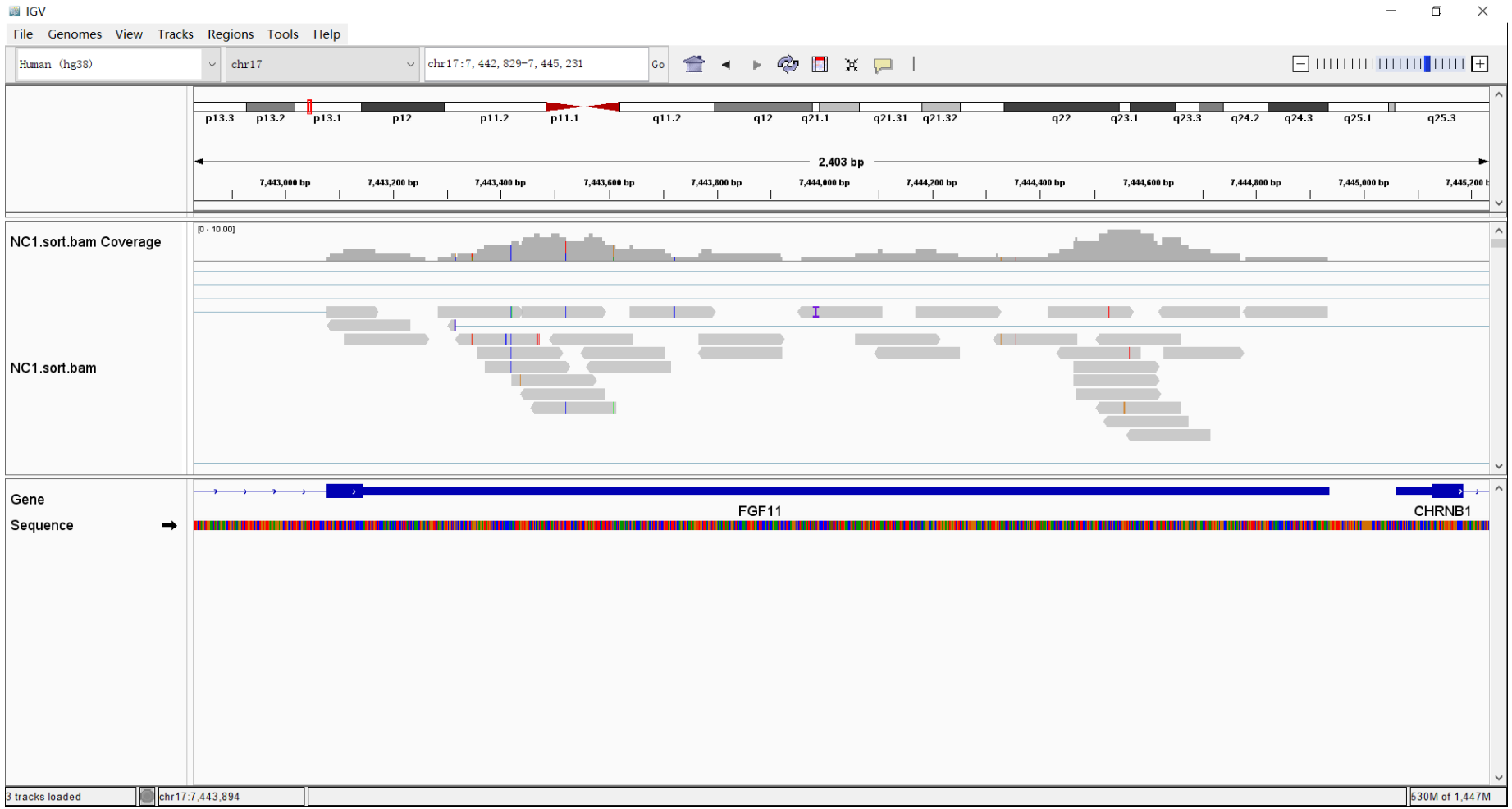
微生信
Wei Sheng Xin

RNA-seq
比对、定量和差异分析

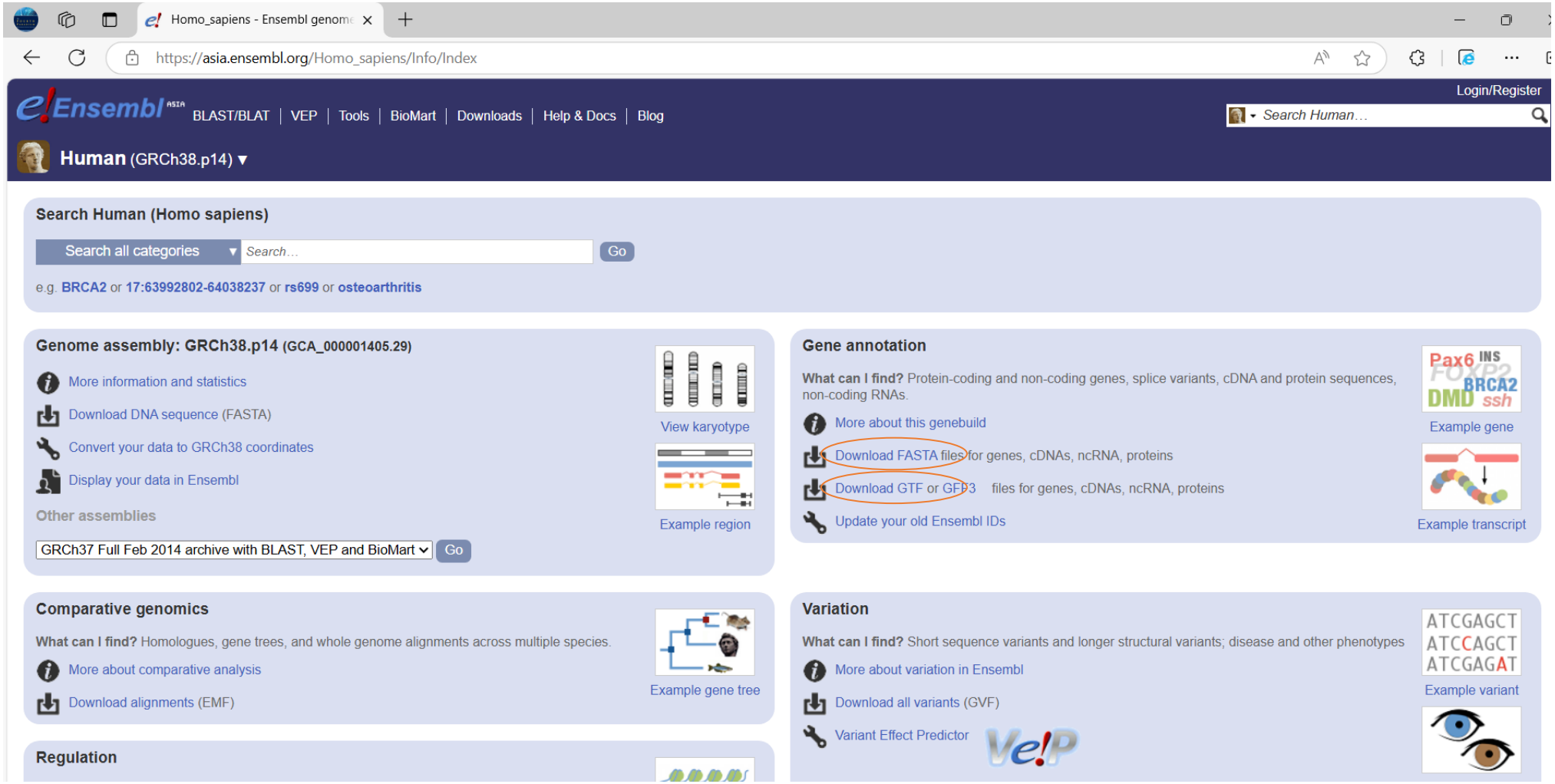
陈明杰 202411

比对和定量

比对/回贴：找到reads在基因组上的来源位置　　定量：统计比对到基因上的reads数



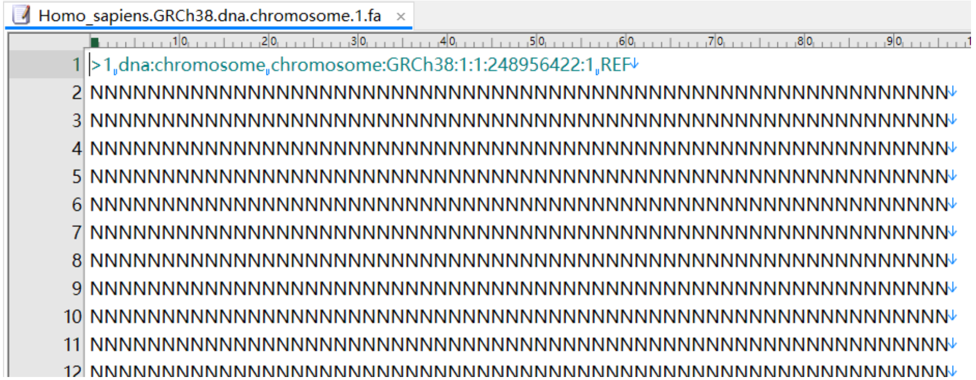
基因组文件

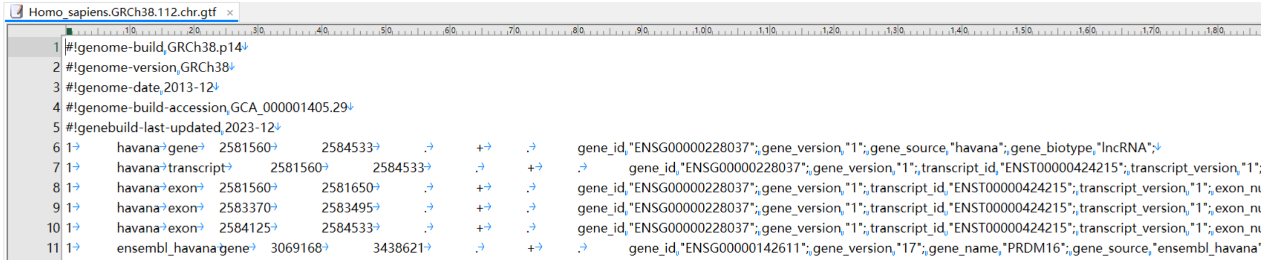


基因组序列文件

基因注释文件

下载





原则：名字匹配
wget分染色体下载fa.gz，然后合并
或者直接下载总的（包含一堆非chr的染色体）
染色体名字：1 vs chr1
建议：chr*

Fa文件

chrM.fa

一个文件一条序列

```
>chrM
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATC
CTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTA
AAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAAT
GTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCA
```

DNA：A腺嘌呤、C胞嘧啶、G鸟嘌呤、T胸腺嘧啶
RNA：A腺嘌呤、C胞嘧啶、G鸟嘌呤、T尿嘧啶

Tips
N表示不确定/未知的碱基（信号太弱，测序错误等）
小写字母：信号强度不够，测序仪不能100%确定

protein.fa

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

```

Protein序列：20种氨基酸

hsa.mature.fa

一个文件多条序列

