SNP Identification Using Next-Generation Sequencing (NGS)

冯婉滢 刘鹤 钱喜静

2025/06/11

北京大学
PEKING UNIVERSITY

目录 CONTENT

北京大学
PEKING UNIVERSITY

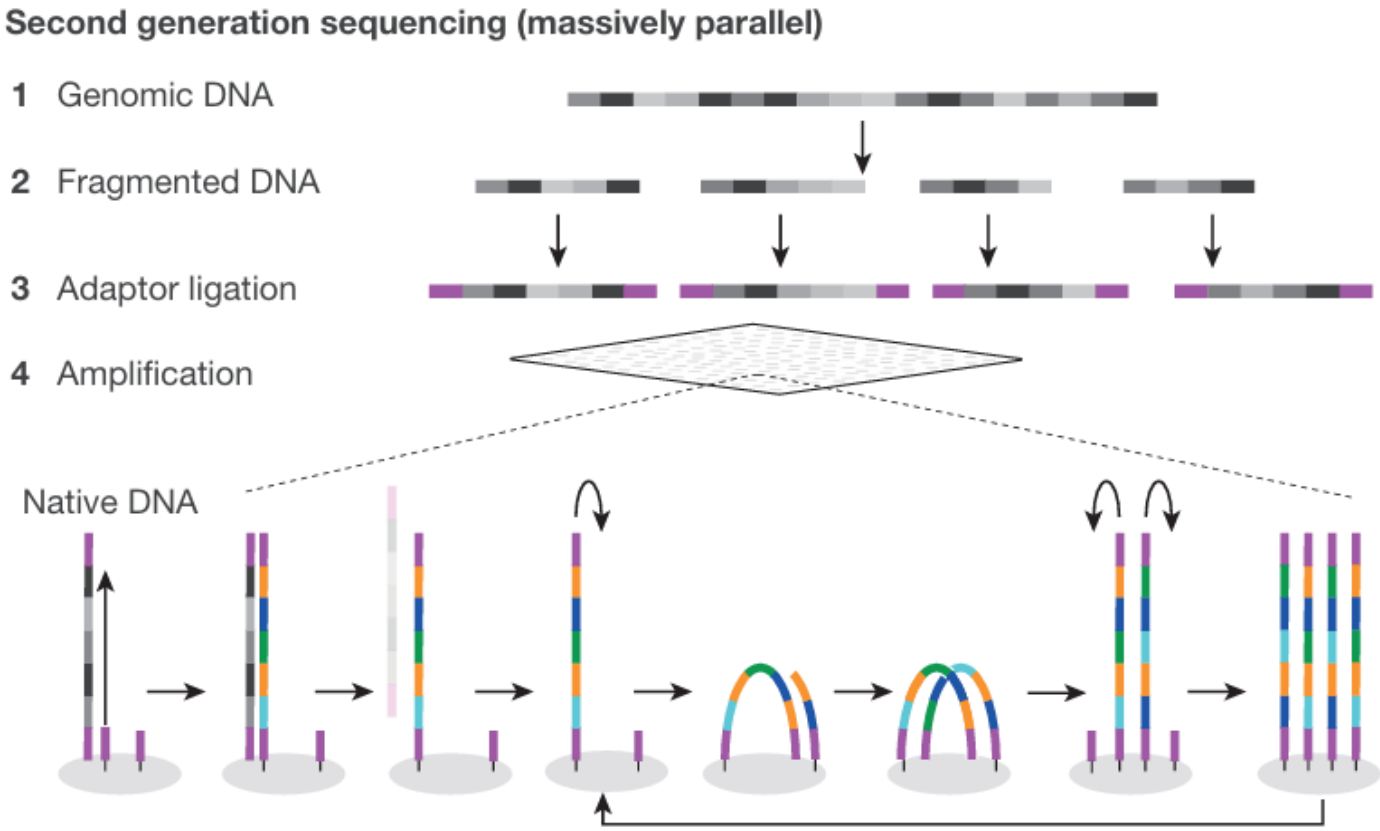
PART 01

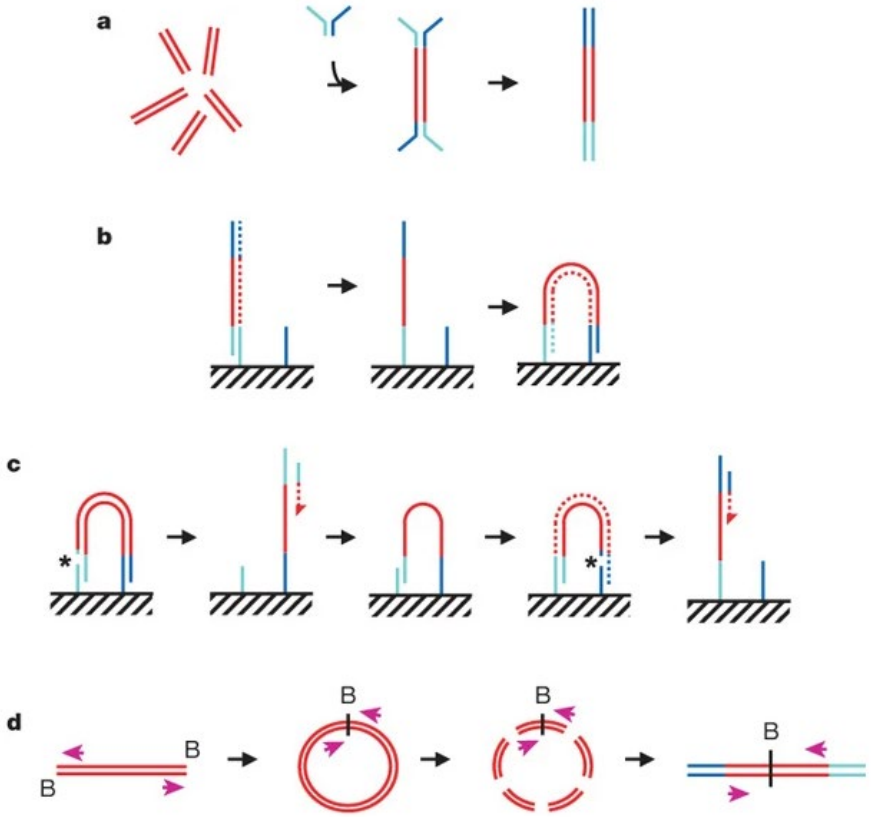
Principle & Workflow

Next-Generation Sequencing

北京大学 PEKING UNIVERSITY

Preparation of samples.



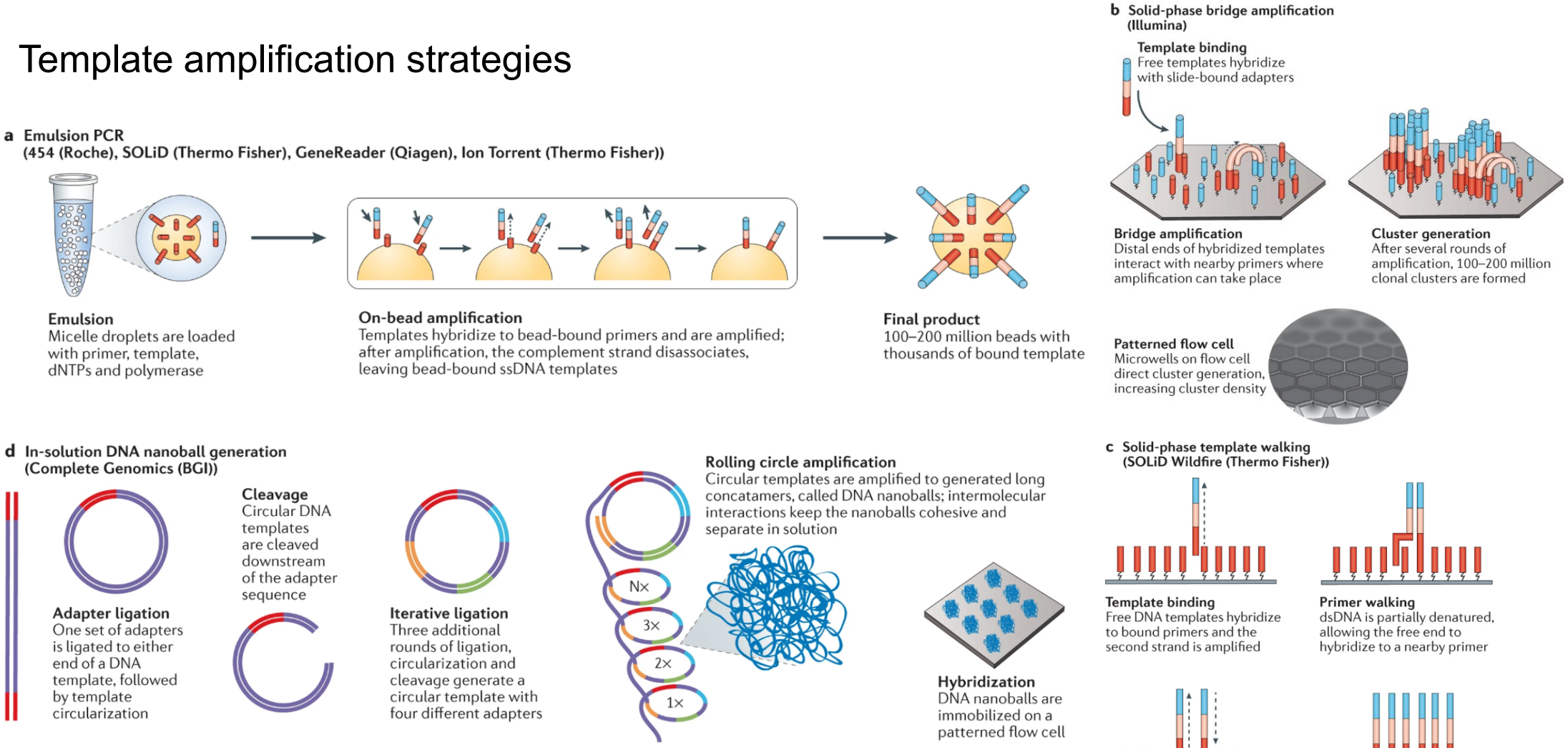


Bentley, D., Balasubramanian, S., Swerdlow, H. et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).

Shendure, J., Balasubramanian, S., Church, G. et al. DNA sequencing at 40: past, present and future. *Nature* 550, 345–353 (2017). https://doi.org/10.1038/nature24286

