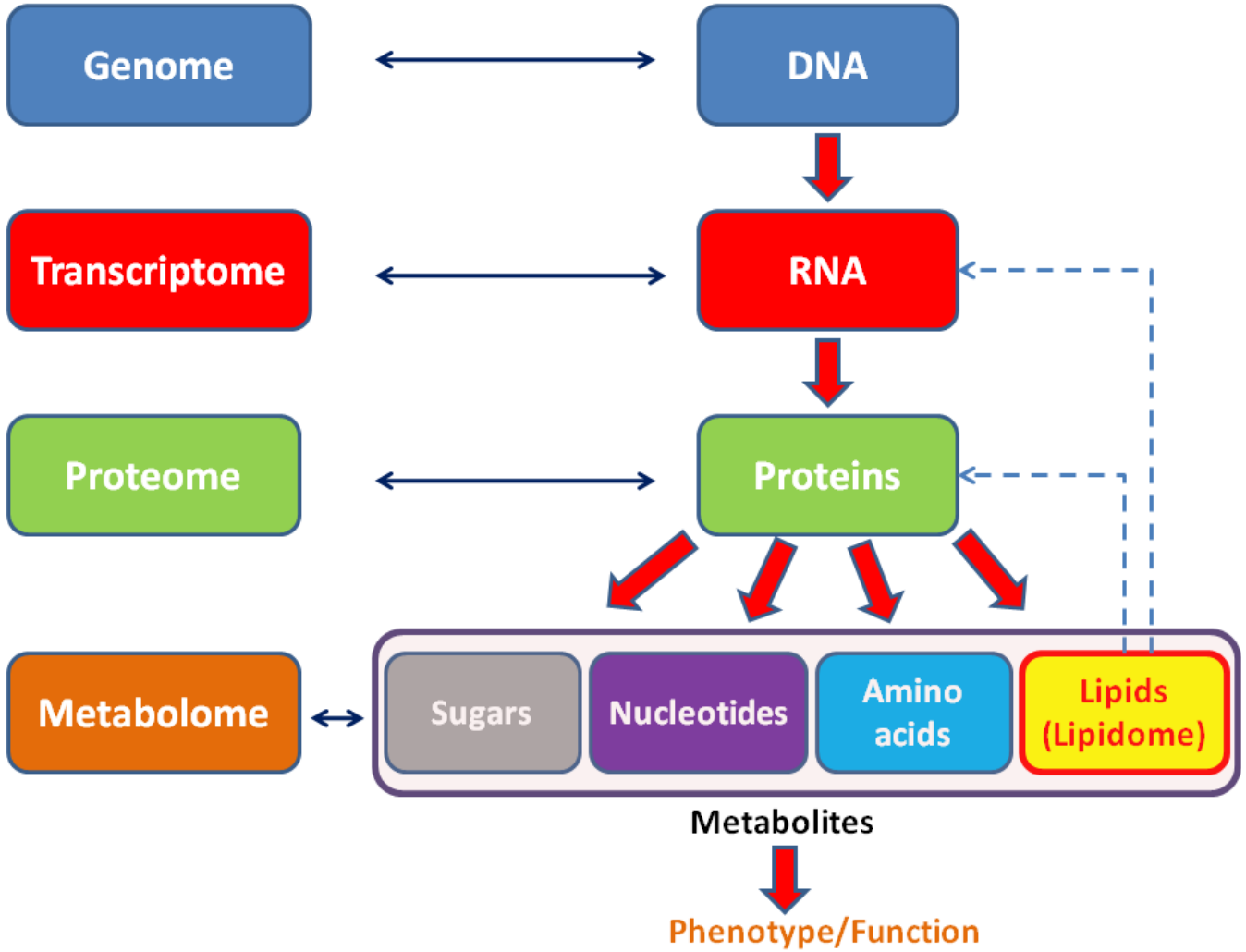
目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

什么是转录组

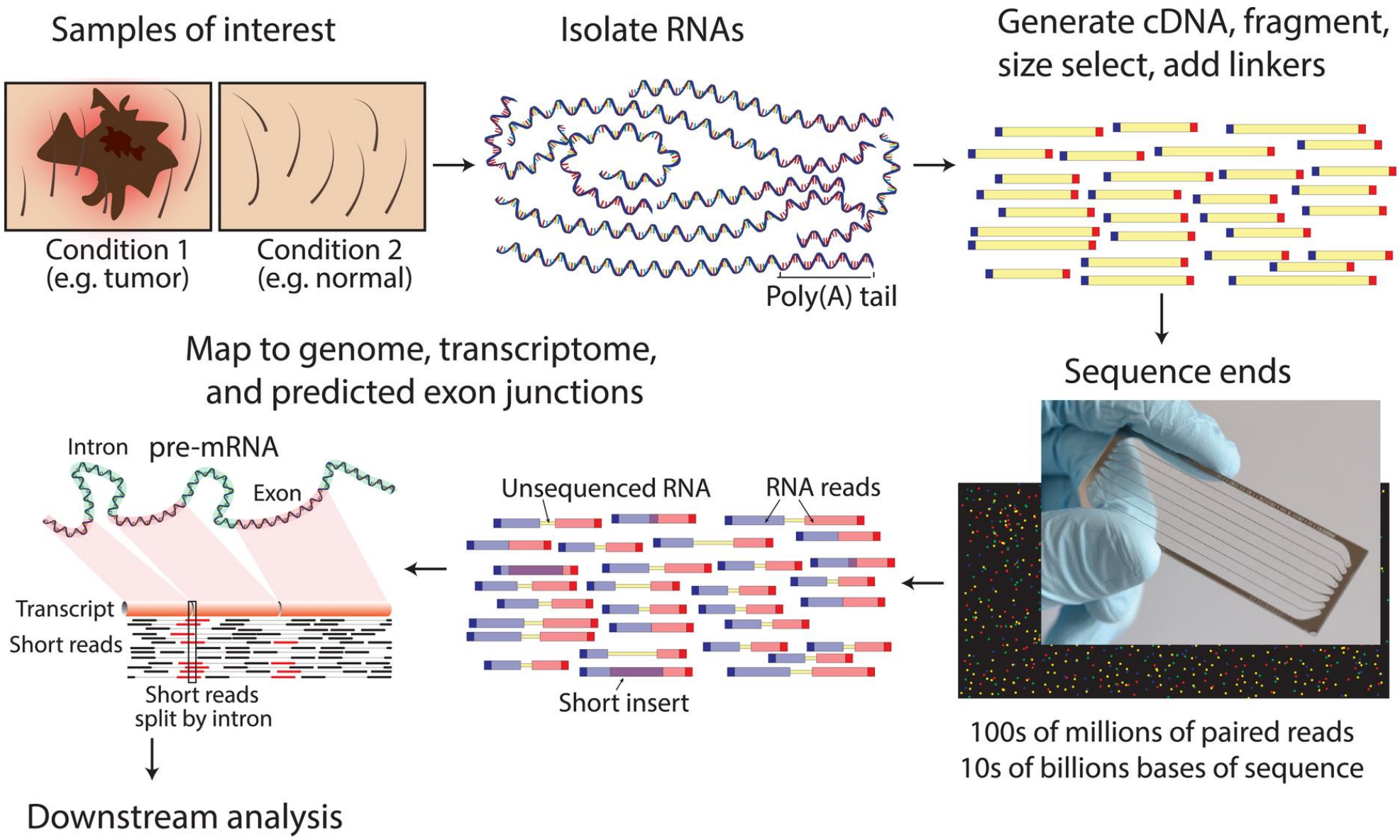
- 广义上指在相同环境（或生理条件）下的在一个细胞、或一群细胞中所能转录出的所有RNA的总和，包括信使RNA（mRNA）、核糖体RNA（rRNA）、转运RNA（tRNA）及非编码RNA

- 狭义上则指细胞所能转录出的所有信使RNA（mRNA）



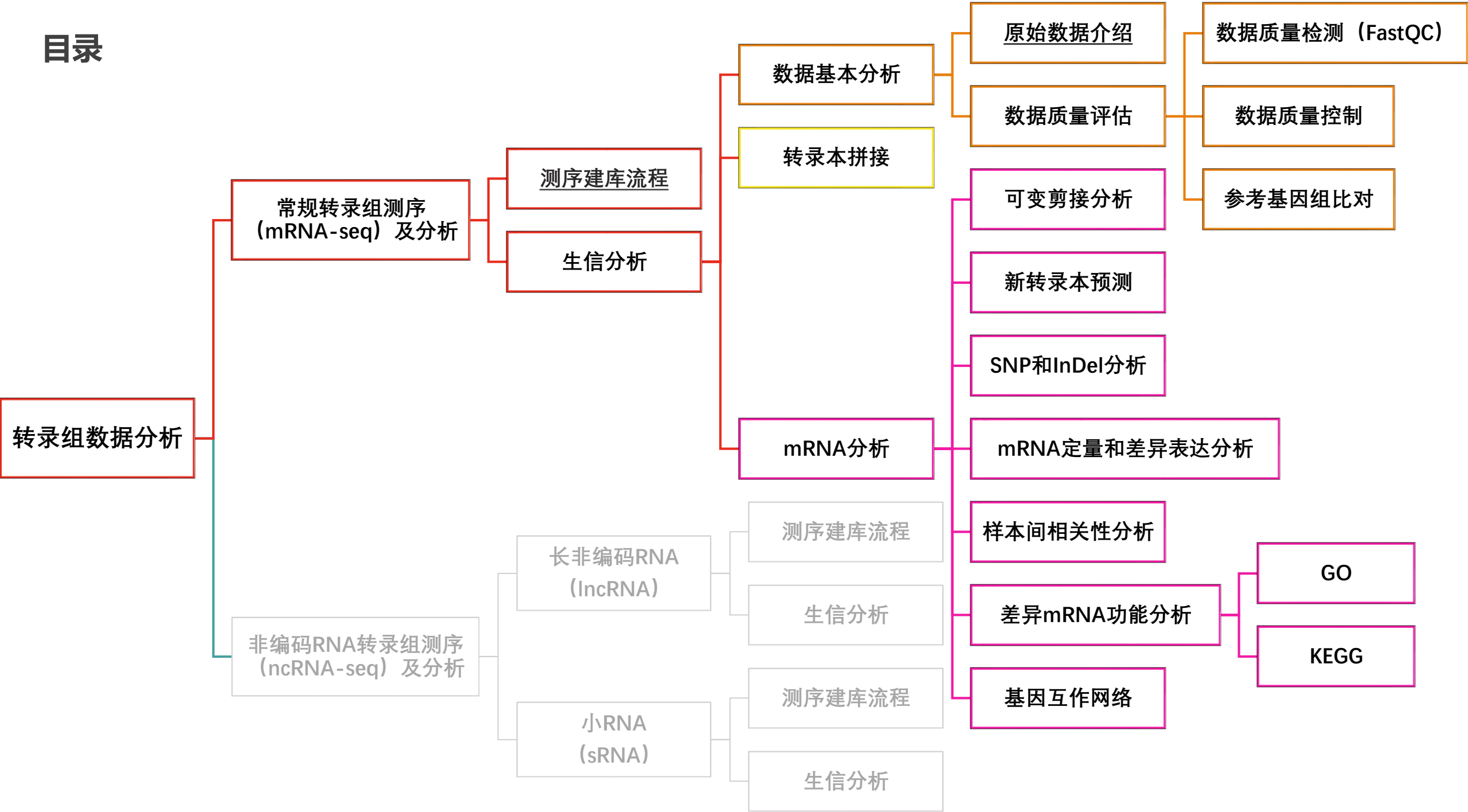
目录

转录组数据的产生（RNA-Seq）



Griffith *et al.*, 2015

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

转录组数据的存储格式——FASTQ

```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133 1:N:0:CTTGTA
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

第1行主要储存序列测序时的坐标等信息；

第2行就是测序得到的序列信息，一般用ATCGN来表示，其中N用于荧光信号干扰无法判断到底是哪个碱基时的代表符号；

第3行以“+”开始，可以储存一些附加信息，但目前的测序fastq文件这一行一般是空的。

第4行储存的是质量信息，与第2行的碱基序列是一一对应的，其中的每一个符号对应的ASCII值是经过换算的phred值，可以简单理解为对应位置碱基的测序质量值，越大说明测序的质量越好。不同的版本对应的phred值范围不同。

测序标识符

```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133 1:N:0:CTTGTA
```

实例	说明
ST-E00126	the unique instrument name
128	run ID
HJFLHCCXX	flowcell ID
2	flowcell lane
1101	tile number within the flowcell lane
7405	'x'-coordinate of the cluster within the tile
1133	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2
N	Y if the read is filtered, N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
CTTGTA	index sequence

测序质量值编码——phred值

Phred值是评估这个bp测序质量的值，测序仪通过判断荧光信号的颜色来判断碱基的种类，ATCG分别对应红黄蓝绿，信号强弱不同，在这种情况下对每个结果的判断的正确性都存在一个概率值，这个值被储存为ASCII码形式，转化方式如下：

- **将该碱基判断错误概率值P取log10之后再乘以-10，得到的结果为Q。**

若P=1%，那么对应的
Q=-10*log10（0.01）=20
（这个计算公式illumina平台使用，Solexa系列测序仪使用不同的公示来计算质量值：Q=-10log(P/1-P)）

- **把这个Q加上33或者64转成一个新的数值，称为Phred，最后把Phred对应的ASCII字符对应到这个碱基。**

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

PS：序列存储格式FASTA

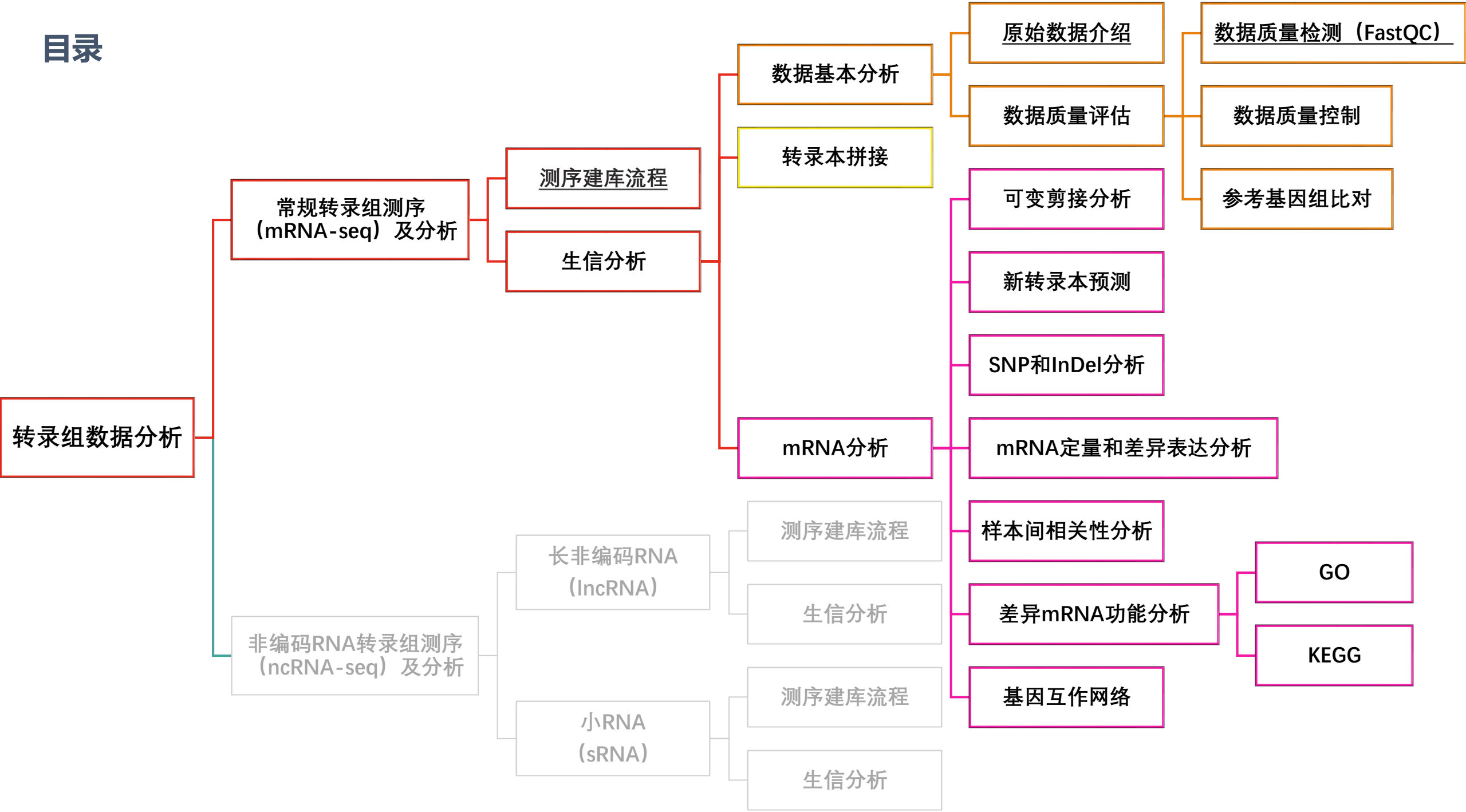
蛋白质fasta文件

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDD
MPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKLASVSTVLTSKYR
```

核酸序列文件（mRNA序列中的U均用T来代替）

```
>gi|13650073|gb|AF349571.1| Homo sapiens hemoglobin alpha-1 globin chain
(HBA1) mRNA, complete cds
CCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGACGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG
GTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCT
ACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGAC
CAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGG
GTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCA
CCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGC
TGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTTG
```

