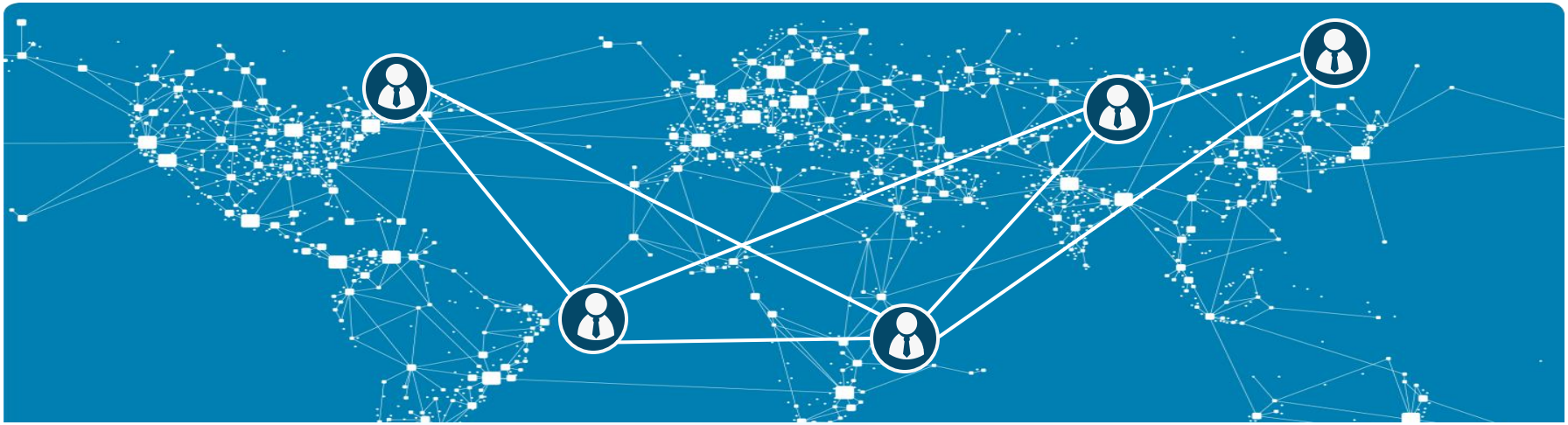
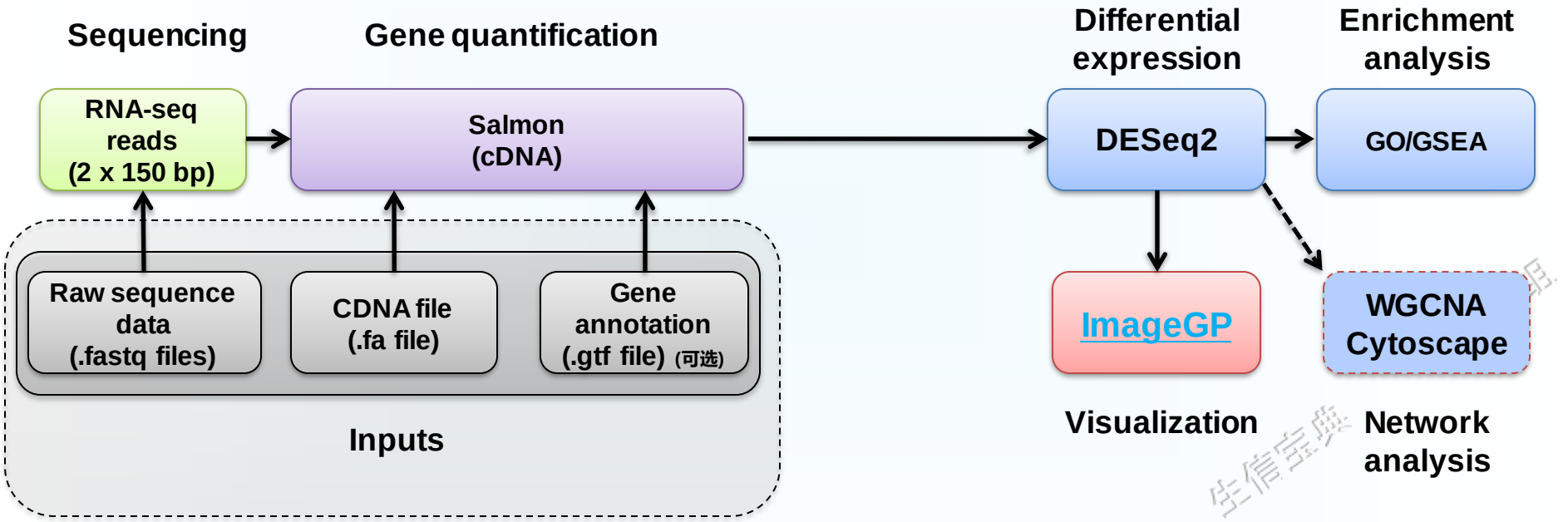
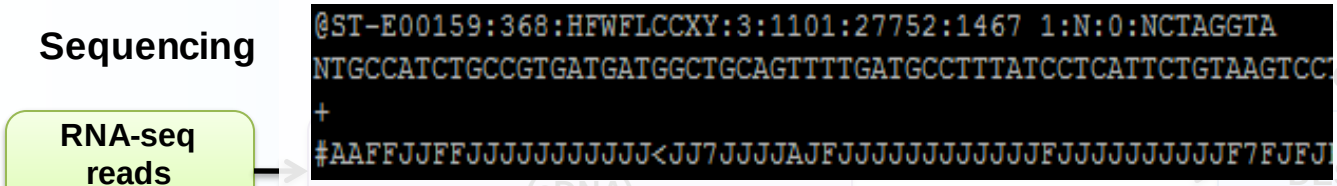


12 转录组有参分析Salmon流程

有参不比对转录组分析流程



FASTQ格式解释 – 四行一组



Raw sequence data (.fastq files)

FASTQ quality scores = ASCII quality score - either 33 or 64

Illumina FASTQ quality scores

Q = -10log10(e)

e= estimated probability of the base call being wrong

Q40: 1 error in 10,000 base calls; Q30: 1 error in 1,000;

Q28: 1.6 error in 1000; Q25: 3 error in 1000; Q20: 1 error in 100

生信宝典

FASTQ质量值 – 不同编码方式的质量值展示

Sequencing

RNA-seq reads (2 x 150 bp)

Raw sequence data (.fastq files)

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                            104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

测序质量值是什么含义？怎么解读？

公共数据从NCBI SRA库下载

SRX3067795: Sample_KTN102_Blood 样品名字
1 ILLUMINA (Illumina HiSeq 4000) run: 118.6M spots, 23.7G bases, 4.4Gb downloads

Design: 100PE

Submitted by: The UT MD Anderson Cancer Center

Study: Adaptive and Acquired Evolution in Response to Chemotherapy in Triple-Negative Breast Cancer
PRJNA396019 • SRP114962 • All experiments • All runs
hide Abstract 头一次接触时，这一段也仔细看下

Triple-negative breast cancer (TNBC) is an aggressive subtype that displays extensive intratumor heterogeneity and frequently develops resistance to neoadjuvant chemotherapy (NAC). An unresolved question is whether resistance is caused by the adaptive selection of rare mutations in pre-existing subclones (adaptive resistance), or alternatively through the acquisition of new mutations induced by the therapeutic agents. To investigate this question, we applied single cell DNA and RNA sequencing, in addition to bulk deep-exome sequencing, to study matched longitudinal samples from 20 TNBC patients during NAC. Deep-exome sequencing identified 11 patients in which NAC led to clonal extinction, and 9 patients in which clones persisted during NAC and established the resistant tumor mass. Single cell DNA sequencing of 1000 cells from 8 patients showed that clones with chemoresistant copy number profiles were pre-existing and selected in response to NAC, following an adaptive resistance model. In contrast, single cell RNA sequencing of 8,500 cells in 8 patients, showed that NAC induced transcriptional reprogramming in response to NAC, including the upregulation of MTORC, interferon response, MYC, EMT and Angiogenesis signatures. Our data suggests chemoresistance evolution is a two-step process that involves both the adaptive selection of genotypes, and acquired resistance of chemoresistant phenotypes.

Sample: Sample_KTN102_Blood
SAMN07457099 • SRS2412441 • All experiments • All runs
Organism: Homo sapiens 物种信息

Library:
Name: KTN102Blood
Instrument: Illumina HiSeq 4000
Strategy: WXS 测序策略，全外显子组，还有转录组等；与Source一起看
Source: GENOMIC
Selection: RANDOM PCR
Layout: PAIRED 双端测序；如果写single，则是单端测序。有时也会有图示，指示单端还是双端

Links:

Runs: 1 run, 118.6M spots, 23.7G bases, 4.4Gb

Run	# of Spots	# of Bases	Size	Published
SRR5906250	118,619,247	23.7G	4.4Gb	2018-04-23

一般是只有一个run，这个SRR号是我们下载需要的。如果这里有多个SRR号，则代表这个样品有多个run，需要全部下载，进行后续分析。

使用NCBI提供的SRA-toolkit中的工具fastq-dump直接下载SRR文件，并转换为FASTQ格式，--split-3参数表示如果是双端测序就自动拆分，如果是单端不受影响。--gzip转换fastq为压缩文件，节省空间。

