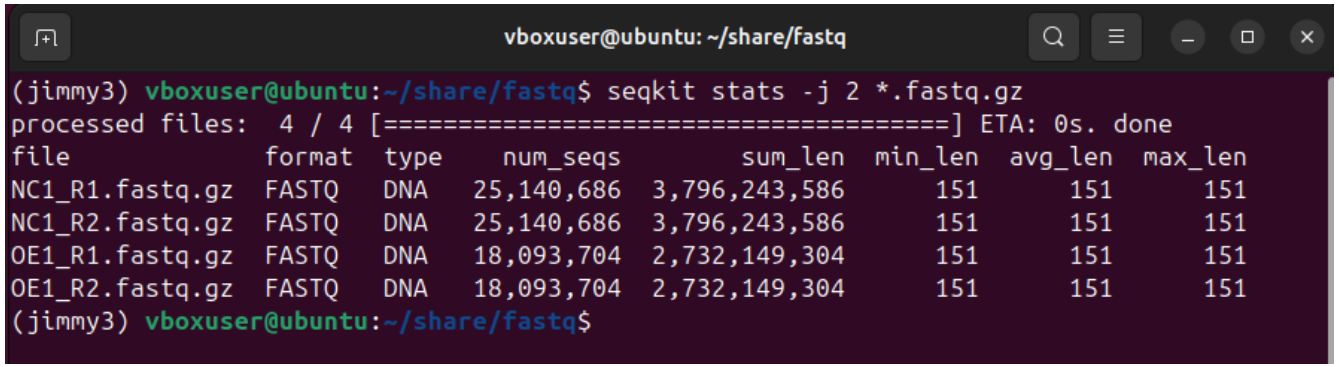
测序下机数据fastq文件介绍

陈明杰 202410

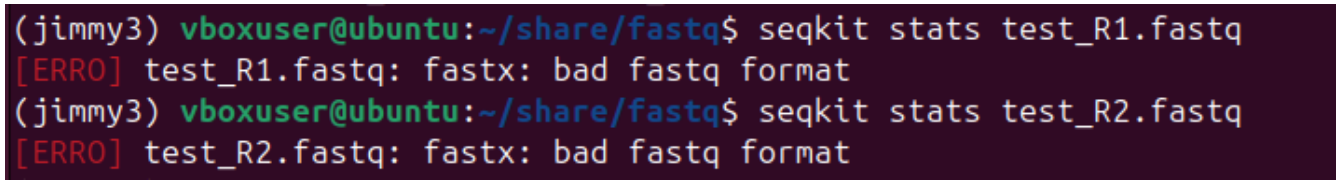
拿到数据第一步：FASTQ文件校验

- seqkit stats -j 2 *.fastq.gz

https://bioinf.shenwei.me/seqkit/
下载，直接放在/usr/local/bin里边，赋予权限
chmod 777 seqkit



校验成功
文件OK



校验失败
文件损坏 → 第一时间问测序公司要数据！！！

实操：文件校验

Fastq文件大小

SRX25765751: PAR-CLIPseq for tG3BP1-259-466 binding RNA
1 ILLUMINA (Illumina HiSeq 1500) run: 12.4M spots, 619.4M bases, 184.3Mb downloads

Design: N/A

Submitted by: Ruijin Hospital Affiliated to Shanghai Jiao Tong university school of medicine

Study: PAR-CLIPseq for C-terminal truncated G3BP1 binding RNA
PRJNA1149923 • SRP527431 • All experiments • All runs
show Abstract

Sample: overexpressing tG3BP1-259-466-Flag
SAMN43265621 • SRS22402420 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: SZG2021-G3BP1-2
Instrument: Illumina HiSeq 1500
Strategy: RIP-Seq
Source: TRANSCRIPTOMIC
Selection: RANDOM PCR
Layout: SINGLE