Next-Generation Sequencing

Template amplification strategies

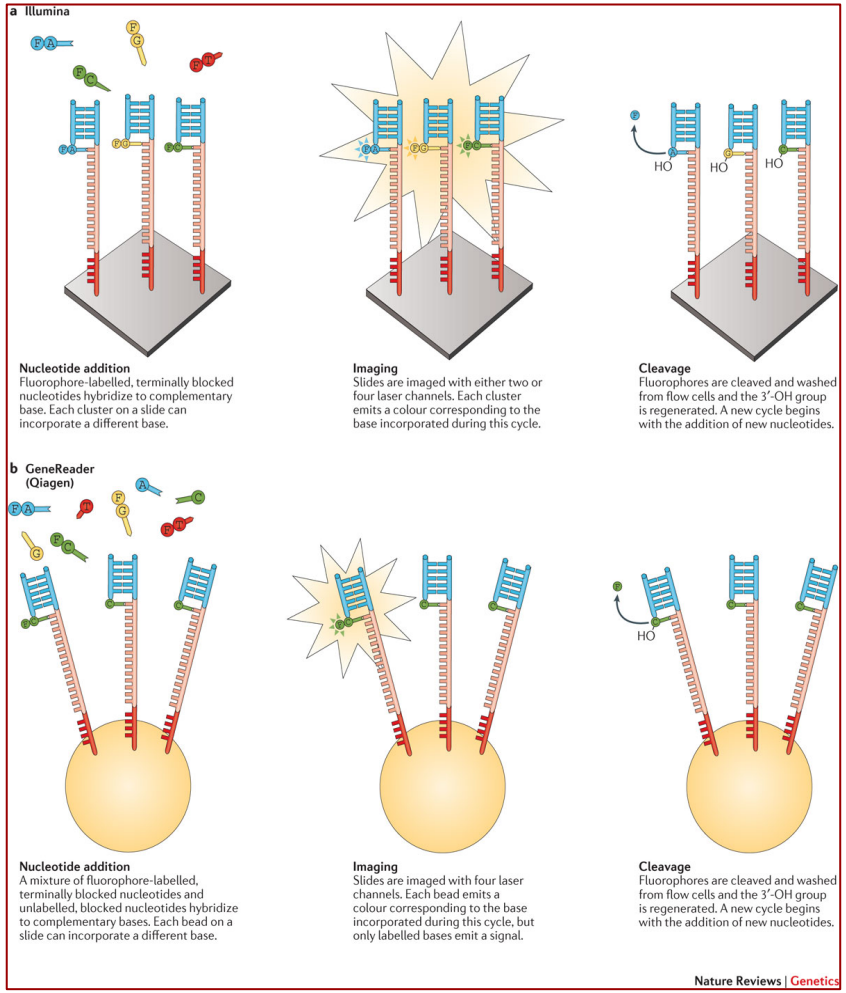


Goodwin, S., McPherson, J. & McCombie, W. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).