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

转录组数据的质量检测——FastQC

运行命令

```
fastqc -t 8 -o path/fastqc
sample1_R1.fq sample1_R2.fq
```

参数

-o --outdir:输出路径

--extract：结果文件解压缩

--noextract：结果文件压缩

-f --format:输入文件格式

-t --threads:线程数

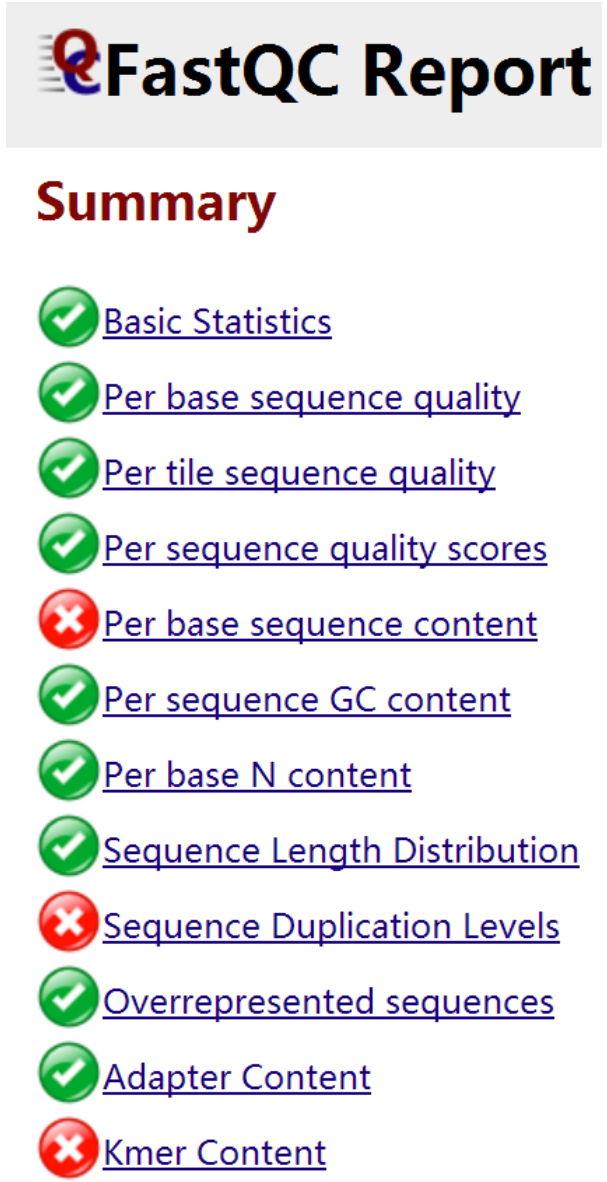
-c --contaminants：制定污染序列

-a --adapters：指定接头序列

-k --kmers：指定kmers长度（2-10bp,默认7bp）

-q --quiet： 安静模式

运行结果（html和zip）

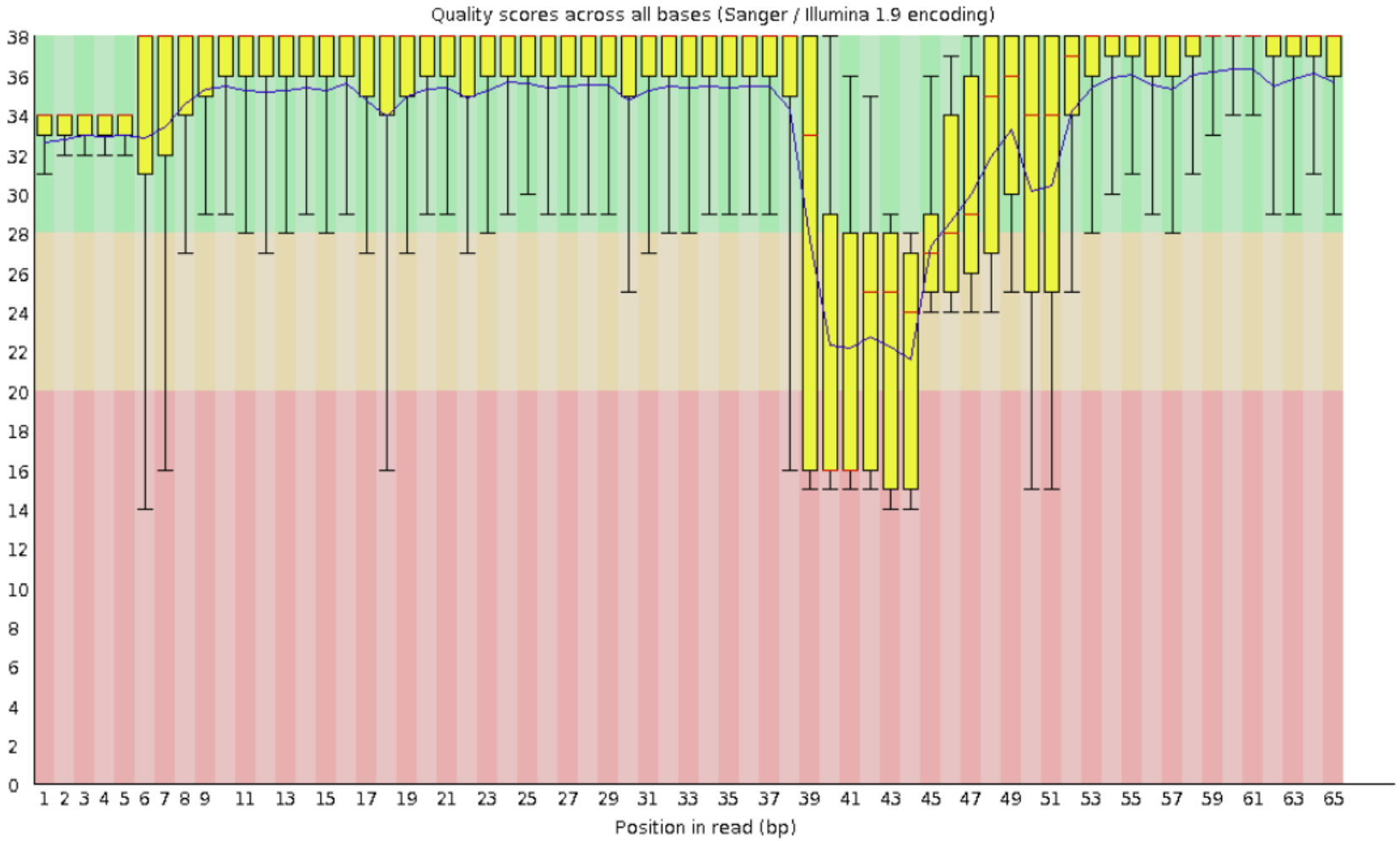


转录组数据的质量检测——FastQC结果解读

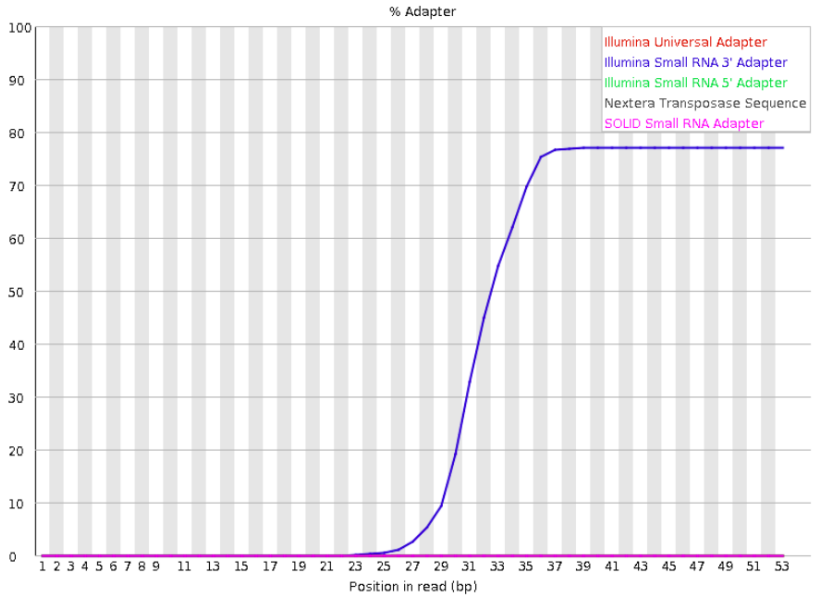
Basic Statistics

Measure	Value
Filename	SRR5345622.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	47613077
Sequences flagged as poor quality	0
Sequence length	65
%GC	64

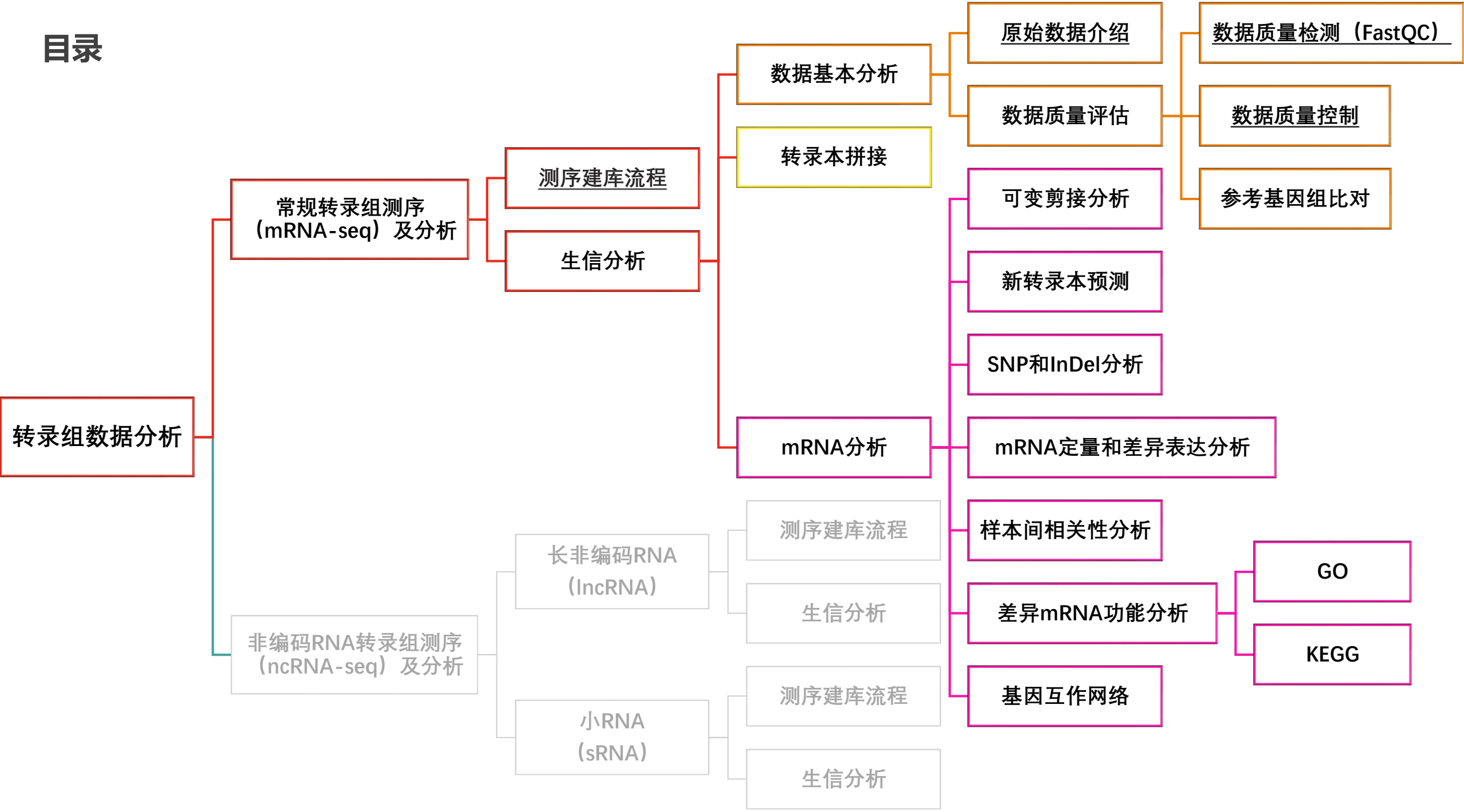
Per base sequence quality



Adapter Content



目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

转录组数据的质量控制——FASTX-Toolkit

（1）fastx_clipper去接头

运行命令

```
fastx_clipper [-h] [-a ADAPTER] [-D] [-l N] [-n] [-d N] [-c] [-C] [-o] [-v] [-z] [-i INFILE] [-o OUTFILE]
```

去接头例子

Clipping Example:

```
>1
ATGTAATGTTTATATATATCGTAAATCCAACACAAT
>2
TATTTTGGAATTCCACGACCCTGTAGGCACCATCAA
>3
ACGTTGTTCGGTGCGTCCTGAACTGTAGGCACCATC
>4
TTTCTTCTTATCTCTTCGAGTCTGTAGGCACCATCA
>5
TGGAACTTGCTGTAGGCACCATCATTATTTATATAA
>6
TTTACCGGAAGCATAACTCTTCTGTAGGCACCATCA
>7
TGTATTAGCGGTGGGGCCCGACTGTAGGCACCATCA
```

```
>2
TATTTTGGAATTCCACGACC
>3
ACGTTGTTCGGTGCGTCCTGAA
>4
TTTCTTCTTATCTCTTCGAGT

>6
TTTACCGGAAGCATAACTCTT
>7
TGTATTAGCGGTGGGGCCCGA
```

In the above example:

- Sequence no. 1 was discarded since it wasn't clipped (i.e. didn't contain the adapter sequence). (**Output** parameter).
- Sequence no. 5 was discarded --- it's length (after clipping) was shorter than 15 nt (**Minimum Sequence Length** parameter).

fastx_clipper参数

[-a ADAPTER] = 接头序列（默认为CCTTAAGG）

[-l N] = 忽略那些碱基数目少于N的reads，默认为5

[-d N] = 保留接头序列后的N个碱基默认 -d 0

[-c] = 只保留包含接头的序列

[-C] = 只保留没有接头的序列

[-k] = 报告只有接头的序列

[-n] = 保留有N多序列，默认不保留

[-v] = 详细-报告序列编号

[-z] = 压缩输出

[-D] = 输出调试结果

[-M N] = 要求最小能匹配到接头的长度N

[-i INFILE] = 输入文件

[-o OUTFILE] = 输出文件

转录组数据的质量控制——FASTX-Toolkit

（2）fastq_quality_filter过滤低质量序列

运行命令

```
fastq_quality_filter [-h] [-v] [-q N] [-p N] [-z] [-i INFILE] [-o OUTFILE]
```

参数

[-q N] = 最小的需要留下的质量值

[-p N] = 每个reads中最少有百分之多少的碱基需要有-q的质量值

[-z] = 压缩输出

[-v] = 详细-报告序列编号，如果使用了-o则报告会直接在STDOUT，如果没有则输入到STDERR

例子

Quality score distribution (of all cycles) is calculated for each read. If it is lower than the quality cut-off value - the read is discarded.

Example:

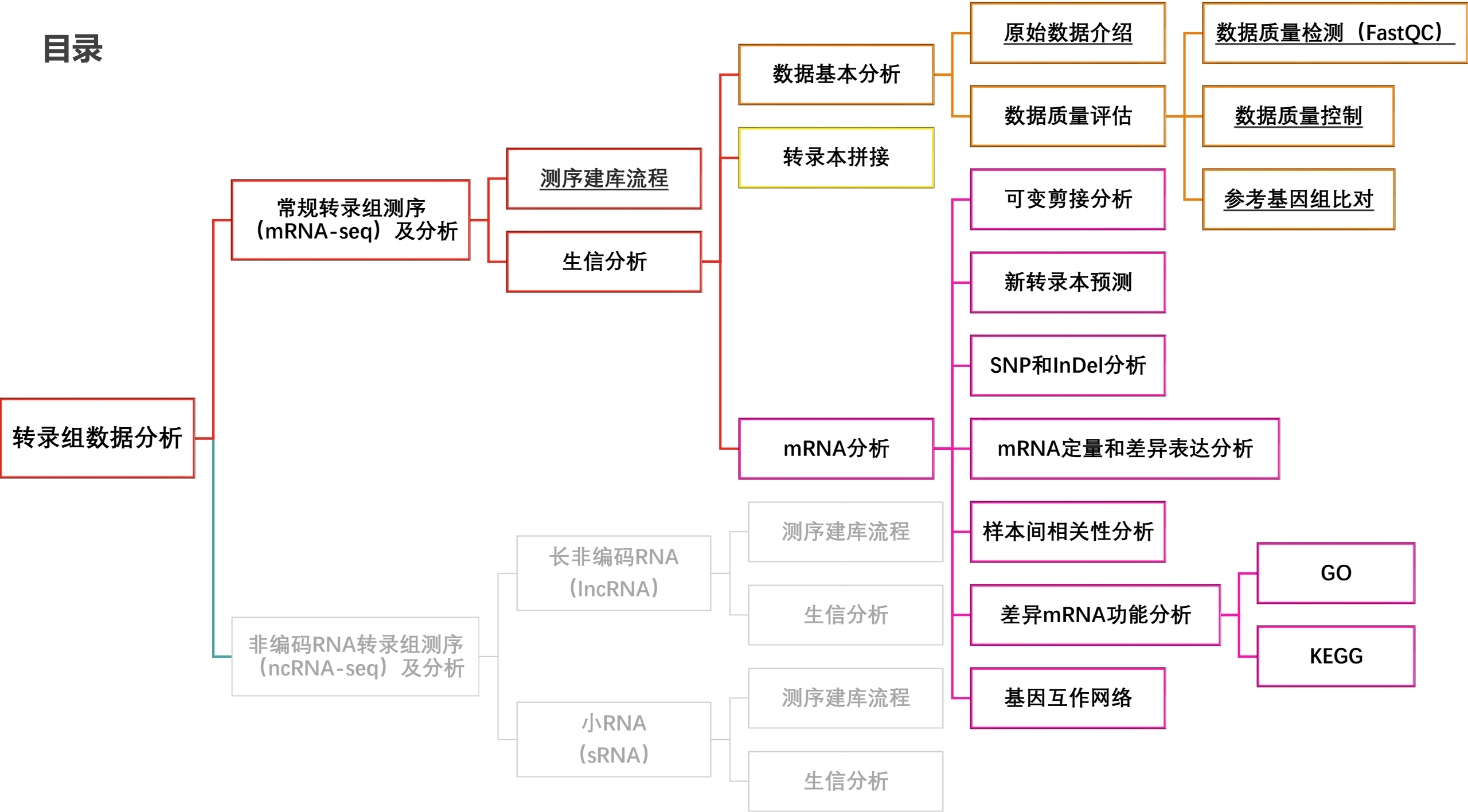
```
@CSHL_4_FC042AG00II:1:2:214:584
GACAATAAAC
+CSHL_4_FC042AG00II:1:2:214:584
30 30 30 30 30 30 30 30 20 10
```

Using **percent** = **50** and **cut-off** = **30** - This read will not be discarded (the median quality is higher than 30).

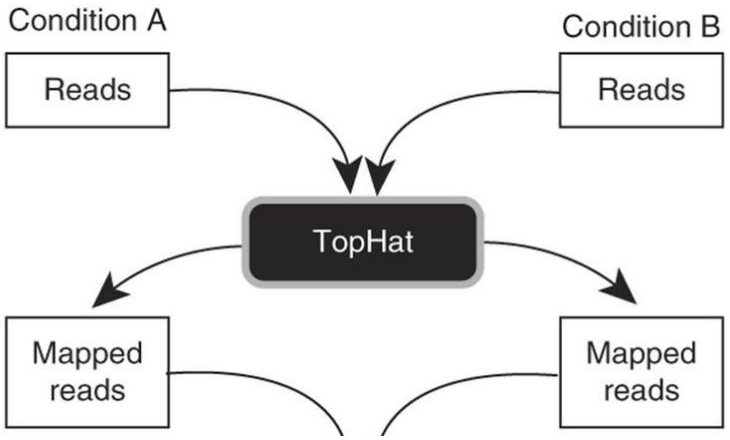
Using **percent** = **90** and **cut-off** = **30** - This read will be discarded (90% of the cycles do no have quality equal to / higher than 30).

Using **percent** = **100** and **cut-off** = **20** - This read will be discarded (not all cycles have quality equal to / higher than 20).

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

参考基因组比对——Tophat（调用bowtie）



输出文件

accepted_hits.bam
align_summary.txt
deletions.bed
insertions.bed
junctions.bed
logs
prep_reads.info
unmapped.bam

genes.gtf为基因组注释文件
.fq为测序文件

建bowtie index库
bowtie-build genome.fa genome

1| Map the reads for each sample to the reference genome:

```
$ tophat -p 8 -G genes.gtf -o C1_R1_thout genome C1_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R2_thout genome C1_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R3_thout genome C1_R3_1.fq C1_R3_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R1_thout genome C2_R1_1.fq C2_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R2_thout genome C2_R2_1.fq C2_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R3_thout genome C2_R3_1.fq C2_R3_2.fq
```

什么是gtf格式?

```
#gtf(General Transfer Format) format
1 Ensembl gene 11869 14409 . + . gene_id "ENSG00000223972";
```

```
#Details
1. seqname - name of the chromosome or scaffold.
2. source - name of the program that generated this feature, or the data source
3. feature - feature type name, e.g. Gene, Variation, Similarity
4. start - Start position of the feature, with sequence numbering starting at 1.
5. end - End position of the feature, with sequence numbering starting at 1.
6. score - A floating point value.
7. strand - defined as + (forward) or - (reverse).
8. frame - One of '0', '1' or '2'.
9. attribute - A semicolon-separated list of tag-value pairs, providing additional
information about each feature.
```

什么是bam格式?

#bam是sam（Sequence Alignment/Map format）文件的二进制格式

```
1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0 CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG 0;==-
==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
```

Col	Field	Brief Description	Alignment
1	QNAME	Query template NAME	1:497:R:-272+13M17D24M
2	FLAG	bitwise FLAG	113
3	RNAME	References sequence NAME	1
4	POS	1- based leftmost mapping Position	497
5	MAPQ	Mapping Quality	37
6	CIGAR	CIGAR String	37M
7	MRNM/RNEXT	Ref. name of the mate/next read	15
8	MPOS/NEXT	Position of the mate/next read	100338662
9	ISIZE/TLEN	observed Template LENgth	0
10	SEQ	segment SEQuence	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
11	QUAL	ASCII of Phred-scaled b	0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>
12	TAGs	TAGs	XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37

参考基因组比对——HISAT（调用BWA）

Extract splice-site and exon information from the gene annotation file

```
$ extract_splice_sites.py chrX_data/genes/chrX.gtf >chrX.ss
$ extract_exons.py chrX_data/genes/chrX.gtf >chrX.exon
```

Build a HISAT2 index

```
$ hisat2-build --ss chrX.ss --exon chrX.exon chrX_data/genome/chrX.fa chrX_tran
```

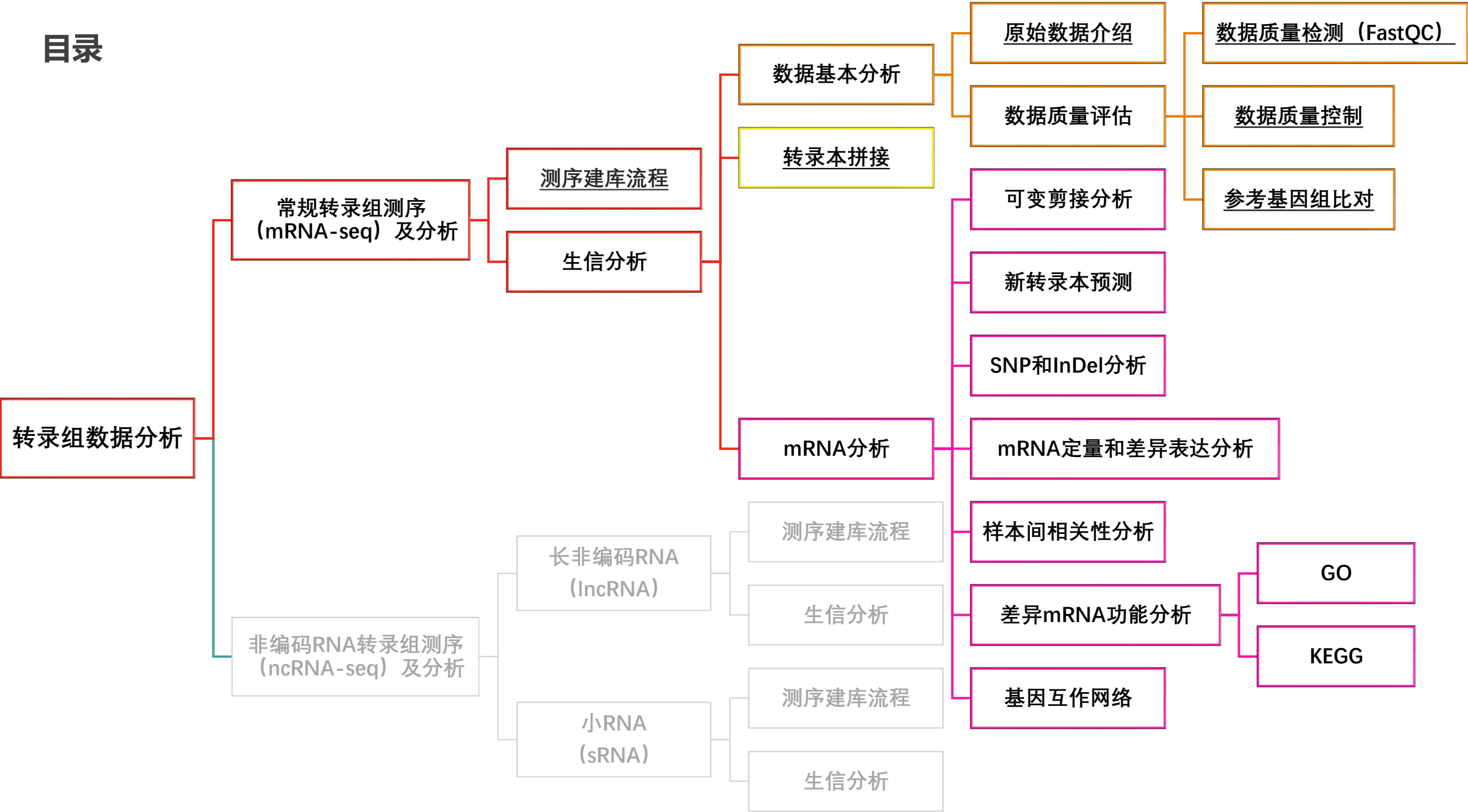
Map the reads for each sample to the reference genome

```
$ hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1
chrX_data/samples/ERR188044_chrX_1.fastq.gz -2
chrX_data/samples/ERR188044_chrX_2.fastq.gz -S ERR188044_chrX.sam
```

