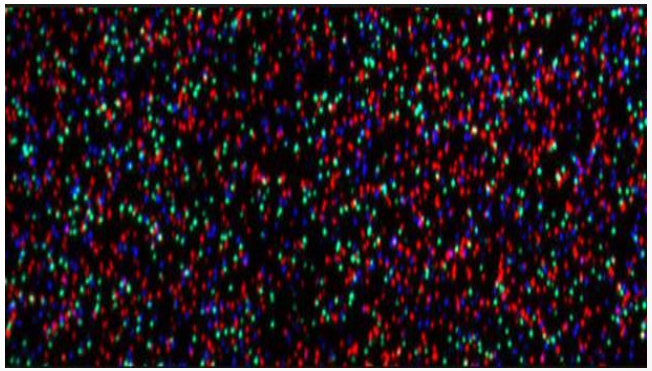
测序数据预处理

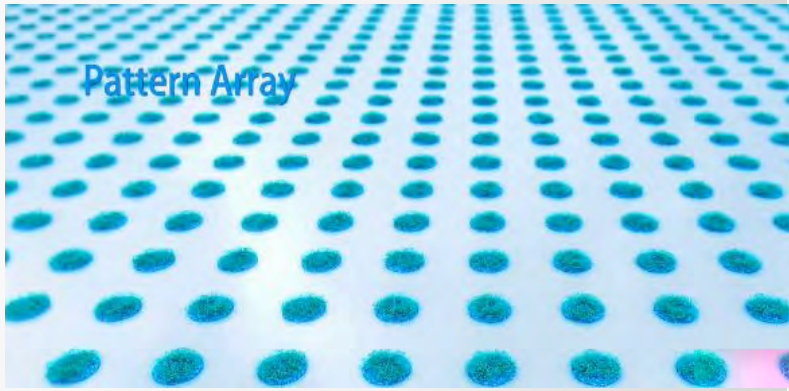
下机数据质量评价

VS

每个点包含的信息	下机文件对应结果
Coordinate 坐标	序列ID(Line 1)
Color 颜色	碱基序列(Line 2)
Intensity profiles 光强	碱基质量值(Line 4)
Signal to noise ratio信噪比	碱基质量值(Line 4)

术语
Read(s)　　sequence(s)





Reading photosignal sequences

文件格式.fastq，简称fq

```
@B819B9ABXX:7:1101:1175:2095#ACAACCCGA/2
NCTTCTGAAACTATTCCAAACAACAGAAAAAGAGGGACTCCTCCCTAACTCATTTTAT
+
BXUUUXVUVXccccccccc_cZXXccaccXUUUY_c_ccccccXX_c_C_cC__ccX
```

```
@CL100030398L1C001R001_0/1
GTACACTATGAGTACATCGTATGTGTAAATATGCATAATAAATGGAATAC
+
FFFFDDFFFF:@DF<DFACEFFFFFFFCECEE7BFFFFEFEFF>CDFBFF
```

下机数据质量评价：二代测序数据格式介绍

DNBSEQ PE测序数据

CL100030398_L01_read_1.fq.gz:

```
@CL100030398L1C001R001_0/1
GTACACTATGAGTACATCGTATGTGTAAATATGCATAATAAATGGAATAC
+
FFFFDDFFFF:@DF<DFACEFFFFFFFCECEE7BFFFFEFEFF>CDFBFF
```

CL100030398_L01_read_2.fq.gz:

```
@CL100030398L1C001R001_0/2
AGAGTTTTTTCTAAAATTCAGGTCATGATGAGTTCTATAGGTTTTCTTGT
+
F<EE8AC<EFE?DFEFBC>DFB<DFD@EFDDC;DDEE>EC4?CC.AF>49
```

二代测序下机数据fastq, 存储了核酸序列和相应质量值信息, 每条测序结果包含四行:

第一行：序列ID信息，此ID唯一

第二行：read 序列，由ATCGN组成

第三行：“+”

第四行：第二行每个碱基对应的质量值

下机数据质量评价：第四行质量值详解

编码	字符	编码	字符	编码	字符	编码	字符
0	NUL	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	TAB	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	DC1	49	1	81	Q	113	q
18	DC2	50	2	82	R	114	r
19	DC3	51	3	83	S	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	ETB	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