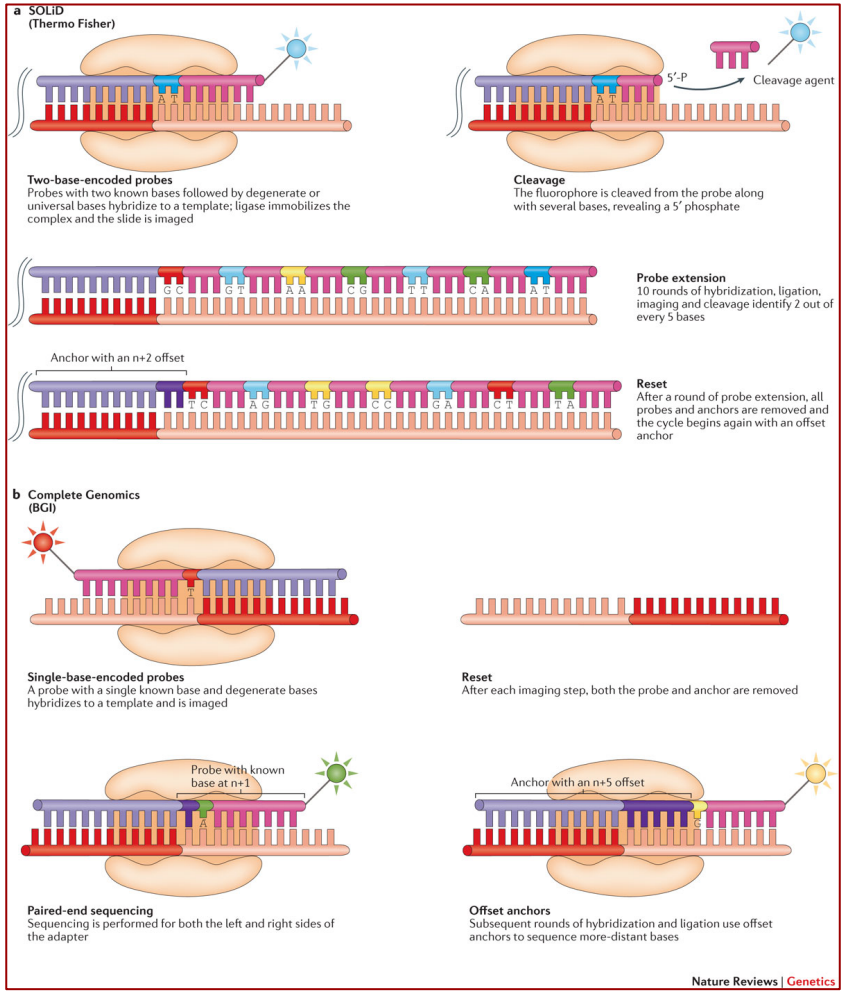
Next-Generation Sequencing

北京大学 PEKING UNIVERSITY

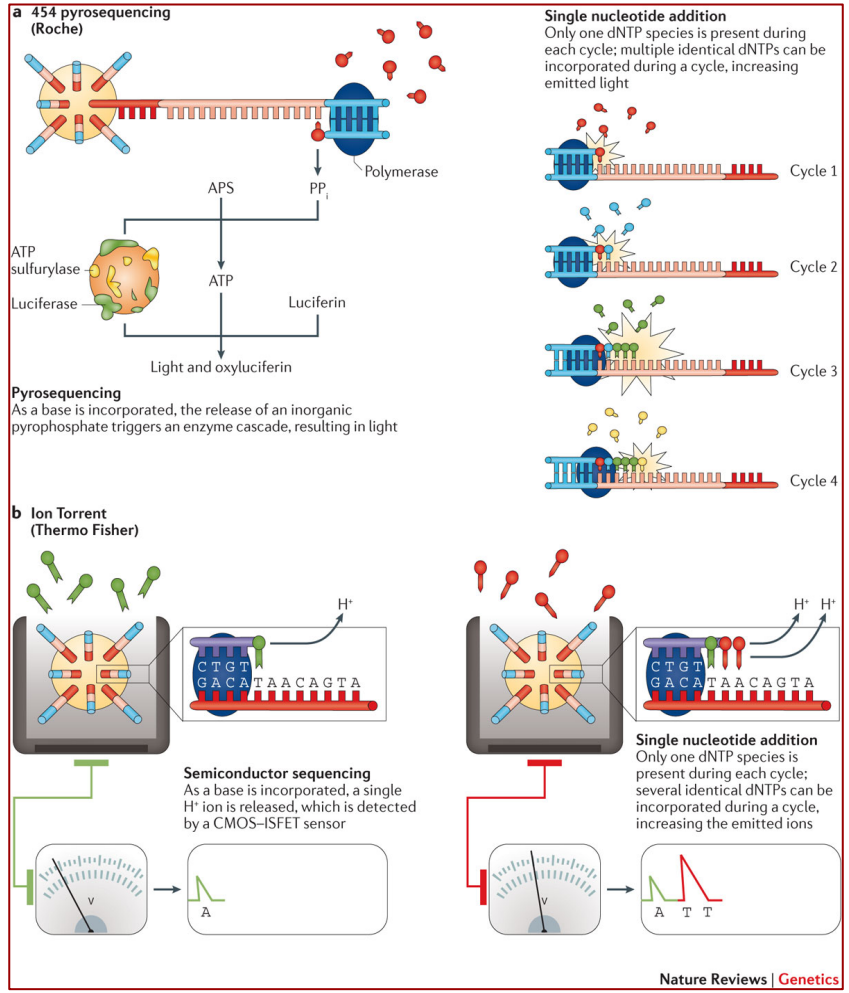
Sequencing by synthesis: cyclic reversible termination approaches



Sequencing by ligation methods



Sequencing by synthesis: single-nucleotide addition approaches



Goodwin, S., McPherson, J. & McCombie, W. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).

Single Nucleotide Polymorphism

