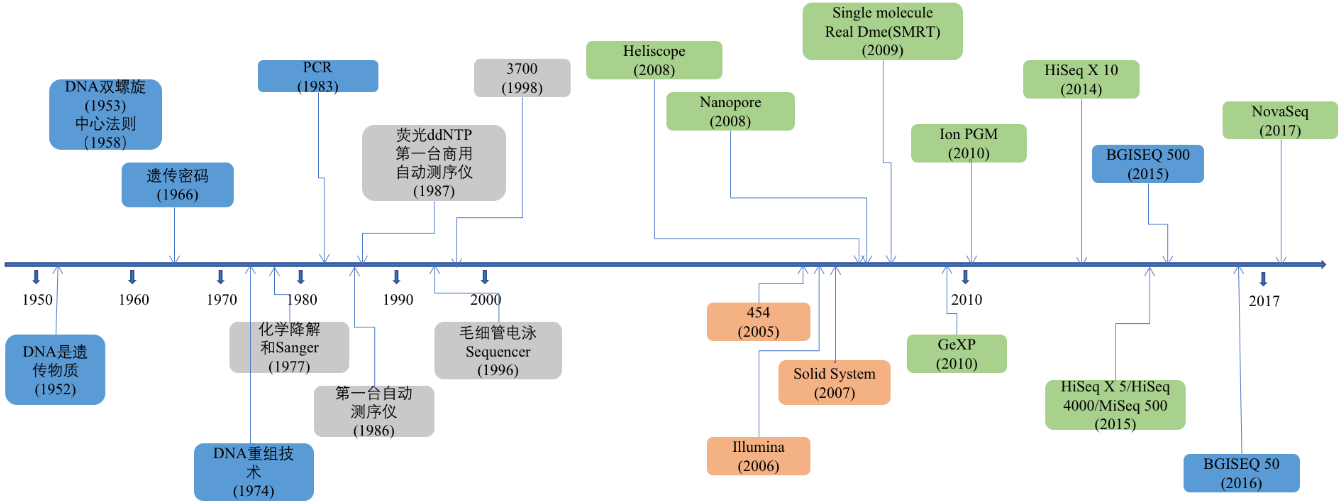
Principles and Applications of NGS-sequencing

新一代测序技术的原理和部分应用

G12：　A伊宗裔　B张小雪　D李楷

20180114

Develop of DNA sequencing technologies



DNA双螺旋 (1953) 中心法则 (1958)

遗传密码 (1966)

PCR (1983)

荧光ddNTP 第一台商用 自动测序仪 (1987)

3700 (1998)

Heliscope (2008)

Nanopore (2008)

Single molecule Real Dme(SMRT) (2009)

Ion PGM (2010)

HiSeq X 10 (2014)

BGISEQ 500 (2015)

NovaSeq (2017)

DNA是遗传物质 (1952)

DNA重组技术 (1974)

化学降解和Sanger (1977)

第一台自动测序仪 (1986)

毛细管电泳 Sequencer (1996)

454 (2005)

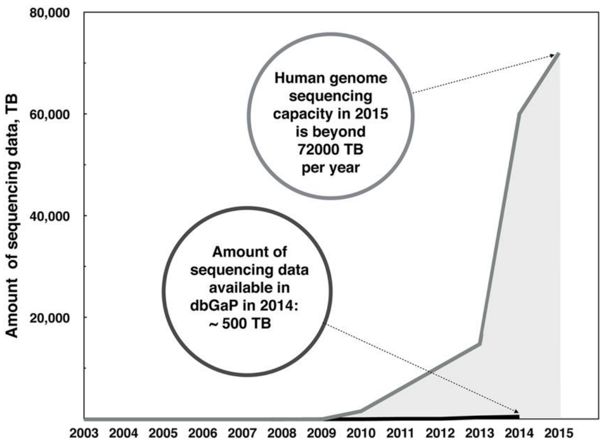
Illumina (2006)

Solid System (2007)

GeXP (2010)

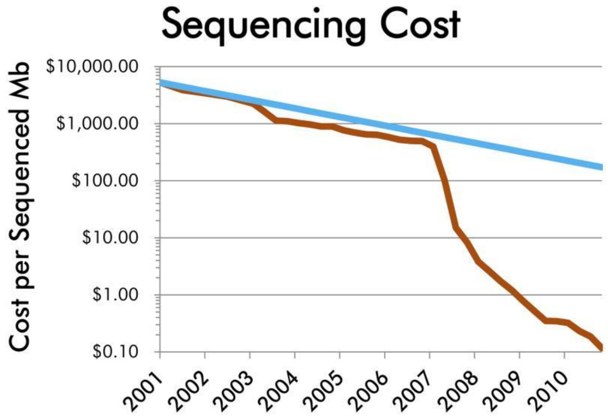
HiSeq X 5/HiSeq 4000/MiSeq 500 (2015)

BGISEQ 50 (2016)



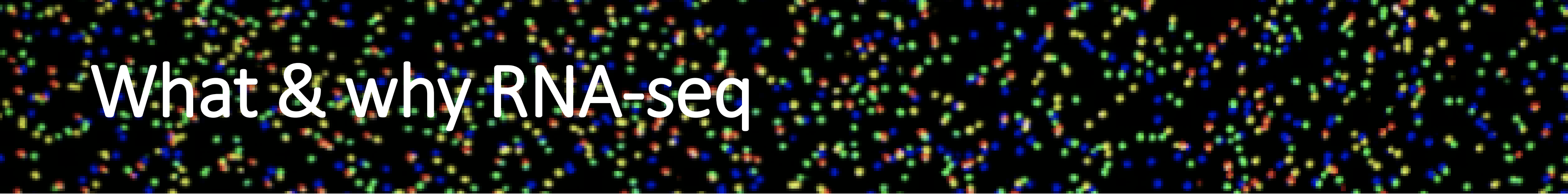
Human genome sequencing capacity in 2015 is beyond 72000 TB per year

Amount of sequencing data available in dbGaP in 2014: ~ 500 TB



Sequencing Cost

What & why RNA-seq



(1) What is RNA-seq

Whole transcriptome sequencing **（全转录组测序）**

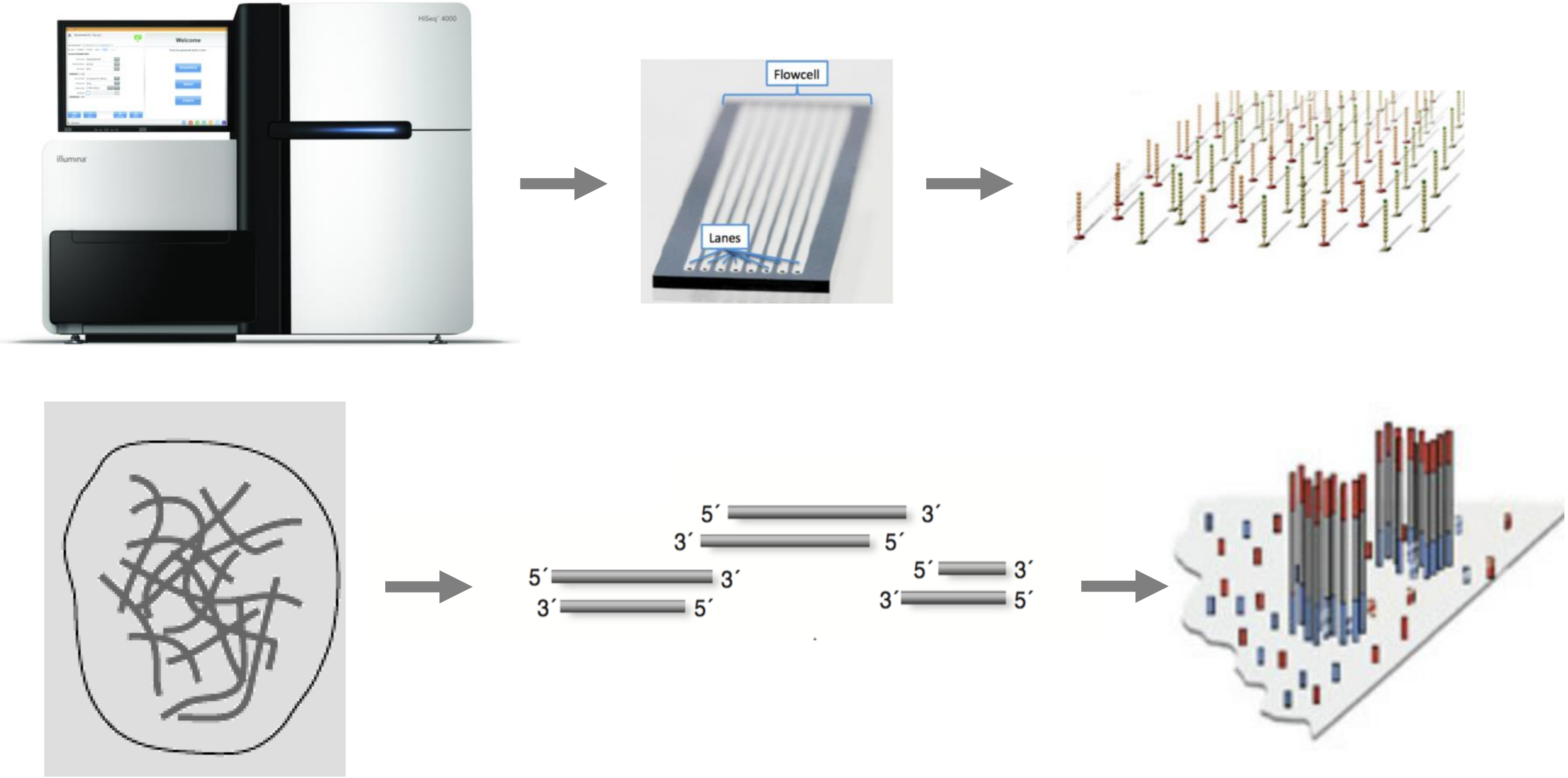
Uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time