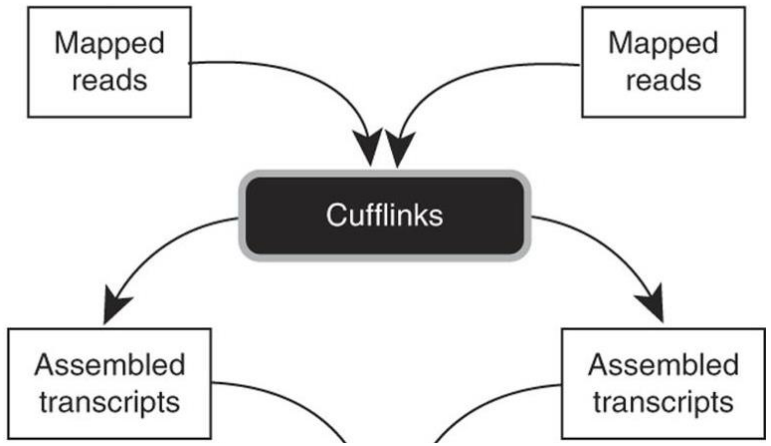
Sort and convert the SAM files to BAM (SAMtools version 1.3 or newer)

```
$ samtools sort -@ 8 -o ERR188044_chrX.bam ERR188044_chrX.sam
```

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

转录本拼接/组装——Cufflinks



2| Assemble transcripts for each sample:

```
$ cufflinks -p 8 -o C1_R1_clout C1_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R2_clout C1_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R3_clout C1_R3_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R1_clout C2_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R2_clout C2_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R3_clout C2_R3_thout/accepted_hits.bam
```

输出文件

genes.fpkm_tracking

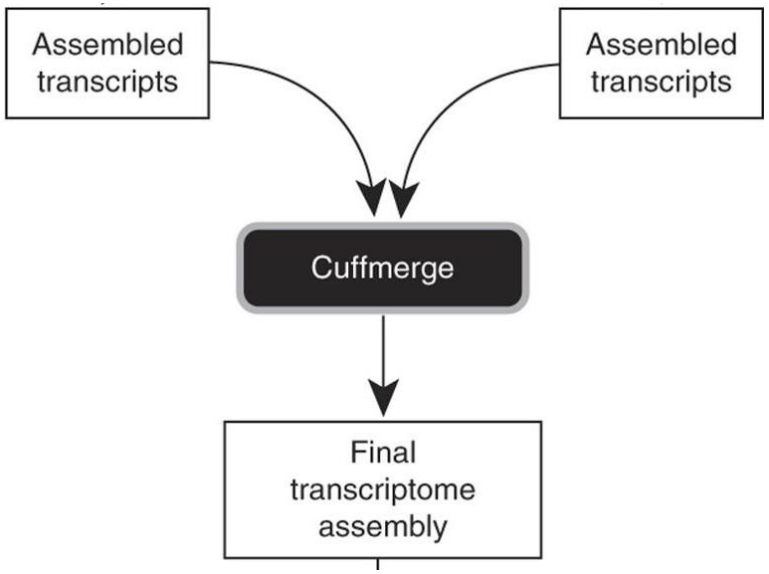
isoforms.fpkm_tracking

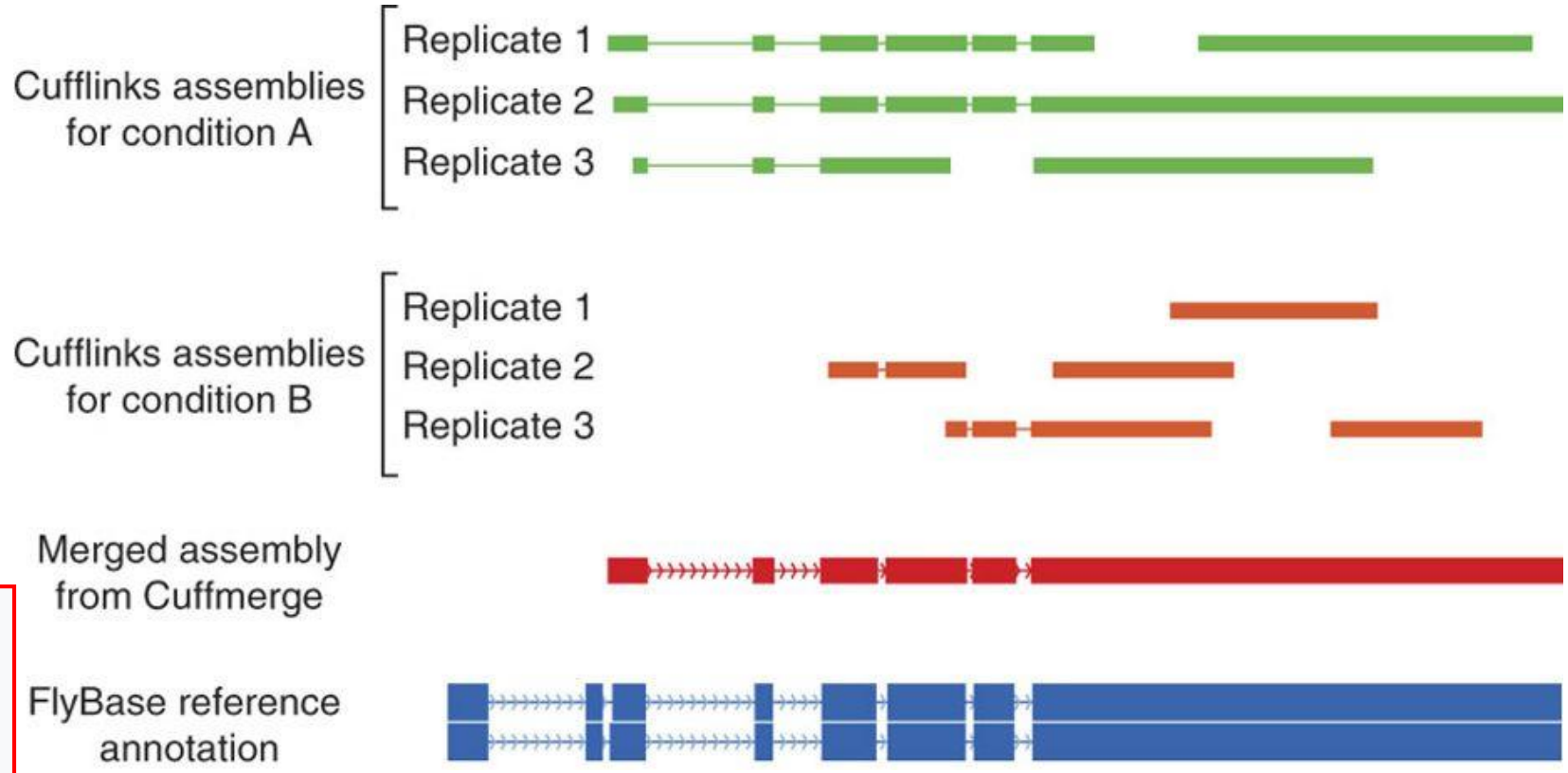
skipped.gtf

transcripts.gtf

```
2L  Cufflinks  exon  73485  73692  1000  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "1";
2L  Cufflinks  exon  74903  75018  1000  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "2";
2L  Cufflinks  exon  75078  76210  1000  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "3";
2L  Cufflinks  transcript  74485  76210  310  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.3"; FPKM "9.24272
2L  Cufflinks  exon  74485  74572  310  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.3"; exon_number "1"; FP
2L  Cufflinks  exon  74903  75018  310  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.3"; exon_number "2"; FP
2L  Cufflinks  exon  75078  76210  310  +  .  gene_id "CUFF.7"; transcript_id "CUFF.7.3"; exon_number "3"; FP
2L  Cufflinks  transcript  102382  105332  1000  +  .  gene_id "CUFF.8"; transcript_id "CUFF.8.1"; FPKM "1
2L  Cufflinks  exon  102382  102906  1000  +  .  gene_id "CUFF.8"; transcript_id "CUFF.8.1"; exon_number "
2L  Cufflinks  exon  103006  103434  1000  +  .  gene_id "CUFF.8"; transcript_id "CUFF.8.1"; exon_number "
2L  Cufflinks  exon  103516  105332  1000  +  .  gene_id "CUFF.8"; transcript_id "CUFF.8.1"; exon_number "
2L  Cufflinks  transcript  102382  106710  1000  -  .  gene_id "CUFF.9"; transcript_id "CUFF.9.1"; FPKM "8
2L  Cufflinks  exon  102382  104947  1000  -  .  gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "
```

转录本拼接/组装——Cuffmerge整合Cufflinks组装的转录本






```
./C1_R1_clout/transcripts.gtf
./C2_R2_clout/transcripts.gtf
./C1_R2_clout/transcripts.gtf
./C2_R1_clout/transcripts.gtf
./C1_R3_clout/transcripts.gtf
./C2_R3_clout/transcripts.gtf
```

输出结果 `merged_asm/merged.gtf`

3| Create a file called assemblies.txt that lists the assembly file for each sample.
4| Run Cuffmerge on all your assemblies to create a single merged transcriptome annotation:

```
cuffmerge -g genes.gtf -s genome.fa -p 8 assemblies.txt
```

转录本拼接/组装——StringTie

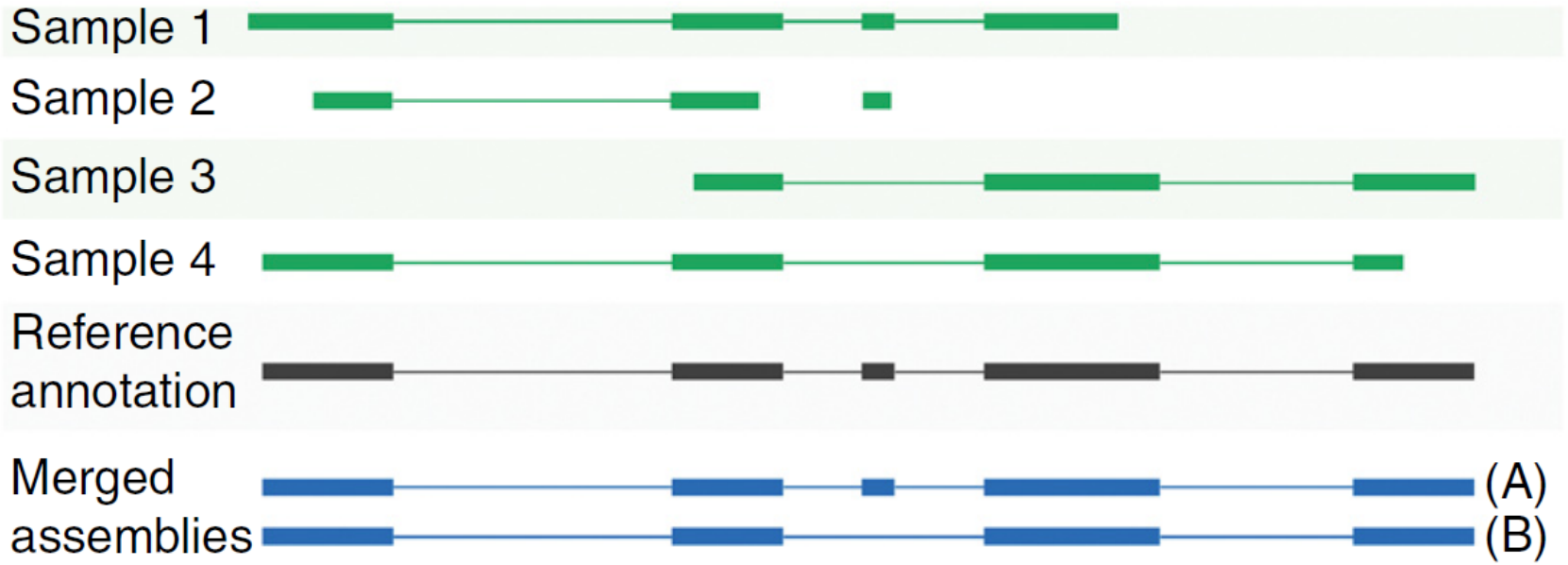
Assemble transcripts for each sample:

```
$ stringtie -p 8 -G chrX_data/genes/chrX.gtf -o ERR188044_chrX.gtf -l ERR188044
ERR188044_chrX.bam
```

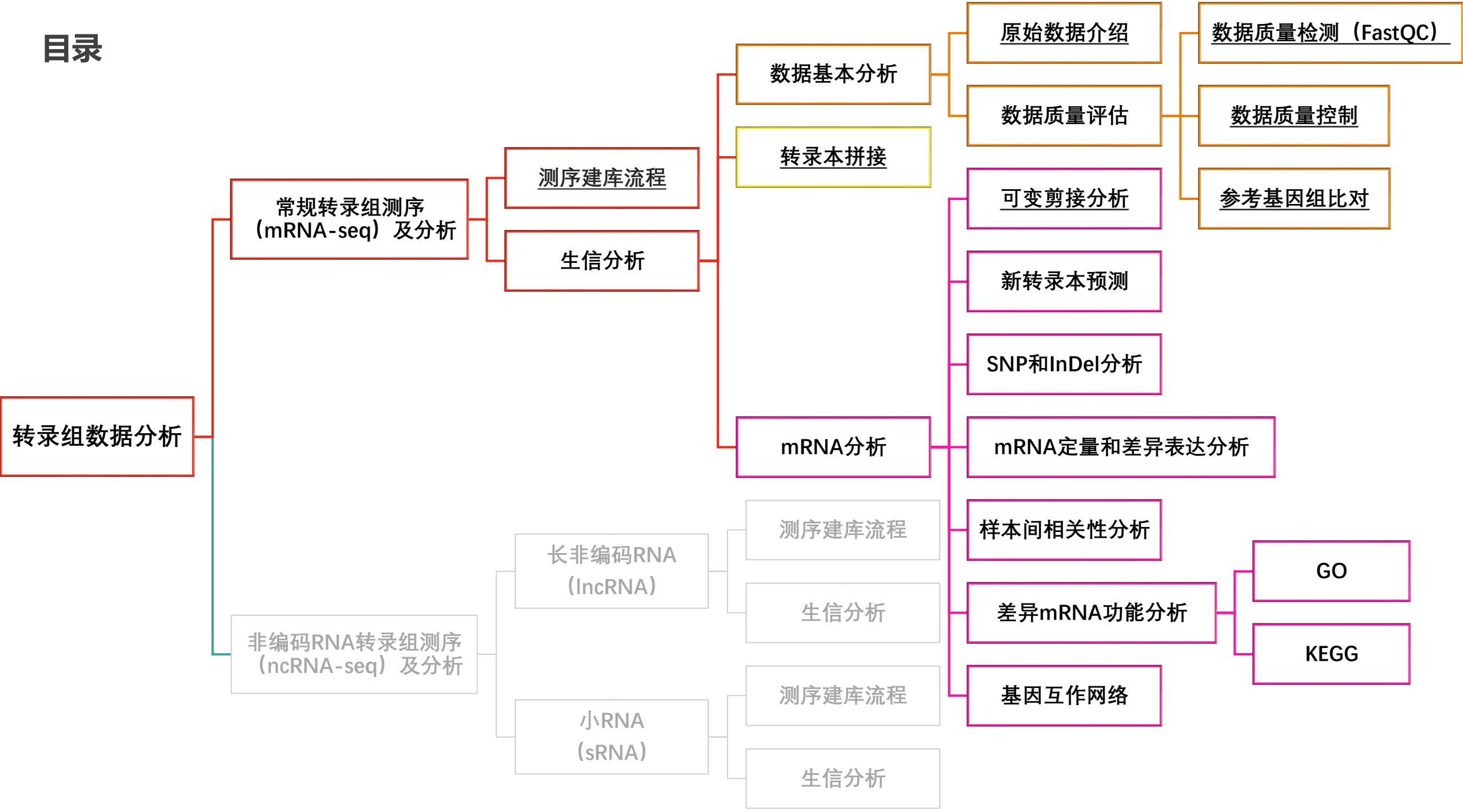
Merge transcripts from all samples

```
$ stringtie --merge -p 8 -G chrX_data/genes/chrX.gtf -o stringtie_merged.gtf
chrX_data/mergelist.txt
```

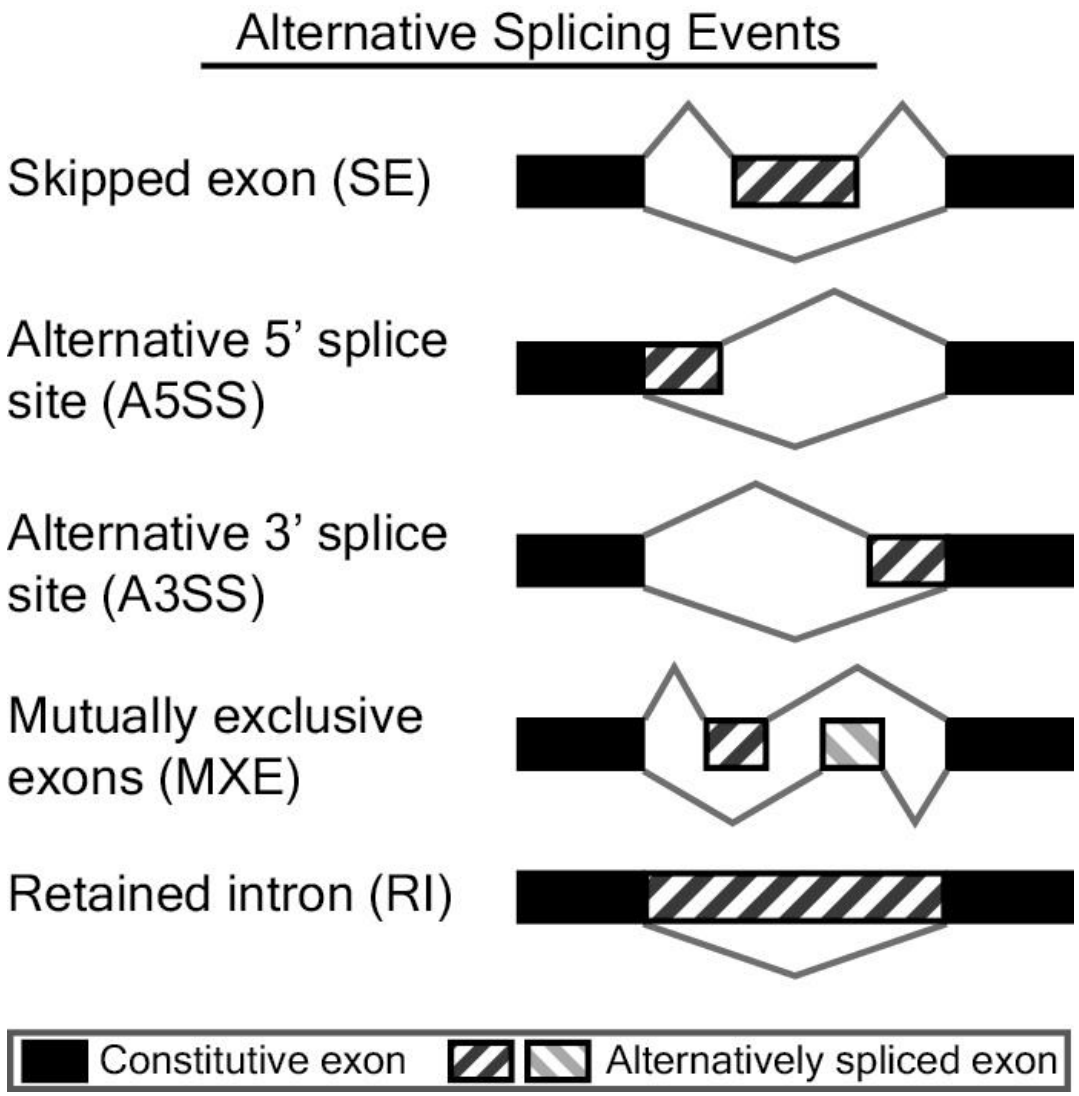
mergelist.txt is a text file that has the names of the gene transfer format (GTF) files created in the previous step, with each file name on a single line (see ChrX_data for an example file).