样本测序
样本 DNA → 建库 → 测序
获得 FASTQ 格式原始序列数据

↓

质量控制
利用FastQC, Trimmomatic, cutadapt等工具去除低质量reads、接头污染等

↓

比对到参考基因组
工具：BWA、Bowtie2；
输出文件格式：SAM/BAM；
后处理：使用 SAMtools 或 Picard 去除重复、排序等

→

变异检测
工具：GATK HaplotypeCaller；FreeBayes, Bcftools, VarScan；
输出：VCF文件，记录 SNP 和 InDel 等变异

↑

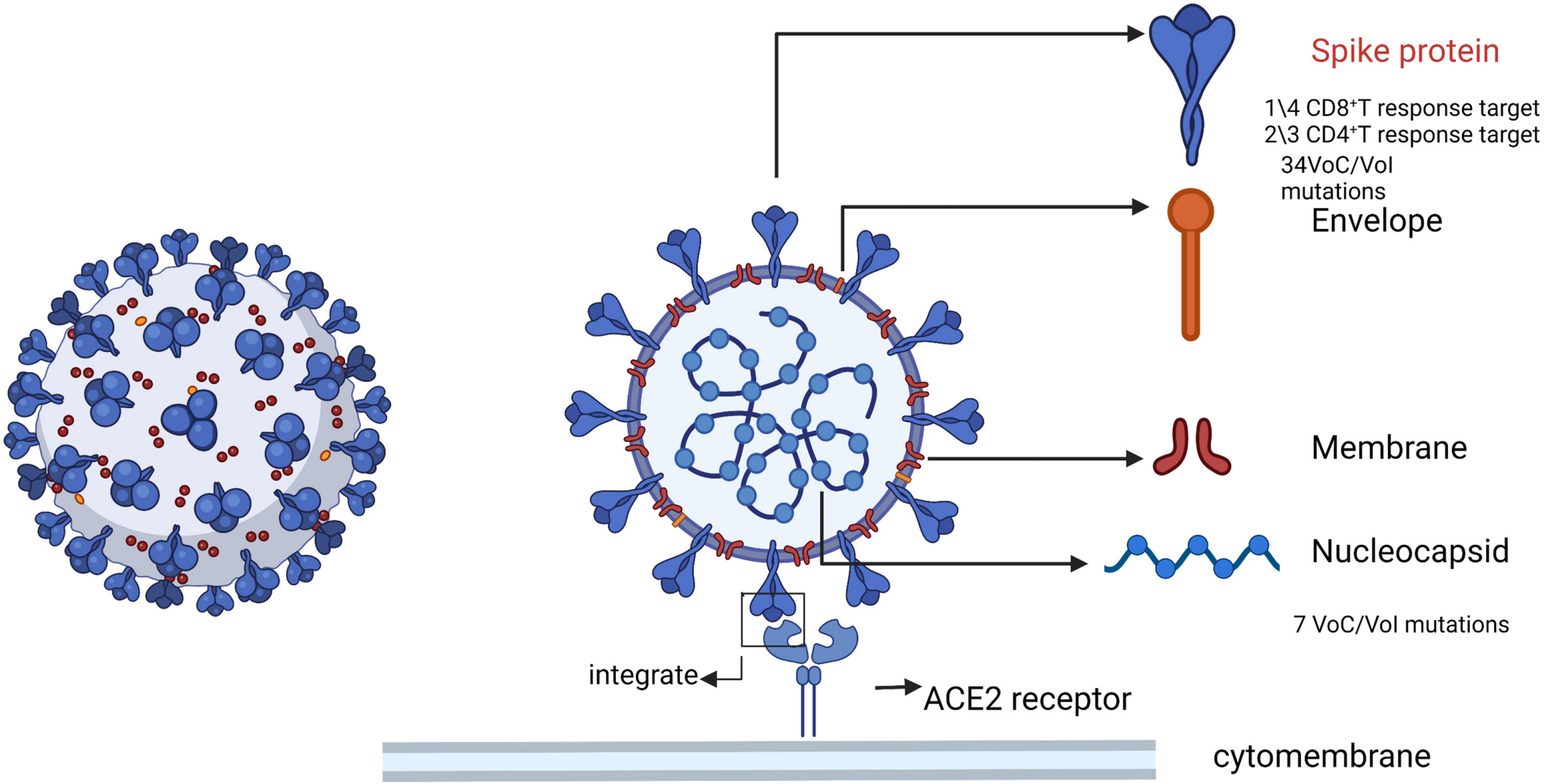
变异过滤与注释
过滤标准：测序深度（DP）、碱基质量（QUAL）、支持比例（AF）等；
工具：GATK VariantFiltration；Bcftools filter；
注释工具：ANNOVAR, VEP, SnpEff