```
>hsa-let-7a-5p
TGAGGTAGTAGGTTGTATAGTT
>hsa-let-7b-5p
TGAGGTAGTAGGTTGTGTGGTT
>hsa-let-7c-5p
TGAGGTAGTAGGTTGTATGGTT
>hsa-let-7d-5p
AGAGGTAGTAGGTTGCATAGTT
>hsa-let-7e-5p
TGAGGTAGGAGGTTGTATAGTT
>hsa-let-7f-5p
TGAGGTAGTAGATTGTATAGTT
```

第1部分：定界符大于号+名字描述信息
第2部分：序列
可以一行
可以多行（一般60-80个碱基一行）

基因组/转录组版本

- 人类基因组的GRCh37（hg19）和GRCh38（hg38）是两个常用的参考基因组版本。GRCh38是在GRCh37之后发布的更新版本，包含了更多的改进，例如填补了序列间隙（gaps）、修正了一些错误组装的区域、增加了着丝粒序列，并在某些区域增加了alternate loci来代表序列的多样性。这些改进使得GRCh38在基因组分析中，尤其是在检测结构变异方面，比GRCh37具有更高的准确性和可靠性。GRCh38相比于GRCh37，减少了一些N（表示序列间隙或未注释区域）的数量，增加了GC含量，并且扩大了外显子组的大小。
- 不同的基因组版本可能会影响基因组数据分析的结果，包括序列比对、变异检测、基因表达分析等。因此，当使用公共数据库中的基因组数据或者进行跨研究比较时，需要注意基因组版本的一致性。

- Ensembl 112: May 2024 (GRCh38.p14) - patched/updated gene set Dec 2023
- Ensembl 111: Jan 2024 (GRCh38.p14) - patched/updated gene set Jul 2023
- Ensembl 110: Jul 2023 (GRCh38.p14) - patched/updated gene set Mar 2023
- Ensembl 109: Feb 2023 (GRCh38.p13) - patched/updated gene set Nov 2022
- Ensembl 108: Oct 2022 (GRCh38.p13) - patched/updated gene set Jul 2022
- Ensembl 107: Jul 2022 (GRCh38.p13) - patched/updated gene set Apr 2022

3-6个月更新一次

实操：找到XX文章里边的基因组版本、转录组注释版本

GTF – 基因注释

Homo_sapiens.GRCh38.104.chr.gtf

1	chr1	ensembl_havana	gene	685679	686673	.	gene_id "ENSG00000284662"; gene_version "1
2	chr1	ensembl_havana	transcript	685679	686673	.	gene_id "ENSG00000284662"; gene_v
3	chr1	ensembl_havana	exon	685679	686673	.	gene_id "ENSG00000284662"; gene_version "1
4	chr1	ensembl_havana	CDS	685719	686654	.	0 gene_id "ENSG00000284662"; gene_version "1
5	chr1	ensembl_havana	start_codon	686652	686654	.	0 gene_id "ENSG00000284662"; gene_v
6	chr1	ensembl_havana	stop_codon	685716	685718	.	0 gene_id "ENSG00000284662"; gene_v
7	chr1	ensembl_havana	five_prime_utr	686655	686673	.	. gene_id "ENSG00000284662"; gene_v
8	chr1	ensembl_havana	three_prime_utr	685679	685715	.	. gene_id "ENSG00000284662"; gene_v

1 2 3 4 5 6 7 8

一行表示一个exon/CDS等子区域，多行联合表示一个gene

```
gene_id "ENSG00000284662"; gene_version "1"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding";
.       gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; gene_name "OR4F16"; gene_source "ensembl
gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; gene_sourc
gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; gene_sourc
0       gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; ger
0       gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; ger
.       gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; gene_name "OR4F16"; gene_source "ensembl
.       gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; gene_name "OR4F16"; gene_source "ensembl
-       .       gene_id "ENSG00000186827"; gene_version "11"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding";
```

实操：下载并打开查看

建索引

nohup hisat2-build genome.fa genome -p 4 &
约30分钟

```
vboxuser@ubuntu: ~/share/database/human/genome
(jimmy3) vboxuser@ubuntu:~/share/database/human/genome$ ls
chromsomes  genome.fa  Homo_sapiens.GRCh38.112.chr.gtf
(jimmy3) vboxuser@ubuntu:~/share/database/human/genome$ conda activate jimmy3
(jimmy3) vboxuser@ubuntu:~/share/database/human/genome$ hisat2-build
No input sequence or sequence file specified!
HISAT2 version 2.2.1 by Daehwan Kim (infphilo@gmail.com, http://www.ccb.jhu.edu/people/infphilo)
Usage: hisat2-build [options]* <reference_in> <ht2_index_base>
    reference_in            comma-separated list of files with ref sequences
    hisat2_index_base       write ht2 data to files with this dir/basename
Options:
    -c                      reference sequences given on cmd line (as
                            <reference_in>)
    --large-index           force generated index to be 'large', even if ref
                            has fewer than 4 billion nucleotides
    -a/--noauto             disable automatic -p/--bmax/--dcv memory-fitting
    -p <int>                number of threads
    --bmax <int>            max bucket sz for blockwise suffix-array builder
    --bmaxdivn <int>        max bucket sz as divisor of ref len (default: 4)
    --dcv <int>             diff-cover period for blockwise (default: 1024)
    --nodc                  disable diff-cover (algorithm becomes quadratic)
    -r/--noref              don't build .3/.4.ht2 (packed reference) portion
    -3/--justref            just build .3/.4.ht2 (packed reference) portion
    -o/--offrate <int>      SA is sampled every 2^offRate BWT chars (default: 5)
    -t/--ftabchars <int>    # of chars consumed in initial lookup (default: 10)
```

```
(jimmy3) vboxuser@ubuntu:~/share/database/human/genome$ ll
total 8932889
drwxrwx--- 1 root vboxsf       8192 Oct  7 12:03 ./
drwxrwx--- 1 root vboxsf          0 Oct  6 14:25 ../
drwxrwx--- 1 root vboxsf       8192 Oct  7 10:31 chromsomes/
-rwxrwx--- 1 root vboxsf  983425079 Oct  7 11:54 genome.1.ht2*
-rwxrwx--- 1 root vboxsf  734413928 Oct  7 11:54 genome.2.ht2*
-rwxrwx--- 1 root vboxsf       9035 Oct  7 11:20 genome.3.ht2*
-rwxrwx--- 1 root vboxsf  734413921 Oct  7 11:20 genome.4.ht2*
-rwxrwx--- 1 root vboxsf 1290699801 Oct  7 12:01 genome.5.ht2*
-rwxrwx--- 1 root vboxsf  747869486 Oct  7 12:01 genome.6.ht2*
-rwxrwx--- 1 root vboxsf         12 Oct  7 11:21 genome.7.ht2*
-rwxrwx--- 1 root vboxsf          8 Oct  7 11:21 genome.8.ht2*
-rwxrwx--- 1 root vboxsf 3139758018 Oct  7 10:49 genome.fa*
```

比对

hisat2 -x /home/vboxuser/share/database/human/genome/genome -1 NC1_R1.clean.gz -2 NC1_R2.clean.gz -p 1 --rna-strandness RF -S NC1.sam>NC1.mapping.txt 2>&1