（使用高通量测序技术快速的定量展示生物样品中所有的RNA表达量）

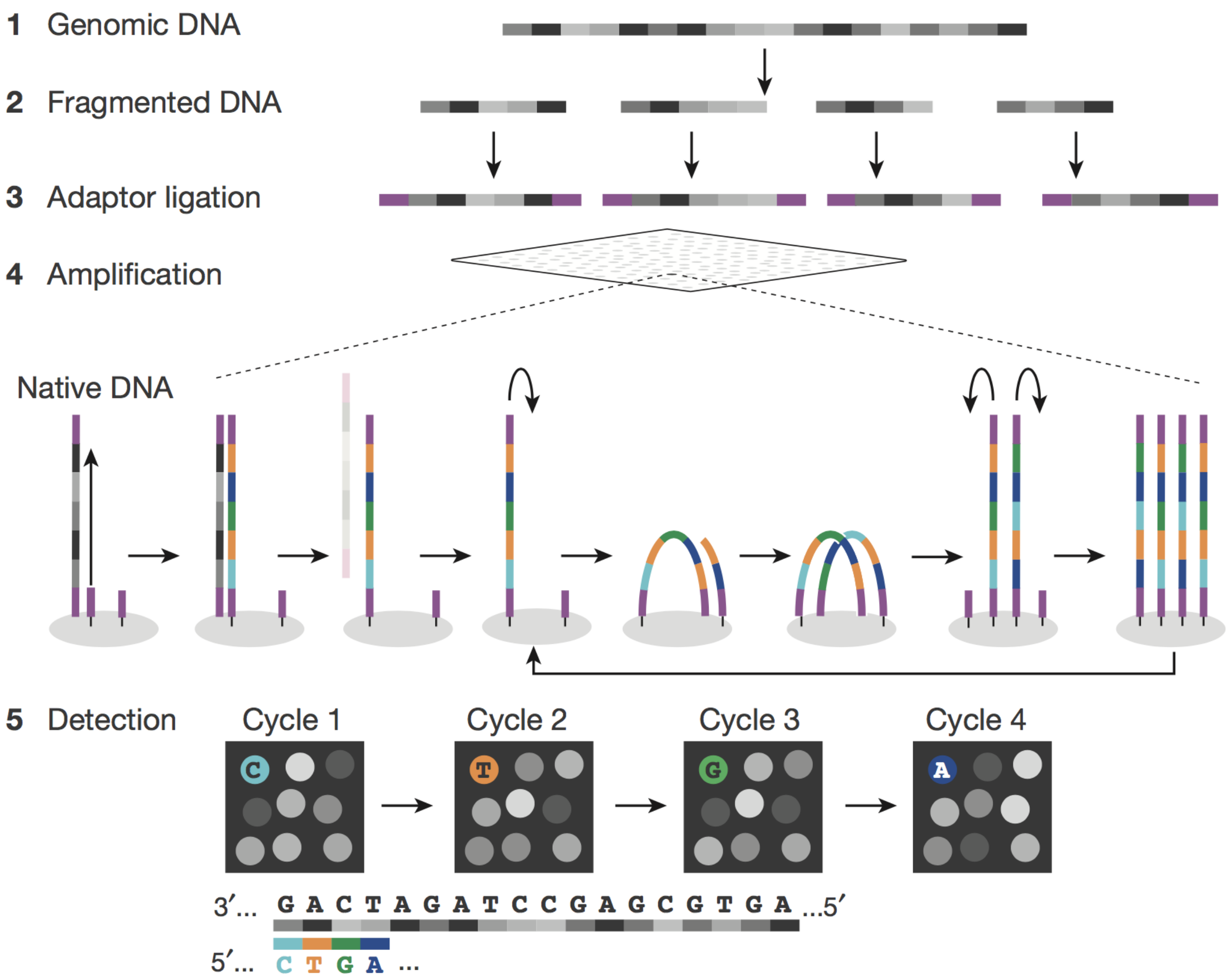
(2) Why RNA-seq

Core technology of NGS sequencing

Optical amplification system（光学放大系统）

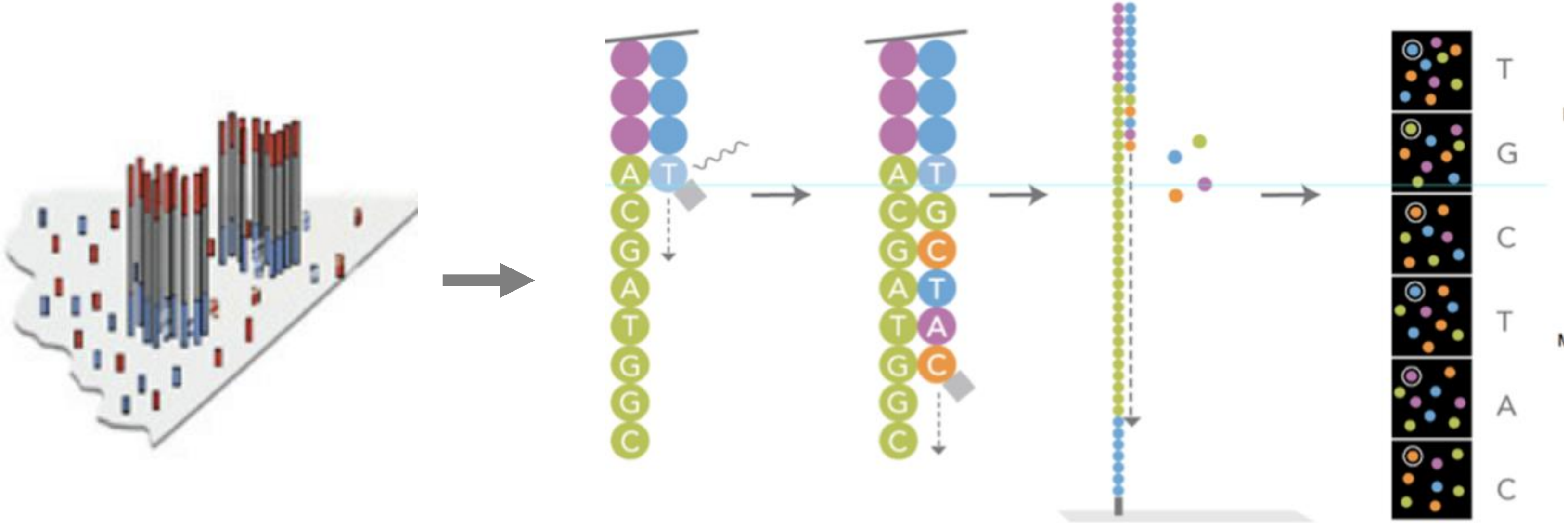


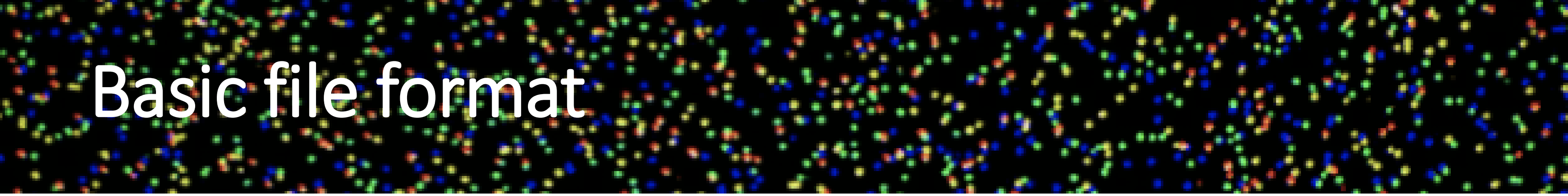
Next generation sequencing



Next generation sequencing

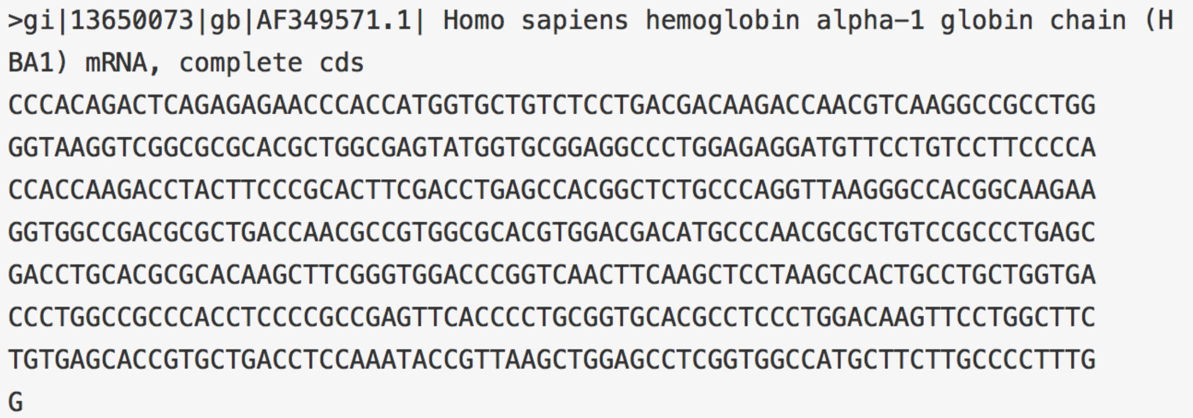
The principle of illumina sequencing
测序原理：边合成边测序



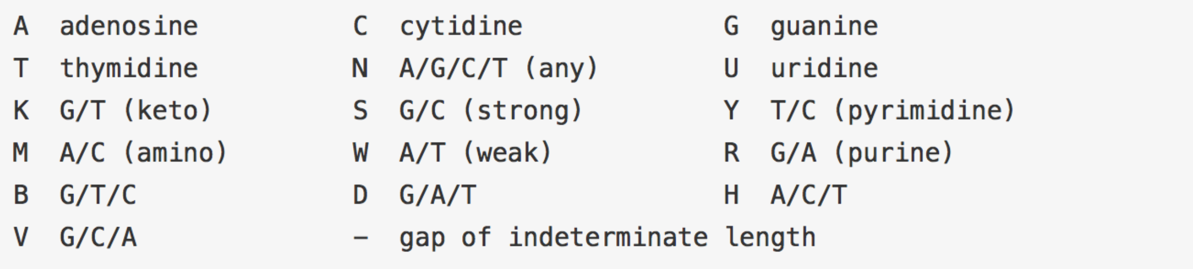


Basic file format

(1) FASTA



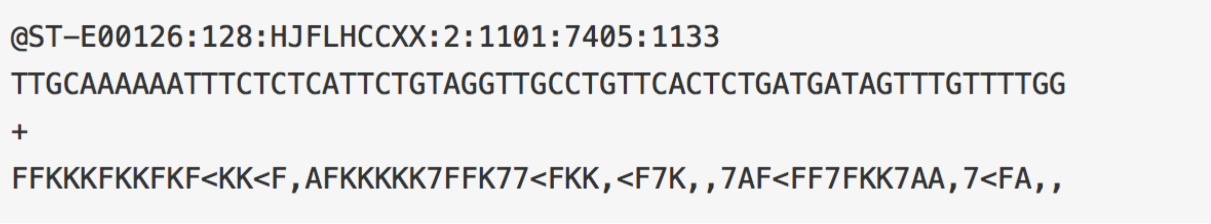
```
>gi|13650073|gb|AF349571.1| Homo sapiens hemoglobin alpha-1 globin chain (H
BA1) mRNA, complete cds
CCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGACGACAAGACCAACGTCAAGGCCGCCTGG
GGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCA
CCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAA
GGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGC
GACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGA
CCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTC
TGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTTG
G
```