北京大学
PEKING UNIVERSITY

PART 02

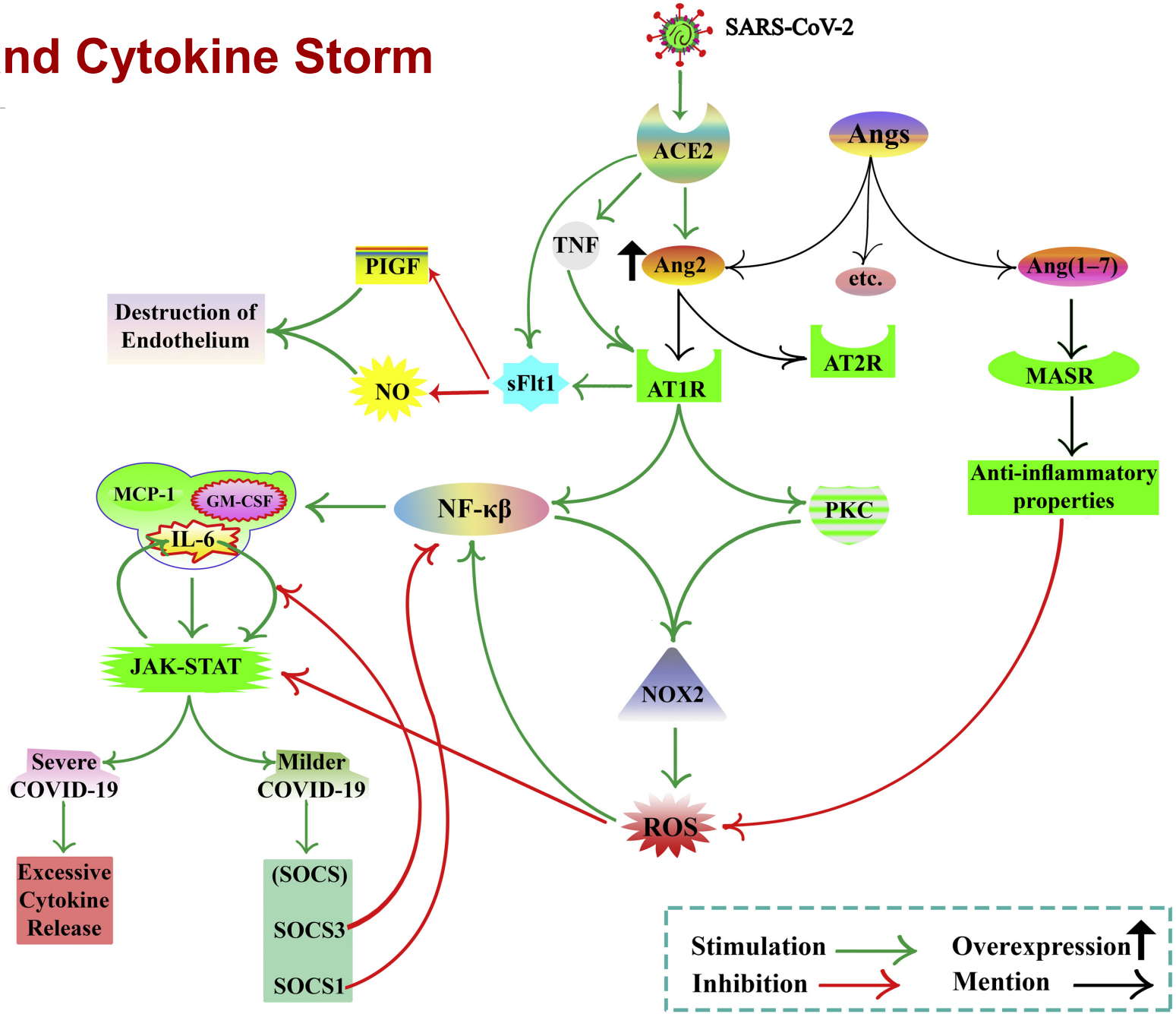
Objective & Framework

COVID-19 and Cytokine Storm



Feng Lin, etc. *Life Sciences.* (2023)