```
@SQ     SN:chr8 LN:145138636
@SQ     SN:chr9 LN:138394717
@SQ     SN:chrM LN:16569
@SQ     SN:chrX LN:156040895
@SQ     SN:chrY LN:57227415
@PG     ID:hisat2       PN:hisat2       VN:2.2.1        CL:"/home/vboxuser/miniconda3/envs/jimmy3/bin/hisat2-align-s --wrapper basic-0 -x
/home/vboxuser/share/database/human/genome/genome -p 4 -S NC1.sam --read-lengths 151 -1 /tmp/3435.inpipe1 -2 /tmp/3435.inpipe2"
LH00157:209:22K3W3LT3:7:1101:32076:1048 99      chr8    22630528        60      151M    =       22630646        269     GNGGATTGGAGAGTAAGT
CCAAGGATATCTTCACGGCAGATTTTGACATGGCTGTGGGAAGCCAAAGCTCGGTGAACCTTCTATGAAGGGGCAGGTTTCACGGCCTTCTCTCCAGGACCCCGTCCTCCTATGCCAGGGGCCTGCACCCTGA I#I9
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIII AS:i:-1 ZS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1 MD:Z:1G149       YS:i:0  YT:Z:CP NH:i:1
LH00157:209:22K3W3LT3:7:1101:32076:1048 147     chr8    22630646        60      151M    =       22630528        -269    CCCGTCCTCCTATGCCAG
GGGCCTGCACCCTGAAGACCTCTTCCCACAGCCACGGCCGTCAAGAACGGCGAAGTACCACACAACCACAAGCTGCTGATCCCCATGGTTTAATGTTCATTCCTGACCCTAACACTGCTCCAGACAATTCTTT IIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIII9IIIIIIII9III
IIII9I9II AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:151YS:i:-1 YT:Z:CP NH:i:1
```

链方向

第一种表示法	第二种表示法(PE)	第二种表示法(SE)	建库方法
fr–unstranded	无	无	Standard Illumina
fr–firststrand	RF	R	dUTP, NSR, NNSR
fr–secondstrand	FR	F	Ligation, Standard SOLiD

```
(jimmy3) vboxuser@ubuntu:~/share/fastq$ cat NC1.mapping.txt
25140686 reads; of these:
  25140686 (100.00%) were paired; of these:
    4589150 (18.25%) aligned concordantly 0 times
    20020738 (79.63%) aligned concordantly exactly 1 time
    530798 (2.11%) aligned concordantly >1 times
    ----
    4589150 pairs aligned concordantly 0 times; of these:
      345136 (7.52%) aligned discordantly 1 time
    ----
    4244014 pairs aligned 0 times concordantly or discordantly; of these:
      8488028 mates make up the pairs; of these:
        6896351 (81.25%) aligned 0 times
        1507261 (17.76%) aligned exactly 1 time
        84416 (0.99%) aligned >1 times
86.28% overall alignment rate
```

Program	Run time (min)	Memory usage (GB)
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
OLego	989.5	3.7
TopHat2	1,170	4.3

Run times and memory usage for HISAT and other spliced aligners to align 109 million 101-bp RNA-seq reads from a lung fibroblast data set. We used three CPU cores to run the programs on a Mac Pro with a 3.7 GHz Quad-Core Intel Xeon E5 processor and 64 GB of RAM.

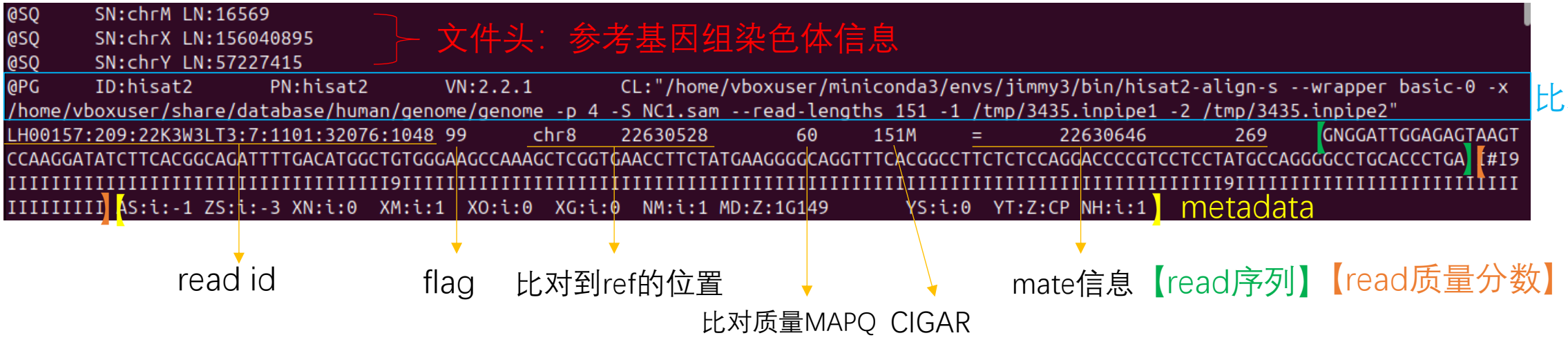
→ 若比对率低，加污染检测

fastq_screen质控（污染检测）

File	Reads processed	Unmapped	One hit / one genome	Multiple hits / one genome	One hit / multiple genomes	Multiple hits / multiple genomes
Human	9,130,699	322,900	2,496,435	1,827,618	479,466	4,004,280
Mouse	9,130,699	5,611,796	483	89	920,472	2,597,859
Rat	9,130,699	5,375,358	808	355	603,664	3,150,514
Drosophila	9,130,699	8,520,285	9	15	12,012	598,378
Worm	9,130,699	8,665,255	23	26	280,074	185,321
Yeast	9,130,699	8,460,966	2	22	886	668,823
Arabidopsis	9,130,699	8,367,130	18	3,678	40,196	719,677
Ecoli_K-12_U00,096	9,130,699	9,124,247	0	0	52	6,400
rRNA	9,130,699	6,747,629	12	0	351,805	2,031,253
MT	9,130,699	8,642,942	3	0	424,719	63,035
PhiX	9,130,699	9,130,677	21	1	0	0
Lambda	9,130,699	9,130,698	0	0	1	0
Vectors	9,130,699	9,091,843	16,890	7,319	2,452	12,195
Adapters	9,130,699	9,130,449	5	2	141	102
支原体 Mycoplasma_sp	9,130,699	9,096,468	28,033	0	5,986	212
AP023,235_Escherichia_coli	9,130,699	9,124,245	0	0	54	6,400
Escherichia_coli_YJ6	9,130,699	9,124,245	0	0	54	6,400

sam → bam → sorted bam

- samtools view -bS NC1.sam >NC1.bam
- samtools sort -@ 2 NC1.bam -o NC1.sort.bam
- samtools index NC1.sort.bam