A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A	-	gap of indeterminate length		

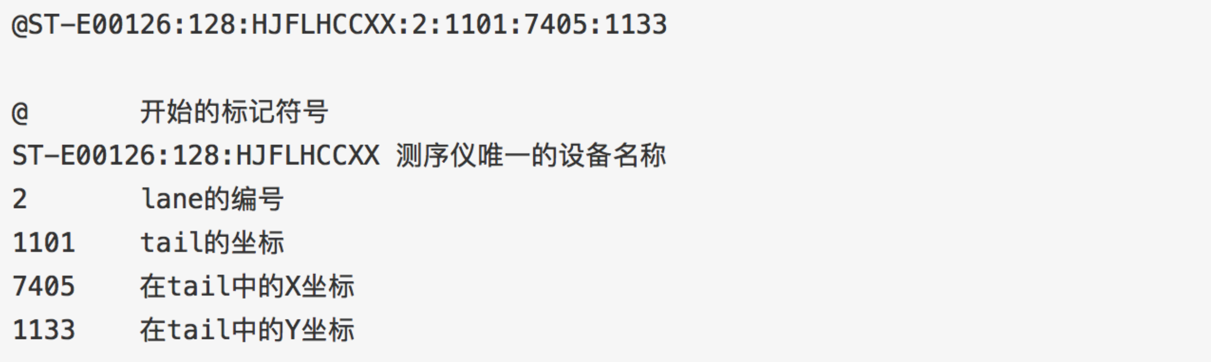
Basic file format

(2) FASTQ



```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133
TTGCAAAAAATTTCTCTCATTCTGTAGGTTGCCTGTTCACTCTGATGATAGTTTGTTTTGG
+
FFKKKFKKFKF<KK<F,AFKKKKK7FFK77<FKK,<F7K,,7AF<FF7FKK7AA,7<FA,,
```

第1行储存序列测序的坐标等信息



```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133

@        开始的标记符号
ST-E00126:128:HJFLHCCXX 测序仪唯一的设备名称
2        lane的编号
1101     tail的坐标
7405     在tail中的X坐标
1133     在tail中的Y坐标
```

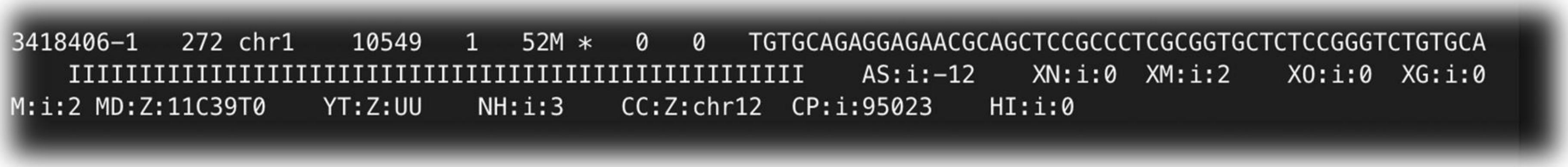
第2行测序得到的序列信息，用ATCGN来表示

第3行以“+”开始，储存附加信息

第4行储存质量信息，与第2行的碱基序列是一一对应，其中的每一个符号对应的ASCII值成为phred值，可以简单理解为对应位置碱基的质量值，越大说明测序的质量越好

Basic file format

(1) SAM



```
3418406-1   272 chr1    10549   1   52M *   0   0   TGTGCAGAGGAGAACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCA
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII    AS:i:-12    XN:i:0  XM:i:2    XO:i:0  XG:i:0
M:i:2 MD:Z:11C39T0    YT:Z:UU    NH:i:3    CC:Z:chr12  CP:i:95023    HI:i:0
```