下载的数据集一般比较大，放入后台不中断下载 (nohup cmd &)。

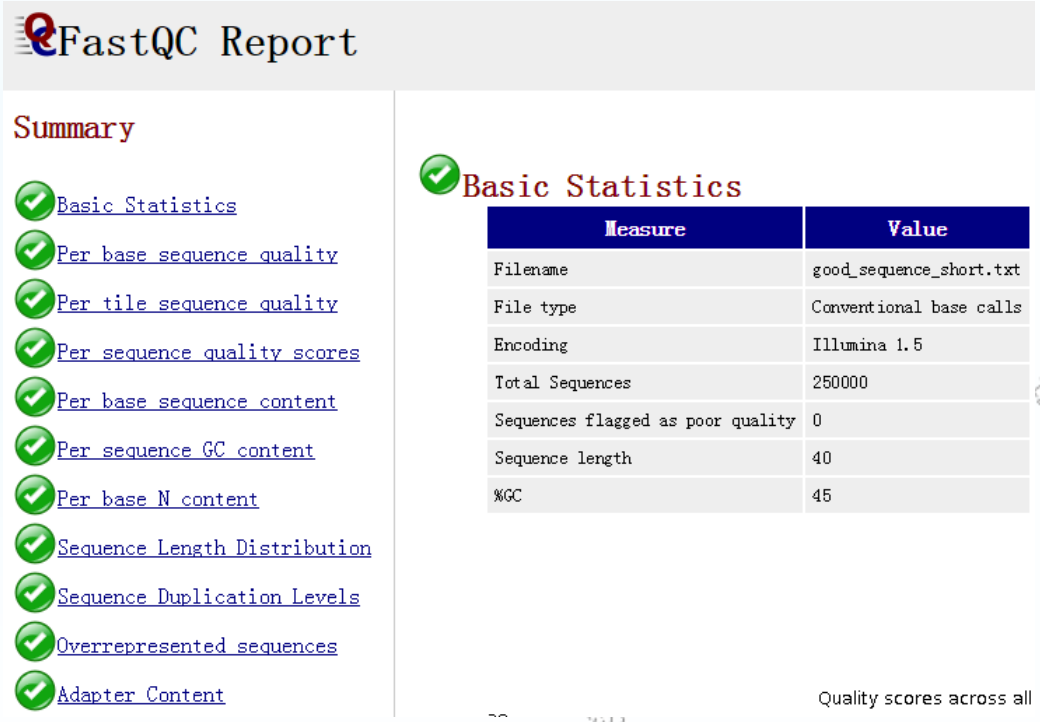
```
nohup fastq-dump -v --split-3 --gzip SRR5908360 &
nohup fastq-dump -v --split-3 --gzip SRR5908361 &
```

更详细描述和下载见：链接

测序数据质控检测

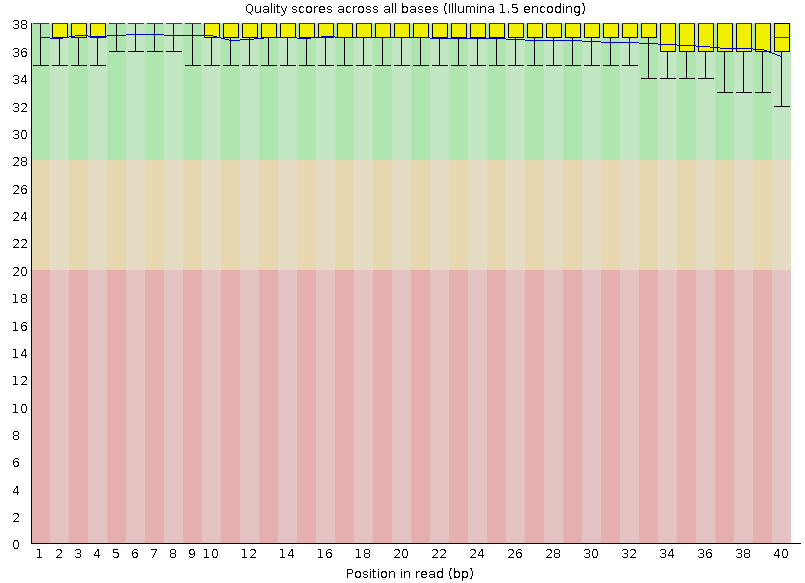
- FastQC

- fastqc sample.fq.gz

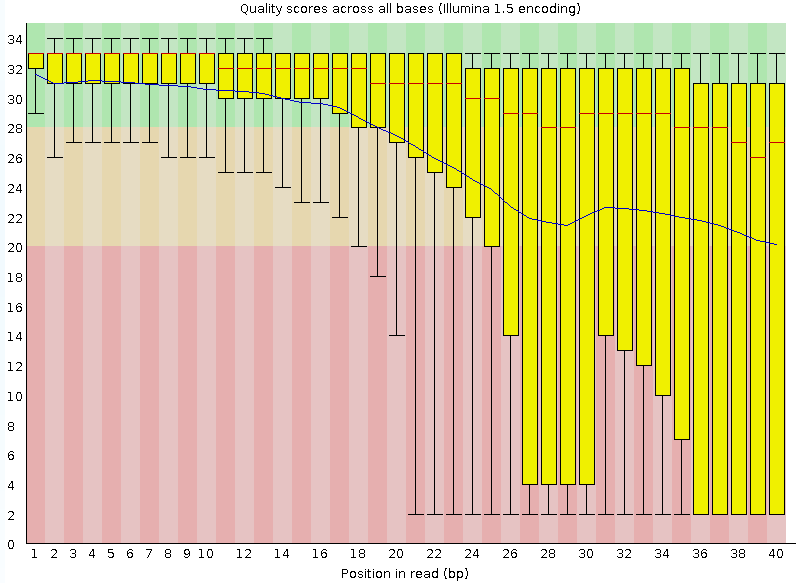


测序碱基质量检测

高的测序质量



低的测序质量



x-axis: Position in read y-axis: Quality scores

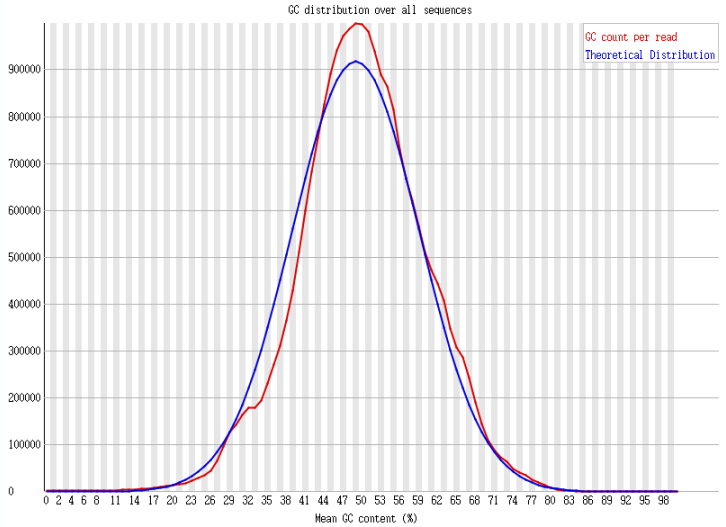
测序质量为什么会从5’-3’逐渐降低?

测序碱基质量检测图的含义

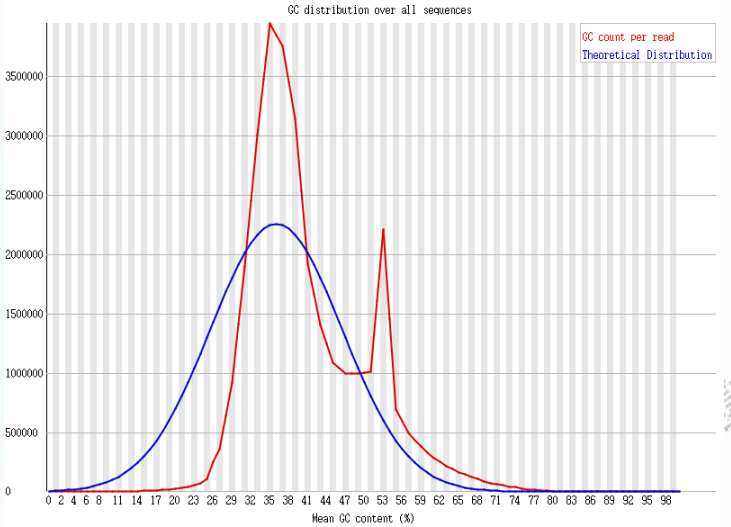
- 单个碱基测序质量的箱线图（箱线图通过最大值、上四分位数、中位数、下四分位数和最小值五处位置来获取一维数据的分布概况)。横坐标表示read中每个碱基的位置，纵坐标表示质量得分。将所有reads的第一位碱基质量得分进行箱线图展示得到第一位碱基质量得分的分布情况，以此类推，获得下图。左图显示每个碱基的中位质量得分(箱线图中间的蓝线)都比较高，而右图的的碱基质量得分变化较大，测序质量逐渐下降，测序质量差。可能需要进行一定的预处理比如移除低质量碱基。

测序Reads GC含量评估

GC含量均一



GC含量双峰 – 可能存在样品污染



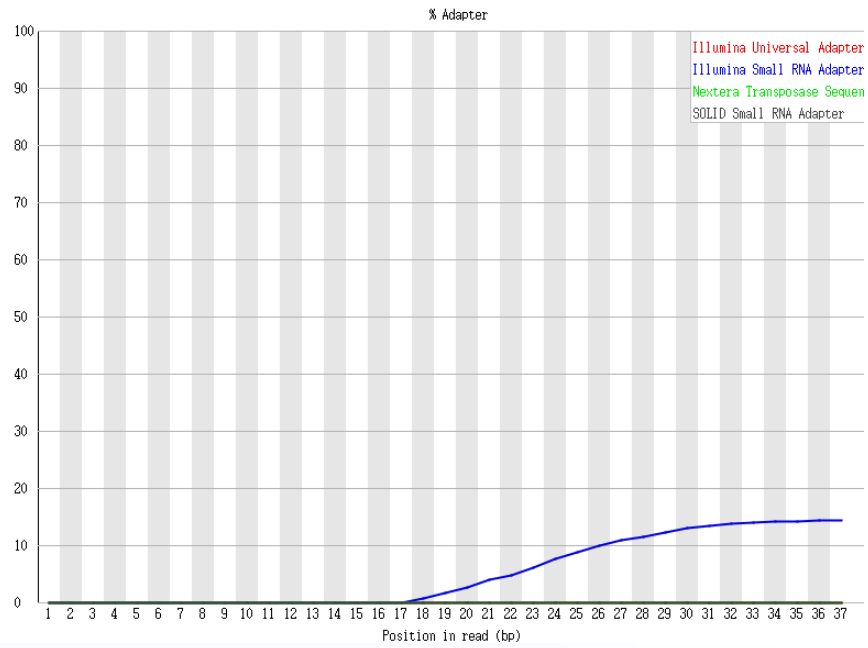
每个read GC含量的分布图。横坐标表示平均GC含量，纵坐标表示reads数。左图显示每个read的GC分布(红线)与理论分布(蓝线)相契合，GC含量均一。右图出现了GC含量双峰，表示测序样品可能存在特定的序列污染如混入了引物二聚体或实验室常见菌，当这一指标异常并且导致后期的序列比对率很低时，需要引起注意。