```
LH00157:209:22K3W3LT3:7:1101:15886:1064 77      *       0       0       *       *0      0       CNTCATTTTGATTTTTTCTCCCCAAATAGTTGATAGCAATAA
TCAATACCAGTTGAGATACCAAAGAGGCAGCCTTTCCCTGATTCCCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGCTTGTATCTAGGGGGCCGGGGTGTG I#III9IIIIIIIIIIIIIIIIIIII
IIIIIIIIIII-IIII9IIII-IIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIII-9IIIIIIIIIIIIIIIIIIII-IIIIIII9IIIIIIIIII-III---I-9I------9YT:Z:UP
LH00157:209:22K3W3LT3:7:1101:15886:1064 141     *       0       0       *       *0      0       TGGGAATCAGGGAAAGGCTGCCTCTTTGGTATCTCAACTGGT
ATTGATTATTGCTATCAACTATTTGGGGAGAAAAAATCAAAATGAAGAGATCGGAAGAGCGTCGTGTAGGGAAAAAGTGTCCTCCAATGTGTAGCTCCCGGTGGTCACC I9IIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIIIIIIIIIIIIIIIIII--III-II9IIIII9III-9I9-I-II---9---I9-9-----YT:Z:UP
```

featureCounts定量

名称		修改日期	类型	大小
NC1_R1.fastq.gz	fastq	2024/4/14 22:14	WinRAR 压缩文...	1,623,317 KB
NC1_R2.fastq.gz		2024/4/14 22:15	WinRAR 压缩文...	1,688,952 KB
NC1_R1_fastqc.html	fastqc	2024/10/11 21:28	HTML 文档	568 KB
NC1_R2_fastqc.html		2024/10/11 21:28	HTML 文档	571 KB
fastp.json		2024/10/12 0:15	JSON 文件	138 KB
NC1.fastp.html		2024/10/12 0:15	HTML 文档	483 KB
NC1_R1.clean.gz	clean	2024/10/12 0:15	WinRAR 压缩文...	1,618,941 KB
NC1_R2.clean.gz		2024/10/12 0:15	WinRAR 压缩文...	1,670,873 KB
NC1.mapping.txt		2024/10/12 1:41	TXT 文件	1 KB
NC1.sam	sam	2024/10/12 1:41	SAM 文件	22,919,196 KB
NC1.bam	bam	2024/10/12 10:01	BAM 文件	4,292,532 KB
NC1.sort.bam	sort.bam	2024/10/12 10:34	BAM 文件	2,353,760 KB
NC1.sort.bam.bai		2024/10/12 10:47	BAI 文件	2,679 KB

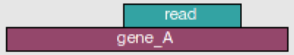
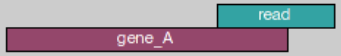
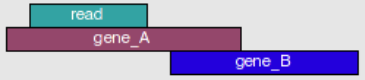
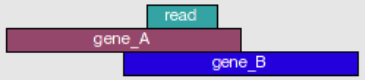
```
(jimmy3) vboxuser@ubuntu:~/share/fastq$ featureCounts -a  /home/vboxuser/share/database/human/genome/Homo_sapiens.GRCh38.112.chr.gtf -p -s 2 -B -C -T  2 -g gene_id -o all.featurecount.mRNA.txt *.sort.bam

        ==========     _____ _    _ ____  _____  ______          _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
            ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
                ====  ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
          v2.0.1

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 2 BAM files                                      ||
||                           o NC1.sort.bam                                   ||
||                           o OE1.sort.bam                                   ||
||                                                                            ||
||             Output file : all.featurecount.mRNA.txt                        ||
```

featureCounts -a /home/vboxuser/share/database/human/genome/Homo_sapiens.GRCh38.112.chr.gtf -p -s 2 -B -C -T 2 -g gene_id -o featurecount.mRNA.txt NC1.sort.bam　　*.sort.bam 全部样品一起跑

-a GTF注释文件，通过这个文件才能区分出哪些区域为外显子
-p 计算fragment（双端），而非reads（单端）
-s 链特异性测序
-B 两端都比对上的reads才计算
-T 线程数
-g 指定注释文件中属性信息，默认为gene_id
-o 输出结果文件，同样会输出一个以该名字结尾的.summary统计文件;

Method	Number of fragments	Time (min)	Memory (MB)
featureCounts	5 392 155	0.9	4
CountOverlaps (whole genome at once)	5 392 155	24.4	7000
CountOverlaps (by chromosome)	5 392 155	36.6	783
htseq-count (union)	4 978 050	36.0	31
htseq-count (intersection-nonempty)	4 993 644	35.7	31
coverageBED	5 366 902	4.4	41

定量的复杂性

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

gene_id

	A	B	C	D	E	F	G	H
1	# Program:featureCounts v2.0.1; Command:"featureCounts" "-a" "/home/vboxuser/share/data/							
2	Geneid	Chr	Start	End	Strand	Length	NC1.sort.bam	OE1.sort.bam
3	ENSG0000	chr1;chr1;	2581560;2	2581650;2	+;+;+	626	1	33
4	ENSG0000	chr1;chr1;	3069168;3	3069296;3	+;+;+;+;+	9767	0	0
5	ENSG0000	chr1;chr1;	5301928;5	5302004;5	-;-;-;-	1225	0	0
6	ENSG0000	chr1;chr1;	2403964;2	2405834;2	-;-;-;-;-;-;	3782	503	143
7	ENSG0000	chr1	5492978	5494674	+	1697	0	0
8	ENSG0000	chr1	10054445	10054781	-	337	0	0
9	ENSG0000	chr1;chr1;	3944547;3	3945141;3	-;-;-;-;-;-;	2168	0	0
10	ENSG0000	chr1	4175528	4175899	-	372	0	0
11	ENSG0000	chr1;chr1;	10472288;	10472580;	+;+;+;+;+	2932	407	570

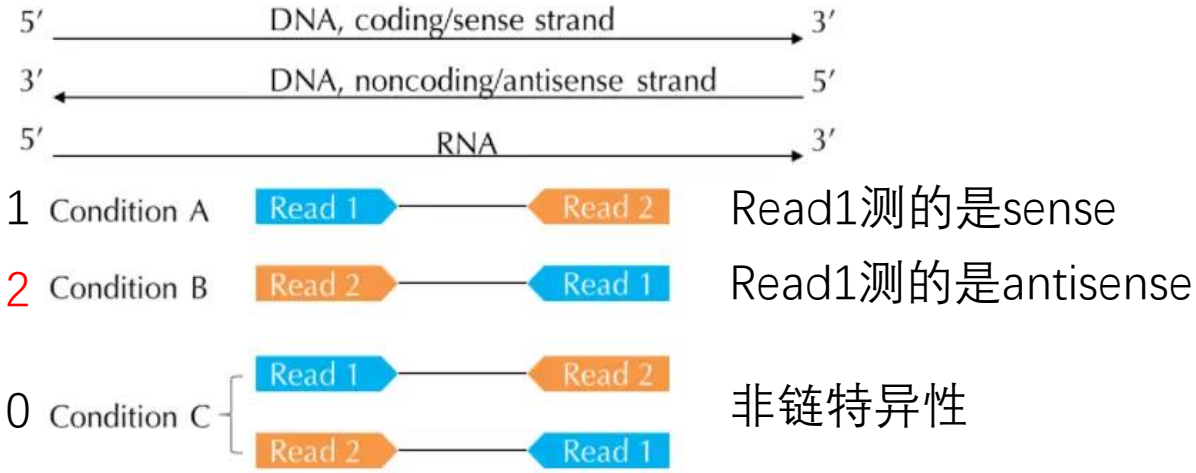
运行的命令

Raw count

链特异性建库 （-s参数）

pip install RSeQC
conda install -c bioconda ucsc-gtftogenepred
gtfToGenePred -genePredExt -geneNameAsName2 /home/vboxuser/share/database/human/genome/Homo_sapiens.GRCh38.112.chr.gtf gene.tmp
awk '{print $2"\t"$4"\t"$5"\t"$1"\t0\t"$3"\t"$6"\t"$7"\t0\t"$8"\t"$9"\t"$10}' gene.tmp > genes_refseq.bed12
infer_experiment.py -r genes_refseq.bed12 -i NC1.bam