SAM是序列比对格式标准，以TAB为分割符。应用于展示测序序列mapping到基因组的结果

第一列：read name
第二列：sum of flags，每个flag用数字来表示
第三列：RNAM比对到参考序列上的染色体号。若是无法比对，则是*
第四列：position，read比对到参考序列上，第一个碱基所在的位置
第五列：Mapping quality，比对的质量分数
第六列：CIGAR值，read比对的具体情况
第七列：MRNM(chr)
第八列：mate position，mate比对到参考序列上的第一个碱基位置
第九列：ISIZE，Inferred fragment size.
第十列：Sequence，就是read的碱基序列
第十一列：ASCII，read质量的ASCII编码
第十二列之后：Optional fields，可选的区域

Story one

感兴趣的Micro RNA X

在果蝇胚胎发育的不同阶段，miR-X 对果蝇胚胎发育的影响？

NGS sequencing applications

RNA-seq analysis of 0-2h development in *D. melanogaster* embryos

	0-2h	4-8h
W1118（Wild Type）	No Replicate	No Replicate
miR-X Knock out （Heterozygous mutants）	Replicate1	Replicate1
	Replicate2	Replicate2

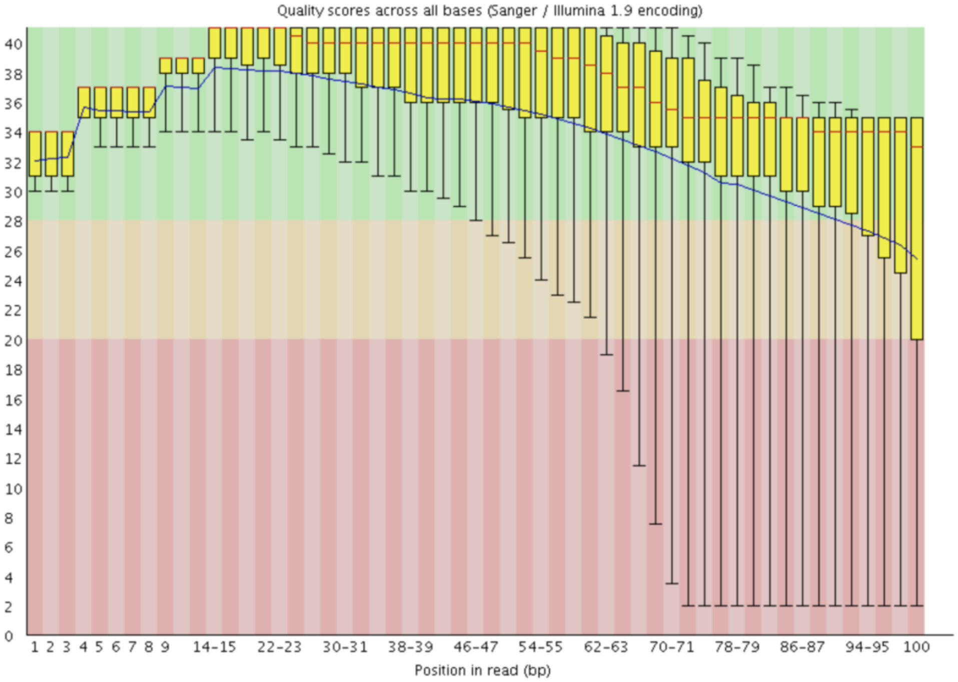
(1) Differential expression gene（寻找差异表达基因）

(2)Gene Ontology analysis

(3)Target Scan analysis

Quality control

(1) Trim the raw data（原始数据预处理）



fastqc

```
trim_galore --length 30 -o ./ $seqData.fastq  ## use trim_galore
```

trim_galore

Remove: low quality; short reads; adapter

fastx_toolkit

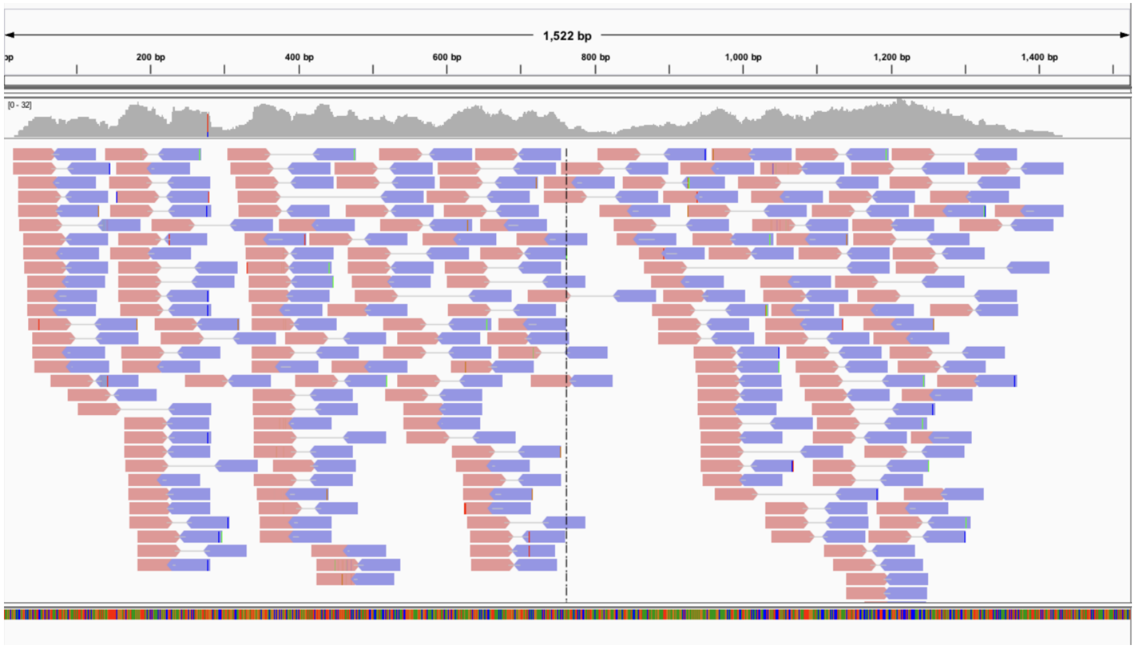
```
fastx_trimmer  -f 6 -l 90 -i seqData.fastq -o seqData.trim.fastq
```

Remove: low quality base

Mapping

(2) Mapping reads to refGenome（序列比对上基因组）

```
tophat2 -p 10 $refGenome.fasta $seqData.fastq –S $mapSeq.sam
```



Bowtie2, Tophat2

Bwa, STAR

igv

Differential expressin gene

(3) Gene expression (FPKM，计算基因表达量)