接头序列信息

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGTGTGAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGA	348471	1.8090753261979748	RNA PCR Primer, Index 1 (100% over 22bp)
GTGTGAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAA	284737	1.4782024362303687	RNA PCR Primer, Index 1 (100% over 23bp)
TGAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACTC	257581	1.3372229872712522	RNA PCR Primer, Index 1 (100% over 26bp)
GTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACTCCAGTC	214605	1.114114547203975	RNA PCR Primer, Index 1 (100% over 31bp)
GTGAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACT	194133	1.0078348565613535	RNA PCR Primer, Index 1 (100% over 25bp)
GAGTGTGAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGG	185873	0.964953347929659	Illumina Small RNA Adapter 2 (100% over 21bp)
AGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACTCCAGT	182710	0.9485327411739628	RNA PCR Primer, Index 1 (100% over 30bp)
GAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACTCC	177563	0.9218122660011622	RNA PCR Primer, Index 1 (100% over 27bp)
AAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACTCCA	172186	0.8938977536630723	RNA PCR Primer, Index 1 (100% over 28bp)
AAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGGAACTCCAG	166132	0.8624686188862831	RNA PCR Primer, Index 1 (100% over 29bp)
CCGTGTGAAAGTAGGACATCGTCAGGCTTGGAATTCTCGGGTGCCAAGG	129104	0.6702390182065748	Illumina Small RNA Adapter 2 (100% over 21bp)

Adapter Content

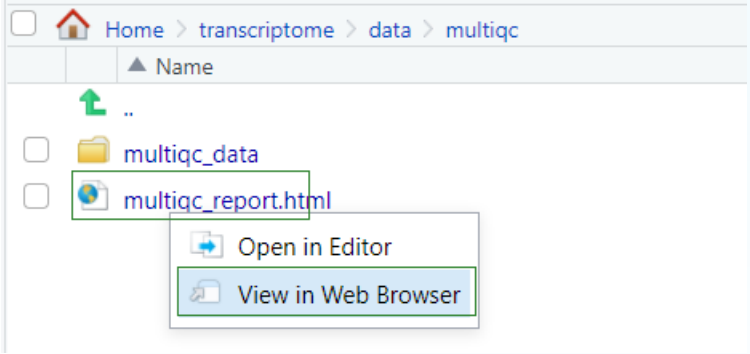


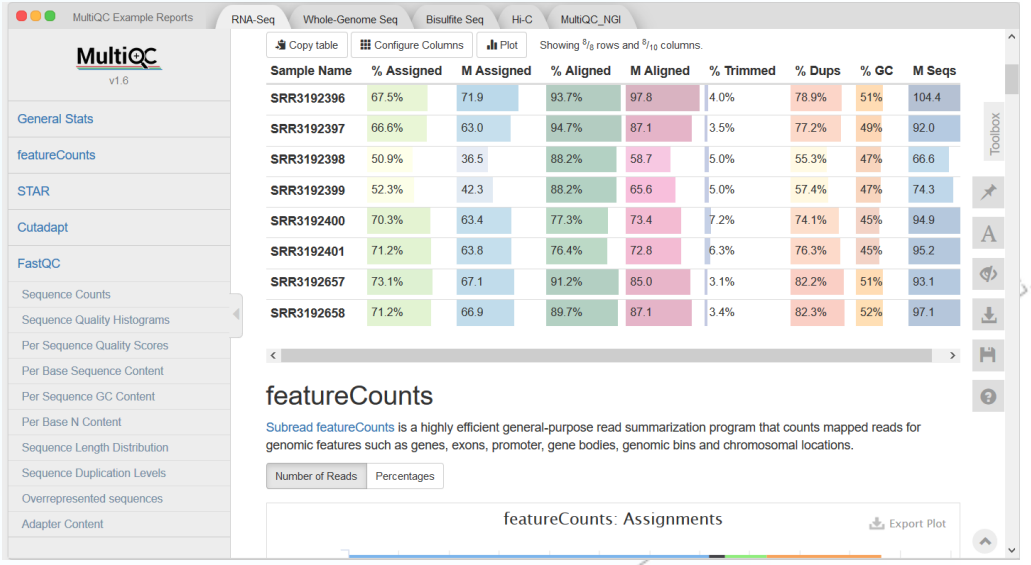
../FastQC/Configuration/adapter_list.txt
contaminant_list.txt
FastQC工具会与这两个文件里的序列进行比对，提醒是否含有非期望序列和接头序列