COVID-19 and Cytokine Storm



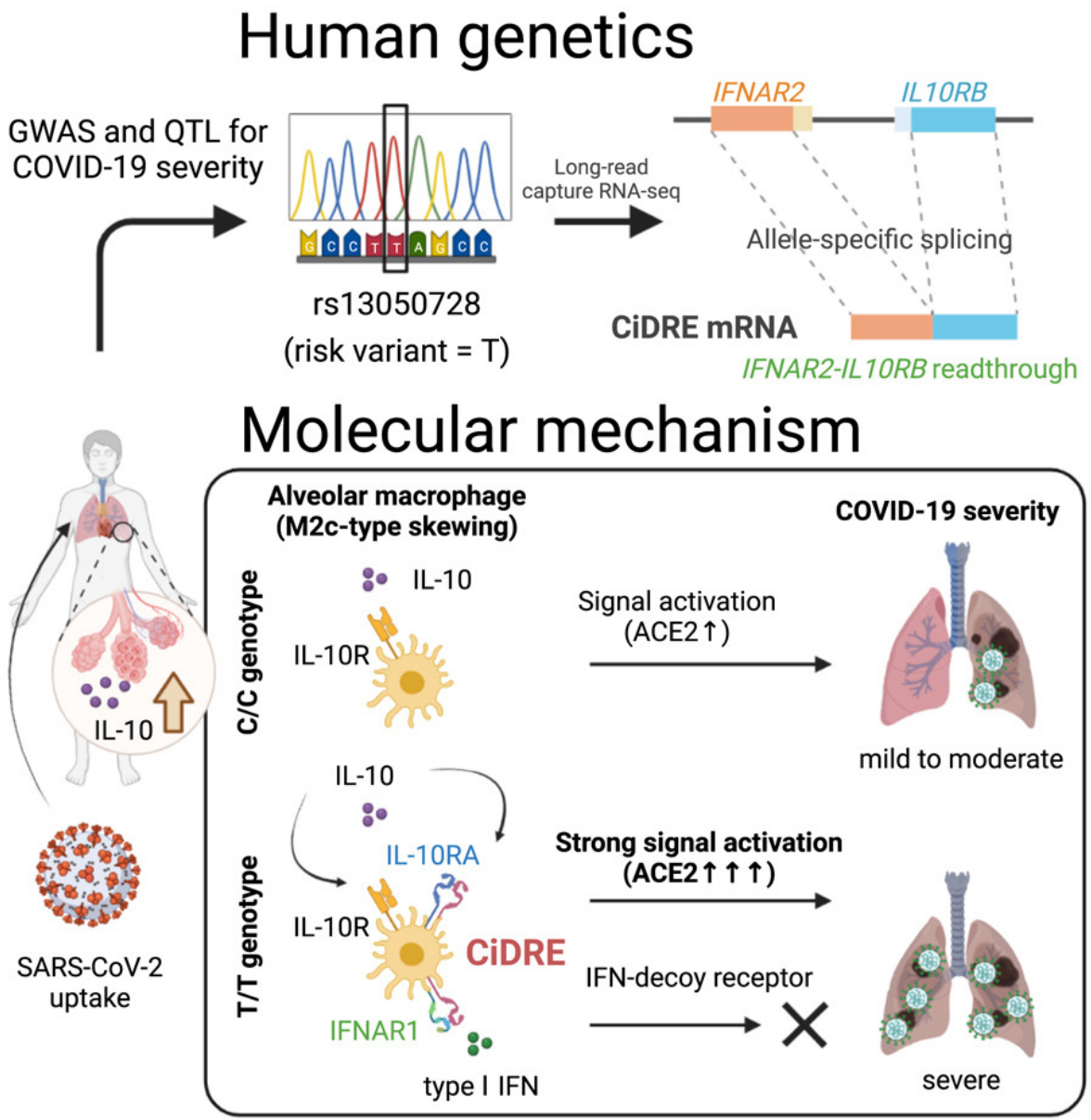
Fahimeh Zamani Rarani, etc. *Molecular Therapy Nucleic Acids.* (2022)

Research Objectives

北京大学 PEKING UNIVERSITY

- To reveale the mechanism by which macrophages acquire susceptibility to SARS-CoV-2 infection.
- To investigated the gene expression pattern differences between healthy hamster lungs and SARS-CoV-2-infected hamster lungs collected 5 days post-infection (dpi) using RNA-seq and its roles in severe COVID-19 patients.
- To elucidate how these genes promote alveolar macrophage mediated infection of SARS-CoV-2.
- To look for a genetic basis (SNP) for the differences in gene expression related to COVID-19 severity in humans.

Research Overview



- SARS-CoV-2 infects IL-10-induced alveolar macrophages to promote a cytokine storm.
- Blockade of IL-10R signaling on alveolar macrophages suppresses COVID-19 pneumonia.
- A readthrough transcript, CiDRE, in COVID-19 was identified by GWAS and QTL analysis.
- CiDRE possesses a dual function in alveolar macrophages, promoting severe COVID-19

北京大学
PEKING UNIVERSITY

PART 03

Code & Visualization

Data Download

北京大学 PEKING UNIVERSITY

Mesocricetus auratus (golden hamster)

Accession: PRJDB14430 ID: 905249

Expression data from whole lung of Syrian hamsters

Whole lungs were extracted from Golden Hamsters infected with SARS-COV-2 or mock for 5 days with or without neutralizing antibodies to the IL-10 receptor, and total RNA was isolated using TRIzol (Invitrogen). More...

See Genome Information for Mesocricetus auratus

NAVIGATE ACROSS

129 additional projects are related by organism.