cufflinks

Cufflinks : assemble transcripts, estimate the abundance of these transcripts

(4) Differential expression gene（寻找差异表达基因）

Cuffdiff: Calculate gene expression differences

cuffdiff

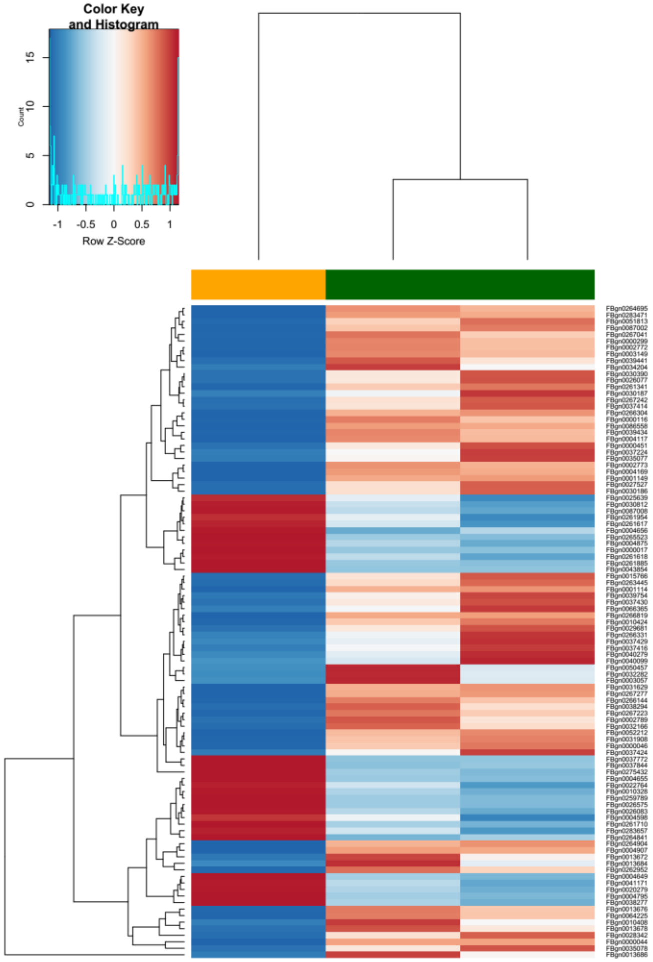
Fdr ≤ 0.05 & fold change ≥ 2

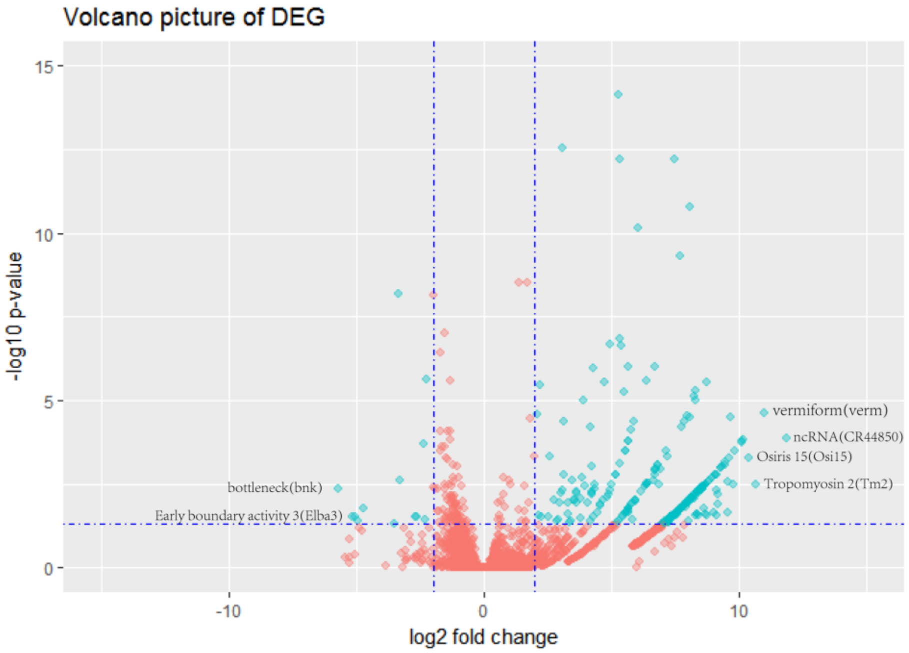
tophat

bowtie

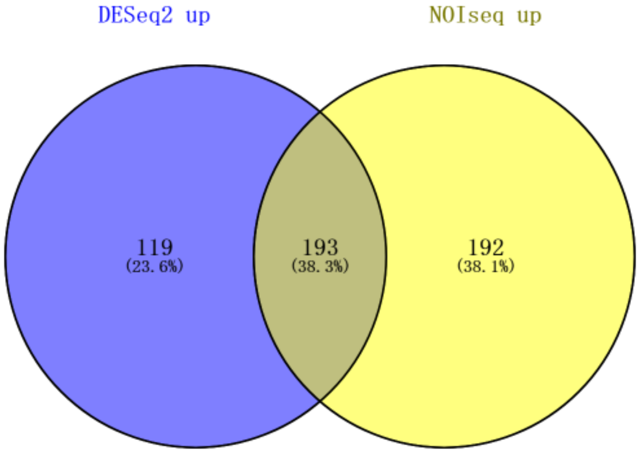
cufflinks

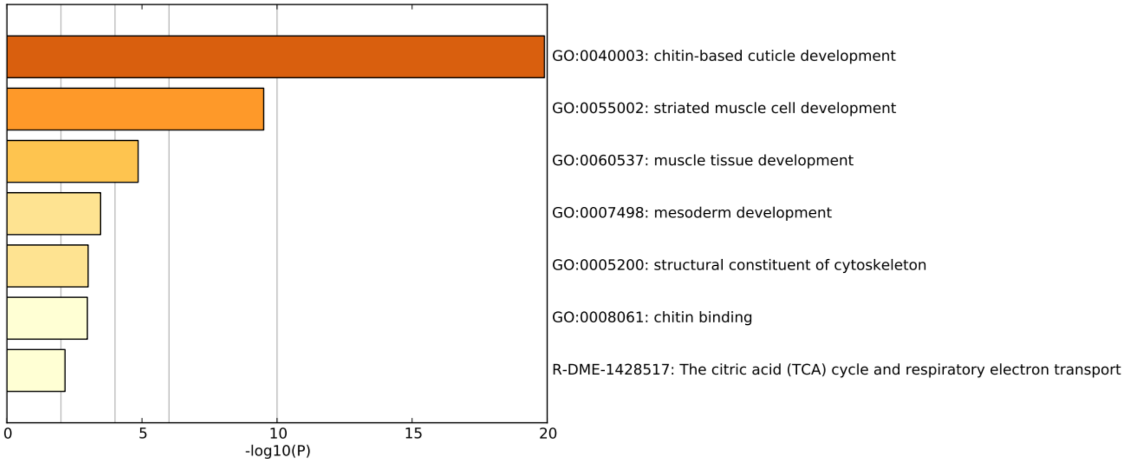
DNA sequencing applications



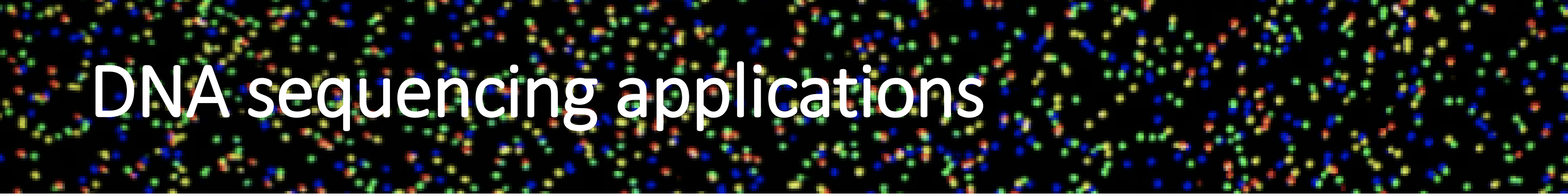


Gene Ontology analysis



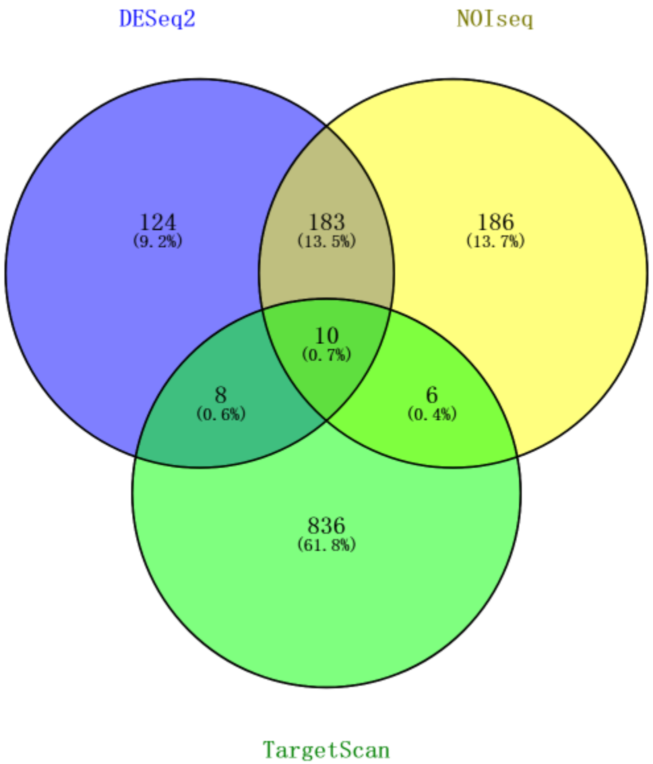


DAVID



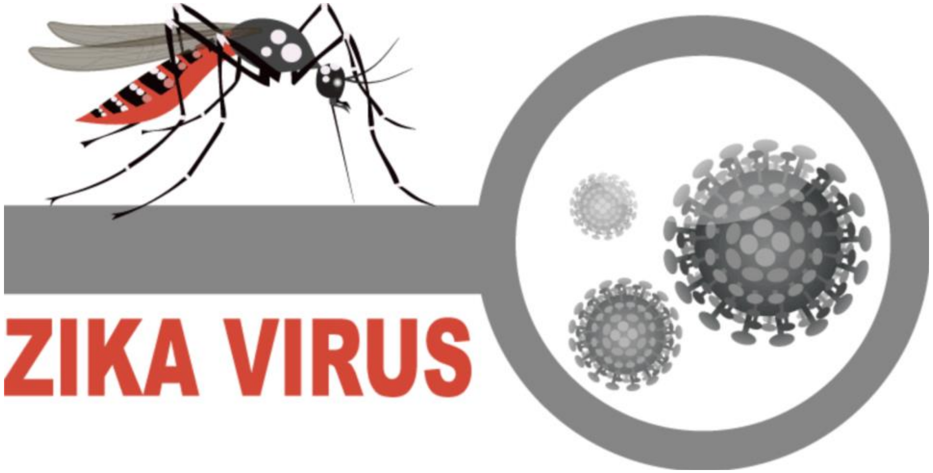
DNA sequencing applications

RNA-seq analysis of 0-2h development in *D. melanogaster* embryos



FLYBASE ID	GENE
FBgn0036600	CG13043 gene product from transcript CG13043-RA(CG13043)
FBgn0037416	Osiris 9(Osi9)