多样品测序数据质控检测整合

- MultiQC

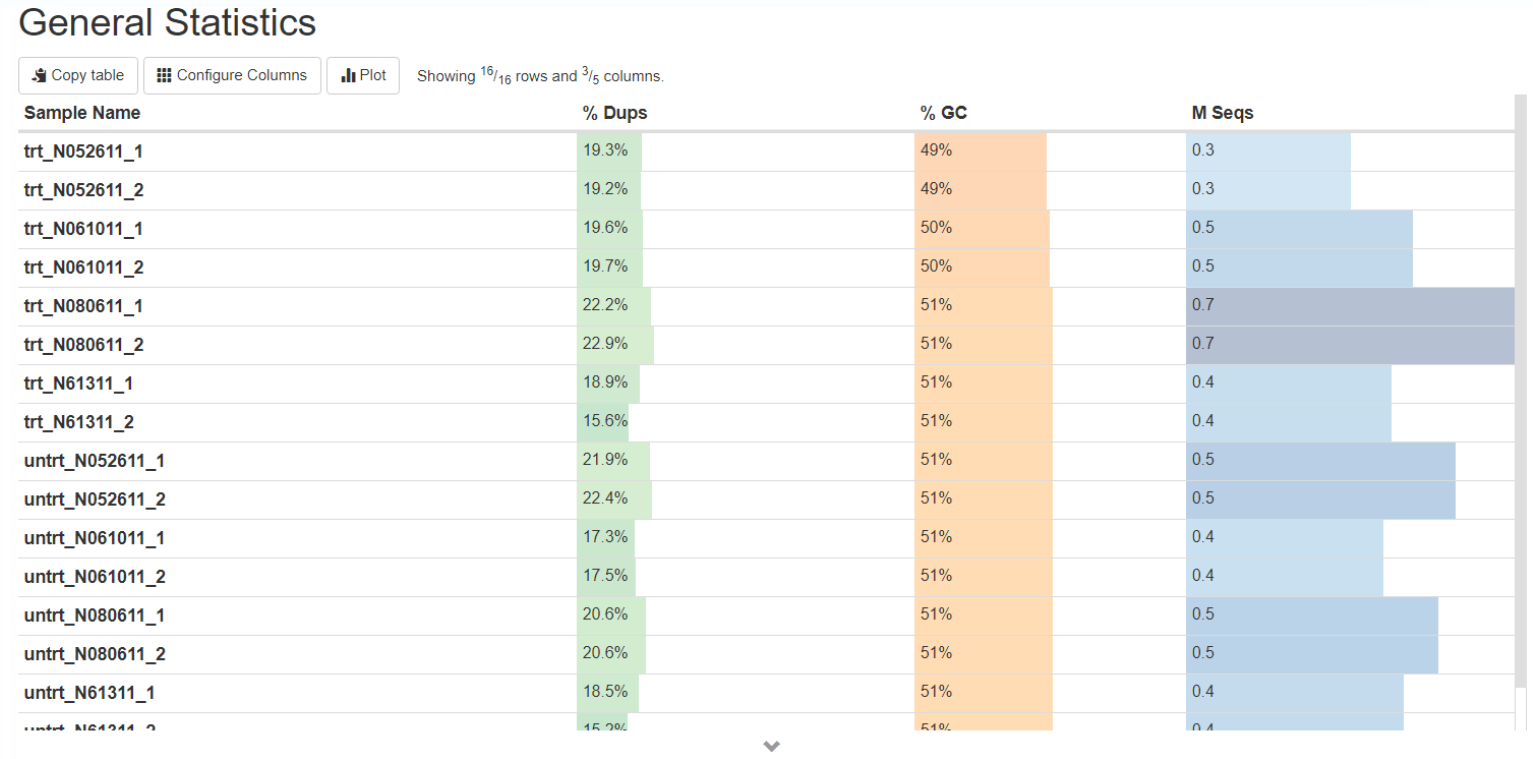
- multiqc –d . –o multiqc





整合QC质控结果神器

测序质量总结 – 测序reads数



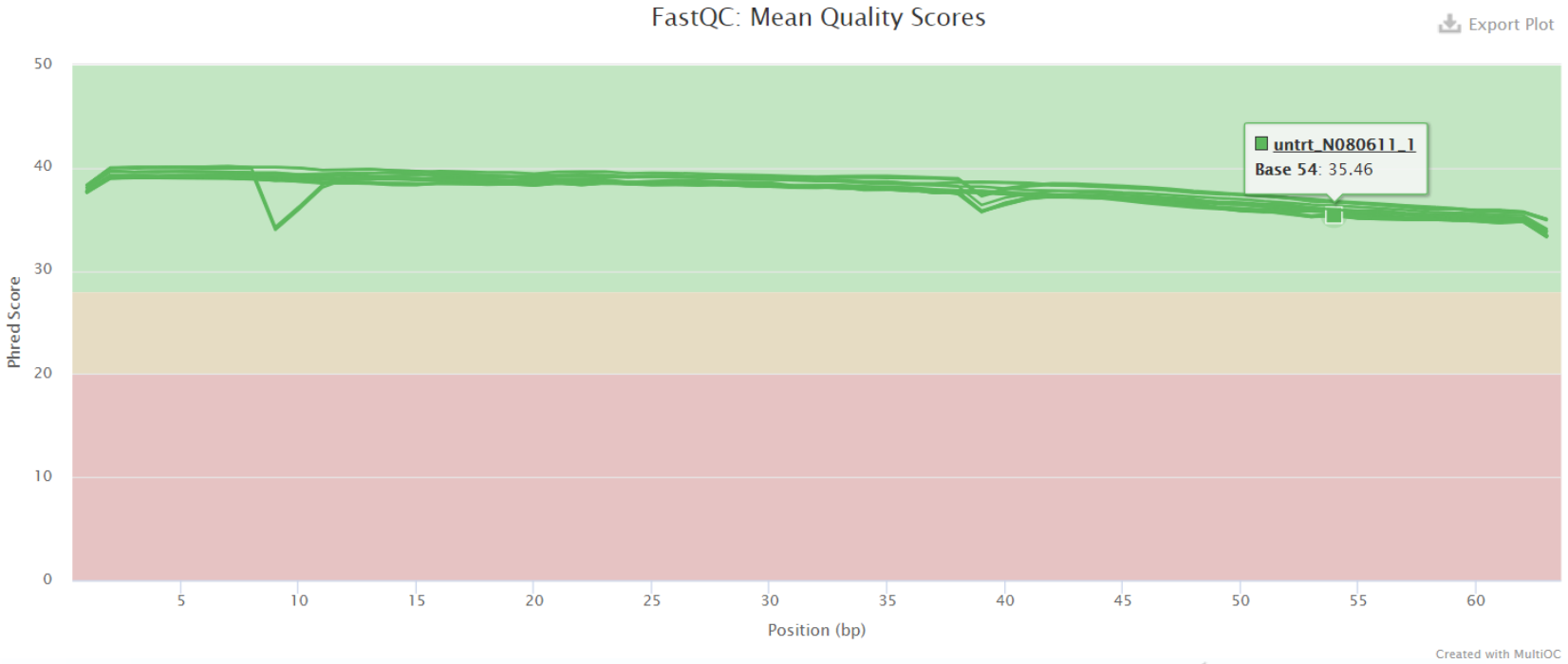
General Statistics

Copy table | Configure Columns | Plot | Showing 16/16 rows and 3/5 columns.

Sample Name	% Dups	% GC	M Seqs
trt_N052611_1	19.3%	49%	0.3
trt_N052611_2	19.2%	49%	0.3
trt_N061011_1	19.6%	50%	0.5
trt_N061011_2	19.7%	50%	0.5
trt_N080611_1	22.2%	51%	0.7
trt_N080611_2	22.9%	51%	0.7
trt_N61311_1	18.9%	51%	0.4
trt_N61311_2	15.6%	51%	0.4
untrt_N052611_1	21.9%	51%	0.5
untrt_N052611_2	22.4%	51%	0.5
untrt_N061011_1	17.3%	51%	0.4
untrt_N061011_2	17.5%	51%	0.4
untrt_N080611_1	20.6%	51%	0.5
untrt_N080611_2	20.6%	51%	0.5
untrt_N61311_1	18.5%	51%	0.4

重复率，GC含量，总reads数(million)

测序质量总结 – 交互式图展示平均质量



重复率，GC含量，总reads数(million)

接头和低质量reads处理 – 根据评估结果判断是否需要做

- Trimmomatic: 链接
  - Commond example:

java -jar trimmomatic-0.30.jar PE --phred33 *input_forward.fq input_reverse.fq output_forward_paired.fq output_forward_unpaired.fq output_reverse_paired.fq output_reverse_unpaired.fq* ILLUMINACLIP:adaptor-PE.fa:2:30:10 LEADING:20 TRAILING:20 MINLEN:36

Two input files and four output files;
Remove adapters: (maximum 2 mismatches in the 'seed' (16 bases) of the adaptor;
palindrome clip threshold 30; simple clip threshold 10)
Remove leading low quality bases (below quality 20);
Remove trailing low quality bases (below quality 20);
Drop reads below the 36 bases long

接头和低质量reads处理 – 根据评估结果判断是否需要做

- Fastp: 链接
  - Commond example:
- 单端测序 single end(SE) fastp –i in.fq –o out.f1
- 双端测序 paired end(PE) fastp –i in_1.fq –I in_2.fq –o out_1.fq –O out_2.fq

iMeta | 引用7000+，海普洛斯陈实富发布新版fastp，更快更好地处理FASTQ数据

现在一般可以不去除接头序列的3个原因

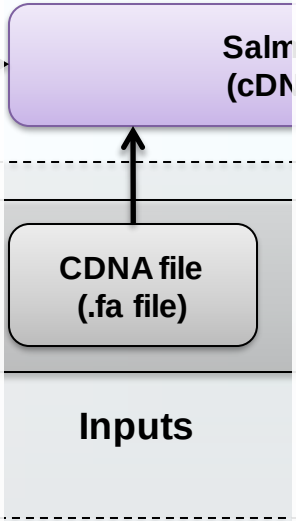
- 测序时插入片段长度一般大于测序的单端读长，接头不会被测到
- Salmon等工具采用K-mer方式估计序列一致性，如果一个K-mer因为质量低或有接头，则匹配不上，不影响同一reads其它kmer的比较
- STAR采用soft-clip机制自动移除末端未比对上的序列（包括接头序列）

cDNA - FASTA格式解释 – 转录本序列

>序列名字
ACTG序列单行或多行
>序列名字2
ACTG序列单行或多行

序列名字中不要有空白字符。

为什么不能有空白字符?



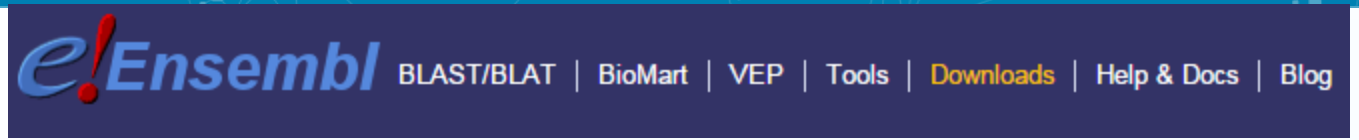

```
>ENST00000608838
ACAGGAATTCATATCGGGGTGATCACTCAGAAGAAAAGGTGAATACCGGATGTTGTAAGCTATTGAACTG
CCACAAGTGATATCTTTACACACCATTCTGCTGTCATTGGGTAGCTTTGAACCCCAAAAATGTTGGAAGA
ATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCT
ATCATGCTGCATTTCTATAATATCACATGAATATACTCGACGACCAGCATTTCCTGTGATTCACCTAGAG
GATATAACATTGGATTATAGTGATGTGGACTCTTTTACTGGTTCCCCAGTATCTATGTTGAATGATCTGA
TAACATTTGACACAACTAAATTTGGAGAAACCATGACACCTGAGACCAATACTCCTGAGACTACTATGCC
ACCATCTGAGGCCACTACTCCCGAGACTACTATGCCACCATCTGAGACTGCTACTTCCGAGACTATGCCA
CCACCTTCTCAGACAGCTCTTACTCATAATTAATTAACATTTACTTCTGGTATGGAACAACTAGAAATAC
TGCTGGAAATAATATCCAAAGAGCTGATTCTACCAATCCAATTTCACCAGGAAAATTCCATCAGGGATTG
GATGACCATGGGGATGGACATAATTGCTACTACCAACACAACAGCCAAGAGAGTTGCCTTACAATTAGAA
ATGTGTAGACAGAAATGTATAGAAGATACAAGGATTCTCTTAATTGGACTTAAATTCTTTATCTGTCTTC
CTCCGATGTACTCAAATATATGAGCTAATTTTTGTCTTAAGTGAACATTTGTATATCTATGTATTTTTCC
ATGCCAAAAACAAAAACGAAGACCATTGTTTGGAGCTGCCTCTTATGACTAAGACAAGAATTTTTACTTT
AACAGTGCCTGGCCCACTACTATCGTATATAGGAGAACATATAAAAGCATATAGAAAGTTCCAGATGAAT
GTTCCCTTCTCTACCCTCCACCTTTTATTGTAAGTTCTGACCCTAAATACTTTTCTGTGTCATGACGTCA
AATTTTGTTTAAGGTTCTAGCTGGTAACTAACAGAGTCAGAAGCTAATTCTTTCATTCAACACAAGCACT
GATCTAACTGGATAGAGATAAAAGTGGGACTTGCCTTGAGAGTACATCATATTAAATTAAAAGCTGCATC
TCAAATTCTA
>ENST00000382410
ATGAATATCCTGATGCTGACCTTCATTATCTGTGGGTTGCTAACTCGGGTGACCAAAGGTAGCTTTGAAC
CCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGTACATACT
TCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATATACTCGACGACCAGCATTT
CCTGTGATTCACCTAGAGGATATAACATTGGATTATAGTGATGTGGACTCTTTTACTGGTTCCCCAGTAT
CTATGTTGAATGATCTGATAACATTTGACACAACTAAATTTGGAGAAACCATGACACCTGAGACCAATAC
TCCTGAGACTACTATGCCACCATCTGAGGCCACTACTCCCGAGACTACTATGCCACCATCTGAGACTGCT
ACTTCCGAGACTATGCCACCACCTTCTCAGACAGCTCTTACTCATAATTAATTAACATTTACTTCTGGTA
TGGAACAACTAGAAATACTGCTGGAAATAATATCCAAAGAGCTGATTCTACCAATCCAATTTCACCAGGA
AAATTCCATCAGGGATTGGATGACCAT
```

cDNA – 不同FASTA格式名字的书写

- 序列名字行包含信息不同会影响程序对序列名字的判断，最保险的方式是使用最简洁名字，无额外信息。
- 序列行可为多行和单行，大部分程序兼容这2种模式，少部分程序只支持序列为单行。

```
>Sox2
ACGTCGACTACGAGACGAGACAAGCCAG
>Sox21 protein coding gene
ACGTCGACTACGAGACGAGACAAGCCAG
ACGTCGACTACGAGACGAGACAAGCCAG
>Pou5f1|Oct4
ACGTCGACTACGAGACGAGACAAGCCAGACGTCGACTACGAGACGAGACAAGCCAG
```

Ensembl (http://www.ensembl.org/index.html?redirect=no)