Runs: 1 run, 12.4M spots, 619.4M bases, 184.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR30305312	12,388,794	619.4M	184.3Mb	2024-08-20

gz压缩包大小：fastq大小约1比4+

- M reads（M表示百万条）
- G 数据量（碱基数：1G=1,000,000,000碱基）
- Gb，Mb：（文件大小：1Gb=1024Mb)

byte

提问：不同测序类型的测序量

Fastq格式

Read 1

xx_R1.fastq

```
1  @LH00348:184:22CLJHLT4:3:1101:2400:1042 1:N:0:GCGGTATT+CCAATCAC
2  ANCGGAACCACAAGAAGGTTTCATCCTTGAGGACACACCAGCGGCGTTTCCAAGAGTTCATCAGGCCACCT
3  +
4  I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
5  @LH00348:184:22CLJHLT4:3:1101:7611:1042 1:N:0:GCGGTATT+CCAATCAC
6  CNAAGATCAAAAAGAAAGAAGACTCTAATAATATCAGAAATGAAAGAGGAGACATTATAGCTGATACCAAA
7  +
8  I#III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
9  @LH00348:184:22CLJHLT4:3:1101:9214:1042 1:N:0:GCGGTATT+CCAATCAC
10 GNCAAACACTTTGAATGCATTGAGTATGGTTTCCTCAGGGTCTGCTCCCTTAAGTTTCTCCCCAAACATTG
11 +
12 I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
13 @LH00348:184:22CLJHLT4:3:1101:11010:1042 1:N:0:GCGGTATT+CCAATCAC
14 TNTCCACCGCGGCCCGCCTGTCTTCCAGGATGTCCTTCCGGCTGTTCCAGGACTCGGCGTCAGGCCGCCCC
15 +
16 I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Read 2

xx_R2.fastq

```
1  @LH00348:184:22CLJHLT4:3:1101:2400:1042 2:N:0:GCGGTATT+CCAATCAC
2  AAGCAGTGGTATCAACGCAGAGTACGGGGGCGTGACTCTGTGTACAGCTGTGTCACTCTGCCATATTTCCACAC
3  +
4  IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
5  @LH00348:184:22CLJHLT4:3:1101:7611:1042 2:N:0:GCGGTATT+CCAATCAC
6  AAGCAGTGGTATCAACGCAGAGTACGGGGGTTTTATGTTTTAGTTTTATGTGTCTAAGAATGTATCCATTTCTT
7  +
8  IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9I9IIIIIIIIIIIIIII
9  @LH00348:184:22CLJHLT4:3:1101:9214:1042 2:N:0:GCGGTATT+CCAATCAC
10 ATGGACCAGAACAGGGATGGCTTCATCGACAAGAACGATCTGAGAGACACCTTTGCTGCCCTTGGGCGTGTGAA
11 +
12 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
13 @LH00348:184:22CLJHLT4:3:1101:11010:1042 2:N:0:GCGGTATT+CCAATCAC
14 GTTTCTTGAAGCAGGATAAGTATGAGTGTCATTTCTTCAACGGGACCGAGCGGGTGCGGTTCCTGCACAGAGAC
15 +
16 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Illumina平台，双端：read1，read2。fastq文件每4行代表一条read

- 第1行：@开头，测序所用的仪器， run号，flow cell编号，lane编号，cluster编号，cluster上的X，Y坐标，index区分样品（拆样）

  @LH00348:184:22CLJHLT4:3:1101:2400:1042 1:N:0:GCGGTATT+CCAATCAC

  华大平台，SRA等信息更少
- 第2行：测序的序列信息，以ATCGN表示，由于荧光信号干扰无法判断是什么碱基时就用N表示
- 第3行：+（古老的文件会携带第一行的信息，现在就是一个+）
- 第4行：与第二行碱基信息一一对应，存储测序碱基的质量值（ASCII码表示。为Q20，Q30来源）