FBgn0037414	Osiris 7(Osi7)
FBgn0000299	Collagen type IV(Cg25C)
FBgn0031629	C-type lectin 27kD(Clect27)
FBgn0032166	CG4619 gene product from transcript CG4619-RA(CG4619)
FBgn0015766	CG10596 gene product from transcript CG10596-RD(Msr-110)
FBgn0035077	CG9083 gene product from transcript CG9083-RB(CG9083)
FBgn0032538	CG16885 gene product from transcript CG16885-RA(CG16885)
FBgn0038294	Myofilin(Mf)

Story two



一甲基腺嘌呤（m1A）是一种RNA上常见的表观遗传修饰

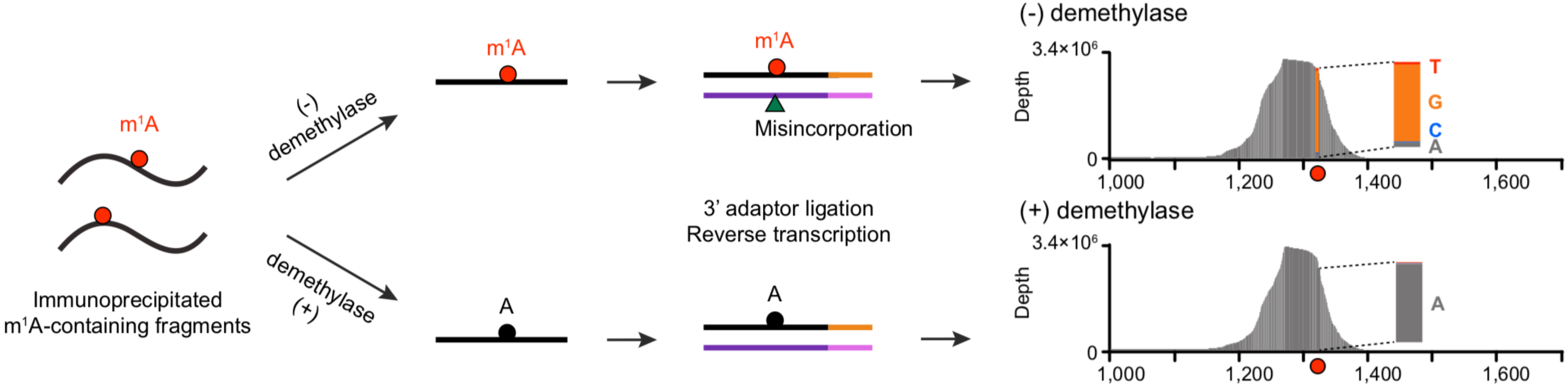
影响基因的转录

寨卡病毒中的m1A分布

对宿主中m1A分布以及基因表达的影响

m1A-map

Base-resolution mapping reveals distinct m(1)A methylome in Zika & host
单碱基分辨率下转录组水平鉴定m1A位点



(1) ChIP-seq call peak

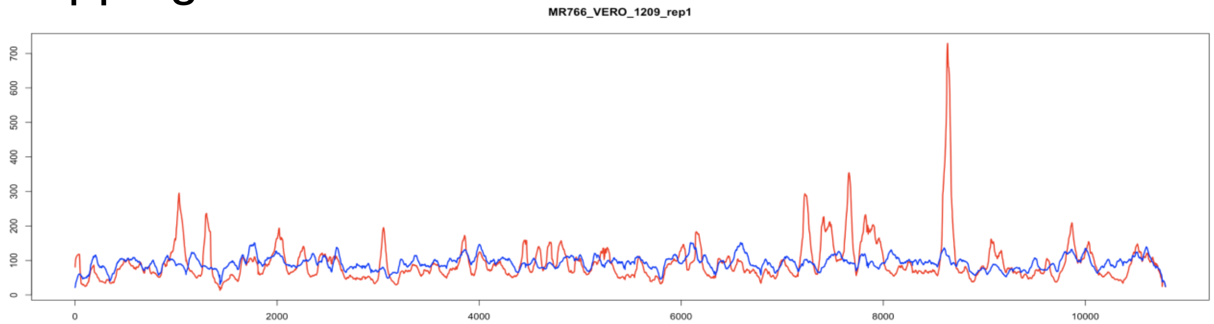
(2) call SNP

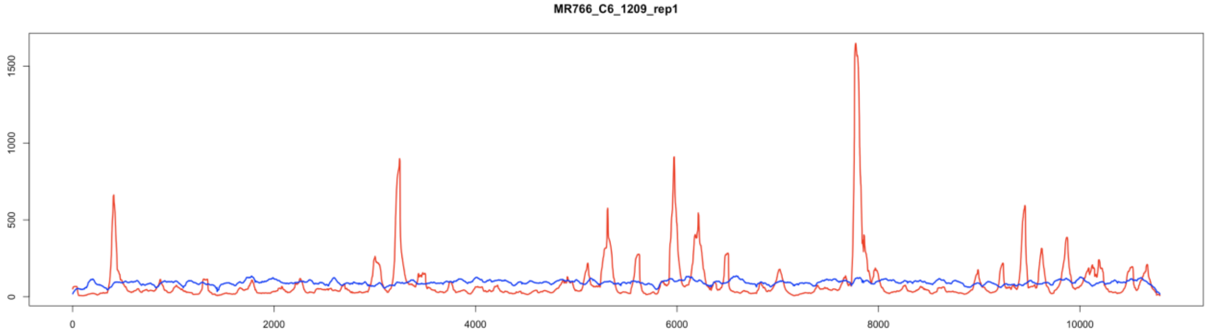
Li X, Xiong X, Zhang M, et al. Molecular cell, 2017, 68(5): 993-1005. e9.