目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

可变剪接分析——rMATS




```
python rmats.py  \
--b1 b1.txt --b2 b2.txt \
--gtf ref.transcript.gtf \
--od out_dir \
-t paired \
--readLength 101 \
--cstat 0.1 \
--libType fr-unstranded
```

b1.txt中保存的是每个样本比对参考基因组的bam文件的路径

```
python rmats.py \
--s1 s1.txt --s2 s2.txt \
--gtf ref.transcript.gtf \
--bi /STARindex/hg19 \
--od out_dir \
-t paired \
--nthread 6 \
--readLength 151
```

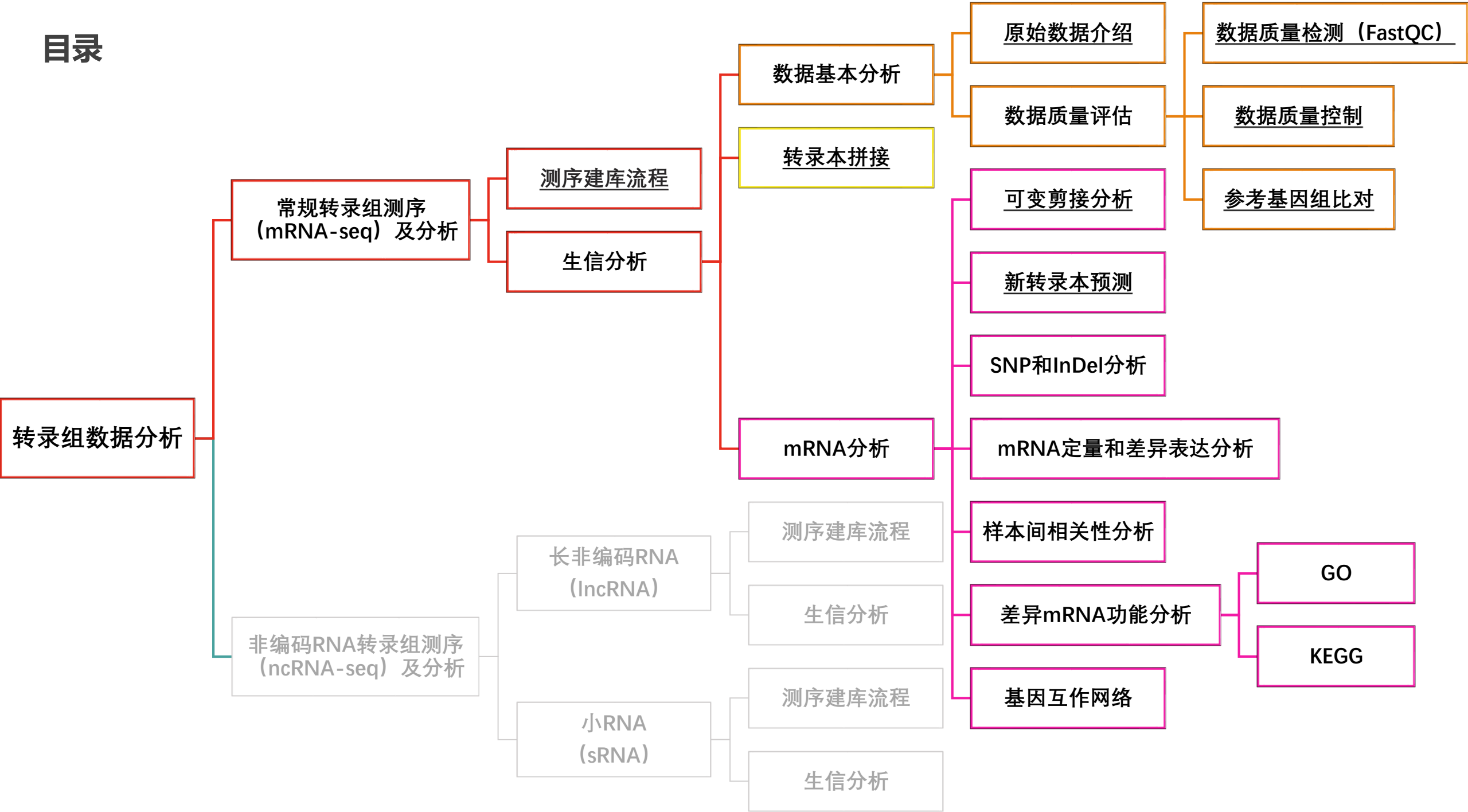
S1.txt中保存的是每个样本fastq文件的路径，rmats会自动调用STAR进行比对，bi参数指定参考基因组STAR的索引

可变剪接分析——rMATS

ID	IJC_SAMPLE_1	SJC_SAMPLE_1	IJC_SAMPLE_2	SJC_SAMPLE_2	IncFormLen	SkipFormLen	PValue	FDR	IncLevel1	IncLevel2	IncLevelDifference
0	706,762,624	0,0,0	801,811,683	2,0,0	198	99	1	1	1.0,1.0,1.0	0.995,1.0,1.0	0.002
1	10,261,009,897	3,4,0	120,914,591,117	2,3,0	198	99	1	1	0.994,0.992,1.(	0.997,0.996,1.0	-0.002
2	490,524,470	0,0,0	647,753,612	2,0,0	198	99	1	1	1.0,1.0,1.0	0.994,1.0,1.0	0.002
3	323,317,318	0,0,0	352,335,353	2,0,0	194	99	1	1	1.0,1.0,1.0	0.989,1.0,1.0	0.004
4	859,804,747	3,6,2	9,141,040,865	2,2,7	194	99	1	1	0.993,0.986,0.9	0.996,0.996,0.984	-0.001
5	529,491,490	0,0,1	608,604,570	0,0,0	198	99	1	1	1.0,1.0,0.996	1.0,1.0,1.0	-0.001
6	302,339,321	0,0,0	455,396,385	1,0,0	147	99	1	1	1.0,1.0,1.0	0.997,1.0,1.0	0.001
7	412,467,422	2,0,4	597,506,459	0,1,0	198	99	1	1	0.99,1.0,0.981	1.0,0.996,1.0	-0.008
8	0,2,10	0,0,3	2,2,3	0,0,0	184	99	1	1	NA,1.0,0.642	1.0,1.0,1.0	-0.179
9	5,1,7	1,0,0	4,6,2	0,2,0	198	99	1	1	0.714,1.0,1.0	1.0,0.6,1.0	0.038
11	87,114,78	884,986,842	88,112,81	10,721,366,971	185	99	0.075639	0.927277	0.05,0.058,0.0	0.042,0.042,0.043	0.009

IJC表示inclusion isoform counts, SJC表示是skipping isoform counts, 生物学重复样本用逗号分隔；IncFormLen代表effective inclusion isoform length, SkipFormLen代表effective inclusion isoform length；lencLevel代表定量的结果，IncIevelDifference就是两组样本中表达量的差值，通过Pvalue和FDR可以对结果进行过滤和筛选。

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

新转录本预测

使用Cuffcompare将cuffmerge得到的转录本与原基因注释文件进行比较

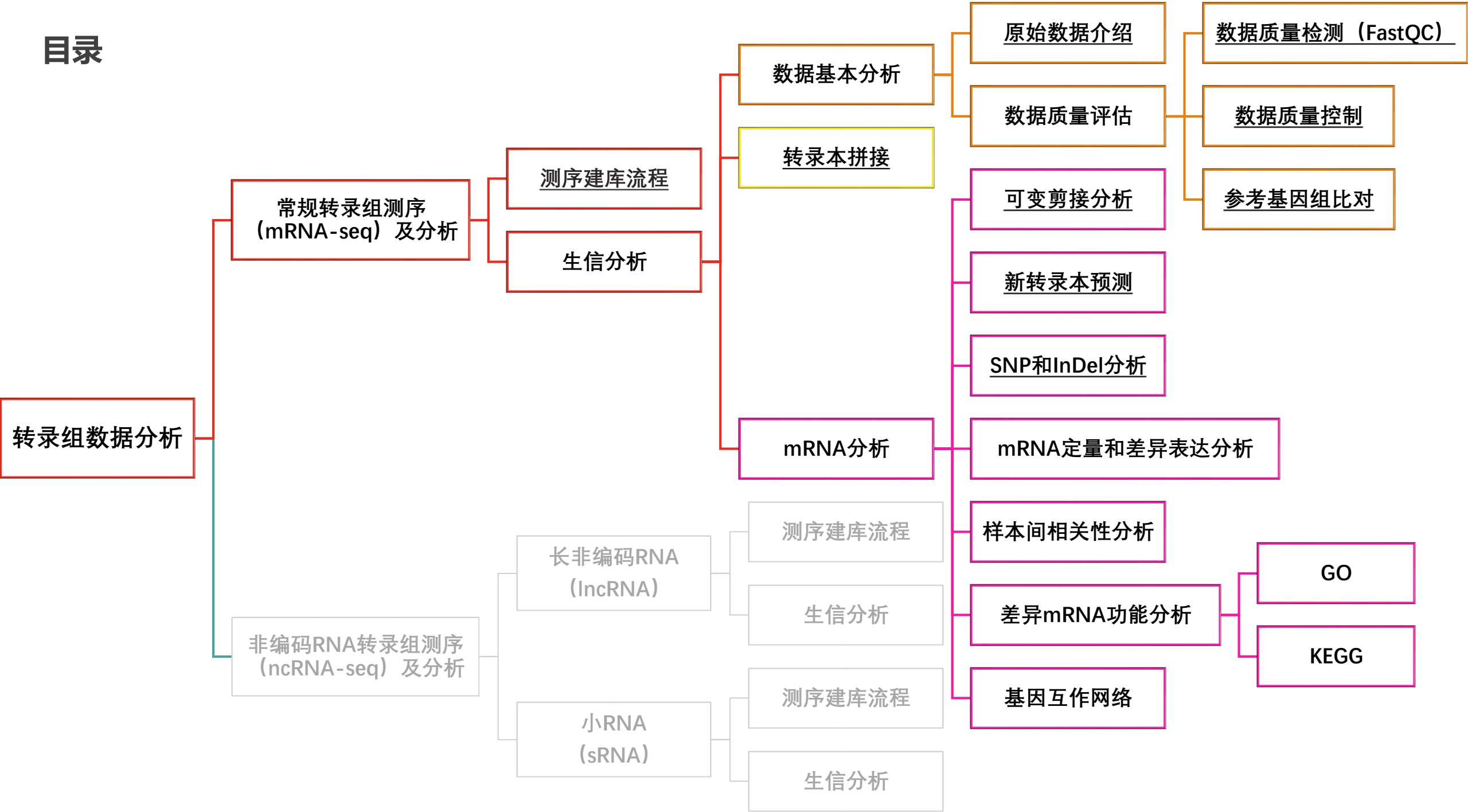
```
cuffcompare -r hg38.gtf -s hg38.fa -C merged.gtf
```

```
gffcompare -r chrX_data/genes/chrX.gtf -G -o merged stringtie_merged.gtf
```

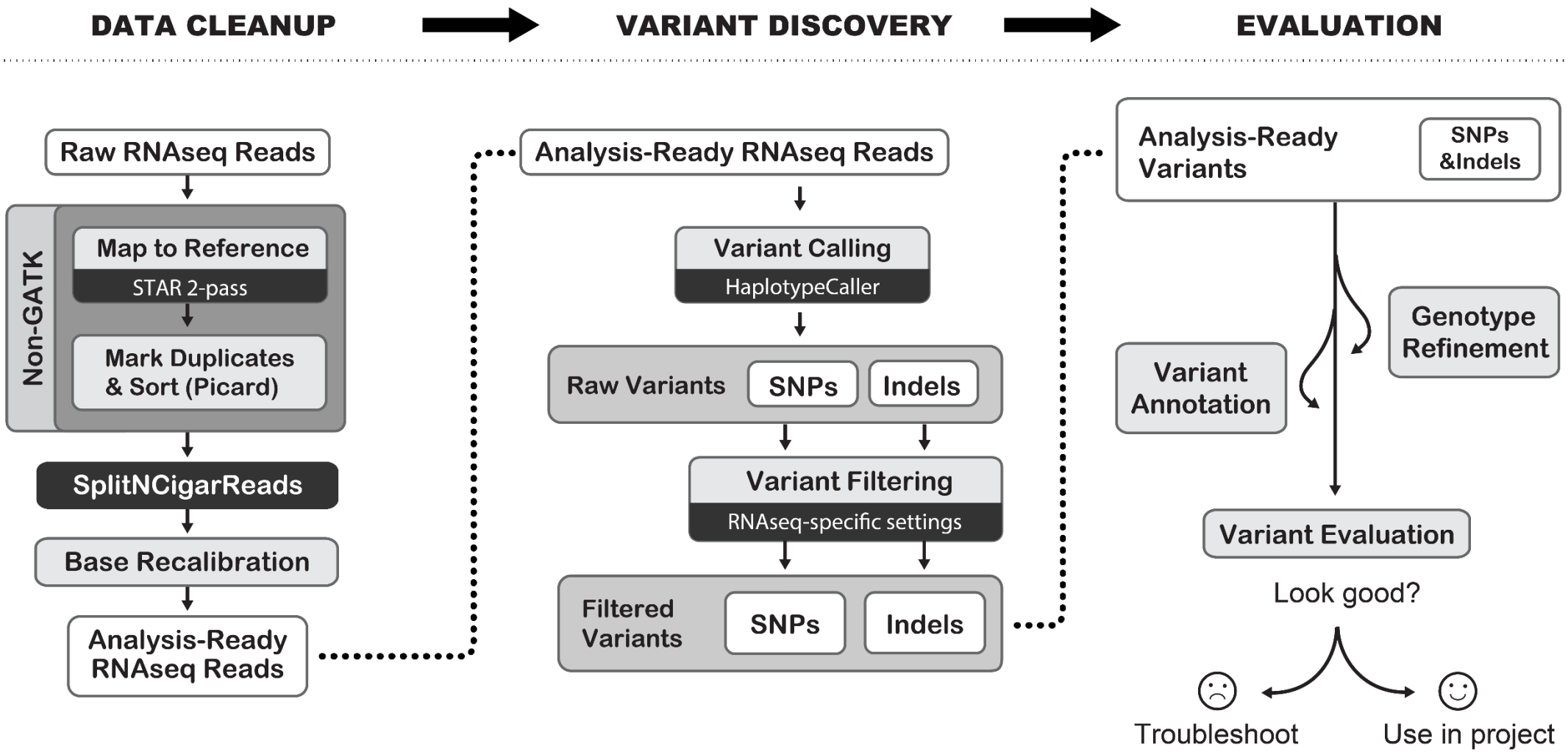
- 发现新的未知基因（相对于原有基因注释文件）
- 发现已知基因新的外显子区域
- 对已知基因的起始和终止位置进行优化

Code	Description
=	Complete match of intron chain
c	Contained
j	Potentially novel isoform (fragment) : at least one splice junction is shared with a reference transcript
e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
i	A trans frag falling entirely within a reference intron
o	Generic exonic overlap with a reference transcript
p	Possible polymerase run-on fragment (within 2k bases of a reference transcripts)
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
u	Unknown, intergenic transcript
x	Exonic overlap with reference on the opposite strand
s	An intron of transfrag overalps a reference intron on the opposite strand
.	.tracking file only, indicates multiple classifications

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

SNP和InDel分析（基于RNA-seq数据）



SNP和InDel分析（基于RNA-seq数据）

Mapping to the Reference
Tools involved: STAR

Data Cleanup
Tools involved: MergeBamAlignment, MarkDuplicates

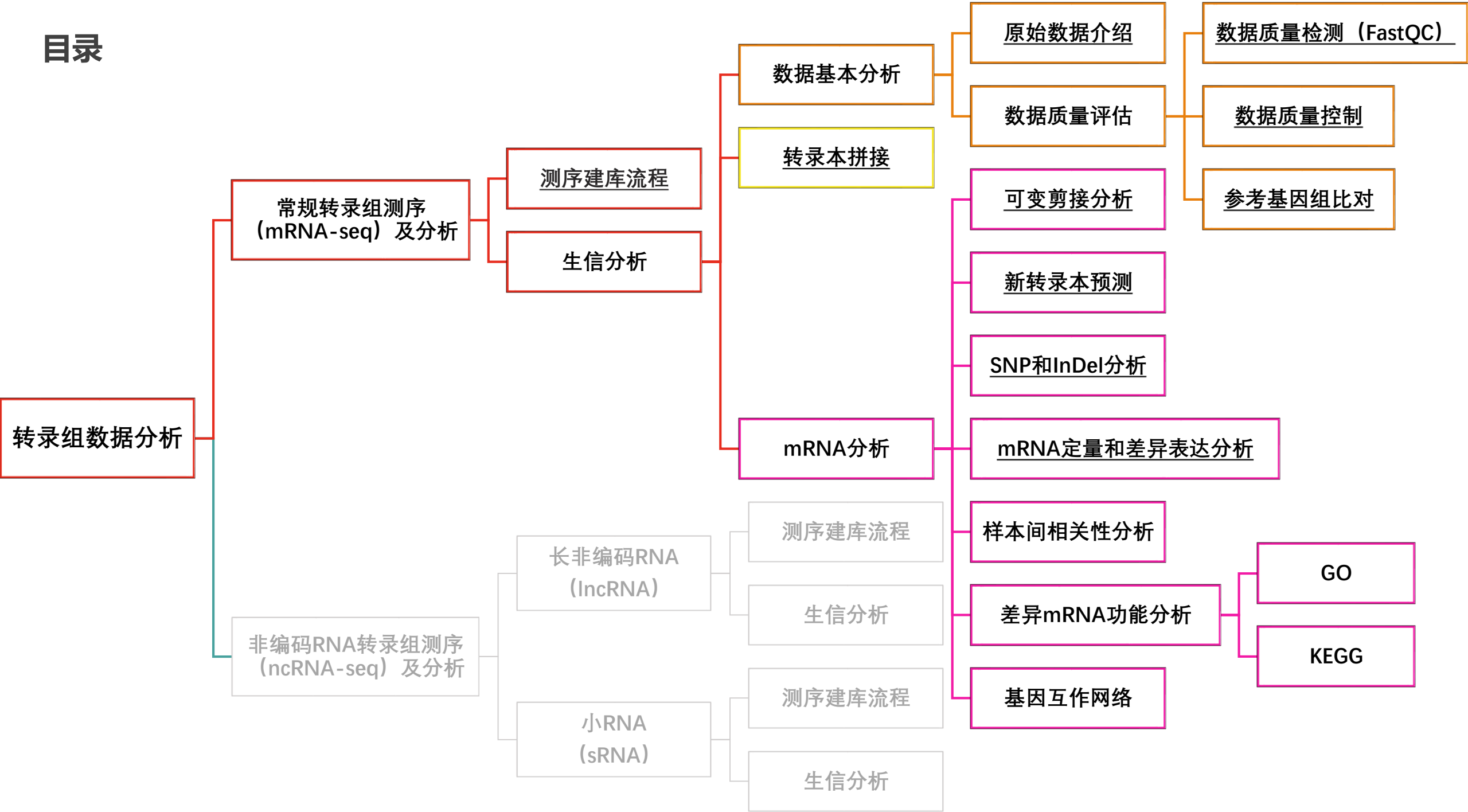
SplitNCigarReads
Tools involved: SplitNCigarReads

Base Quality Recalibration
Tools involved: BaseRecalibrator, Apply Recalibration, AnalyzeCovariates (optional)