  不同测序仪用的字符集不同，所以可以用来大概判断使用的测序仪

实操：解压fastq.gz，截取10行

避坑：判断你的一批数据是怎么测的？

质量分数Q计算方法

- Q=−10 log10(P)
- P是碱基识别的错误概率，来自碱基识别算法（base calling algorithm）并依赖于多少信号被捕获

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

ASCII码表

ASCII 值	控制字符	ASCII 值	控制字符	ASCII 值	控制字符	ASCII 值	控制字符
0	NUT	32	(space)	64	@	96	、
1	SOH	33	!	65	A	97	a
2	STX	34	”	66	B	98	b
3	ETX	35	#	67	C	99	c
4	EOT	36	$	68	D	100	d
5	ENQ	37	%	69	E	101	e
6	ACK	38	&	70	F	102	f
7	BEL	39	,	71	G	103	g
8	BS	40	(	72	H	104	h
9	HT	41	)	73	I	105	i
10	LF	42	*	74	J	106	j
11	VT	43	+	75	K	107	k
12	FF	44	,	76	L	108	l
13	CR	45	-	77	M	109	m
14	SO	46	.	78	N	110	n
15	SI	47	/	79	O	111	o
16	DLE	48	0	80	P	112	p
17	DCI	49	1	81	Q	113	q
18	DC2	50	2	82	R	114	r
19	DC3	51	3	83	X	115	s
20	DC4	52	4	84	T	116	t
21	NAK	53	5	85	U	117	u
22	SYN	54	6	86	V	118	v
23	TB	55	7	87	W	119	w
24	CAN	56	8	88	X	120	x
25	EM	57	9	89	Y	121	y
26	SUB	58	:	90	Z	122	z
27	ESC	59	;	91	[	123	{
28	FS	60	<	92	/	124	|
29	GS	61	=	93	]	125	}
30	RS	62	>	94	^	126	~
31	US	63	?	95	—	127	DEL

实操：编程语言将字母转成ASCII码

提问：为什么不直接用数字表示碱基质量

GEO下载原始数据

- sratoolkit下载（https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit）
- prefetch -X 200G SRR1234656 -o SRR1234656
- fastq-dump --split-files -F SRR1234656

实操：SRA文件下载与格式转换

Illumina vs 华大

xx_R1.fastq

```
1  @LH00348:184:22CLJHLT4:3:1101:2400:1042 1:N:0:GCGGTATT+CCAATCAC
2  ANCGGAACCACAAGAAGGTTTCATCCTTGAGGACACACCAGCGGCGTTTCCAAGAGTTCATCAGGCCACCT
3  +
4  I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
5  @LH00348:184:22CLJHLT4:3:1101:7611:1042 1:N:0:GCGGTATT+CCAATCAC
6  CNAAGATCAAAAAGAAAGAAGACTCTAATAATATCAGAAATGAAAGAGGAGACATTATAGCTGATACCAAA
7  +
8  I#III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
9  @LH00348:184:22CLJHLT4:3:1101:9214:1042 1:N:0:GCGGTATT+CCAATCAC
10 GNCAAACACTTTGAATGCATTGAGTATGGTTTCCTCAGGGTCTGCTCCCTTAAGTTTCTCCCCAAACATTG
11 +
12 I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
13 @LH00348:184:22CLJHLT4:3:1101:11010:1042 1:N:0:GCGGTATT+CCAATCAC
14 TNTCCACCGCGGCCCGCCTGTCTTCCAGGATGTCCTTCCGGCTGTTCCAGGACTCGGCGTCAGGCCGCCCC
15 +
16 I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

huada_R1.fastq

```
1  @E250026663L1C001R00100000637/1
2  CGAGAACGTATTCACCGTAGCGTAGCTGATCTACGATTACTAG
3  +
4  DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
5  @E250026663L1C001R00100000646/1
6  CGGATGACTTGTGAATAGCGGTGAAATTCCAATCGAACCTGGA
7  +
8  DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
9  @E250026663L1C001R00100000797/1
10 GGGGTTGCGCTCGTTGCAGGACTTAACCGAACATCTCACGACA
11 +
12 DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
13 @E250026663L1C001R00100000848/1
14 TCGCCATATTGCTTCTCTTTGTACCGACCATTGTAGCACGTGT
15 +
16 DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
```

项目	Illumina	华大
期刊接受度	*****	****
测序价格	*****	***

避坑：用华大平台测序，收Illumina平台价格