Accession	PRJDB14430
Data Type	Transcriptome or Gene expression
Scope	Multiisolate
Organism	**Mesocricetus auratus** [Taxonomy ID: 10036] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Cricetidae; Cricetinae; Mesocricetus; Mesocricetus auratus
Submission	Registration date: 24-Nov-2022 **Tokyo Medical and Dental university, 1 Chome-5-45 Yushima, Bunkyo City, Tokyo, Japan 113-8510**

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	8
OTHER DATASETS	
BioSample	8

SRA Data Details

Parameter	Value
Data volume, Gbases	9
Data volume, Mbytes	2826

https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB14430a

Quality Control

wget http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip
unzip Trimmomatic-0.39.zip

java -jar /lustre/home/2100012177/group/software/Trimmomatic-0.39/trimmomatic-0.39.jar SE -phred33 DRR424117.fastq.gz output_DRR424117.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 > trim.log 2>&1

LEADING：从 reads 的开头切除质量值低于阈值的碱基
TRAILING：从 reads 的末尾切除质量值低于阈值的碱基
SLIDINGWINDOW：滑窗过滤，去除碱基质量均值低于阈值的窗口
MINLEN：修剪完成后，丢弃长度小于阈值的reads
threads：线程数

```
TrimmomaticSE: Started with arguments:
 -phred33 DRR424117.fastq.gz output_DRR424117.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
Automatically using 2 threads
Input Reads: 17028652 Surviving: 17023269 (99.97%) Dropped: 5383 (0.03%)
TrimmomaticSE: Completed successfully
```

Align

wget https://jaist.dl.sourceforge.net/project/bowtie-bio/bowtie2/2.5.4/bowtie2-2.5.4-source.zip
unzip bowtie2-2.5.4-source.zip
make
vim ~/.bashrc
export PATH=/lustre/home/2100012177/group/software/bowtie2-2.5.4/:$PATH
source ~/.bashrc
bowtie2 -h
wget
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/017/639/785/GCF_017639785.1_BCM_Maur_2.0/GCF_017639785.1_BCM_Maur_2.0_genomic.fna.gz
gunzip GCF_017639785.1_BCM_Maur_2.0_genomic.fna.gz

bowtie2-build GCF_017639785.1_BCM_Maur_2.0_genomic.fna ref_genome
bowtie2 -x ref/ref_genome -p 4 -U output_DRR424117.fastq | samtools view -Sb > sample.bam
-x 参考基因组索引
-U -1 -2 输入fastq文件路径
-S 输出sam文件路径
-p 线程数　　更多参数可参考官方文档 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

```
17028652 reads; of these:
  17028652 (100.00%) were unpaired; of these:
    2195711 (12.89%) aligned 0 times
    11228194 (65.94%) aligned exactly 1 time
    3604747 (21.17%) aligned >1 times
87.11% overall alignment rate
```

Bam

```
1 DRR424117.1     0       NW_024429191.1  38419846        42      101M    *       0       0
1       GNGTGTTCTGGAAGTAATGGCGCCATAGCGGTCTGATCTTGTCCTGGCCGCCCACATCCCACACGGTGAAGCTGATATTCTTGTAC
1 TCAACTGTCTCCACG   ???????????????????????????????????????????????????????????????????????
1 ????????????????????????   AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:1G99
1 YT:Z:UU
```

对应上图	列序号	名称	内容描述
read name	1	QNAME	fq的read ID（QNAME中的Q表示Query）
flags	2	FLAG	比对信息位（bitwise FLAG），是一个由16位整数
position	3	RNAME	参考序列染色体名称（RNAME中的R表示Reference）
	4	POS	比对位置，从对应染色体的第1位开始往后计算
MAPQ	5	MAPQ	比对质量值（Mapping Quality）
CIGAR	6	CIGAR	比对信息
Mate information	7	RNEXT	配对read所比对到的染色体（仅Pair end 测序的数据才有）
	8	PNEXT	配对read所比对到的位置（仅Pair end 测序的数据才有）
	9	TLEN	插入片段长度（仅Pair end 测序的数据才有）
Read sequence	10	SEQ	read序列
Quality scores	11	QUAL	read质量值
metadata	12	Metadata	元信息，从第12列开始往后都是metadata，一般会包括RG信息，missmatch信息，二次比对信息等

Bam

```
1 DRR424117.1     0       NW_024429191.1  38419846        42      101M    *       0       0
1       GNGTGTTCTGGAAGTAATGGCGCCATAGCGGTCTGATCTTGTCCTGGCCGCCCACATCCCACACGGTGAAGCTGATATTCTTGTAC
1 TCAACTGTCTCCACG   ???????????????????????????????????????????????????????????????????????
1 ??????????????????????   AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:1G99
1 YT:Z:UU
```

M（匹配）：alignment match (can be a sequence match or mismatch)
表示read可mapping到第三列的序列上，则read的碱基序列与第三列的序列碱基相同，表示正常的mapping结果，M表示完全匹配，但是无论reads与序列的正确匹配或是错误匹配该位置都显示为M
I（插入）：insertion to the reference
表示read的碱基序列相对于第三列的RNAME序列，有碱基的插入
D（删除）：deletion from the reference
表示read的碱基序列相对于第三列的RNAME序列，有碱基的删除
N：skipped region from the reference　表示可变剪接位置
P：padding (silent deletion from padded reference)
S：soft clipping (clipped sequences present in SEQ) 这部分没比对上但保留了
H：hard clipping (clipped sequences NOT present in SEQ) 这部分没比对上且未保留