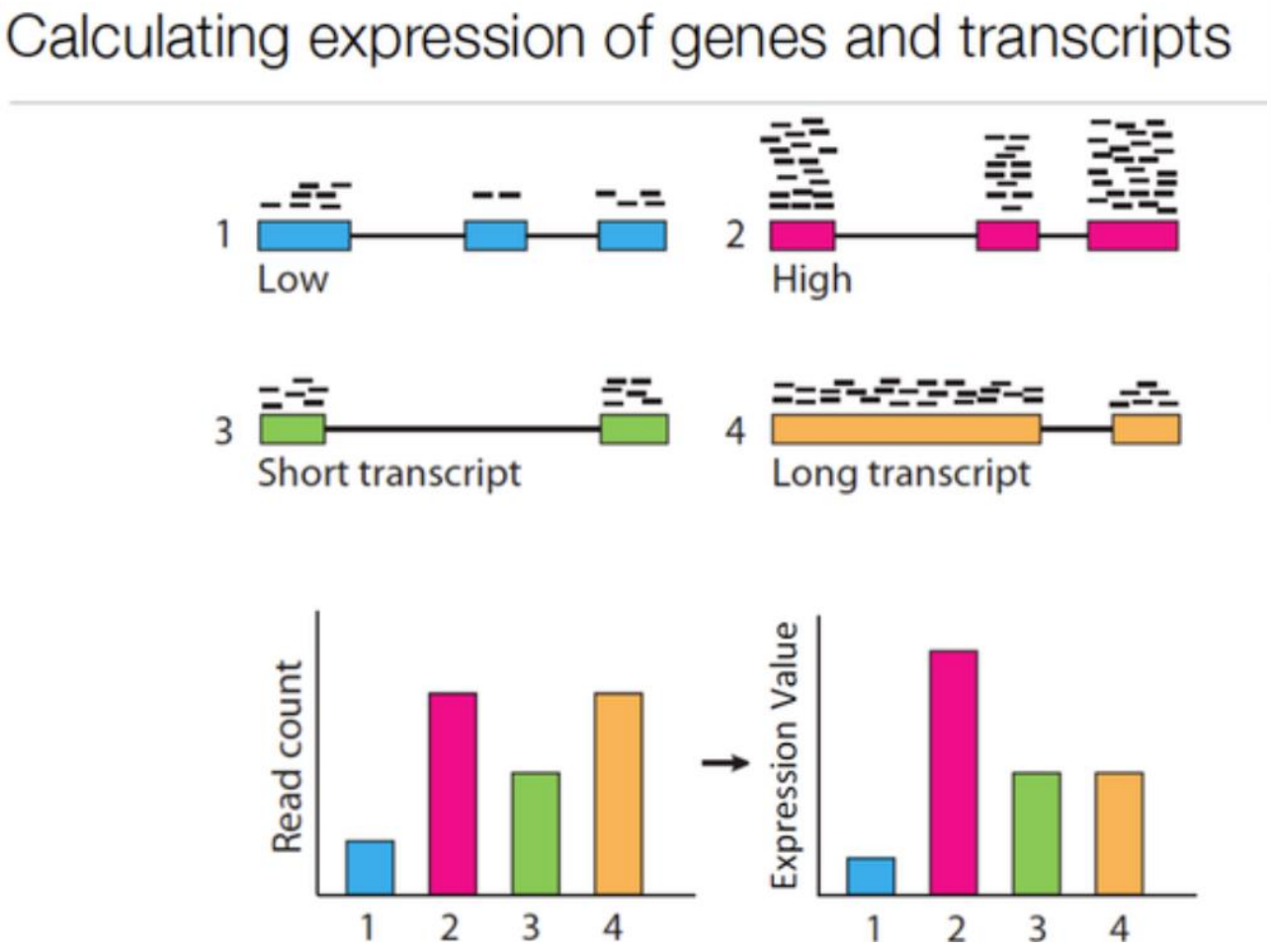
Variant Calling
Tools involved: HaplotypeCaller

Variant Filtering
Tools involved: VariantFiltration

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

mRNA定量——FPKM/RPKM、TPM、read count



$$TPM = \frac{X_i}{L_i} * \frac{1}{\sum_j \frac{X_i}{L_j}} = \frac{\frac{X_i}{L_i}}{\sum_j \frac{X_i}{L_j}}$$

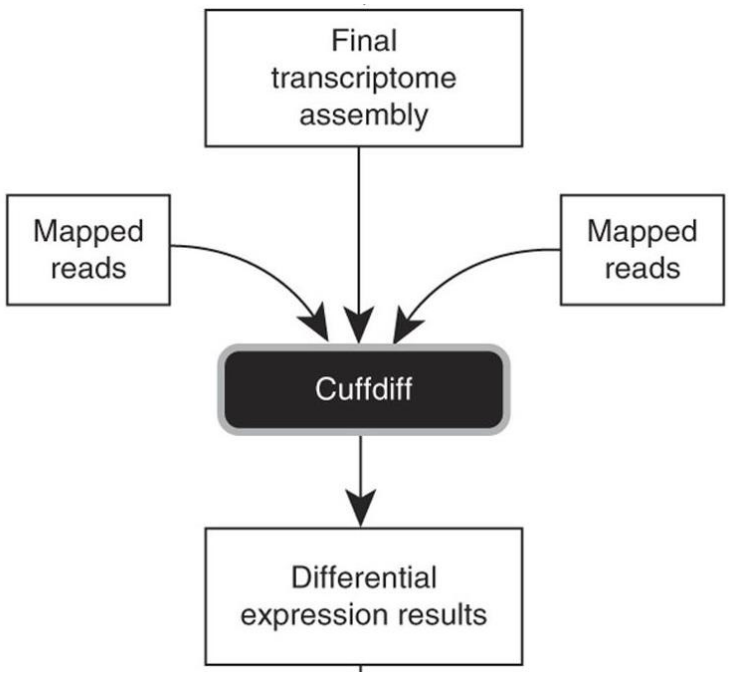
$$FPKM = \frac{X_i}{L_i} * \frac{1}{\sum_j X_j} = \frac{\frac{X_i}{L_i}}{\sum_j X_j}$$

- DESeq2和edgeR不用RPKM/FPKM或TPM做均一化，而是直接用原始的read counts做均一化处理。

- Xi表示比对到某基因上的Fragment数目，单位是Millions Fragments；Li表示这个基因外显子的长度，单位是Kb。

- 二者分母不同，TPM的分母中考虑了长度的影响，可以理解成总的转录本丰度；而FPKM中，没有考虑长度影响，可以理解成总的Fragment数目

差异表达分析——Cuffdiff

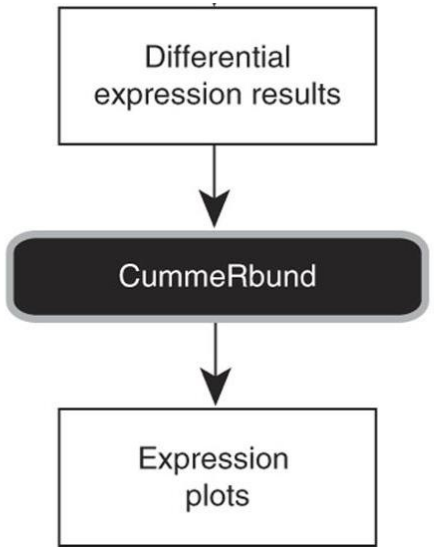


输出文件	diff_out/isoforms.count_tracking
diff_out/bias_params.info	diff_out/isoforms.fpkm_tracking
diff_out/cds.count_tracking	diff_out/isoforms.read_group_tracking
diff_out/cds.diff	diff_out/promoters.diff
diff_out/cds.fpkm_tracking	diff_out/read_groups.info
diff_out/cds.read_group_tracking	diff_out/run.info
diff_out/cds_exp.diff	diff_out/splicing.diff
diff_out/**gene_exp.diff**	diff_out/tss_group_exp.diff
diff_out/genes.count_tracking	diff_out/tss_groups.count_tracking
diff_out/genes.fpkm_tracking	diff_out/tss_groups.fpkm_tracking
diff_out/genes.read_group_tracki	diff_out/tss_groups.read_group_tracking
diff_out/isoform_exp.diff	diff_out/var_model.info

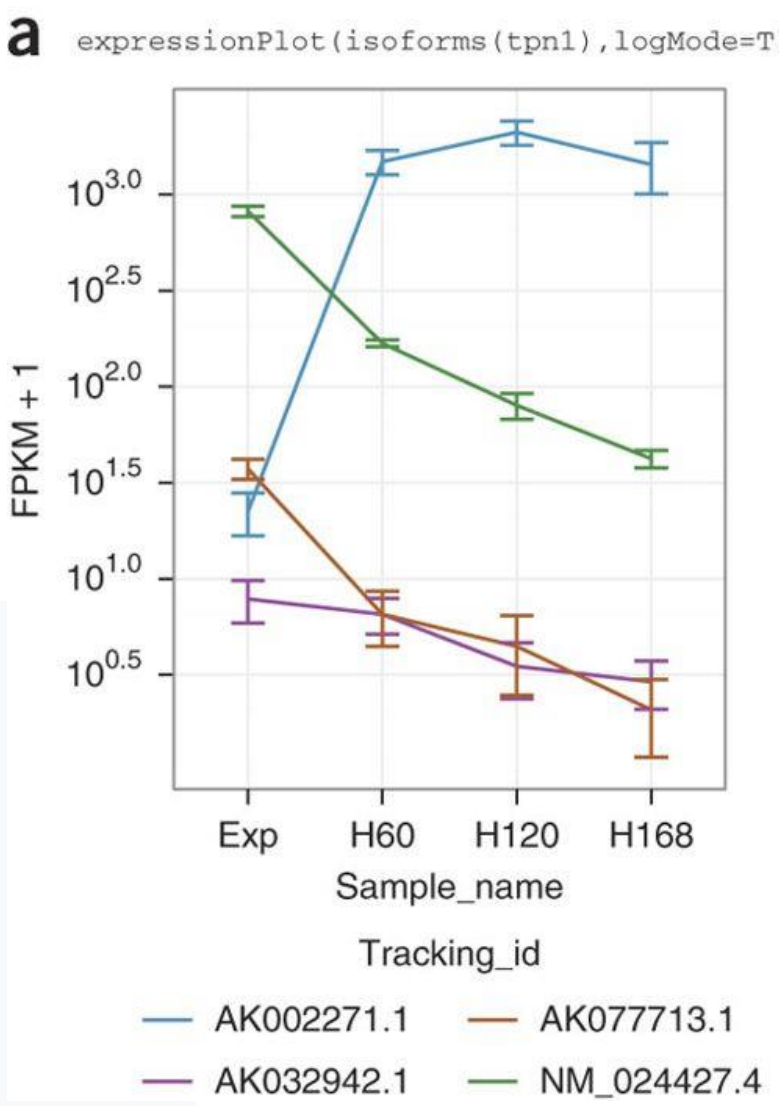
5| Run Cuffdiff by using the merged transcriptome assembly along with the BAM files from TopHat for each replicate:

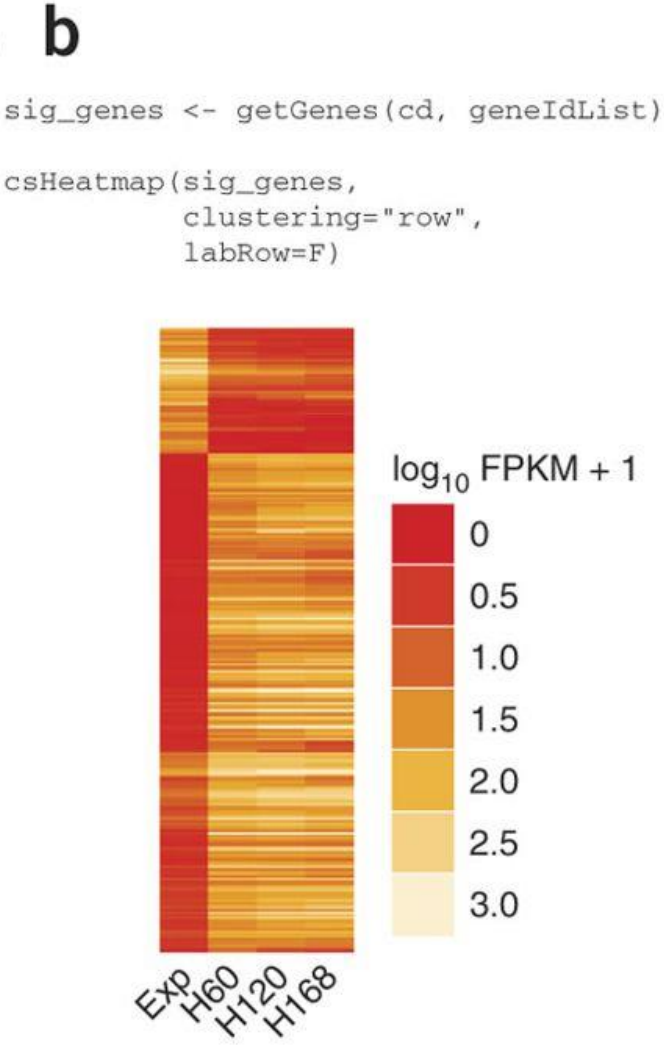
```
$ cuffdiff -o diff_out -b genome.fa -p 8 -L C1,C2 -u merged_asm/merged.gtf
./C1_R1_thout/accepted_hits.bam,./C1_R2_thout/accepted_hits.bam,./C1_R3_thout/accepted_hits.bam
./C2_R1_thout/accepted_hits.bam,./C2_R3_thout/accepted_hits.bam,./C2_R2_thout/accepted_hits.bam
```

差异表达可视化——CummeRbund

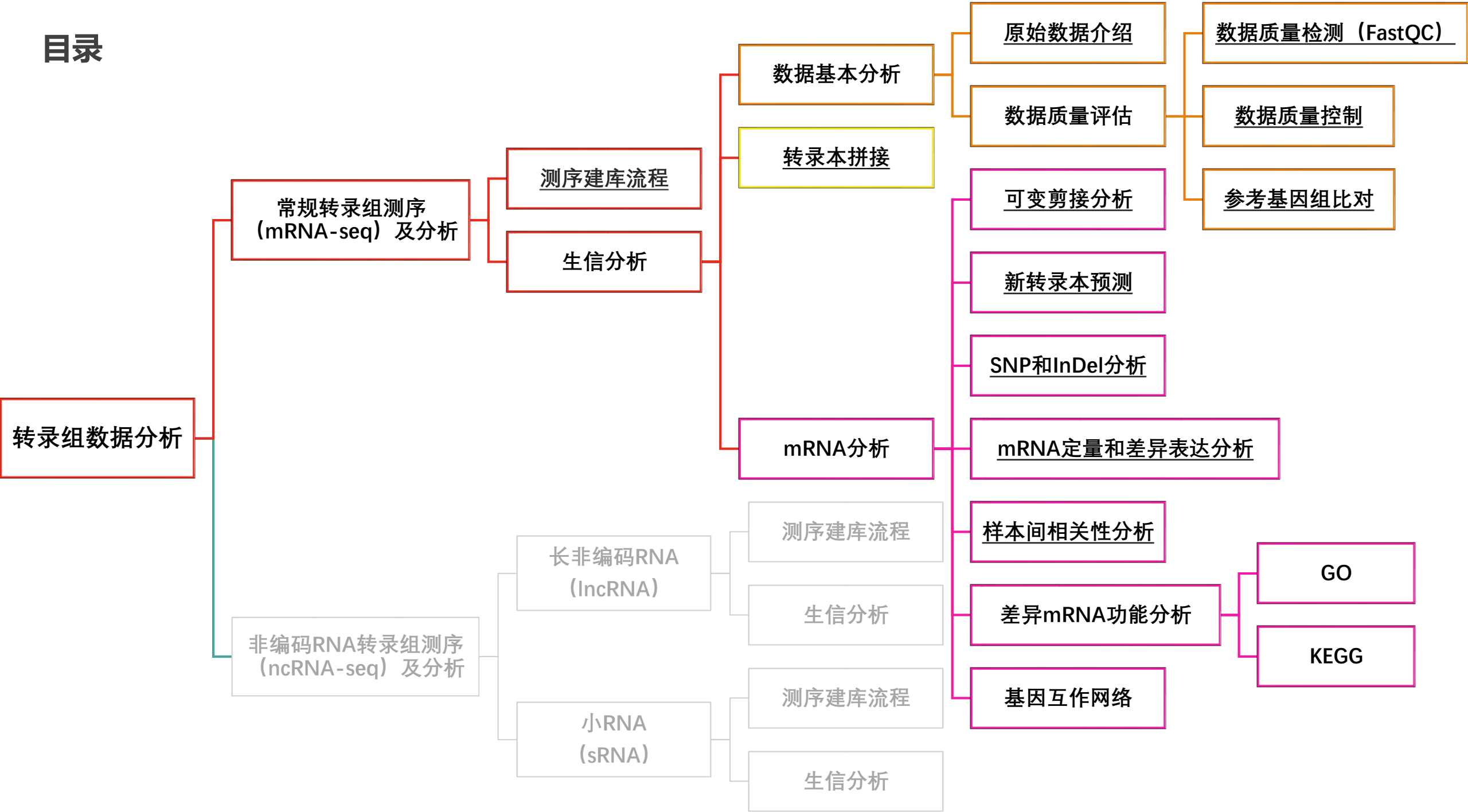



```
library(cummeRbund)
cuff_data < - readCufflinks('diff_out')
csDensity(genes(cuff_data))
mygene < - getGene(cuff_data,'regucalcin')
expressionBarplot(mygene)

expressionPlot(isoforms(tpn1),logMode=T)

sig_genes <- getGenes(cd, geneIdList)
csHeatmap(sig_genes, clustering="row", labRow=F)
```