Show 10 entries | Show/hide columns | Filter

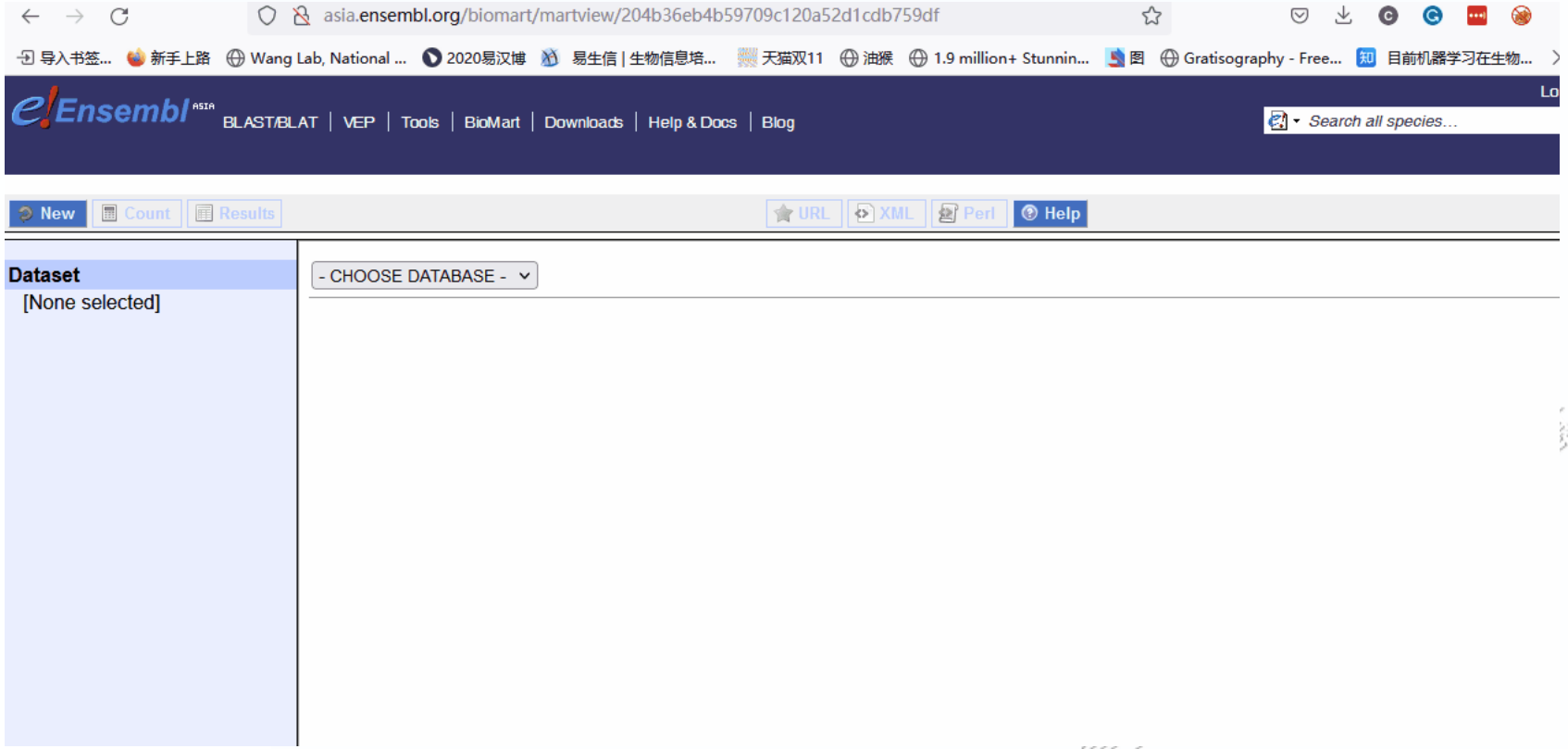
★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM/BigWig
Y	Human *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BigWig
Y	Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BigWig
Y	Zebrafish *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	-	-	BAM/BigWig
	Alpaca *Vicugna pacos*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	-	-	VEP	-	-	-
	Amazon molly *Poecilia formosa*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	-	-	VEP	-	-	BAM/BigWig
	Anole lizard *Anolis carolinensis*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	-	-	VEP	-	-	BAM/BigWig
	Armadillo *Dasypus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	-	-	VEP	-	-	BAM/BigWig

数据下载的trick?

如何选取合适的基因组组装

Ensembl cDNA指编码基因的cDNA，**不包括非编码基因**

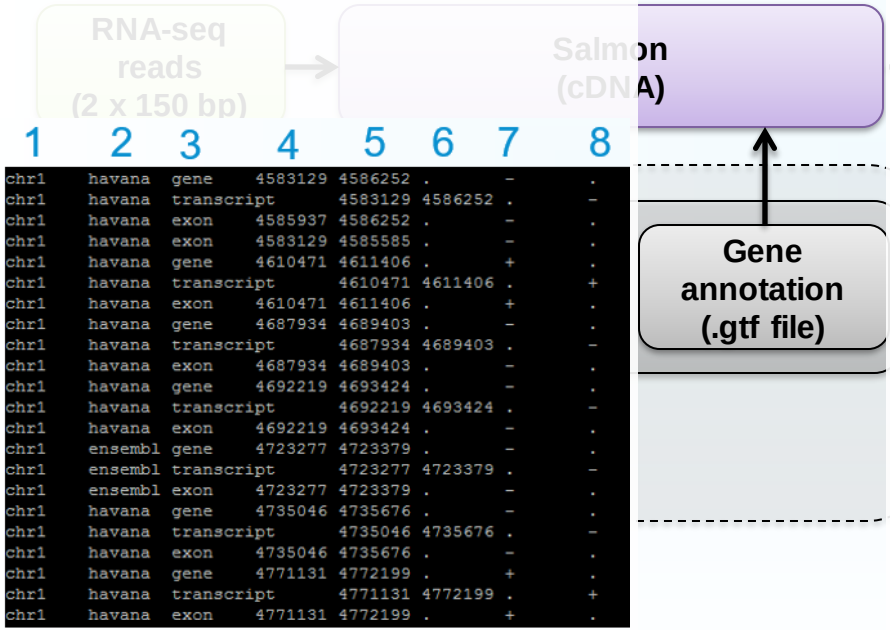
Idmap: 不同数据库ID转换



Ensembl Biomart

基因注释 - GTF格式解释

- GTF (Gene Transfer Format, GTF2.2) is an extension to, and backward compatible with, GFF2 (General Feature Format).



基因注释 - GTF格式解释

- **1-seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
- **2-source** - name of the program that generated this feature, or the data source (database or project name)
- **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'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
- **9-attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

中文解释

基因注释 - GTF格式解释 – 属性列

gene_id	ENSG00000178591
gene_version	6
gene_name	DEFB125
gene_source	ensembl_havana
gene_biotype	protein_coding

gene_id	ENSG00000178591
gene_version	6
transcript_id	ENST00000608838
transcript_version	1
gene_name	DEFB125
gene_source	ensembl_havana
gene_biotype	**protein_coding**
transcript_name	DEFB125-202
transcript_source	havana
transcript_biotype	*processed_transcript*
transcript_support_level	2

一些基因可以同时转录编码基因和非编码转录本。

GTF中的趣味知识

- 从ENSEMBL的注释来看，人基因组中包含60,676个注释的基因，*19968*个蛋白编码基因。
- 编码转录本最多的是肢带型肌营养不良相关基因SGCE，可以编码98条转录本。
- 可变剪接调控基因RBFOX1以2.7 million的长度超过之前文献报道的最长基因CNTNAP2（智力语言损伤相关基因）。
- T细胞受体相关基因TRDD1作为最短的基因，长度只有8 nt,编码的小肽序列包含2个氨基酸 EI。
- 外显子长度最长的蛋白编码基因是NFIA，一个转录因子，其外显子长度超*4万 nt*。
- 包含外显子数目最多的转录本是ENST00000589042，共有*363*个外显子。其对应的基因是TTN，横纹肌发育相关基因。

bed格式 – 至少3列，另外9列可选，0-start，前闭后开

生信分析过程中这些常见文件的格式以及查看方式你都知道吗？

UCSC

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade									
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand. Either "." (=no strand) or "+" or "-".
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

基因注释 – bed格式，cds和外显子位置

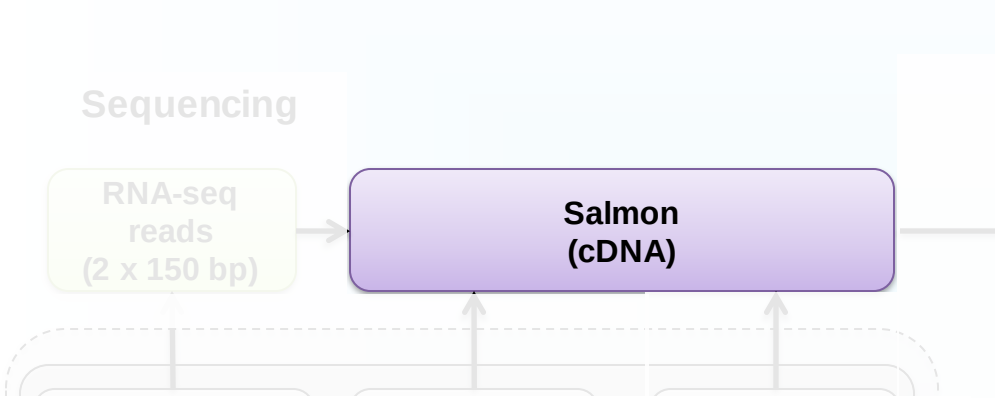
染色体	转录起始	终止	名字	0	链	翻译起始	翻译终止	0	外显子数目	外显子长度	外显子相对起始
chr20	87249	97094	ENST00000608838	0	+	97094	97094	0	2	110,1090,	0,8755,
chr20	87709	96533	ENST00000382410	0	+	87709	96417	0	2	58,529,	0,8295,
chr20	142368	145751	ENST00000382398	0	+	142628	145692	0	2	318,337,	0,3046,
chr20	142633	145749	ENST00000542572	0	+	145749	145749	0	3	53,74,171,	0,2781,2945,
chr20	157469	159163	ENST00000382388	0	+	157544	159024	0	2	124,390,	0,1304,
chr20	187852	189681	ENST00000334391	0	-	187885	189623	0	2	266,107,	0,1722,
chr20	227257	229886	ENST00000246105	0	+	227288	229771	0	2	89,609,	0,2020,
chr20	257735	261096	ENST00000382376	0	+	257778	259306	0	2	101,2020,	0,1341,
chr20	267185	268857	ENST00000608495	0	+	268857	268857	0	1	1672,	0,