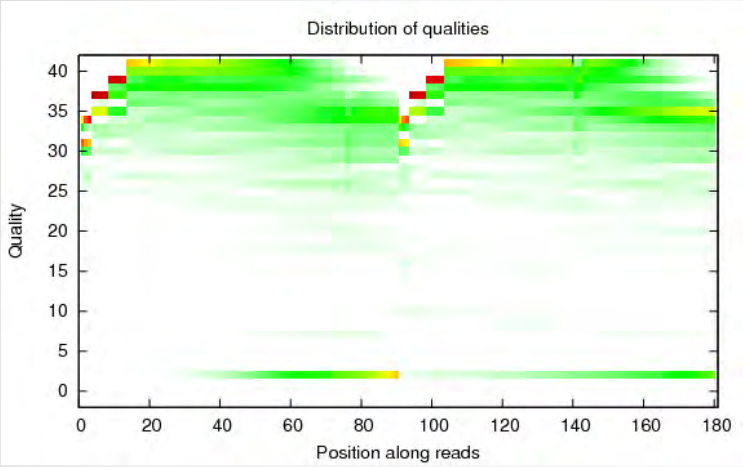
质量值	错误率	准确度
10	0.1	90%
20	0.01	99%
30	0.001	99.9%
40	0.0001	99.99%
50	0.00001	99.999%

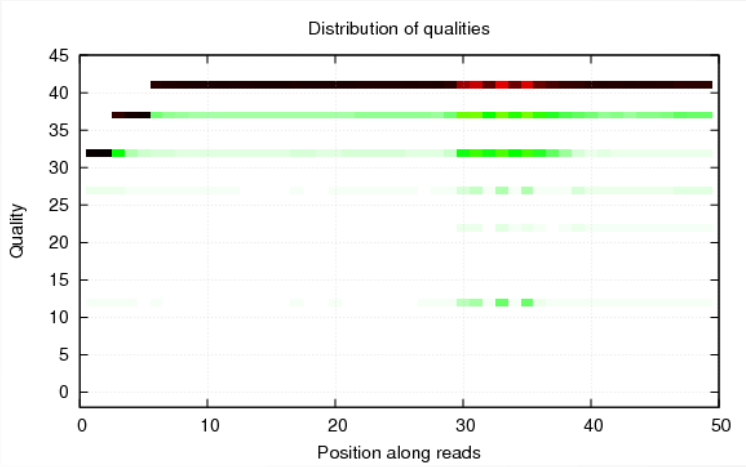
Q=-10logP，其中P代表该碱基被测错的概率。

若P=0.001，则Q=30，那么质量值编码字符为：

30+33=63对应的ASCII码，即“？”

下机数据质量评价






```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&' ()*+,-./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)
```

Table 1: Q Scores Based upon an Optimized 8-level Mapping

Old Quality Score	New Quality score
N (no call)	N (no call)
2–9	6
10–19	15
20–24	22
25–29	27
30–34	33
35–39	37
≥ 40	40

下机数据质量评价

测序技术	错误率	读长	linkage	通量
Sanger	极低，0.01%	不定，~600bp	SE	极低
Roche/454	低，~2%	不定，~400bp	SE，PE	高
Illumina/solexa	低，~1%	定长，25 -300bp	SE，PE	高→极高
ABI/SOLiD	很低，~0.1%	定长，25 -35bp	SE	高
Ion torrent	低，~1%	定长，100 -200bp	SE	高
MGI/CG	低，~0.5%	定长，25 –150bp	SE，PE	高→极高
Helicos	高，未知	不定，20-100bp	SE	低
PacBio	高，~15%	不定，可达40kb	SE	中→高
Nanopore	极高→高，40%→4%	不定，可达10Mb	SE	中→高

测序数据质量评估实践

1. 进入自己工作目录:
$ cd

2. 查看当前文件夹:
$ ls

3. 创建分析流程需要的全部文件夹
$ mkdir 00_ref 01_clean 02_align 03_SNPCalling 04_annotation raw_data

4. 拷贝你的测序原始文件到raw_data文件夹下
$ mv ~/*.fq.gz ~/raw_data

```
LY@ecoli-ThinkCentre-M930q-N000:~$ mv *.fq.gz ./raw_data
LY@ecoli-ThinkCentre-M930q-N000:~$ ll
```

测序数据质量评估实践

5. 将以下路径目录中所有文件（/home/00_ref）拷贝至自己个人的新建目录00_ref下

```
$ cd
$ cd 00_ref
$ cp /home/00_ref/*.* ./
```

```
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ cd
LY@ecoli-ThinkCentre-M930q-N000:~$ cd 00_ref
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ ll
total 8
drwxr-xr-x  2 LY student 4096 10月  24 16:17 ./
drwxr-xr-x 13 LY student 4096 10月  24 16:17 ../
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ cp /home/00_ref/*.* ./
LY@ecoli-ThinkCentre-M930q-N000:~/00_ref$ ll
total 12804
drwxr-xr-x  2 LY student    4096 10月  24 16:28 ./
drwxr-xr-x 13 LY student    4096 10月  24 16:17 ../
-rw-r--r--  1 LY student    2331 10月  24 16:28 alleles_all.vcf.gz
-rw-r--r--  1 LY student    3143 10月  24 16:28 alleles_all.vcf.gz.tbi
-rw-r--r--  1 LY student  448559 10月  24 16:28 MGI358.SNP.fa
-rw-r--r--  1 LY student  891846 10月  24 16:28 MGI358.SNP.fa.0123
-rw-r--r--  1 LY student      13 10月  24 16:28 MGI358.SNP.fa.amb
-rw-r--r--  1 LY student    7087 10月  24 16:28 MGI358.SNP.fa.ann
-rw-r--r--  1 LY student 5351187 10月  24 16:28 MGI358.SNP.fa.bwt.2bit.64
-rw-r--r--  1 LY student 6243027 10月  24 16:28 MGI358.SNP.fa.bwt.8bit.32
-rw-r--r--  1 LY student    6861 10月  24 16:28 MGI358.SNP.fa.fai
-rw-r--r--  1 LY student  111482 10月  24 16:28 MGI358.SNP.fa.pac
-rw-r--r--  1 LY student    4661 10月  24 16:28 target.358.SE50.subSNP.bed
-rw-r--r--  1 LY student    4957 10月  24 16:28 vcf2geno_free.pl
```