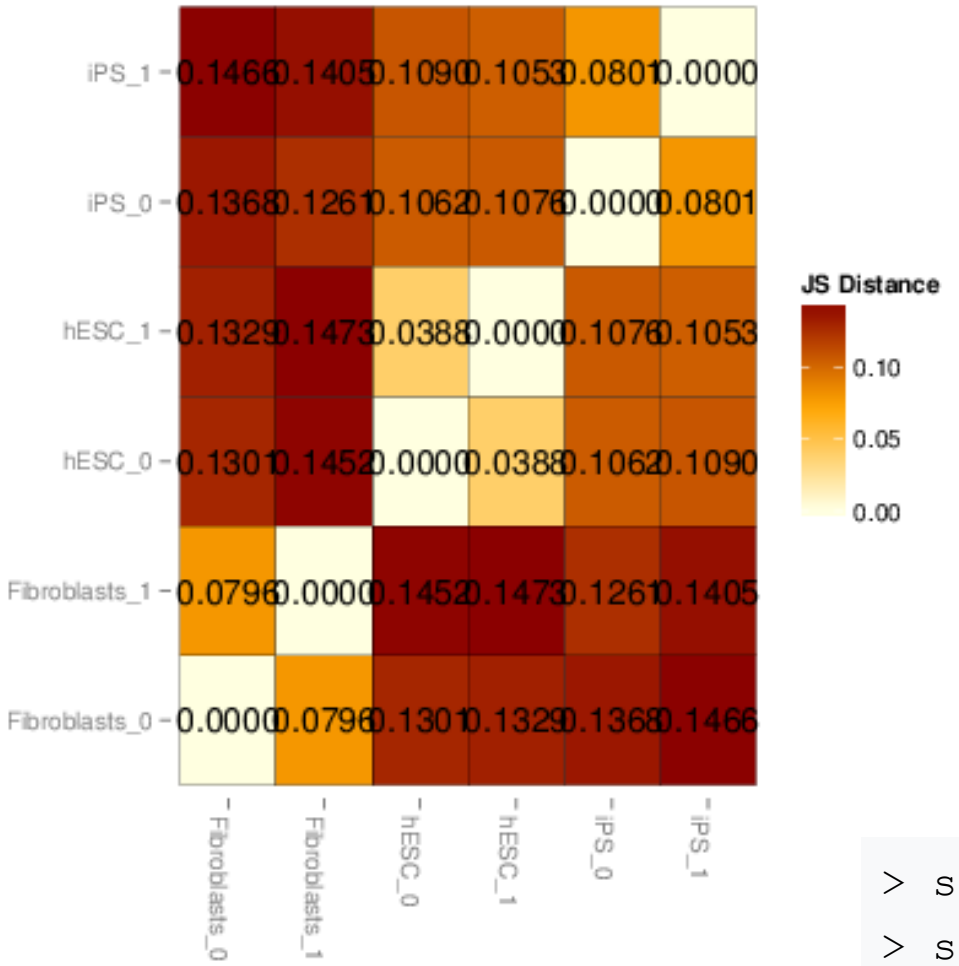


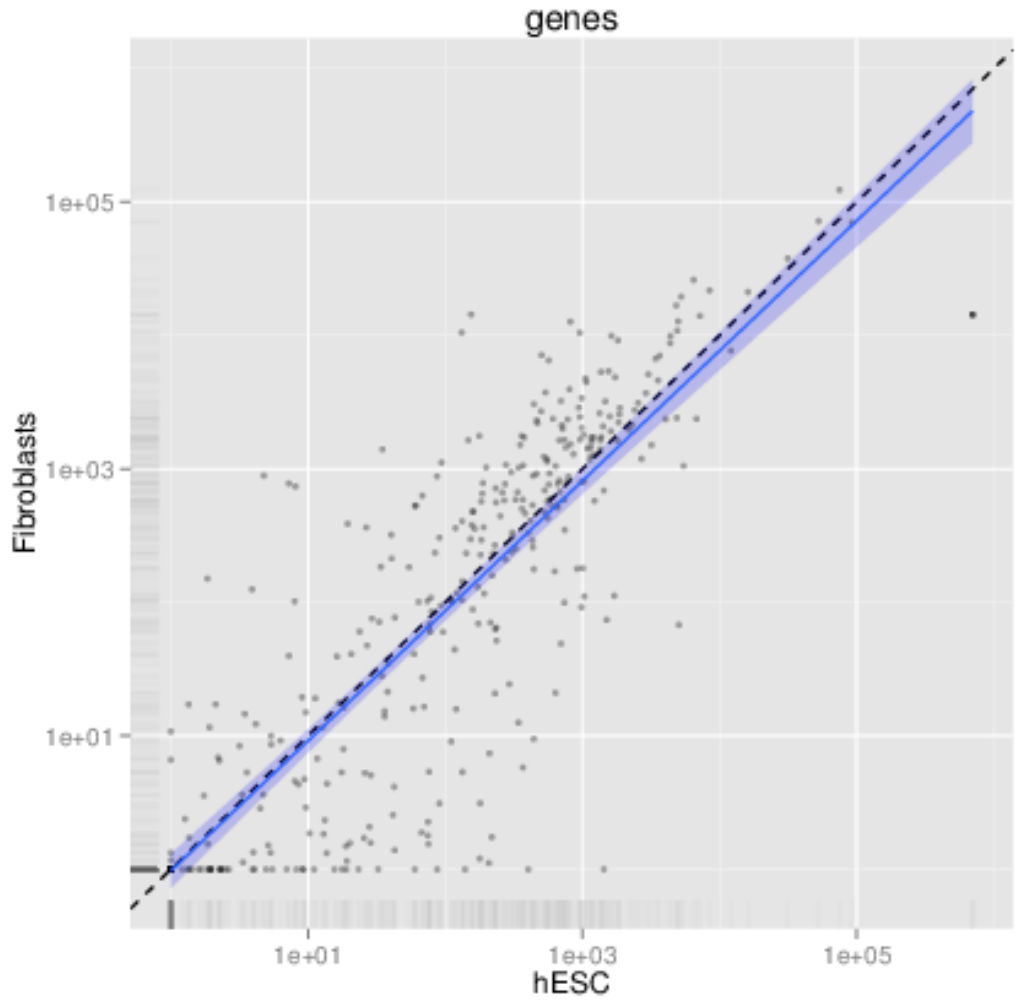


目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

样本间相关性分析（cummeRbund）

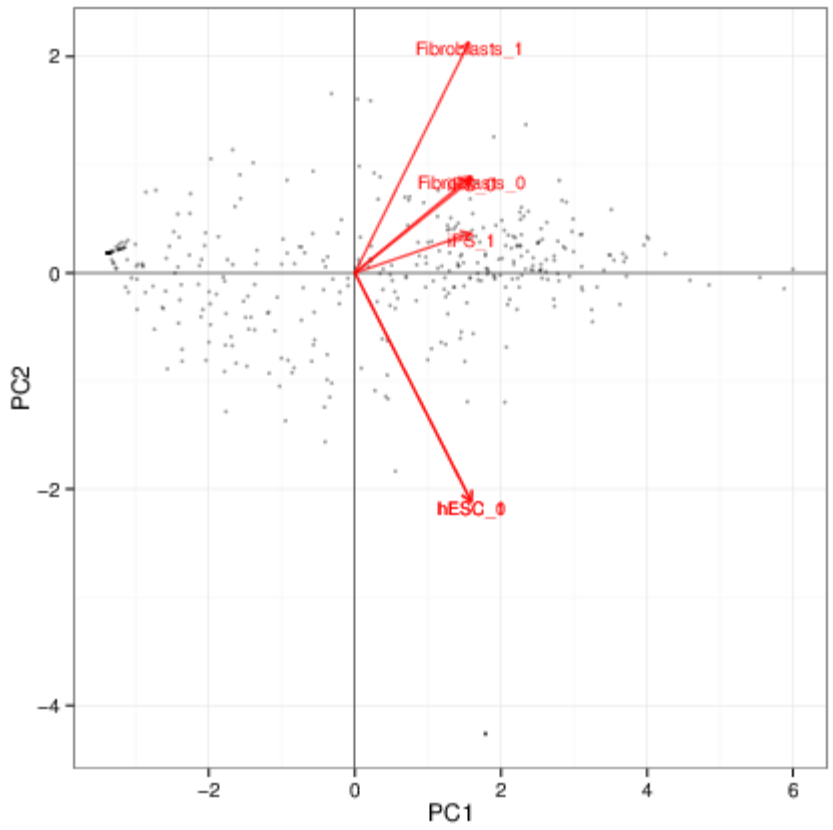


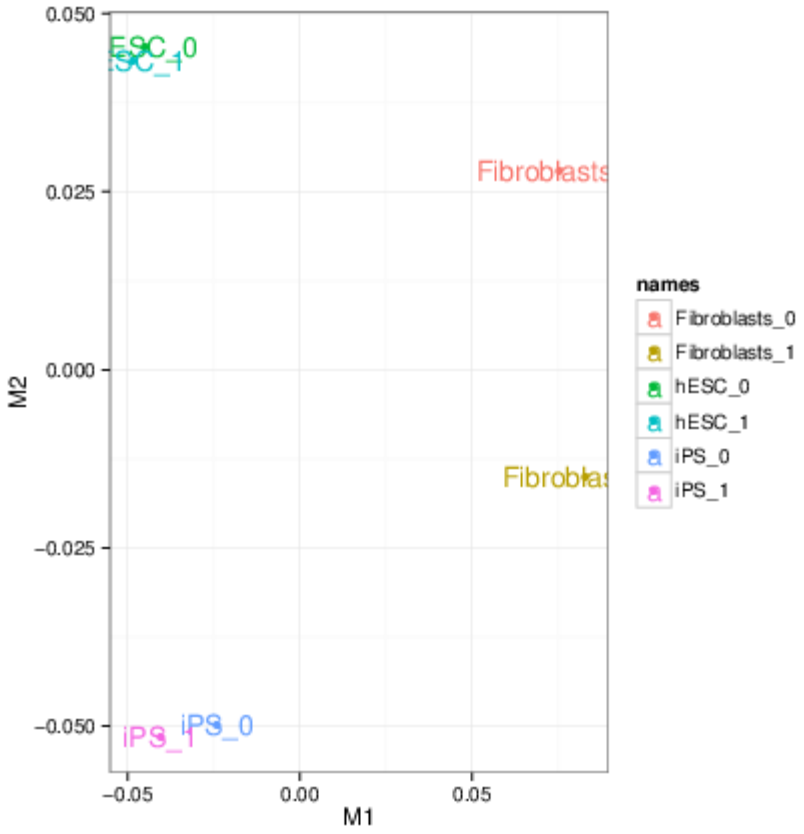



```
> s<-csScatter(genes(cuff),"hESC","Fibroblasts",smooth=T)
> s
```

```
> myRepDistHeat<-csDistHeat(genes(cuff),replicates=T)
> myRepDistHeat
```

样本间相关性分析（cummeRbund）

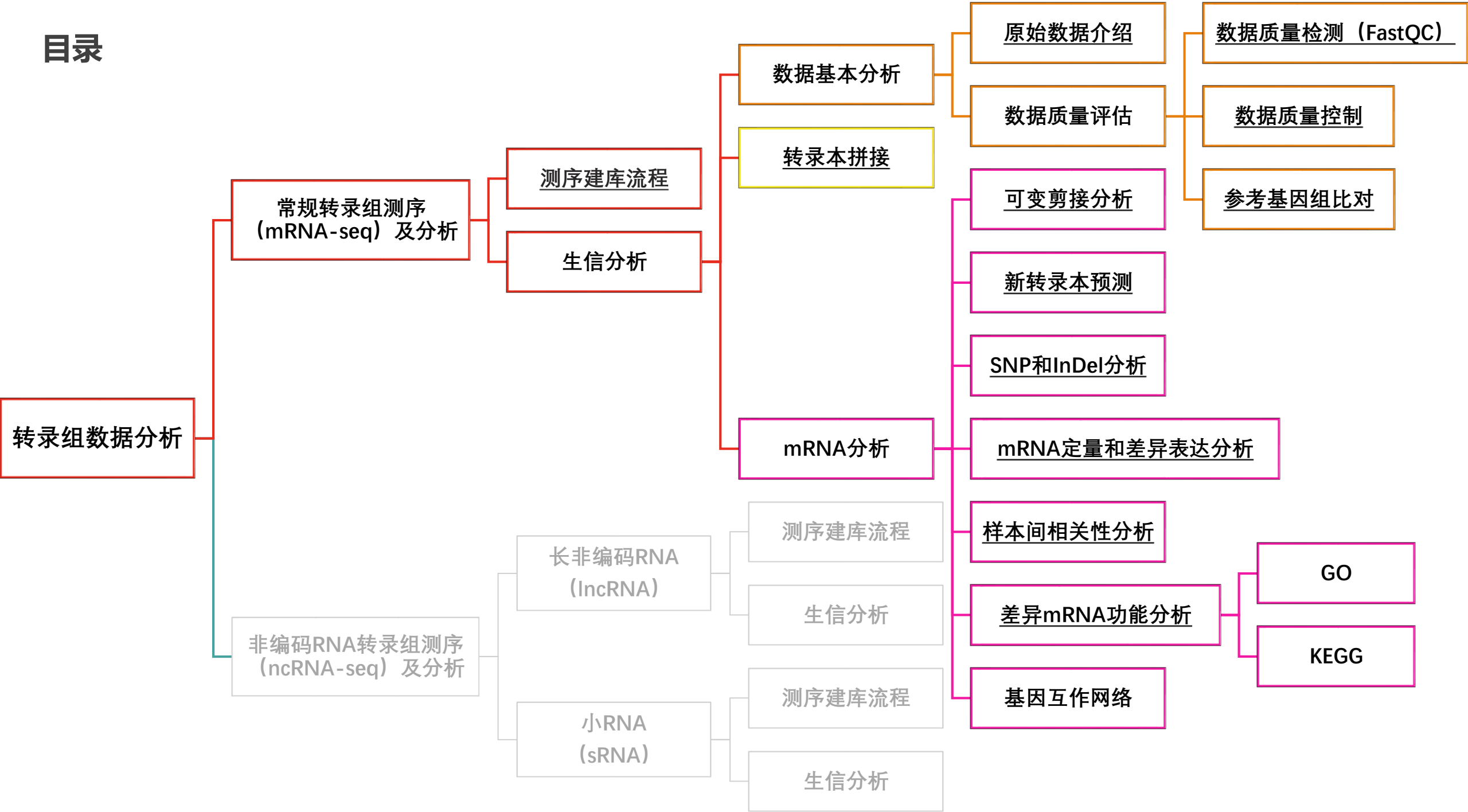





```
> genes.PCA.rep<-PCAplot(genes(cuff),"PC1","PC2",replicates=T)
> genes.PCA.rep
```

```
> genes.MDS.rep<-MDSplot(genes(cuff),replicates=T)
> genes.MDS.rep
```

目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

功能分析——GO

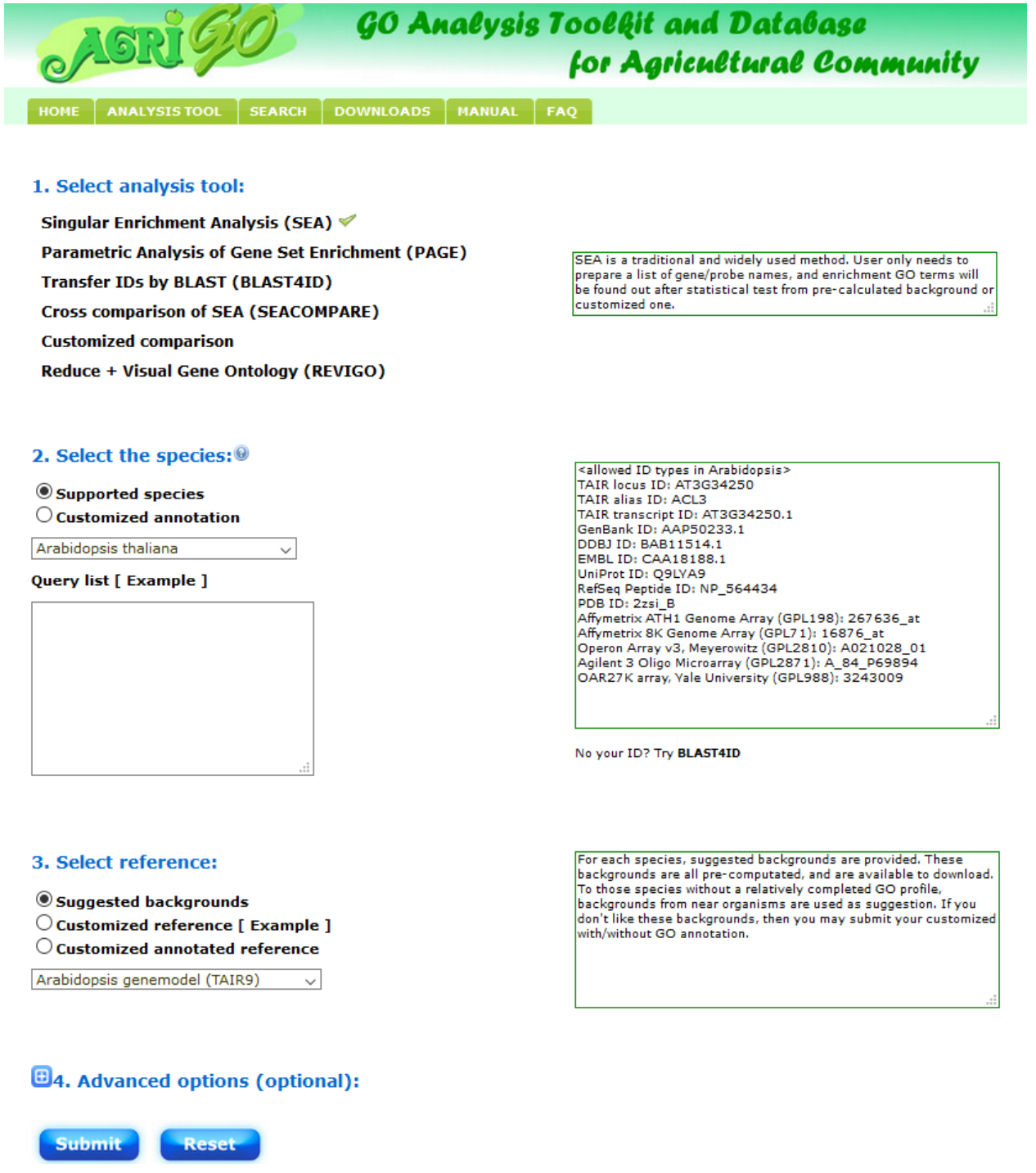
GO（Gene Ontology，基因本体论）数据库将基因功能分为三大类：

分子功能 (Molecular Function)、

生物学过程 (Biological Process)、

细胞组分 (Cellular Component)。

在数据分析中，研究者可以找出哪些变化基因属于一个共同的**GO**功能分支，并用统计学方法检定结果是否具有统计学意义，从而得出一组基因主要参与了哪些生物功能。



GO功能分类——AGRIGO结果

Analysis Brief Summary

Job ID: 286731400 [Useful within 7 days]
Job Name: 286731400
Species: *Arabidopsis thaliana*
GO type: Completed GO
Background/Reference: Affymetrix ATH1 Genome Array (GPL198)
Annotated number in query list: 168 [Download]
Annotated number in background/reference: 22479
Significant GO terms: 20 [Details]

Graphical Results

Select Category

◉ Biological Process ○ Cellular Component ○ Mollecular Function

Advanced Parameter Settings

Graphic result format: ◉ PNG ○ PDF ○ JPEG ○ GIF ○ SVG

Graph rank direction: ◉ Top to Bottom ○ Left to Right ○ Bottom to Top ○ Right to Left

Graph font size (pt): ○ 7 ○ 8 ○ 9 ◉ 10 ○ 11 ○ 12

Generate Image

GO flash Chart

Select Category

☑ Biological Process ☐ Cellular Component ☐ Mollecular Function

Advanced Parameter Settings

Bar style: ◉ Glass Bar ○ Filled Bar ○ 3D Bar ○ Cylinder Bar

Query bar color: #1A8CFF Bg/Ref bar color: #66FF33 [HEX format only] [default]

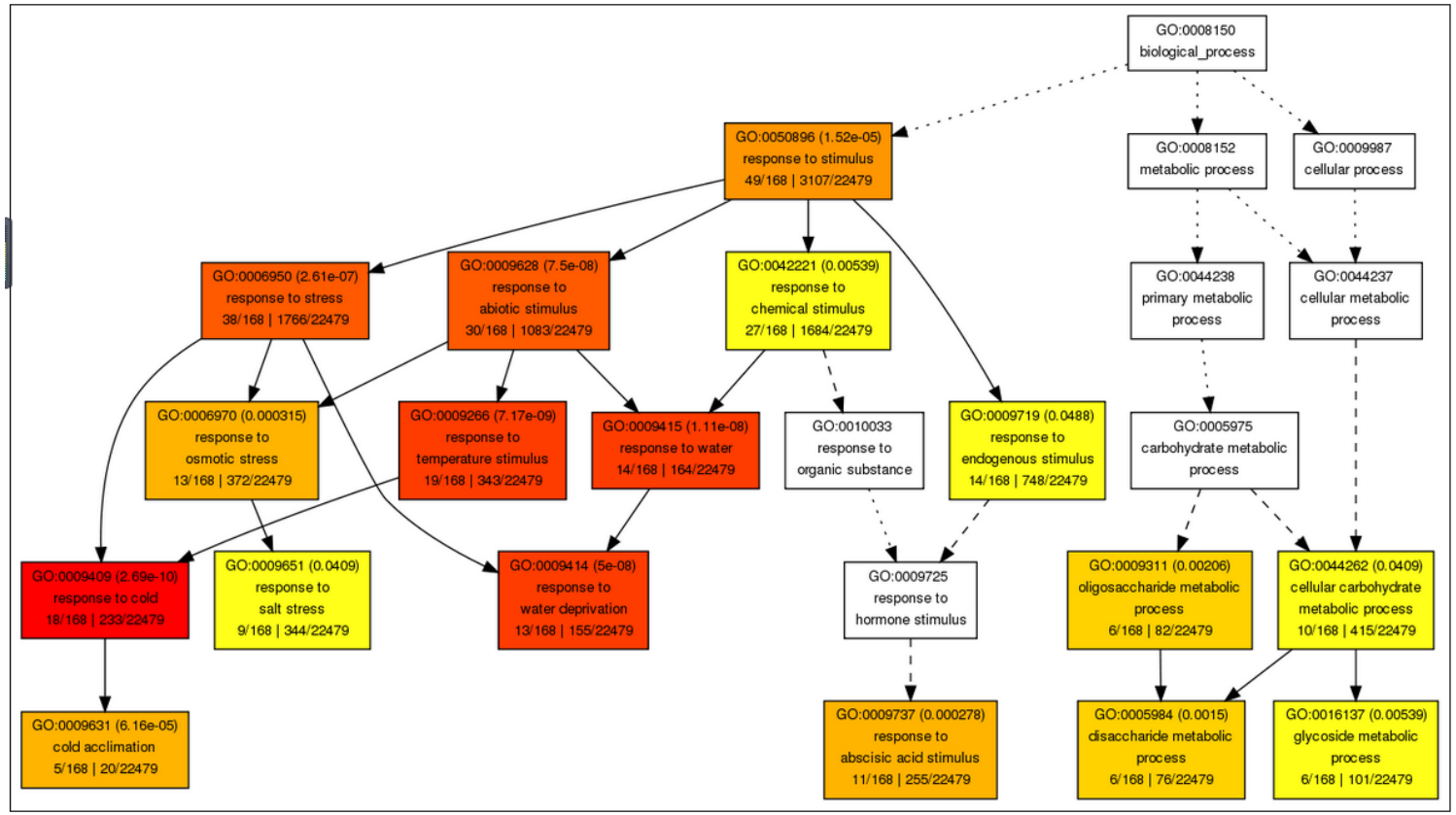
X legend content: ◉ GO annotation ○ GO accession font: 14

X legend rotation: 300 [270 to 315 is suggested]

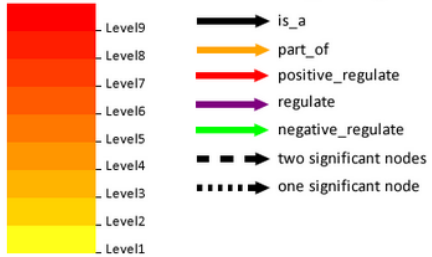
Generate Chart

Detail information

You can [Browse in tree traversing mode] [Browse all GO terms] [Download] [REVIGO]
Or select from following significant terms to [Draw graphical results] [Create bar chart]