gtf转bed12格式

gtfToGenePred -ignoreGroupsWithoutExons GRCh38.gtf GRCh38.gtf.50505050.pred

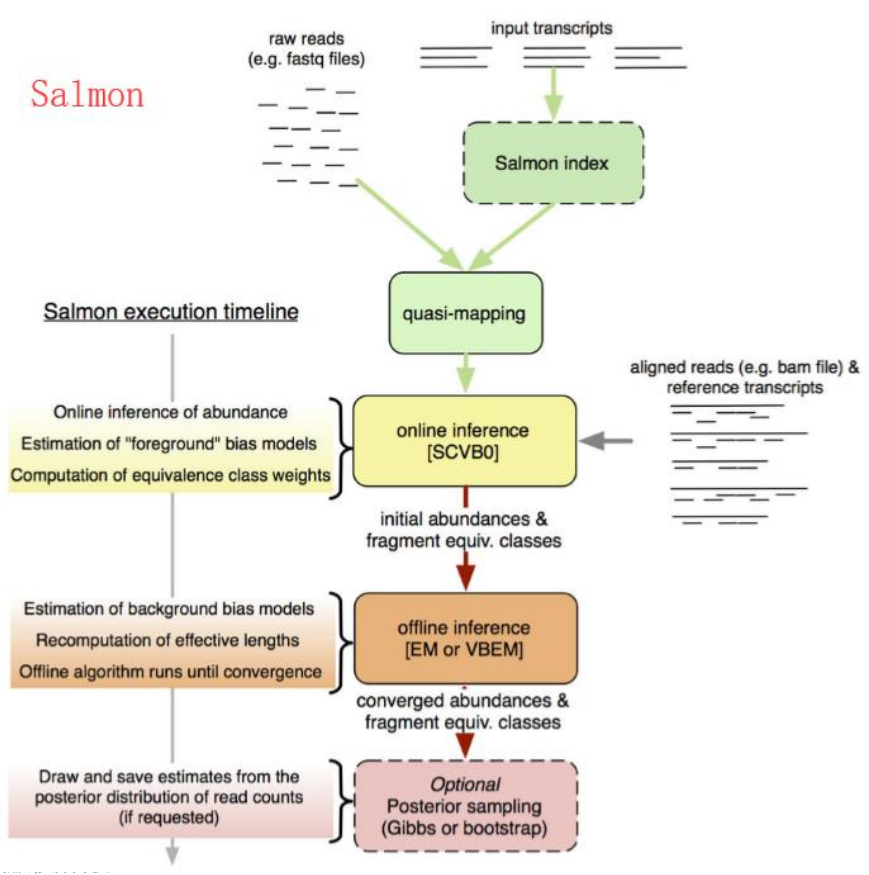
genePredToBed GRCh38.gtf.50505050.pred GRCh38.gtf.bed12

基因定量工具 - Salmon




```
salmon index -t transcripts.fa -i transcripts_index --type quasi -k 31

salmon quant -i transcripts_index -l <LIBTYPE> -1 reads1.fq -2 reads2.fq -o transcripts_quant
```



Salmon的优势

- 定量时考虑到不同样品中基因长度的改变（比如不同isoform的使用）
- 速度快、需要的计算资源和存储资源小
- 敏感性高，不会丢弃匹配到多个基因同源区域的reads
- 可以直接校正GC-bias
- 自动判断文库类型

Salmon构建转录本+基因组的双索引

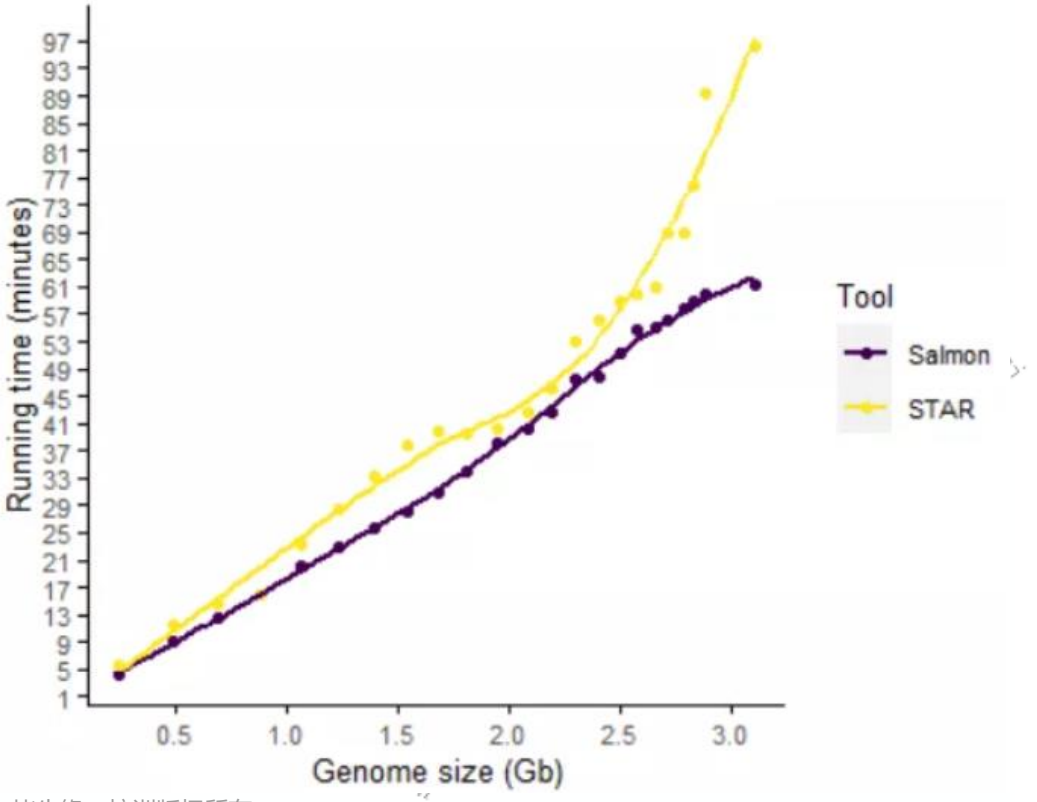
Specifically, there are 3 possible ways in which the salmon index can be created:

- cDNA-only index : salmon_index - https://combine-lab.github.io/salmon/getting_started/. This method will result in the smallest index and require the least resources to build, but will be the most prone to possible spurious alignments.
- SA mashmap index: salmon_partial_sa_index - (regions of genome that have high sequence similarity to the transcriptome) - Details can be found in this README and using this script. While running mashmap can require considerable resources, the resulting decoy files are fairly small. This will result in an index bigger than the cDNA-only index, but still mucch smaller than the full genome index below. It will confer many, though not all, of the benefits of using the entire genome as a decoy sequence.
- SAF genome index: salmon_sa_index - (the full genome is used as decoy) - The tutorial for creating such an index can be found here. This will result in the largest index, but likely does the best job in avoiding spurious alignments to annotated transcripts.