```
(jimmy3) vboxuser@ubuntu:~/share/fastq$ infer_experiment.py -r genes_refseq.bed
2 -i NC1.bam

Reading reference gene model genes_refseq.bed12 ... Done
Loading SAM/BAM file ...  Total 200000 usable reads were sampled


This is PairEnd Data
Fraction of reads failed to determine: 0.1445
Fraction of reads explained by "1++,1--,2+-,2-+": 0.0229
Fraction of reads explained by "1+-,1-+,2++,2--": 0.8326
```



Read1测的是sense

Read1测的是antisense

非链特异性

1++,1--,2+-,2-+： read1比对上的链和基因所在的链一样，read2比对上的链和基因所在的链是相反的。SOLiD测序方法
1+-,1-+,2++,2--： read1比对上的链和基因所在的链是相反的，read2比对上的链和基因所在的链一样。Illumina Truseq 方法

https://www.jianshu.com/p/670cd485d673
https://www.jianshu.com/p/109997345a67

Count，FPKM，TPM

statquest视频

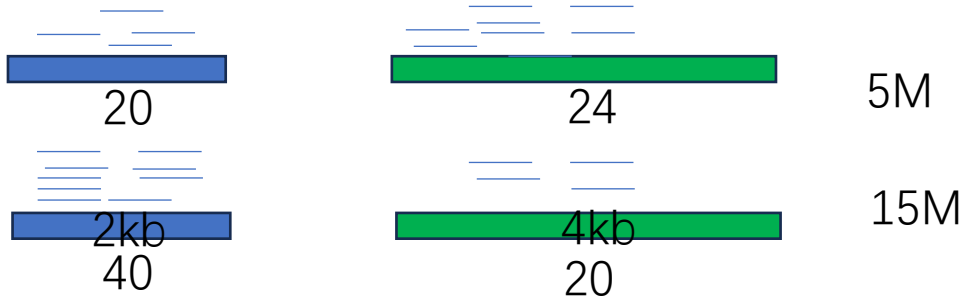
How NCBI generates RNA-seq count data

Briefly, SRA runs where the organism is Homo sapiens and type is Transcriptomic are aligned to genome assembly GCA_000001405.15 using HISAT2. Runs that pass a 50% alignment rate are further processed with Subread featureCounts which outputs a raw count file for each run. For Human data, the *Homo sapiens* Annotation Release 109.20190905 was used for gene annotation. GEO further processes these SRR raw count files into GEO Series raw counts matrices. Data derived from single cell samples are skipped. In cases where there is more than one SRA run per GEO Sample, the raw counts are summed. Values in the raw count matrices are rounded so that they are compatible input for common differential expression analysis software. Using the raw counts as input, GEO then computes FPKM(RPKM) and TPM normalized values.

GSE164073_raw_counts_GRCh38.p13_NCBI.tsv

	GeneID	GSM4996084	GSM4996085	GSM4996086	GSM4996087			