Significance levels and Arrow types Diagram

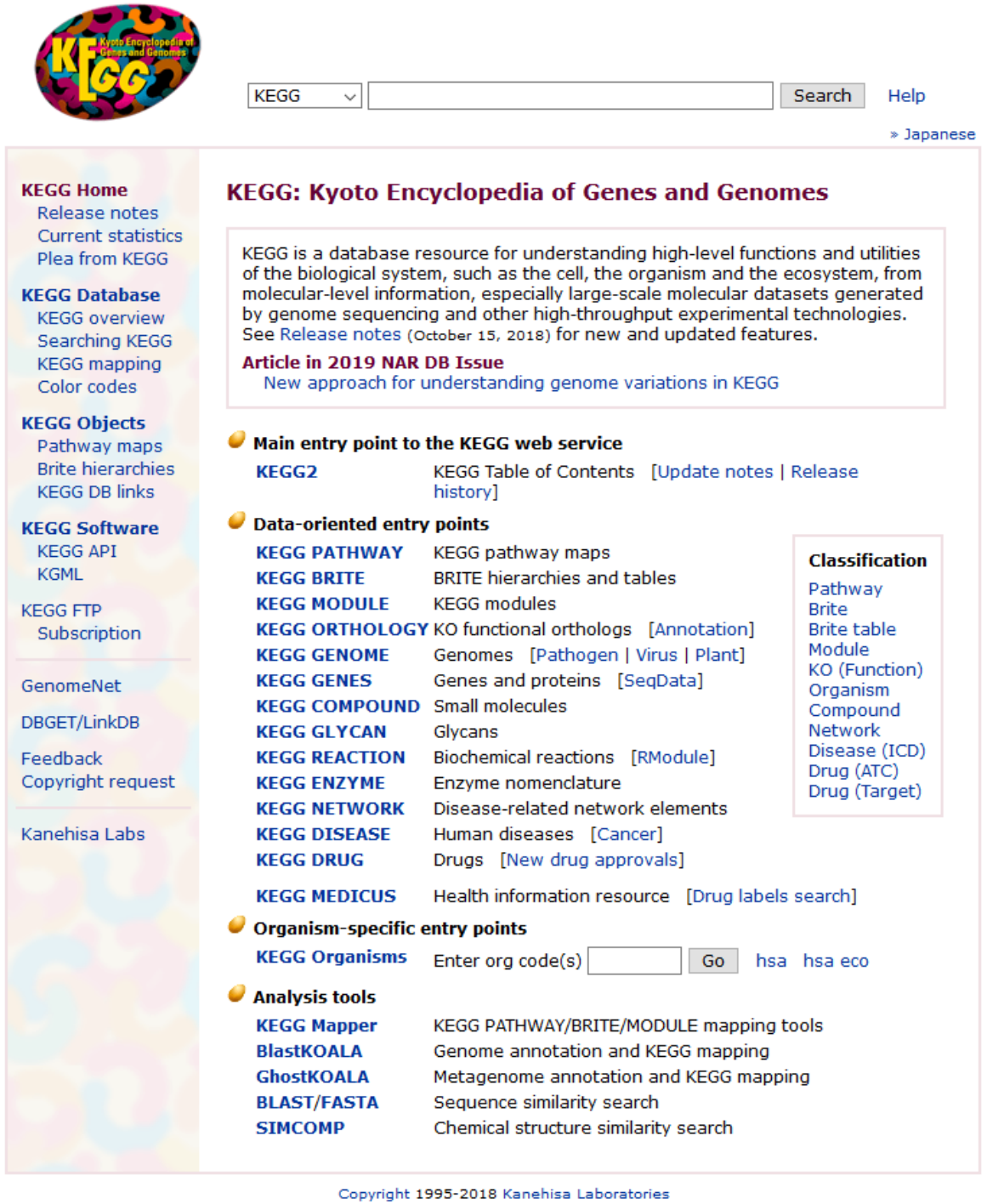


is_a
part_of
positive_regulate
regulate
negative_regulate
two significant nodes
one significant node

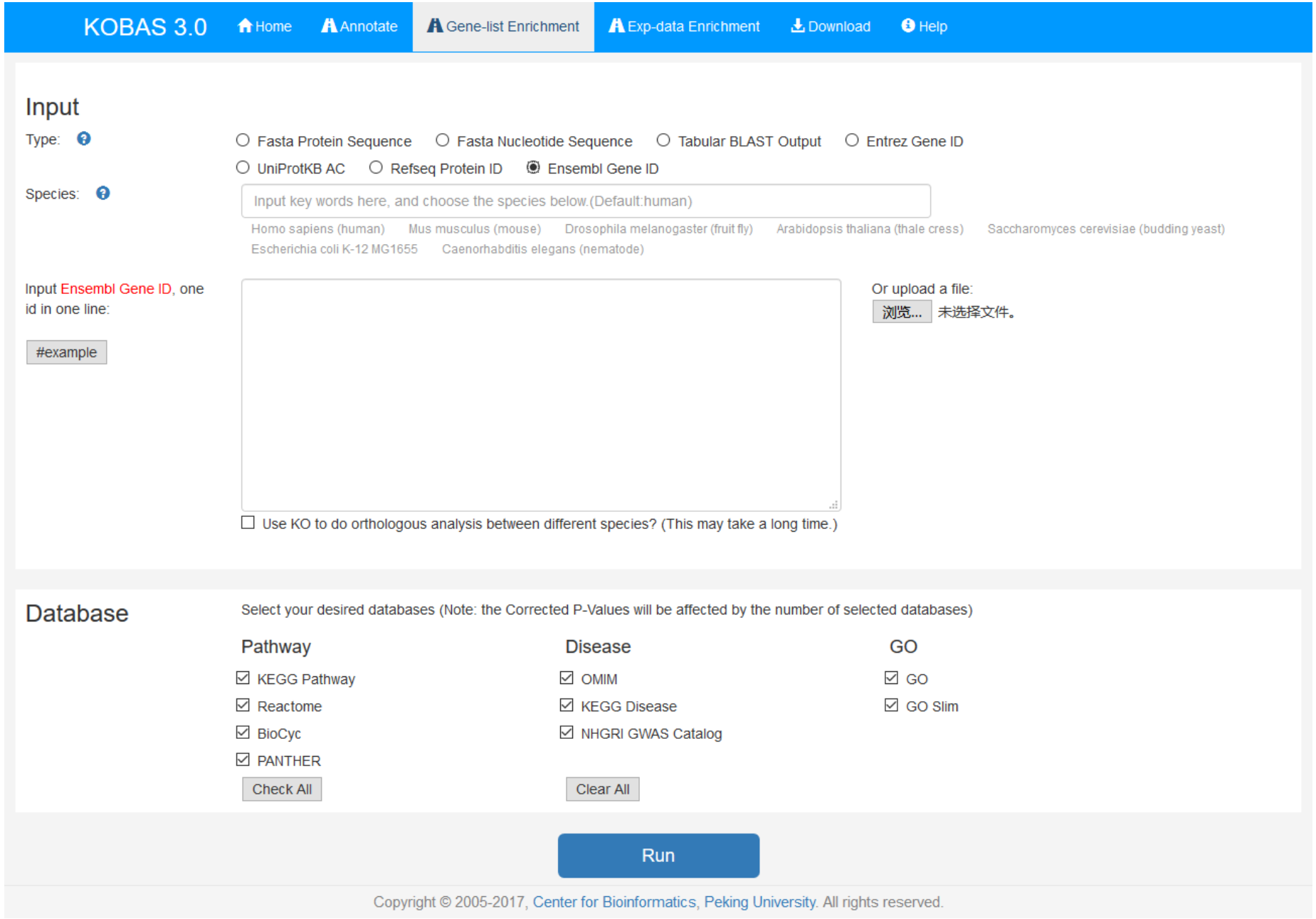
功能分析——KEGG

全称京都基因与基因组百科全书。它是关于基因、蛋白质、生化反应以及通路的综合生物信息数据库。由多个子库构成。

这些子库中，KEGG PATHWAY 数据库包含了大量物种的代谢与生物信号传导通路信息。Pathway 数据库下又分为7 个部分：1）Metabolism，2）Genetic Information Processing，3）Environmental Information Processing，4）Cellular Processes，5）Organismal Systems，6）Human Diseases，7）Drug Development。

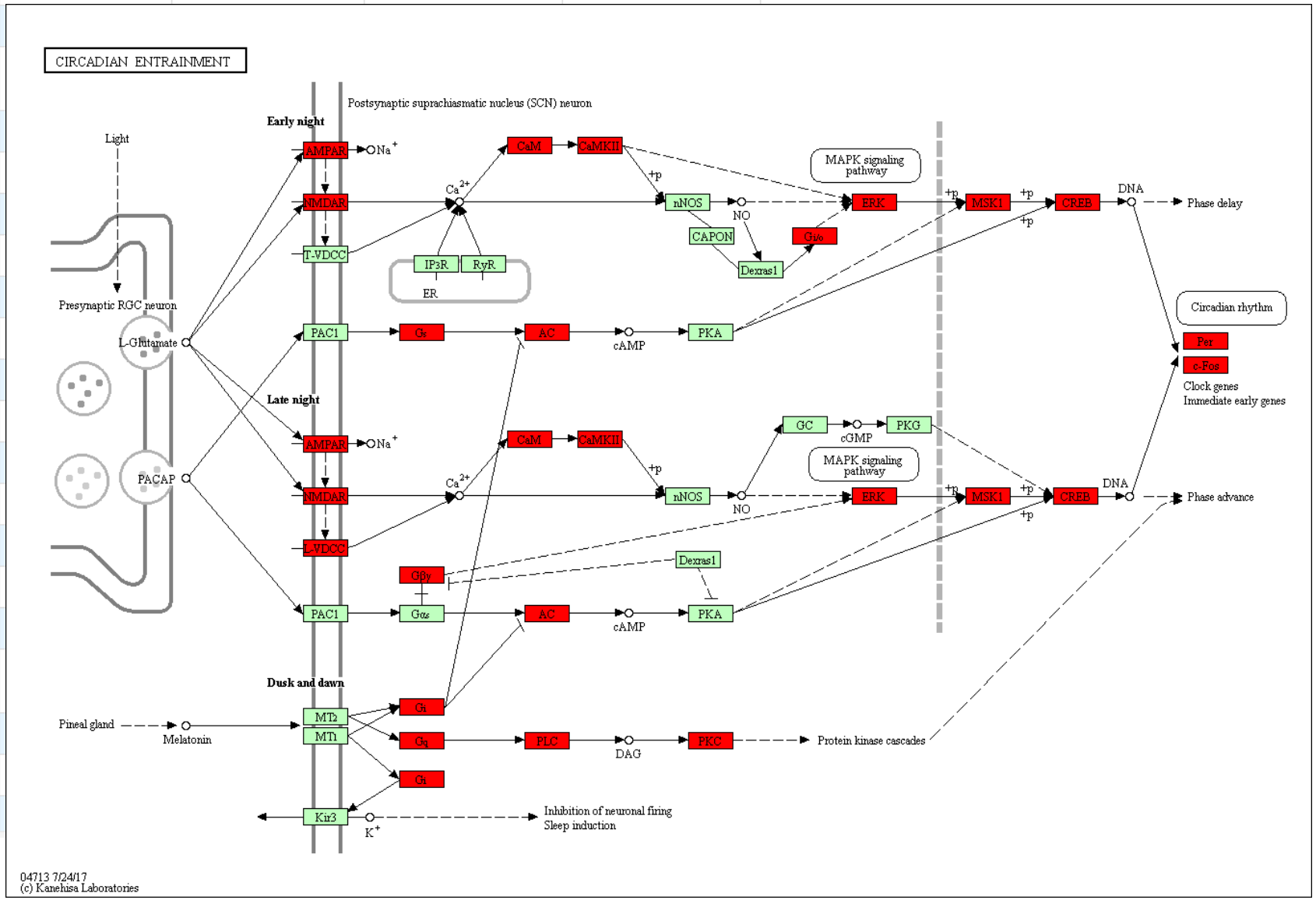


KEGG分析——KOBAS界面



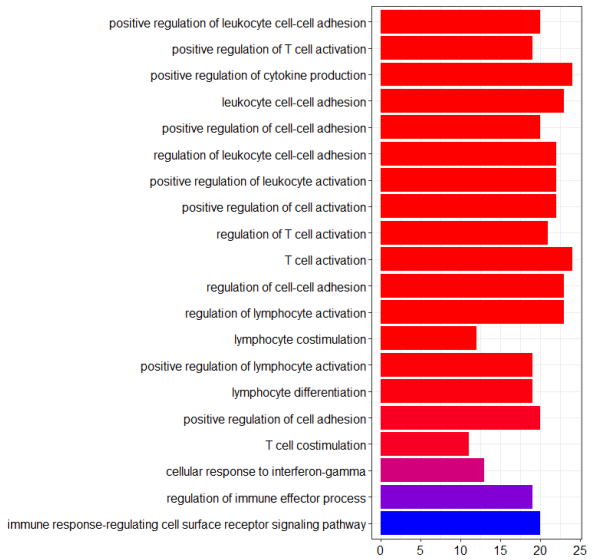
KEGG分析——KOBAS结果

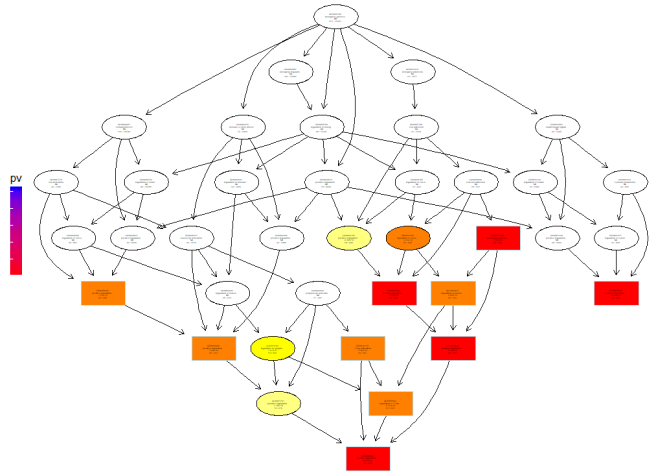
Term	Database	ID	Input number	Background number	P-Value	Corrected P-Value
Circadian entrainment	KEGG PATHWAY	hsa04713	23	95	1.27808899232e-19	3.15687981103e-17
Dopaminergic synapse	KEGG PATHWAY	hsa04728				
Retrograde endocannabinoid signaling	KEGG PATHWAY	hsa04723				
Pathways in cancer	KEGG PATHWAY	hsa05200				
Amphetamine addiction	KEGG PATHWAY	hsa05031				
Estrogen signaling pathway	KEGG PATHWAY	hsa04915				
cAMP signaling pathway	KEGG PATHWAY	hsa04024				
Serotonergic synapse	KEGG PATHWAY	hsa04726				
Glutamatergic synapse	KEGG PATHWAY	hsa04724				
MAPK signaling pathway	KEGG PATHWAY	hsa04010				
GABAergic synapse	KEGG PATHWAY	hsa04727				
Gap junction	KEGG PATHWAY	hsa04540				
Alzheimer's disease	KEGG PATHWAY	hsa05010				
Cholinergic synapse	KEGG PATHWAY	hsa04725				
HIF-1 signaling pathway	KEGG PATHWAY	hsa04066				
Cocaine addiction	KEGG PATHWAY	hsa05030				
Neuroactive ligand-receptor interaction	KEGG PATHWAY	hsa04080				
HTLV-I infection	KEGG PATHWAY	hsa05166				
Alcoholism	KEGG PATHWAY	hsa05034				
Oxytocin signaling pathway	KEGG PATHWAY	hsa04921				

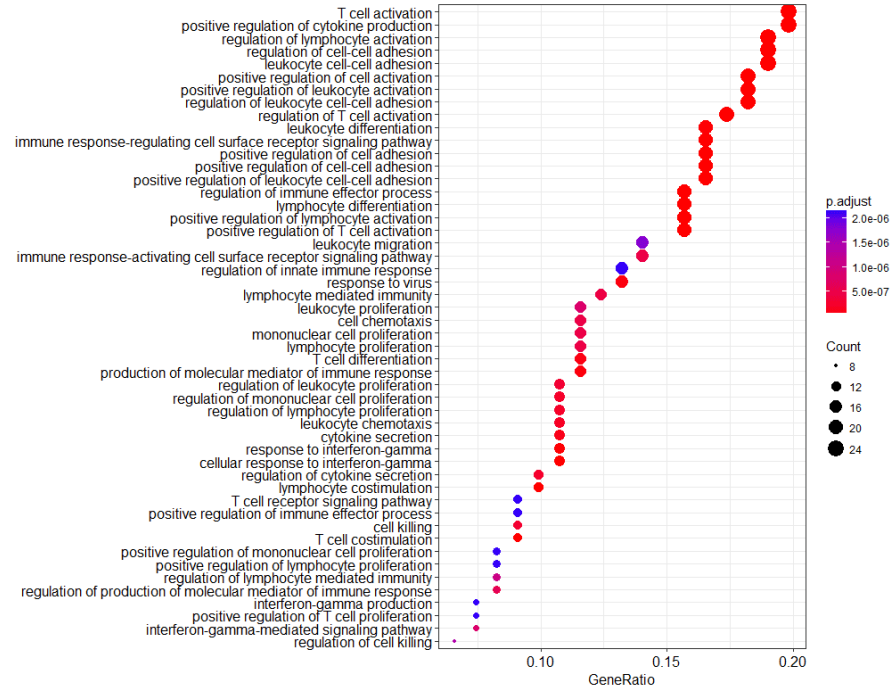


clusterProfiler——GO分析

```
enrichGO Usage:
  enrichGO(gene, OrgDb, keyType = "ENTREZID", ont =
"MF", pvalueCutoff = 0.05, pAdjustMethod = "BH",
universe, qvalueCutoff = 0.2, minGSSize = 10,
maxGSSize = 500, readable = FALSE, pool = FALSE)

> go <- enrichGO(genelist, OrgDb = org.Hs.eg.db,
ont='ALL',pAdjustMethod = 'BH',pvalueCutoff = 0.05,
qvalueCutoff = 0.2,keyType = 'ENTREZID’)

# 可视化
> barplot(go,showCategory=20,drop=T)
> dotplot(go,showCategory=50)

# 绘制GO的网络关系图，注意只能是富集一个GO通路（BP、CC或MF
）的数据
> go.BP <- enrichGO(genelist, OrgDb = org.Hs.eg.db,
ont='BP',pAdjustMethod = 'BH',pvalueCutoff = 0.05,
qvalueCutoff = 0.2,keyType = 'ENTREZID')
> plotGOgraph(go.BP)
```