Salmon

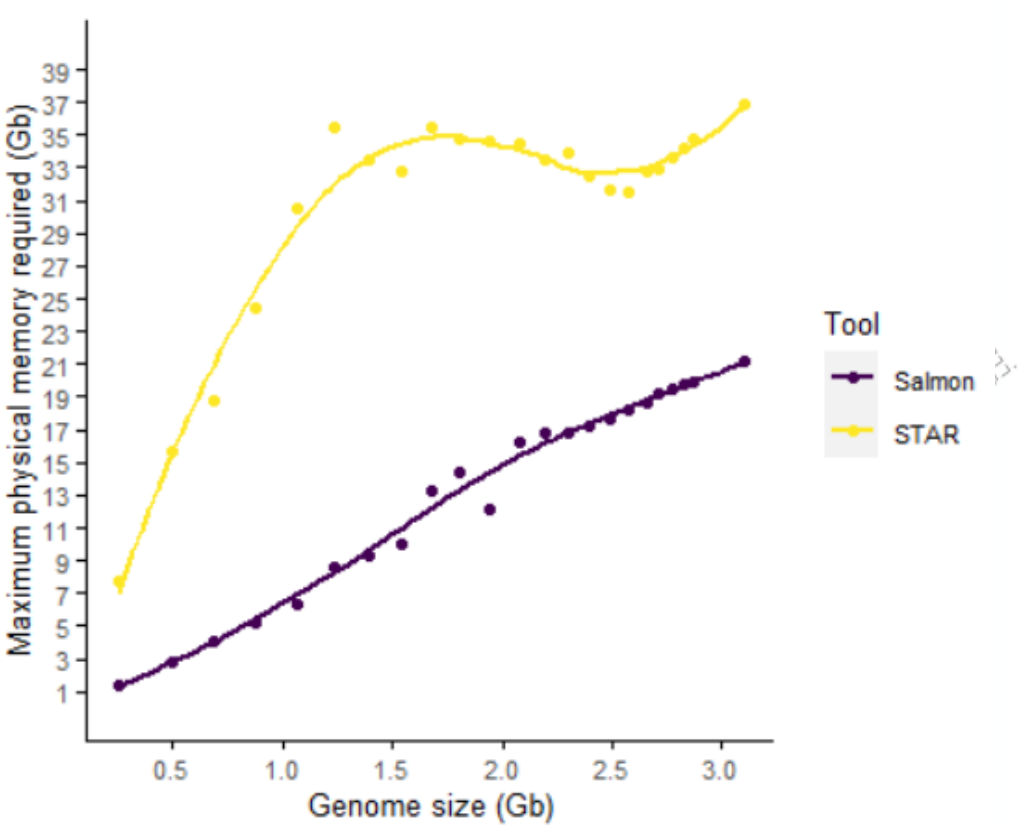
同样基因组大小，给定相同线程数时，Salmon速度快于STAR

- 同样基因组大小，给定相同线程数时，Salmon速度快于STAR。
- 数据量更大时，salmon的速度优势更明显。



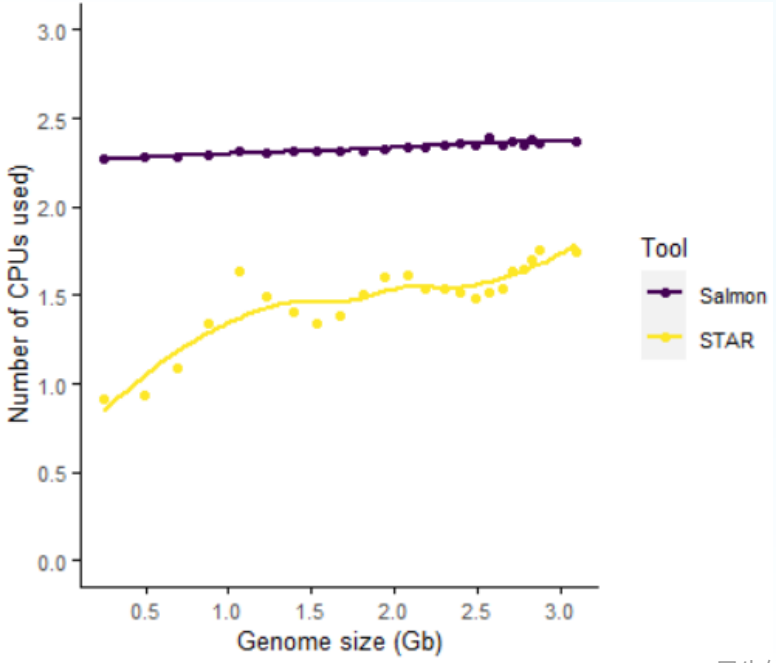
对人类3G大小的基因组，Salmon的内存需求只占STAR的一半

- Salmon对内存的需求明显小于STAR的需求。
- 对人类3G大小的基因组，Salmon的内存需求只占STAR的一半
- 如果用Salmon构建索引时不考虑基因组信息所需内存更少

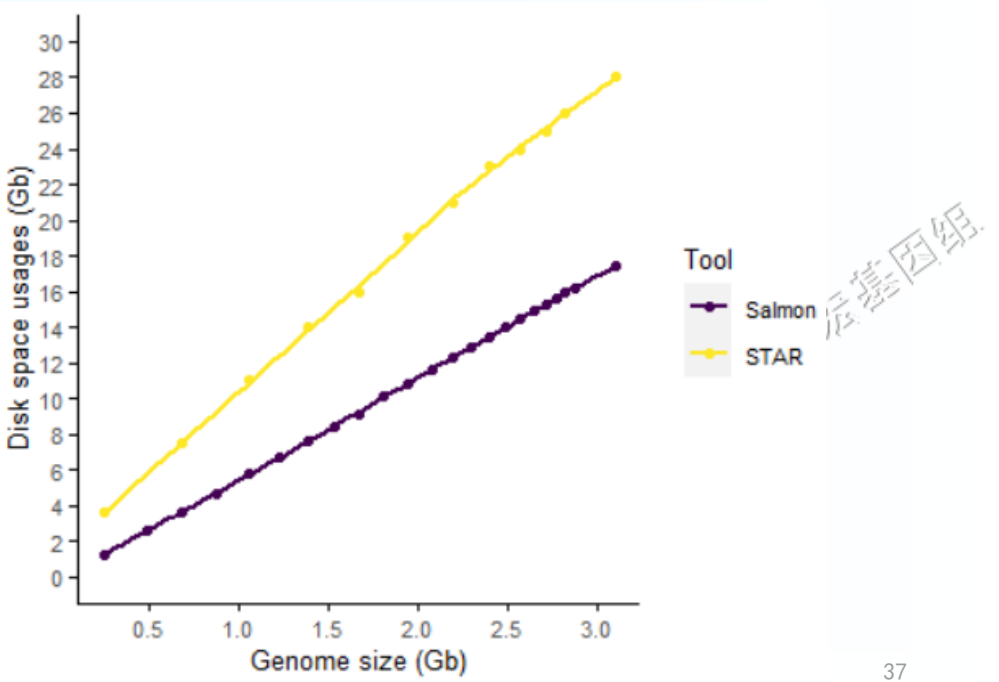


Salmon性能优于STAR

Salmon的CPU利用率更稳定，且明显高于STAR



Salmon构建的索引占用磁盘空间更小
基因组增大时，Salmon所需磁盘空间增速小于STAR



Salmon定量适用三代测序和无参转录组

- 三代获得的转录本可作为salmon定量所需的cDNA，然后结合二代测序数据进行定量。
- 无参采用Trinity拼装获得的转录本可作为salmon定量所需的cDNA，然后结合二代测序数据进行定量。

基因定量工具 – Salmon输出结果 effective length

EffectiveLength — This is the computed *effective* length of the target transcript. It takes into account all factors being modeled that will effect the probability of sampling fragments from this transcript, including the fragment length distribution and sequence-specific and gc-fragment bias (if they are being modeled).

Name	Length	EffectiveLength	TPM	NumReads	
ENST00000608838	1200	1043.763	0.000000	0.000	
ENST00000382410	587	431.213	0.000000	0.000	
ENST00000382398	655	499.101	0.000000	0.000	
ENST00000542572	298	148.609	0.000000	0.000	
ENST00000382388	514	358.450	0.000000	0.000	
ENST00000334391	373	219.458	0.000000	0.000	
ENST00000246105	698	542.049	0.000000	0.000	
ENST00000382376	2121	1964.763	0.000000	0.000	
ENST00000608495	1672	1515.763	3.336665	2.071	
ENST00000382369	1420	1263.763	46.887137	24.263	
ENST00000360321	1575	1418.763	104.546732	60.737	
ENST00000400269	1608	1451.763	0.000000	0.000	
ENST00000500893	3354	3197.763	307.007551	402.000	
ENST00000414676	750	594.002	213.718181	51.983	
ENST00000442637	762	605.997	32.308779	8.017	
ENST00000342665	4630	4473.763	313.880372	575.000	
ENST00000609179	578	422.227	28.261349	4.886	
ENST00000492242	846	689.947	92.365434	26.095	
ENST00000382291	2088	1931.763	1034.542929	818.339	
ENST00000609504	582	426.220	0.000000	0.000	
ENST00000382285	1047	890.763	92.380787	33.696	
ENST00000608467	504	348.508	0.000000	0.000	
ENST00000470439	697	541.052	53.445125	11.841	
ENST00000608736	712	556.028	25.826886	5.880	
ENST00000621012	1852	1695.763	932.148633	647.263	
ENST00000608875	941	784.907	0.000000	0.000	
ENST00000615226	877	720.931	0.000000	0.000	
ENST00000217233	2499	2342.763	182.142169	174.731	

基因定量工具 – Salmon输出结果 TPM

$$RPKM = \frac{\text{Read counts} * 10^6 * 1000}{\text{Total reads} * \text{Gene length}}$$

$$RPK = \frac{\text{Read counts} * 1000}{\text{Gene length}}$$

$$TPM = \frac{RPK * 10^6}{\text{sum (RPK)}}$$

Name	Length	EffectiveLength	TPM	NumReads
ENST00000608838	1200	1043.763	0.000000	0.000
ENST00000382410	587	431.213	0.000000	0.000
ENST00000382398	655	499.101	0.000000	0.000
ENST00000542572	298	148.609	0.000000	0.000
ENST00000382388	514	358.450	0.000000	0.000
ENST00000334391	373	219.458	0.000000	0.000
ENST00000246105	698	542.049	0.000000	0.000
ENST00000382376	2121	1964.763	0.000000	0.000
ENST00000608495	1672	1515.763	3.336665	2.071
ENST00000382369	1420	1263.763	46.887137	24.263
ENST00000360321	1575	1418.763	104.546732	60.737
ENST00000400269	1608	1451.763	0.000000	0.000
ENST00000500893	3354	3197.763	307.007551	402.000
ENST00000414676	750	594.002	213.718181	51.983
ENST00000442637	762	605.997	32.308779	8.017
ENST00000342665	4630	4473.763	313.880372	575.000
ENST00000609179	578	422.227	28.261349	4.886
ENST00000492242	846	689.947	92.365434	26.095
ENST00000382291	2088	1931.763	1034.542929	818.339
ENST00000609504	582	426.220	0.000000	0.000
ENST00000382285	1047	890.763	92.380787	33.696
ENST00000608467	504	348.508	0.000000	0.000
ENST00000470439	697	541.052	53.445125	11.841
ENST00000608736	712	556.028	25.826886	5.880
ENST00000621012	1852	1695.763	932.148633	647.263
ENST00000608875	941	784.907	0.000000	0.000
ENST00000615226	877	720.931	0.000000	0.000
ENST00000217233	2499	2342.763	182.142169	174.731

视频讲解RPKM

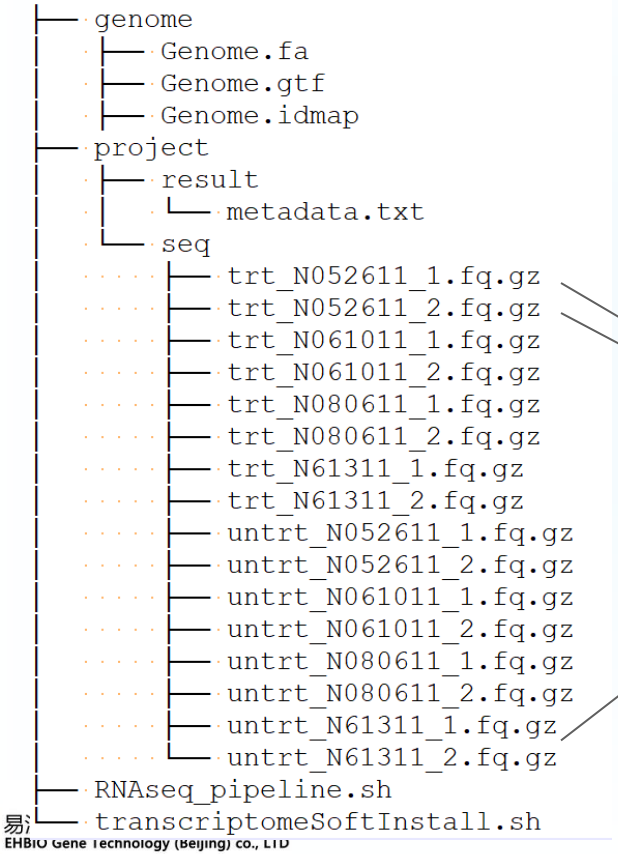
基因的表达量和转录本的表达量

- Salmon默认输出转录本的表达量
- 可以指定参数提供转录本-基因的映射关系，输出基因的表达量
- 可以分析差异基因，也可以分析差异转录本
- R包tximport可以转换转录本表达量为基因表达量，结果传递给DESeq2进行差异基因分析