2	100287102	2	5	3	2	5	2	4
3	653635	244	236	337	266	317	226	303	196
4	102466751	25	17	34	22	24	19	23
5	107985730	1	1	1	0	0	0	1
6	100302278	0	0	0	0	0	0	0
7	645520	0	0	0	0	1	0	0	0
8	79501	0	0	0	0	0	0	0	0
9	100996442	39	38	59	35	61	37	25
10	729737	51	52	111	73	125	46	51	52

fpkm = 20/(2*5)=2　　fpkm = 24/(4*5)=1.2



fpkm = 40/(2*15)=1.3　　fpkm = 20/(4*15)=0.33

gene1/(gene1+gene2+…+geneN)*1000000

tpm是fpkm的占比

TPM相当于重新标准化的文库，保证每个样本中所有TPM的总和相同，即1000000

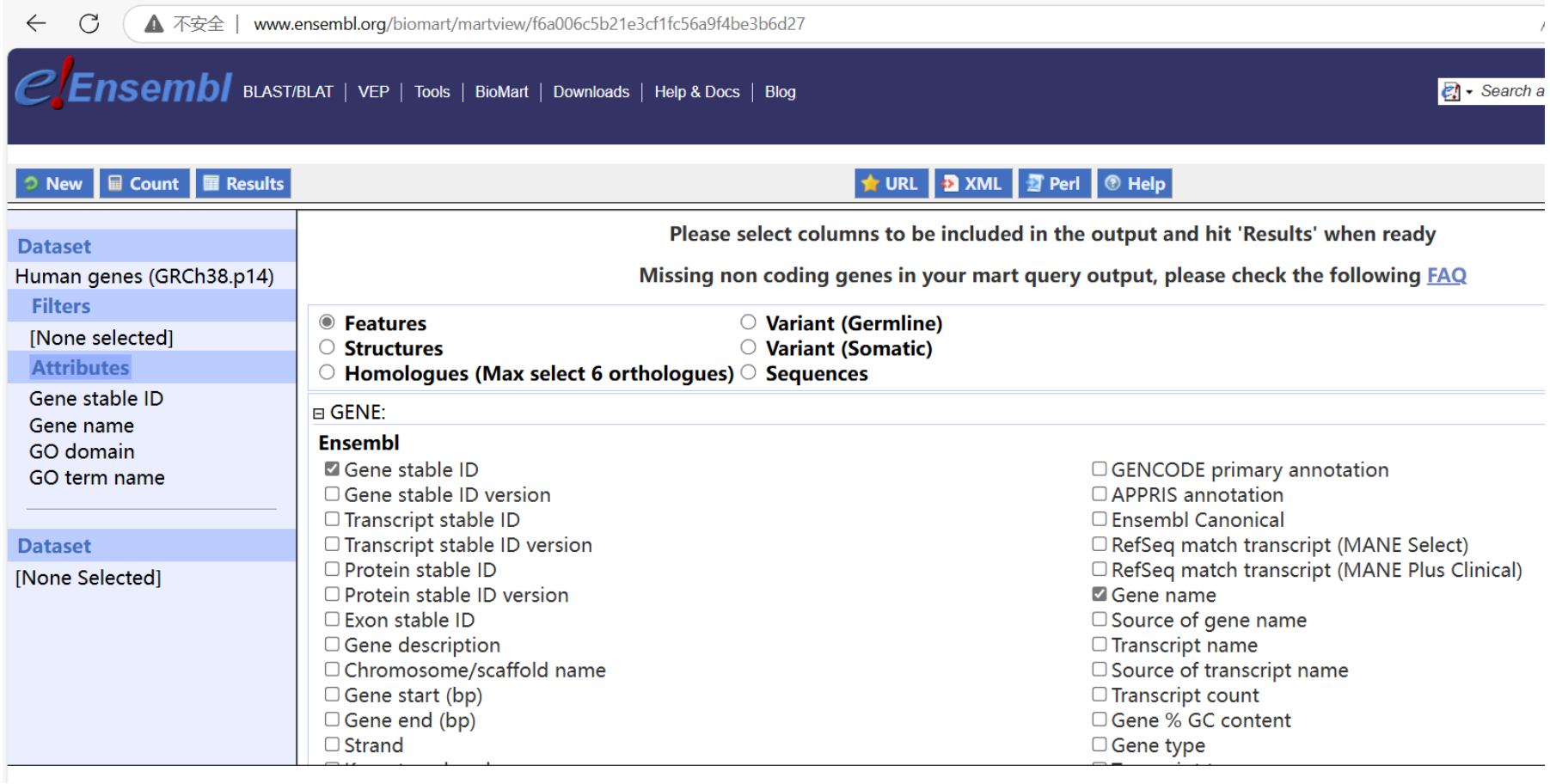
Count转FPKM、TPM

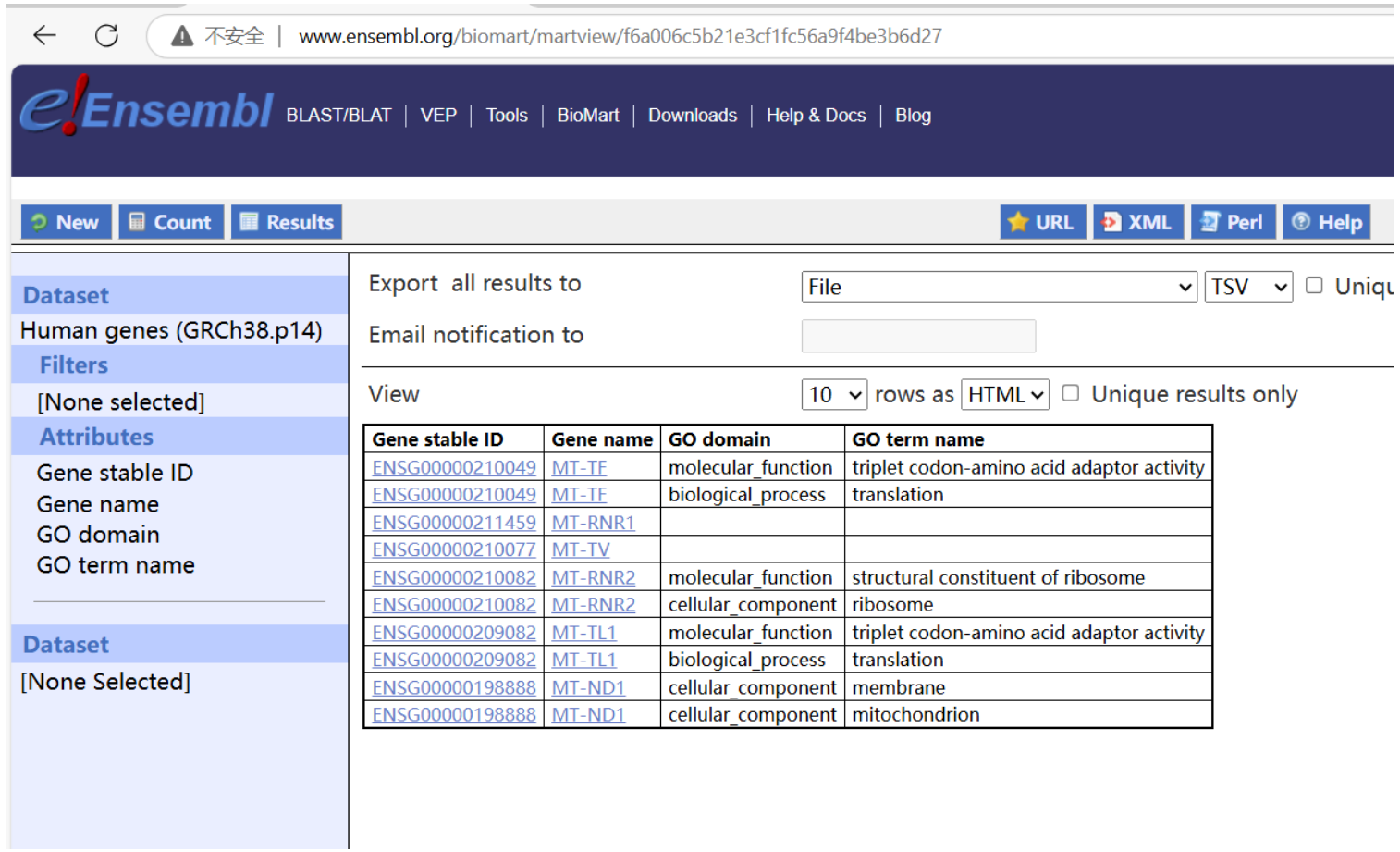
- 安装：pip install rnanorm
- 行为样品（需要转置）
- rnanorm tpm count.txt --gtf ensembl.gtf > tpm.txt
- rnanorm fpkm count.txt --gtf ensembl.gtf > fpkm.txt

答疑：两家公司的fpkm为何有差异?
注释版本，基因长度等可能不完全相同

	A	B	C	D	E
1	gene	mean	median	longest_isoform	merged
2	ENSG00000242268.3	1709	1709	2708	2750
3	ENSG00000270112.4	1924	1875	4685	10683
4	ENSG00000280143.1	5326	5326	5326	5326
5	ENSG00000146083.12	1239	816	4122	5627
6	ENSG00000263642.1	80	80	80	80
7	ENSG00000225275.4	282	282	282	282
8	ENSG00000158486.13	3769	2362	12394	14960

Ensembl ID转symbol，加各种注释





Gene stable ID	Gene name	GO domain	GO term name
ENSG00000210049	MT-TF	molecular_function	triplet codon-amino acid adaptor activity
ENSG00000210049	MT-TF	biological_process	translation
ENSG00000211459	MT-RNR1		
ENSG00000210077	MT-TV		
ENSG00000210082	MT-RNR2	molecular_function	structural constituent of ribosome
ENSG00000210082	MT-RNR2	cellular_component	ribosome
ENSG00000209082	MT-TL1	molecular_function	triplet codon-amino acid adaptor activity
ENSG00000209082	MT-TL1	biological_process	translation
ENSG00000198888	MT-ND1	cellular_component	membrane
ENSG00000198888	MT-ND1	cellular_component	mitochondrion

DESeq2标准化和差异分析

DESeq2.r