Finding m1A peak

(1) Quality assessment

(2) Reads mapping





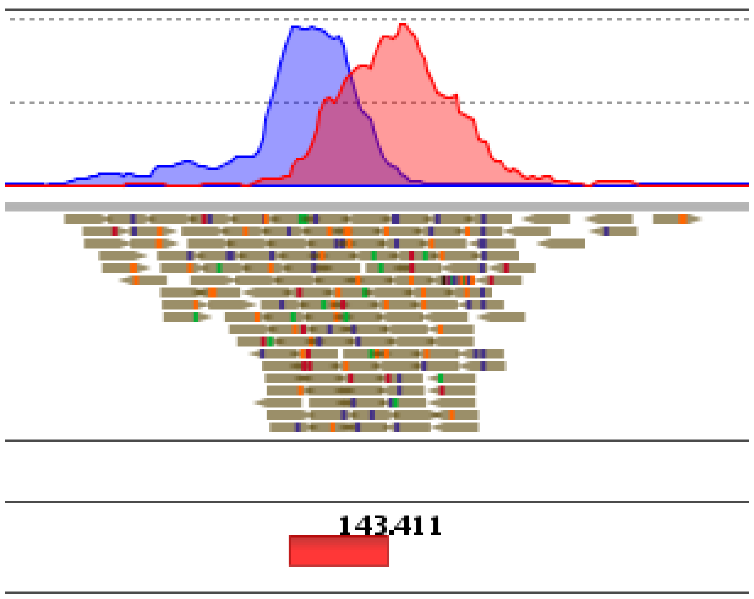
Call peak

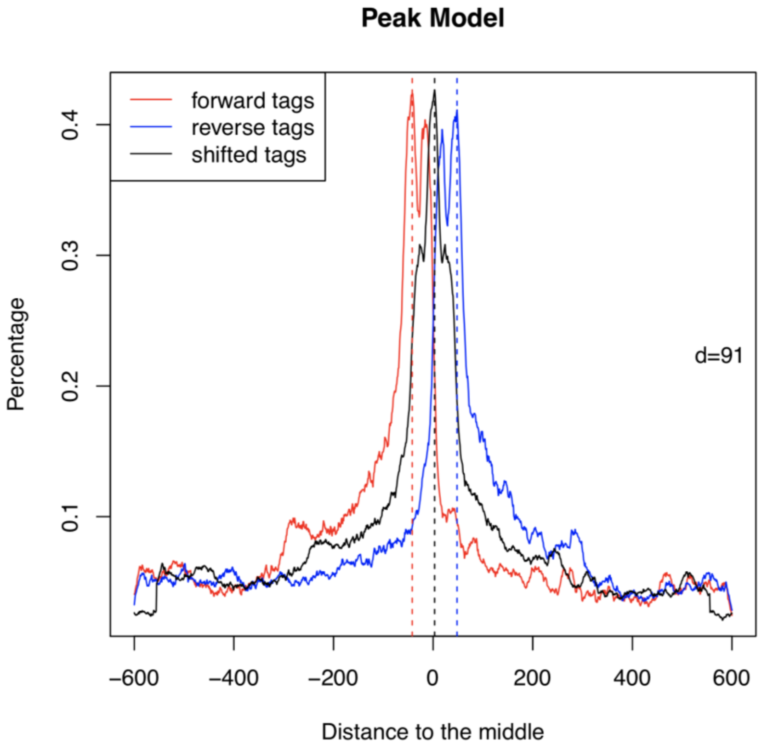
(3) Call peaks（寻找peak）

```
macs2 -t $IP.bam -c $input.bam -f BAM -g hs -n homo_peaks_2   ##use masc2 call peaks
```

study genome-wide protein-DNA interactions; identifying transcript factor binding sites

Peak: the Regina with high reads coverage; statistical test (poisson's distribution)



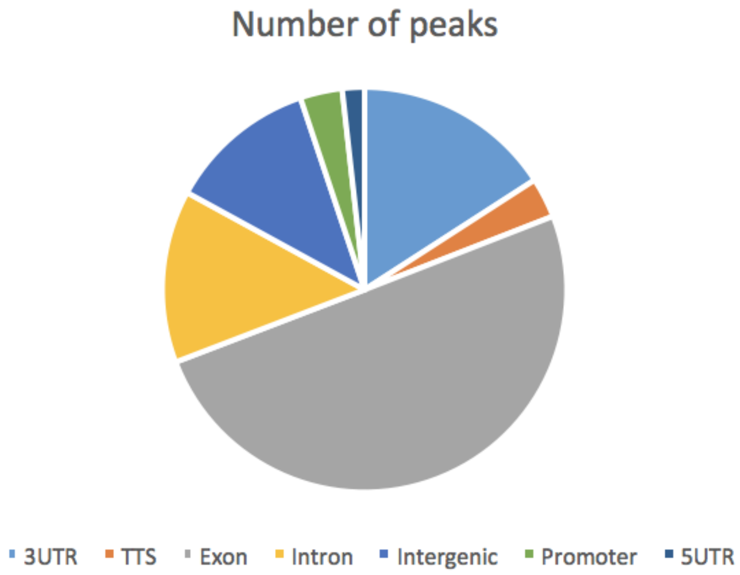


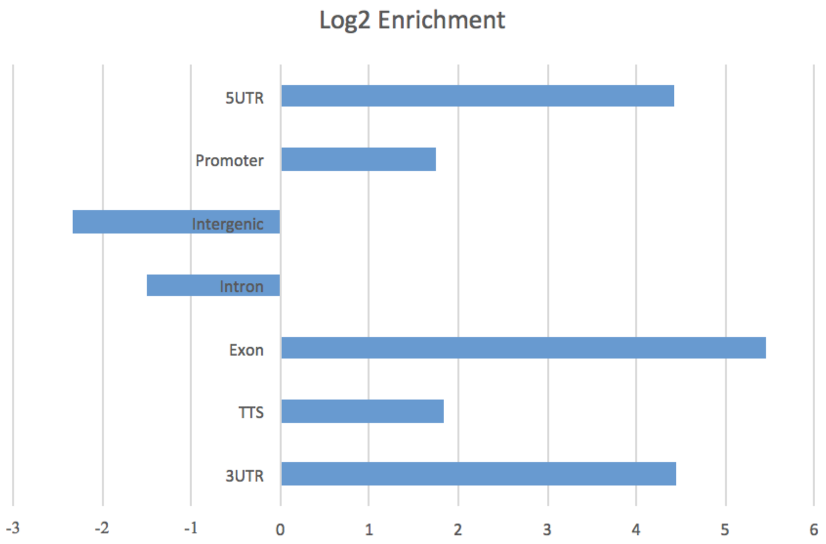
Annotate

(3) Annotate peaks (注释)

```
annotatePeaks.pl $peaks.bed hg19 > $peak_anno.txt  ## use homer annotate peaks
```

Annotate peaks base the annotation file (gtf/gff)





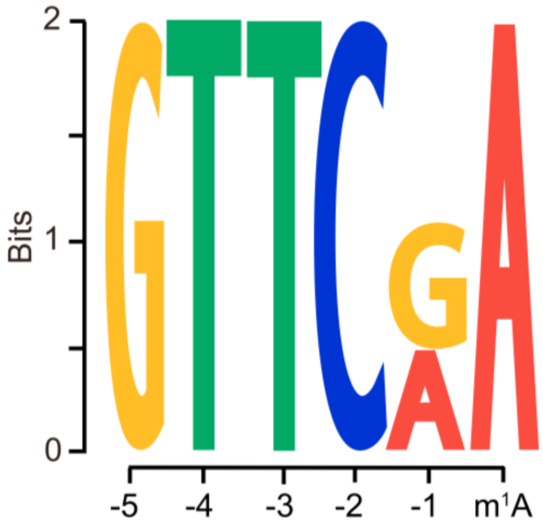
Predict motif

(3) Motif predict（预测motif）

```
meme $peak.fasta -dna -o ${output_file}  ##  use meme predict motif
```

Motif: protein banding short reads

statistical test; kmer



Call SNP

(1) Sort bam_file（将比对结果排序）

```
samtools sort -@ 10 IP.sam $IP.sam.sort  ## use samtools sort
```

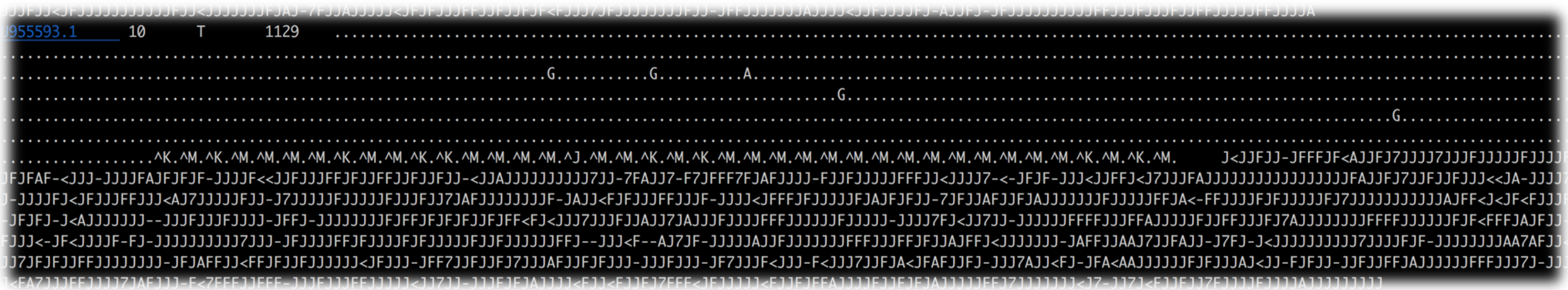
(2) Strand separate（分正负链）

```
samtools view -@ 10 -F 20 -hb -o $IP.sort.fwd.bam $IP.sam.sort

samtools view -@ 10 -f 16 -F 4 -hb -o $IP.sort.rvs.bam $IP.sam.sort
```