样品分组和其它属性信息

Samp	Conditions	Batch	Sex
untrt_N61311	untrt	A	F
untrt_N052611	untrt	A	M
untrt_N080611	untrt	B	M
untrt_N061011	untrt	B	F
trt_N61311	trt	A	F
trt_N052611	trt	A	F
trt_N080611	trt	B	M
trt_N061011	trt	B	F

转录组数据目录结构和metadata与原始数据对应关系



metadata.txt

Samp	conditions
untrt_N61311	untrt
untrt_N052611	untrt
untrt_N080611	untrt
untrt_N061011	untrt
trt_N61311	trt
trt_N052611	trt
trt_N080611	trt
trt_N061011	trt

样品分组和其它属性信息

```
# TAB键分割的样品分组信息
# 这里是为了在流程中，方便操作
# 实际上可以用任何文本编辑工具或者excel导出一个TAB键分割的文件，
# 第一列是样本名字
# 后面的列是样本分组或其它熟悉信息，记录越全越好
# cat <<END 输入内容 END 输入结束
# sed 's/|/\t/g': 把文件中所有的 | 替换为 \t
# sed 's/original/replace/g'
cat <<END | sed 's/|/\t/g' >sampleFile
Samp|conditions
untrt_N61311|untrt
untrt_N052611|untrt
untrt_N080611|untrt
untrt_N061011|untrt
trt_N61311|trt
trt_N052611|trt
trt_N080611|trt
trt_N061011|trt
END
```

```
Samp    conditions
untrt_N61311    untrt
untrt_N052611   untrt
untrt_N080611   untrt
untrt_N061011   untrt
trt_N61311      trt
trt_N052611     trt
trt_N080611     trt
trt_N061011     trt
```

批量Salmon定量

- 结合sampleFile，用for循环代替手写命令进行批量定量

```
for i in `tail -n +2 sampleFile | cut -f 1`
do
  salmon quant -l A -1 ${i}_1.fq.gz -2 ${i}_2.fq.gz  -i genome/GRCh38.salmon \
    -o ${i}/${i}.salmon.count -p 4
done
```

- echo “” 展示for循环实际进行的操作，用于调试

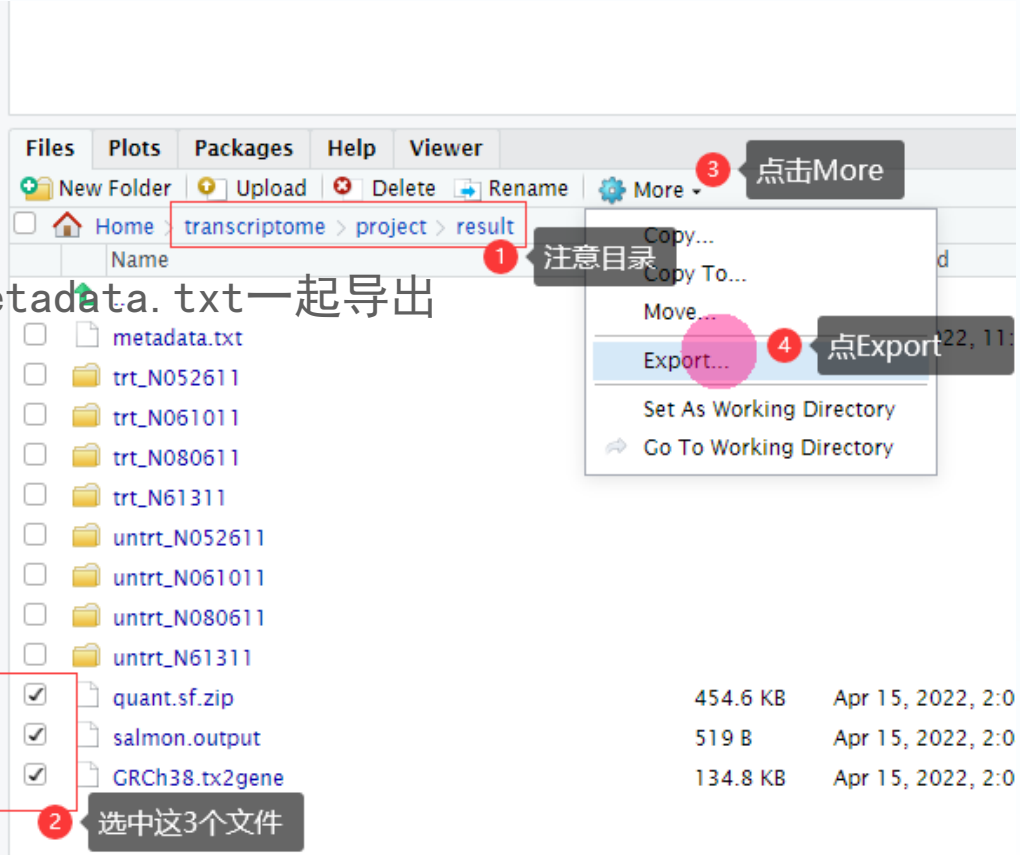
```
RNA_1110@localhost:~/transcriptome/data$ for i in `tail -n +2 sampleFile | cut -f 1`
> do
>   echo "salmon quant -l A -1 ${i}_1.fq.gz -2 ${i}_2.fq.gz  -i genome/GRCh38.salmon -o ${i}/${i}.salmon.count -p 4"
> done
salmon quant -l A -1 untrt_N61311_1.fq.gz -2 untrt_N61311_2.fq.gz  -i genome/GRCh38.salmon -o untrt_N61311/untrt_N61311.salmon.count -p 4
salmon quant -l A -1 untrt_N052611_1.fq.gz -2 untrt_N052611_2.fq.gz  -i genome/GRCh38.salmon -o untrt_N052611/untrt_N052611.salmon.count -p 4
salmon quant -l A -1 untrt_N080611_1.fq.gz -2 untrt_N080611_2.fq.gz  -i genome/GRCh38.salmon -o untrt_N080611/untrt_N080611.salmon.count -p 4
salmon quant -l A -1 untrt_N061011_1.fq.gz -2 untrt_N061011_2.fq.gz  -i genome/GRCh38.salmon -o untrt_N061011/untrt_N061011.salmon.count -p 4
salmon quant -l A -1 trt_N61311_1.fq.gz -2 trt_N61311_2.fq.gz  -i genome/GRCh38.salmon -o trt_N61311/trt_N61311.salmon.count -p 4
salmon quant -l A -1 trt_N052611_1.fq.gz -2 trt_N052611_2.fq.gz  -i genome/GRCh38.salmon -o trt_N052611/trt_N052611.salmon.count -p 4
salmon quant -l A -1 trt_N080611_1.fq.gz -2 trt_N080611_2.fq.gz  -i genome/GRCh38.salmon -o trt_N080611/trt_N080611.salmon.count -p 4
salmon quant -l A -1 trt_N061011_1.fq.gz -2 trt_N061011_2.fq.gz  -i genome/GRCh38.salmon -o trt_N061011/trt_N061011.salmon.count -p 4
```

从服务器通过Rstudio导出用于差异分析的文件


```
print FNR, $0}'
```

quant.sf.zip文件

也勾选metadata.txt一起导出



MultiQC展示Salmon比对结果

- multiqc -f -d . -o multiqc/

General Statistics

Copy table | Configure Columns | Plot | Showing 8/24 rows and 2/7 columns.

Sample Name	% Aligned	M Aligned
trt_N052611 | trt_N052611.salmon.count | aux_info | trt_N052611.salmon.count	95.0%	0.3
trt_N061011 | trt_N061011.salmon.count | aux_info | trt_N061011.salmon.count	94.8%	0.4
trt_N080611 | trt_N080611.salmon.count | aux_info | trt_N080611.salmon.count	94.3%	0.6
trt_N61311 | trt_N61311.salmon.count | aux_info | trt_N61311.salmon.count	94.1%	0.4
untrt_N052611 | untrt_N052611.salmon.count | aux_info | untrt_N052611.salmon.count	93.9%	0.5
untrt_N061011 | untrt_N061011.salmon.count | aux_info | untrt_N061011.salmon.count	93.6%	0.4
untrt_N080611 | untrt_N080611.salmon.count | aux_info | untrt_N080611.salmon.count	93.6%	0.5
untrt_N61311 | untrt_N61311.salmon.count | aux_info | untrt_N61311.salmon.count	93.7%	0.4

整理Salmon结果，便于读入DESeq2进行差异基因分析

```
# 列出salmon的输出文件
find . -name quant.sf
# ./trt_N080611/trt_N080611.salmon.count/quant.sf
# ./trt_N061011/trt_N061011.salmon.count/quant.sf
# ./untrt_N61311/untrt_N61311.salmon.count/quant.sf

# 生成一个两列文件方便R导入
# xargs接收上一步的输出，按批次提供给下游程序作为输入
# -i: 用{}表示传递的值
cut -f 1 sampleFile | xargs -i echo -e "{}\t{}/{}.salmon.count/quant.sf" >salmon.output
head salmon.output
# Samp    Samp/Samp.salmon.count/quant.sf
# untrt_N61311    untrt_N61311/untrt_N61311.salmon.count/quant.sf
# untrt_N052611   untrt_N052611/untrt_N052611.salmon.count/quant.sf


# 注意修改$14, $10为对应的信息列，
# tx2gene为一个两列文件，第一列是转录本没名字，第二列是基因名字。
sed 's/"/\t/g' genome/GRCh38.gtf | \
  awk 'BEGIN{OFS=FS="\t"}{if($3=="transcript") print $14, $10}' >genome/GRCh38.tx2gene
head genome/GRCh38.tx2gene
# ENST00000608838 ENSG00000178591
# ENST00000382410 ENSG00000178591
# ENST00000382398 ENSG00000125788
# ENST00000542572 ENSG00000125788
```

登录服务器操作 pipelineSalmon.sh

Sequencing costs a lot and gains more

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识