```r
# 参考ncbi GEO2R
# DESeq2官网

############################ 简单版，不过滤
library(DESeq2)

# 读取表达矩阵
rawcount <- read.table("DESeq2_input.txt", header=T, row.names=1, sep="\t", quote="")

# 给样品分配组名，不要带特殊符号
condition <- factor(c("mock","mock","mock","cov2","cov2","cov2"))
sample_info <- data.frame(Group = condition, row.names = colnames(rawcount))

# 构建DESeq数据集
ds <- DESeqDataSetFromMatrix(countData=rawcount, colData=sample_info, design= ~Group)

# 建模，计算，具体参数 ?DESeq2
ds <- DESeq(ds)

# 比较并导出结果 #Group名字必需跟12行的保持一致
de <- results(ds, contrast=c("Group", "cov2", "mock"))

# 读取注释并添加注释
annodata = as.matrix(data.table::fread("Human.GRCh38.p13.annot.tsv", header=T))
merged_columns <- cbind(rawcount, de, counts(ds, normalized=T), annodata)

# 合并的结果写到txt里边
write.table(merged_columns, file="out2.txt", row.names=T, sep="\t", quote = FALSE)
```

每个基因在每个样品中的原始counts

唯一id

	A	B	C	D	E	F	G
1	GeneID	GSM4996084	GSM4996085	GSM4996086	GSM4996087	GSM4996088	GSM4996089
2	100287102	2	5	3	2	5	2
3	653635	244	236	337	266	317	226
4	102466751	25	17	34	22	24	19
5	107985730	1	1	1	0	0	0
6	100302278	0	0	0	0	0	0
7	645520	0	0	0	0	1	0
8	79501	0	0	0	0	0	0

3个mock　3个cov2

比较方式：
实验组在前，对照组在后

唯一id

GeneID	Symbol	Descript	Synonyms	GeneType	Ensembl(	Status	ChrAcc	ChrStart	ChrStop
1E+08	DDX11L1	DEAD/H-box helica	pseudo	ENSG0000(	active	NC_00000	11874	14409	
653635	WASH7P	WASP fam	FAM39F|W	pseudo		active	NC_00000	14362	29370
1.02E+08	MIR6859-	microRNA	hsa-mir-	ncRNA	ENSG0000(	active	NC_00000	17369	17436
1.08E+08	MIR1302-	MIR1302-2 host ge		ncRNA		active	NC_00000	29926	31295
1E+08	MIR1302-	microRNA	MIRN1302-	ncRNA	ENSG0000(	active	NC_00000	30366	30503
645520	FAM138A	family w	F379|FAM	ncRNA	ENSG0000(	active	NC_00000	34611	36081
79501	OR4F5	olfactory receptor		protein	ENSG0000(	active	NC_00000	69091	70008

唯一id　差异　Raw　Normalized　Annotations

其他问题

- 倍数变化
  - 上调为正，下调为负
  - 上下调都是正，上调>1，0<下调<1。不直观：例如0.154倍，需要取倒数才能知道大概多少倍
- log2FC
  - 上调为正，下调为负
- P值还是padj
  - 建议用padj，但是若padj筛选后差异太少，只能用p了
- NA
  - P的NA转成1画图

实操作业

- 1，从GEO下载3 vs 3的数据
- 2，SRA转成fastq
- 3，fastqc质控
- 4，fastp去接头
- 5，hisat2比对
- 6，featurecounts定量
- 7，DESeq2差异分析，并与GEO2R结果比较
- 8，PCA，heatmap，火山图
- 9，GO，Pathway分析
- 10，GSEA分析
- 11，PPI
- 12，微生信云平台绘图，AI编辑和拼图