(3) Mpileup（查看基因组每个碱基的比对情况）

```
samtools mpileup -BQ0 -d 10000000 -o $IP.sam.sort.pileup.xls -f Zika_FSS.fa $IP.sort.fwd.bam
```

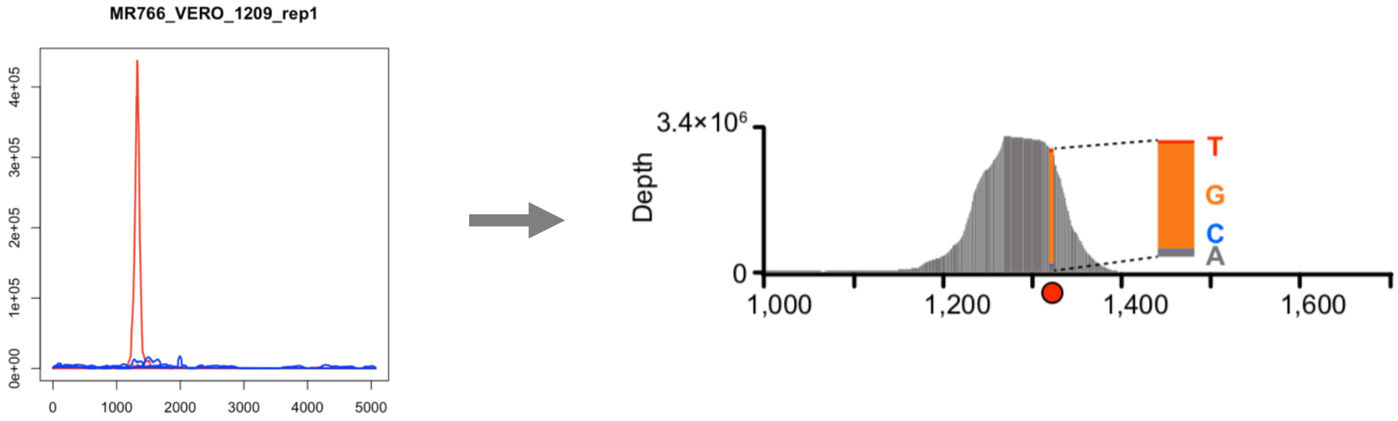


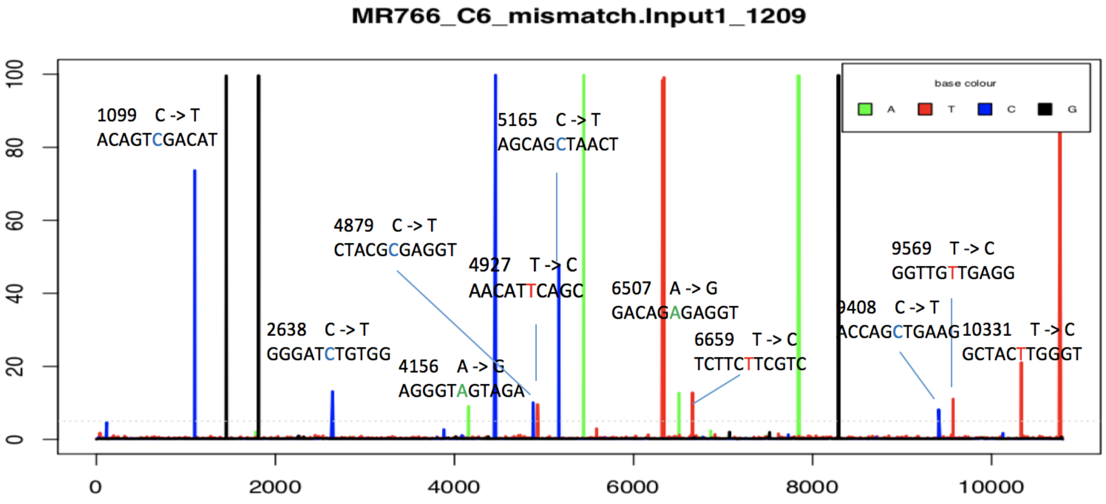
```
JJFJJ<JFJJJJJJJJJJJFJJ<JJJJJJJFJAJ-7FJJAJJJJJ<JFJFJJJFFJJFJJFJF<FJJJ7JFJJJJJJJJJFJJ-JFFJJJJJJJAJJJJ<JJFJJJJFJ-AJJFJ-JFJJJJJJJJJJFFJJJFJJJFJJFFJJJJJFFJJJJA
U955593.1      10      T       1129    ...........................................................................................................
.................................................................................................................................................................
..................................................................G.........G.........A..........................................................................
...............................................................................................G..................................................................
..............................................................................................................................................G...................
.................................................................................................................................................................
................^K.^M.^K.^M.^M.^M.^M.^K.^M.^M.^K.^K.^M.^M.^M.^M.^J.^M.^M.^K.^M.^K.^M.^M.^M.^M.^M.^M.^M.^M.^M.^M.^M.^M.^M.^M.^K.^M.^K.^M.     J<JJFJJ-JFFFJF<AJJFJ7JJJJ7JJJFJJJJJFJJJJ
JFJFAF-<JJJ-JJJJFAJFJFJF-JJJJF<<JJFJJJFFJFJJFFJJFJJFJJ-<JJAJJJJJJJJJJJ7JJ-7FAJJ7-F7JFFF7FJAFJJJJ-FJJFJJJJJFFFJJ<JJJJ7-<-JFJF-JJJ<JJFFJ<J7JJJFAJJJJJJJJJJJJJJJJJFAJJFJ7JJFJJFJJJ<<JA-JJJJ7
J-JJJJFJ<JFJJJFFJJJ<AJ7JJJJJFJJ-J7JJJJJFJJJJJFJJJFJJ7JAFJJJJJJJJF-JAJJ<FJFJJJFFJJJF-JJJJ<JFFFJFJJJJJFJAJFJFJJ-7JFJJAFJJFJAJJJJJJJFJJJJJFFJA<-FFJJJJFJFJJJJJFJ7JJJJJJJJJJJAJFF<J<JF<FJJJF
-JFJFJ-J<AJJJJJJJ--JJJFJJJFJJJJ-JFFJ-JJJJJJJJFJFFJFJFJFJJFJFF<FJ<JJJ7JJJFJJAJJ7JAJJJFJJJJFFFJJJJJJFJJJJJ-JJJJ7FJ<JJ7JJ-JJJJJJFFFFJJJFFAJJJJJFJJFFJJJFJ7AJJJJJJJJFFFFJJJJJJFJF<FFFJAJFJJ-
FJJJ<-JF<JJJJF-FJ-JJJJJJJJJJ7JJJ-JFJJJJFFJFJJJJFJFJJJJJFJJFJJJJJJFFJ--JJJ<F--AJ7JF-JJJJAJJFJJJJJJJFFFJJJFFJFJJAJFFJ<JJJJJJJ-JAFFJJAAJ7JJFAJJ-J7FJ-J<JJJJJJJJJJ7JJJJFJF-JJJJJJJJAA7AFJJ<
JJ7JFJFJJFFJJJJJJJJ-JFJAFFJJ<FFJFJJFJJJJJJ<JFJJJ-JFF7JJFJJFJ7JJJAFJJFJFJJJ-JJJFJJJ-JF7JJJF<JJJ-F<JJJ7JJFJA<JFAFJJFJ-JJJ7AJJ<FJ-JFA<AAJJJJJJFJFJJJAJ<JJ-FJFJJ-JJFJJFFJAJJJJJJFFFJJJ7J-JJJ
J<FA7JJJFFJJJJ7JAFJJJ-F<7FFFJJFFF-JJJFJJJFFJJJJJ<JJ7JJ-JJJFJFJAJJJJ<FJJ<FJJFJ7FFF<JFJJJJJ<FJJFJFFAJJJJFJJFJFJAJJJJJFFJ7JJJJJJJJ<J7-JJ7J<FJJFJJ7FJJJJFJJJJAJJJJJJJJJ
```

m1A-map

(3) Find mismatch in zika genome（寻找错配位点）

```
perl mismatch_derRNA_new.pl $IP.sam.sort.pileup.xls ${file}.bwa.mismatch.xls
```





Acknowledgement

小组全体成员

农科院 李欣

罗老师

把知识和希望留给我们的那些人们