测序数据质量评估实践

6. 在集群分析路径下使用ls命令检查数据文件拷贝是否齐全：

$ cd

$ ls raw_data

$ ls 00_ref/

```
LY@ecoli-ThinkCentre-M930q-N000:~$ cd
LY@ecoli-ThinkCentre-M930q-N000:~$ ls raw_data
test.fq.gz
LY@ecoli-ThinkCentre-M930q-N000:~$ ls 00_ref
alleles_all.vcf.gz      MGI358.SNP.fa       MGI358.SNP.fa.amb  MGI358.SNP.fa.bwt.2bit.64  MGI358.SNP.fa.fai  target.358.SE50.subSNP.bed
alleles_all.vcf.gz.tbi  MGI358.SNP.fa.0123  MGI358.SNP.fa.ann  MGI358.SNP.fa.bwt.8bit.32  MGI358.SNP.fa.pac  vcf2geno_free.pl
```

7. 使用以下命令查看下机数据基本格式，查看完毕，单击q即可退出：

$ less ~/raw_data/test.fq.gz

```
@FS2000L1C002R004000021
CGAAAACCTTTTTCCAAGGACAAATCAGAGAAAAAGTCTTTAACTCCACC
+
IIII=CII<IIIIIIIIIIIIIIIHIIIIIIII0<IHIIIIII12IIIID
@FS2000L1C002R004000025
ACTCTATTAGTCAACGTTCTCCAGAGAAACAGAACCAATAGCATAGATAG
+
IIIIII;IIIDIIIIGIIIIIDIHIII>IIIIIIIIIIII'IIIHIIIC
@FS2000L1C002R004000032
TTATGTTCAGCCCTTGGATTGTCTCAGGATGTTGCAGGCGCAACTTTCAT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

测序数据质量评估实践

8. 使用SOAPnuke生成数据质量评估表，并按指定标准对低质量数据进行过滤，向01_clean中生成clean数据：

```
$ /Pipeline/FIS.Traits/tools/SOAPnuke filter -1 ~/raw_data/test.fq.gz -l 10 -q 0.5 -n 0.01 -T 1 -o ~/01_clean/ -C test_clean.fq.gz
$ ls 00_ref/
```

9. 可在01_clean路径下使用less命令逐一查看测序数据质量评估表，我们简单展示过滤前后数据总体情况如下：

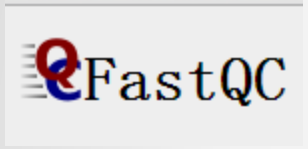
```
$ less ~/01_clean/Basic_Statistics_of_Sequencing_Quality.txt
```

Item	raw reads(fq1)	clean reads(fq1)
Read length	50.0	50.0
Total number of reads	473994 (100.00%)	473598 (100.00%)
Number of filtered reads	396 (0.08%)	-
Total number of bases	23699700 (100.00%)	23679900 (100.00%)
Number of filtered bases	19800 (0.08%)	-
Number of base A	6753420 (28.50%)	6747827 (28.50%)
Number of base C	5643559 (23.81%)	5638836 (23.81%)
Number of base G	4954525 (20.91%)	4950542 (20.91%)
Number of base T	6347790 (26.78%)	6342695 (26.79%)
Number of base N	406 (0.00%)	0 (0.00%)
Q20 number	23171219 (97.77%)	23152679 (97.77%)
Q30 number	22639917 (95.53%)	22622162 (95.53%)

(查看完毕，单击q即可退出)

测序数据质量评估实践

10. 使用FASTQC工具对数据质量情况进行绘图，通过图形化展示了解数据质量：
$ /Pipeline/FIS.Traits/tools/anaconda/bin/fastqc ~/01_clean/test_clean.fq.gz



A quality control tool for high throughput sequence data.

FastQC 简单快速对下机数据进行质控，提供11种质控指标。
可输入BAM/SAM/FASTQ文件格式，直接给出质控结果，
HTML格式报告，一键式操作。

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

测序数据质量评估实践

11. FASTQC结果展示