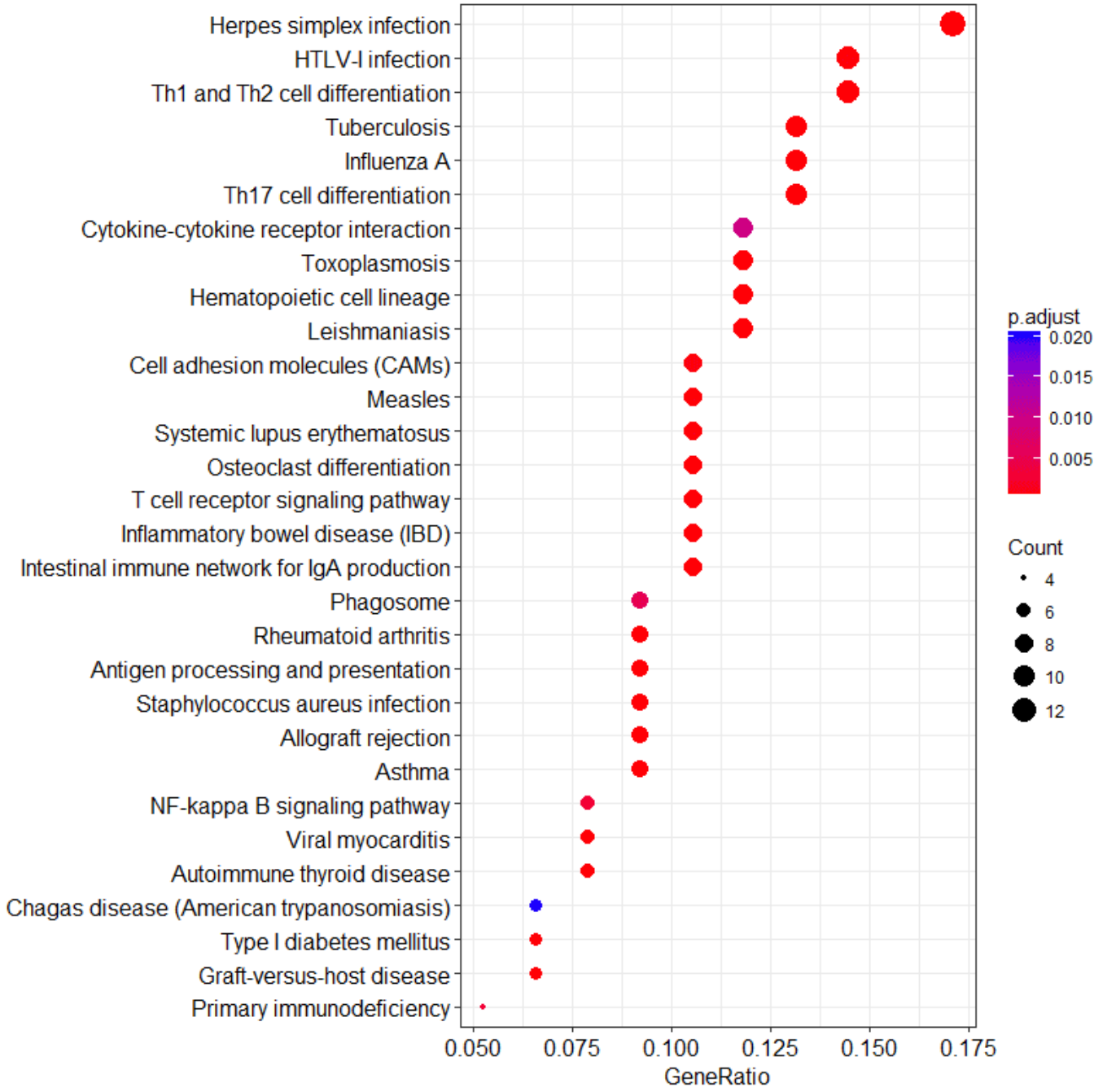




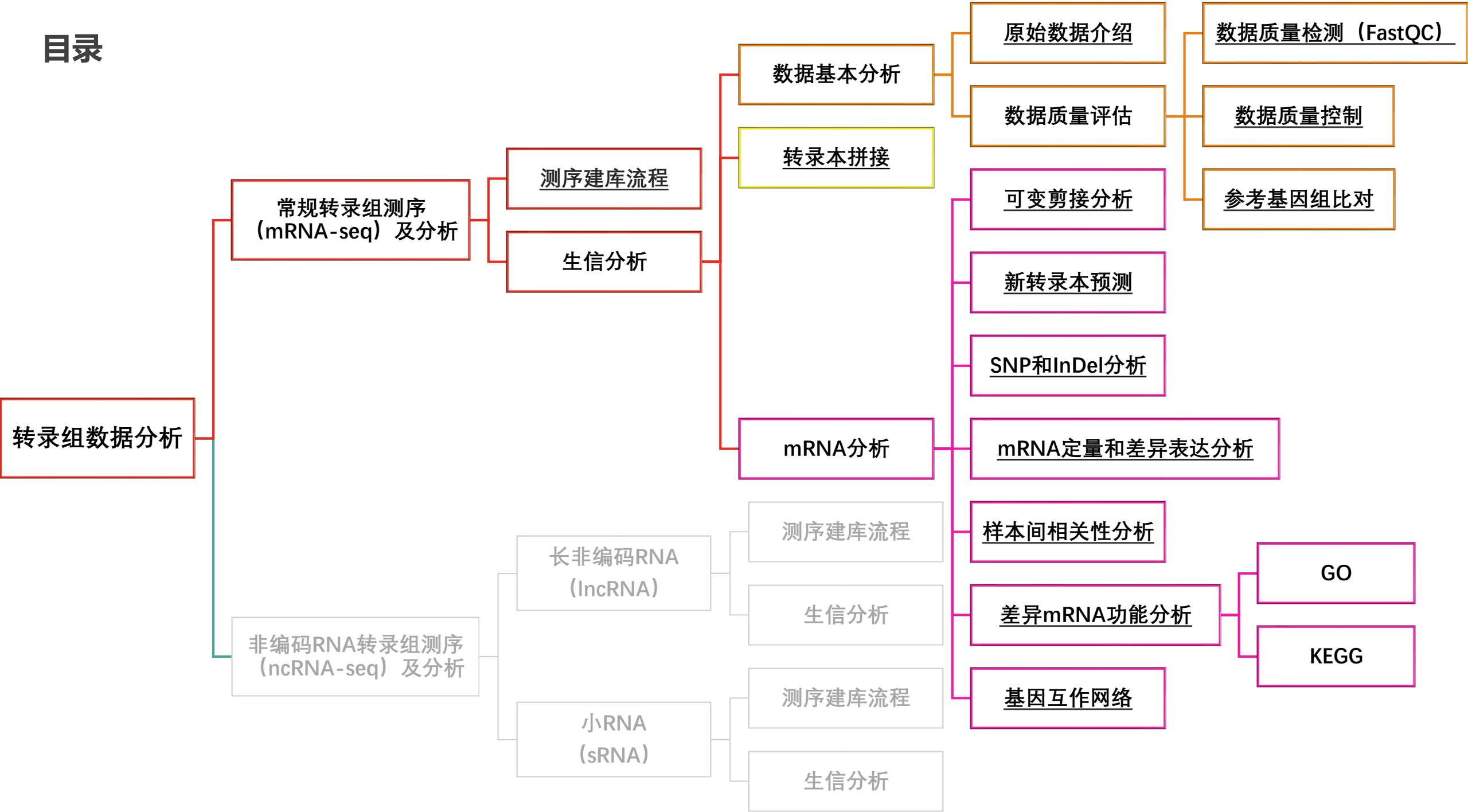


clusterProfiler——KEGG分析

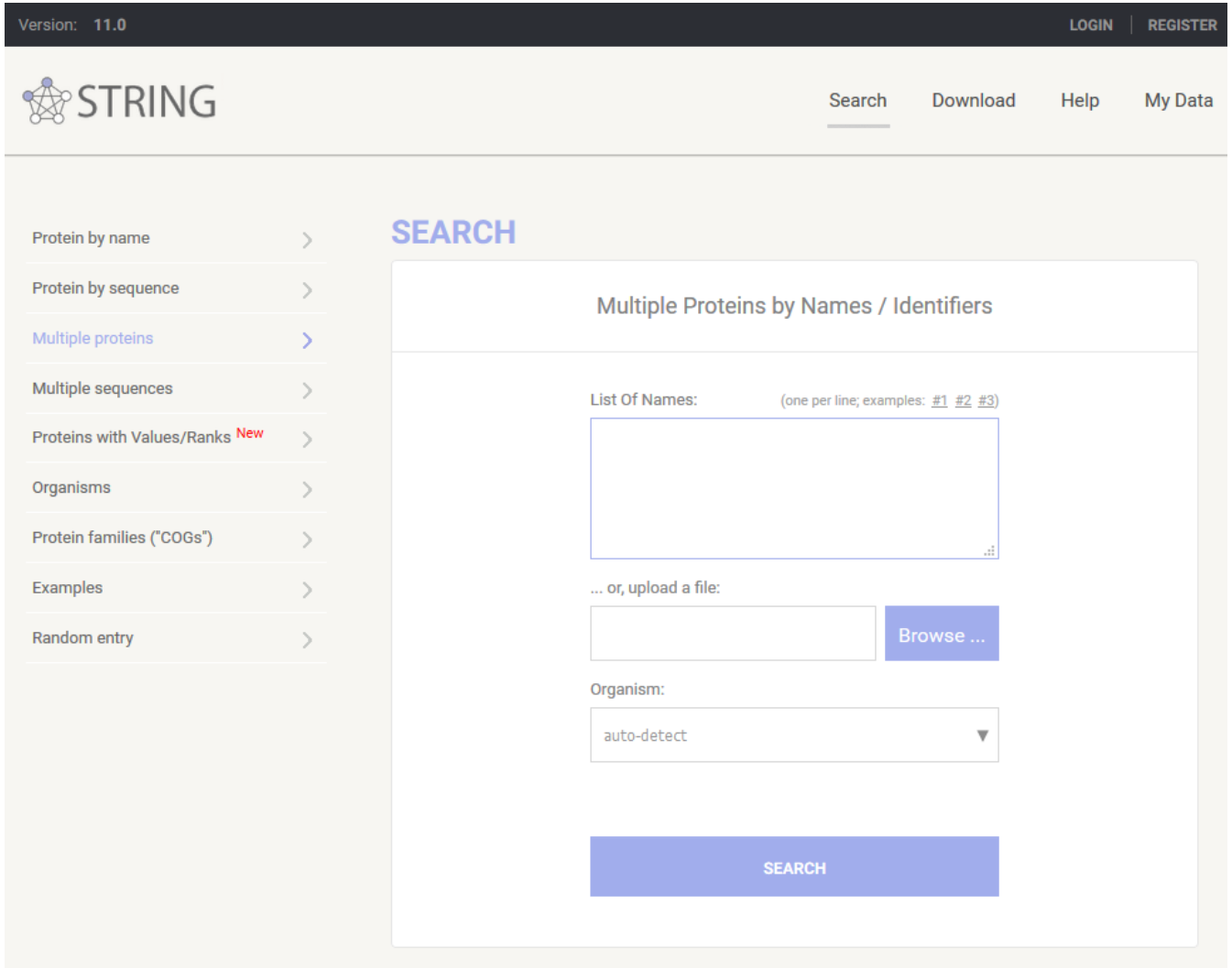
```
enrichKEGG Usage:
  enrichKEGG(gene, organism = "hsa",
keyType = "kegg", pvalueCutoff =
0.05,pAdjustMethod = "BH", universe,
minGSSize = 10, maxGSSize = 500,
qvalueCutoff = 0.2, use_internal_data =
FALSE)

> kegg <- enrichKEGG(genelist, organism =
'hsa', keyType = 'kegg', pvalueCutoff =
0.05,pAdjustMethod = 'BH', minGSSize =
10,maxGSSize = 500,qvalueCutoff =
0.2,use_internal_data = FALSE)

# 可视化
> dotplot(kegg, showCategory=30)
```



目录


转录组数据分析
常规转录组测序（mRNA-seq）及分析
测序建库流程
生信分析
数据基本分析
原始数据介绍
数据质量检测（FastQC）
数据质量评估
数据质量控制
参考基因组比对
转录本拼接
mRNA分析
可变剪接分析
新转录本预测
SNP和InDel分析
mRNA定量和差异表达分析
样本间相关性分析
差异mRNA功能分析
GO
KEGG
基因互作网络
非编码RNA转录组测序（ncRNA-seq）及分析
长非编码RNA（lncRNA）
测序建库流程
生信分析
小RNA（sRNA）
测序建库流程
生信分析

基因互作网络——STRING



基因互作网络——STRING

Multiple Proteins by Names / Identifiers

List Of Names: (one per line; examples: #1 #2 #3)

smoothened
patched
hedgehog
cubitus interruptus

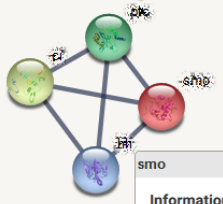
... or, upload a file:

Browse ...

Organism:

Drosophila melanogaster

SEARCH



Viewers | Legend | Settings | Analysis

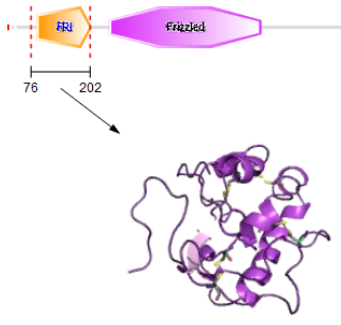
smo

Information
Protein smoothened; Segment polarity protein required for correct patterning of every segment. G protein-coupled receptor that associates with the patched protein (ptc) to transduce the hedgehog (hh) signal through the activation of an inhibitory G-protein. In the absence of hh, ptc represses the constitutive signaling activity of smo through fused (fu). Essential component of a hh-signaling pathway which regulates the Duox-dependent gut immune response to bacterial uracil; required to activate Cad99C-dependent endosome formation, norpA-dependent Ca2+ mobilization and p38 MAPK, which a [...]

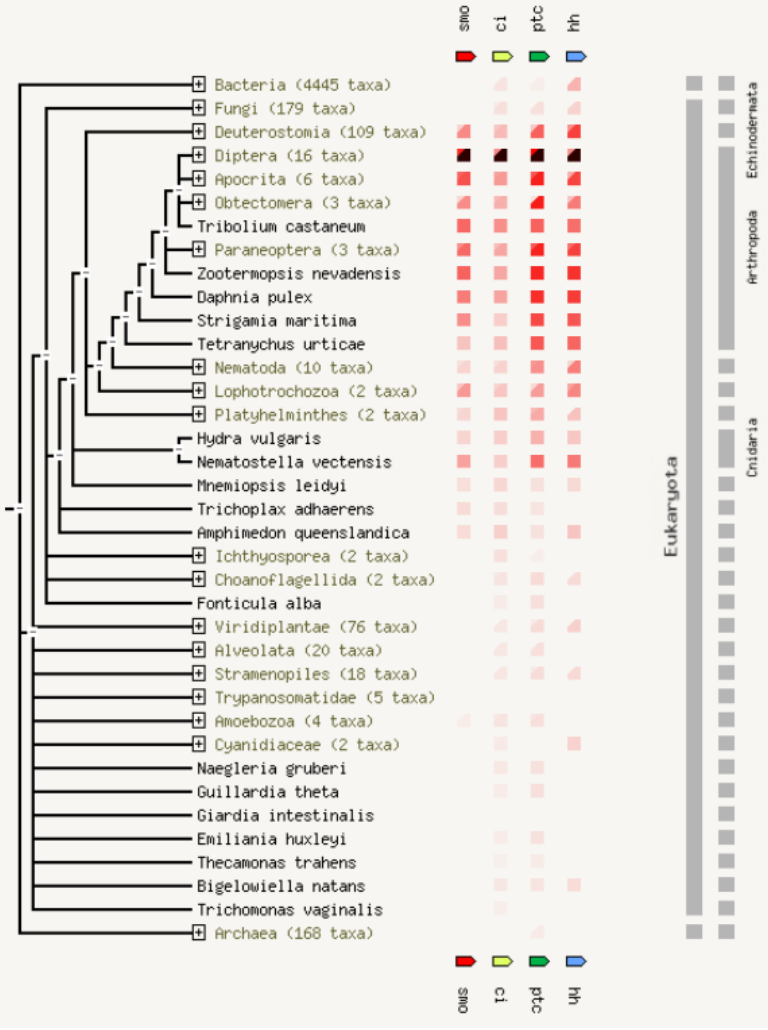
Identifier: FBpp0077788, smo
Organism: Drosophila melanogaster

FlyBase RefSeq UniProt

Actions
- re-center network on this node
- remove this node from input nodes
- show protein sequence
- homologs among STRING organisms



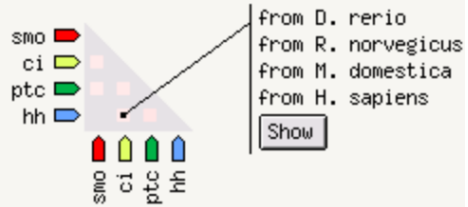
GENE COOCCURRENCE



GENE COEXPRESSION

observed Coexpression in Drosophila melanogaster:

observed Coexpression in other organisms (transferred):



[click on the heatmap elements for details]

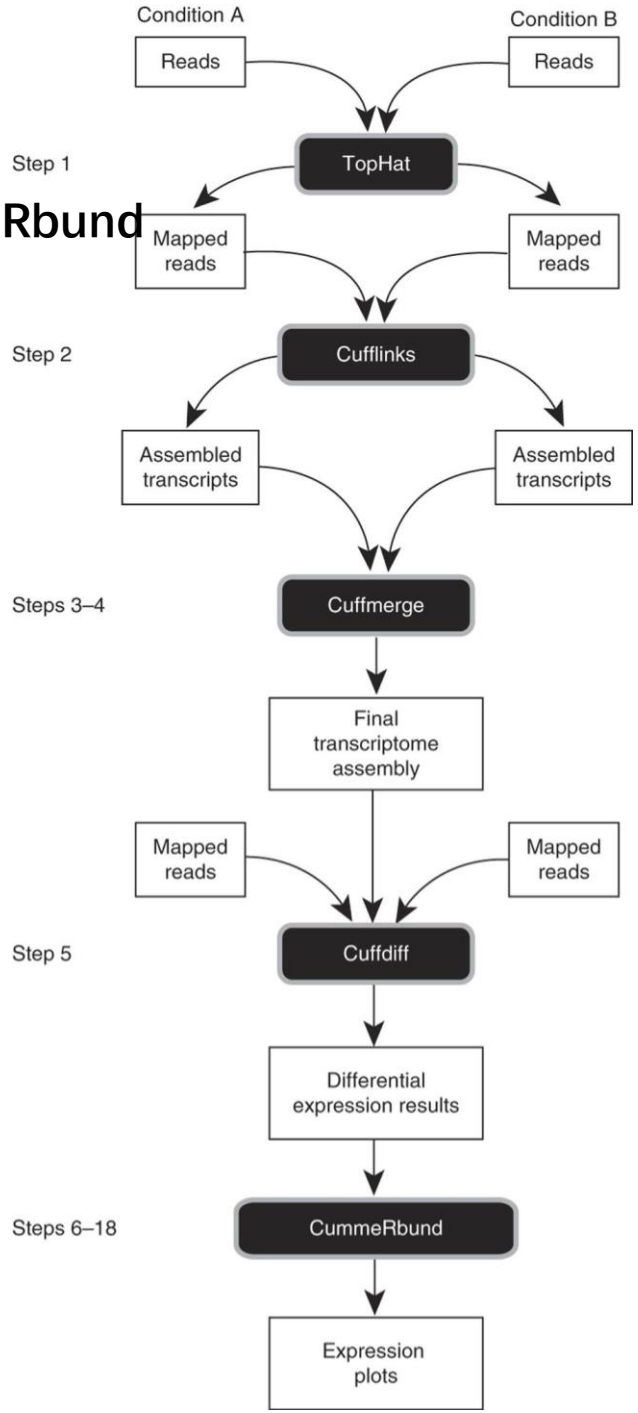
转录组数据分析小结

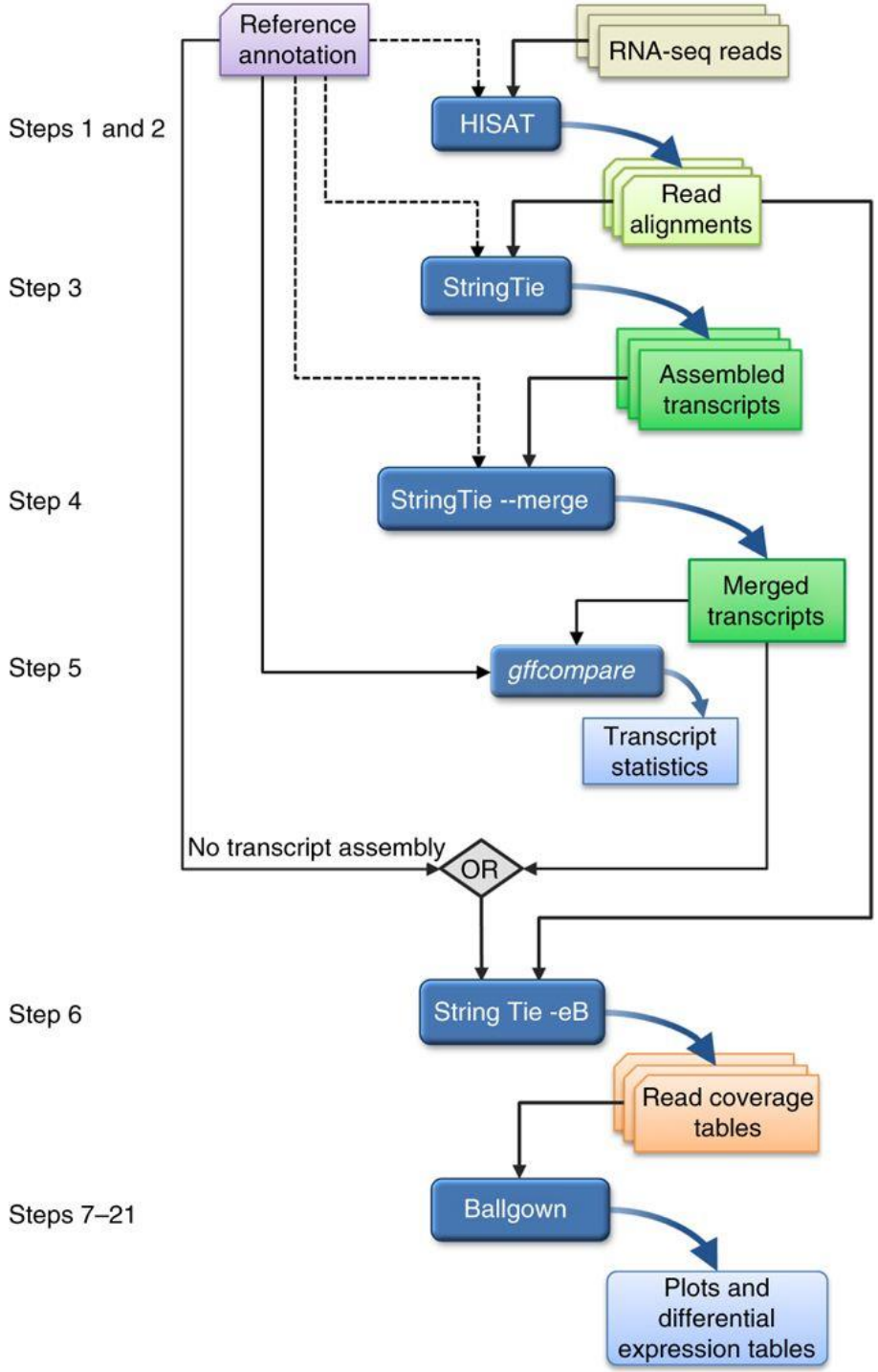
- Tophat、Cufflinks、Cuffdiff、CummeRbund
- HISAT、StringTie、Ballgown

HISAT（Hierarchical Indexing for Spliced Alignment of Transcripts）该软件和 tophat 开发于同一个单位，它相当于 tophat 的升级版，比对速度比 tophat2 快50倍。

StringTie相对于其他拼接软件（Cufflinks, IsoLasso, Scripture，Traph等），能够拼接出更完整、更准确的基因，并且StringTie采用拼接和定量同步进行，相对于其他方法，其定量结果更加准确。

Ballgown是一个R脚本，用来分析差异结果的，差异分布的结果是基于F-test来检测的；就测试内存而言，cuffdiff用148G6 9个小时，而ballgown用18秒8G，就差异结果而言，cuffdiff更保守一些。





谢谢！