Sort

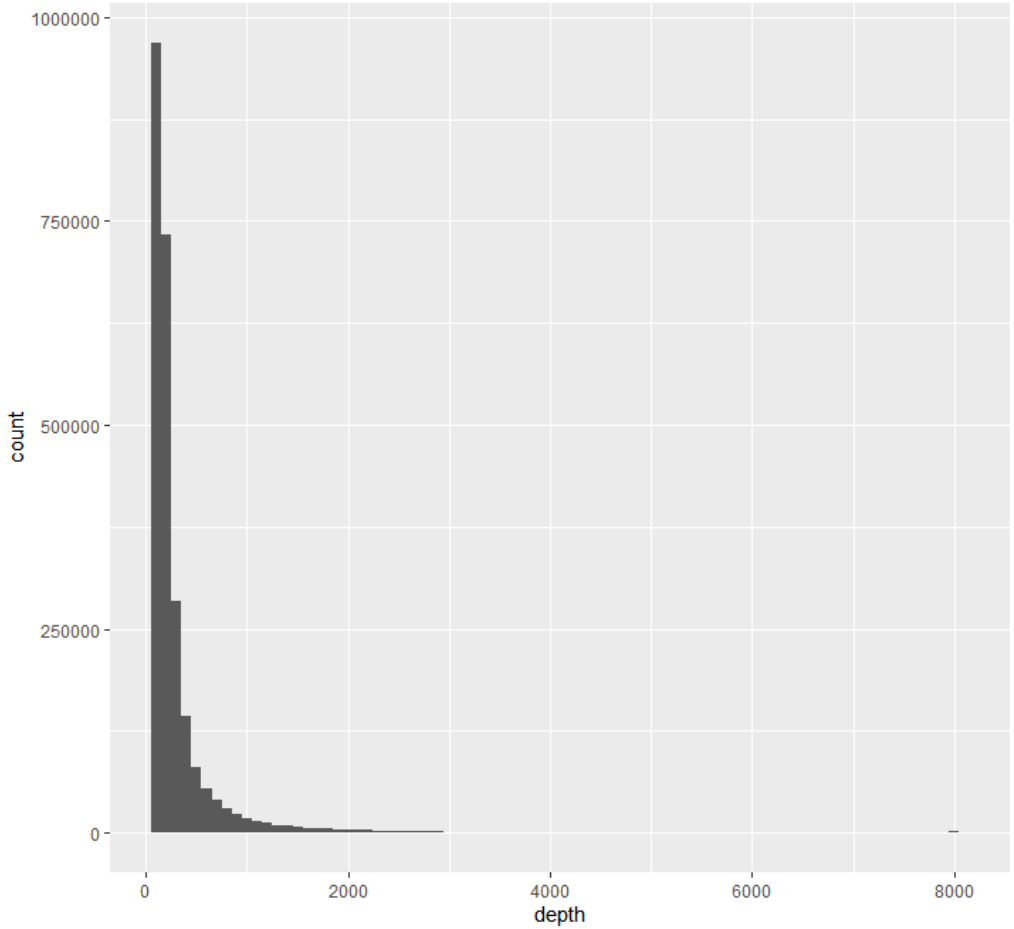
samtools sort sample.bam -o sample.sort.bam
samtools index sample.sort.bam
samtools faidx ref/GCF_017639785.1_BCM_Maur_2.0_genomic.fna
samtools depth sample.sort.bam > sample.sort.depth.txt

```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.9 (using htslib 1.9)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     fqidx          index/extract FASTQ
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags
     markdup        mark duplicates

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
     merge          merge sorted alignments
     mpileup        multi-way pileup
     sort           sort alignment file
     split          splits a file by read group
     quickcheck     quickly check if SAM/BAM/CRAM file appears intact
     fastq          converts a BAM to a FASTQ
     fasta          converts a BAM to a FASTA

  -- Statistics
     bedcov         read depth per BED region
     depth          compute the depth
     flagstat       simple stats
     idxstats       BAM index stats
     phase          phase heterozygotes
     stats          generate stats (former bamcheck)

  -- Viewing
     flags          explain BAM flags
     tview          text alignment viewer
     view           SAM<->BAM<->CRAM conversion
     depad          convert padded BAM to unpadded BAM
```

Sequencing depth

```
library("ggplot2")
data<-read.table("D:/Downloads/PKU/Linux基础/Final/sample.sort.depth.txt", header = FALSE)
colnames(data)[3] <- 'depth'
data_no1<-data[data$depth>100,]
ggplot(data = data_no1,aes(x=depth))+geom_histogram(binwidth = 100)
library(dplyr)
all_max_dplyr <- data %>% filter(depth == max(depth))
```

	V1	V2	depth
1	NW_024429180.1	50672298	8069
2	NW_024429180.1	50672300	8069



Variant

bcftools mpileup -Ou -f ref/GCF_017639785.1_BCM_Maur_2.0_genomic.fna sample.sort.bam | bcftools call -mv > var.vcf
less -SN var.vcf
grep -v "^##" var.vcf|less -SN

```
1 #CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample.sort.bam
2 NW_024429180.1  85763   .       T       C       3.22451 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=42   GT:PL:AD        1/1:30,3,0:0,1
3 NW_024429180.1  116721  .       A       G       3.22451 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=42   GT:PL:AD        1/1:30,3,0:0,1
4 NW_024429180.1  674853  .       G       A       3.83885 .       DP=7;VDB=0.02;SGB=-0.453602;RPBZ=0;MQBZ=0;BQBZ=0;SCBZ=0;MQ0F=0;AC=1;AN=2;DP4=5,0,2,0;MQ=42      GT:PL:AD        0/1:35,0,88:5,2
5 NW_024429180.1  680754  .       T       C       15.5373 .       DP=4;VDB=0.0257451;SGB=-0.556411;MQ0F=0;AC=2;AN=2;DP4=0,0,4,0;MQ=13     GT:PL:AD        1/1:45,12,0:0,4
```

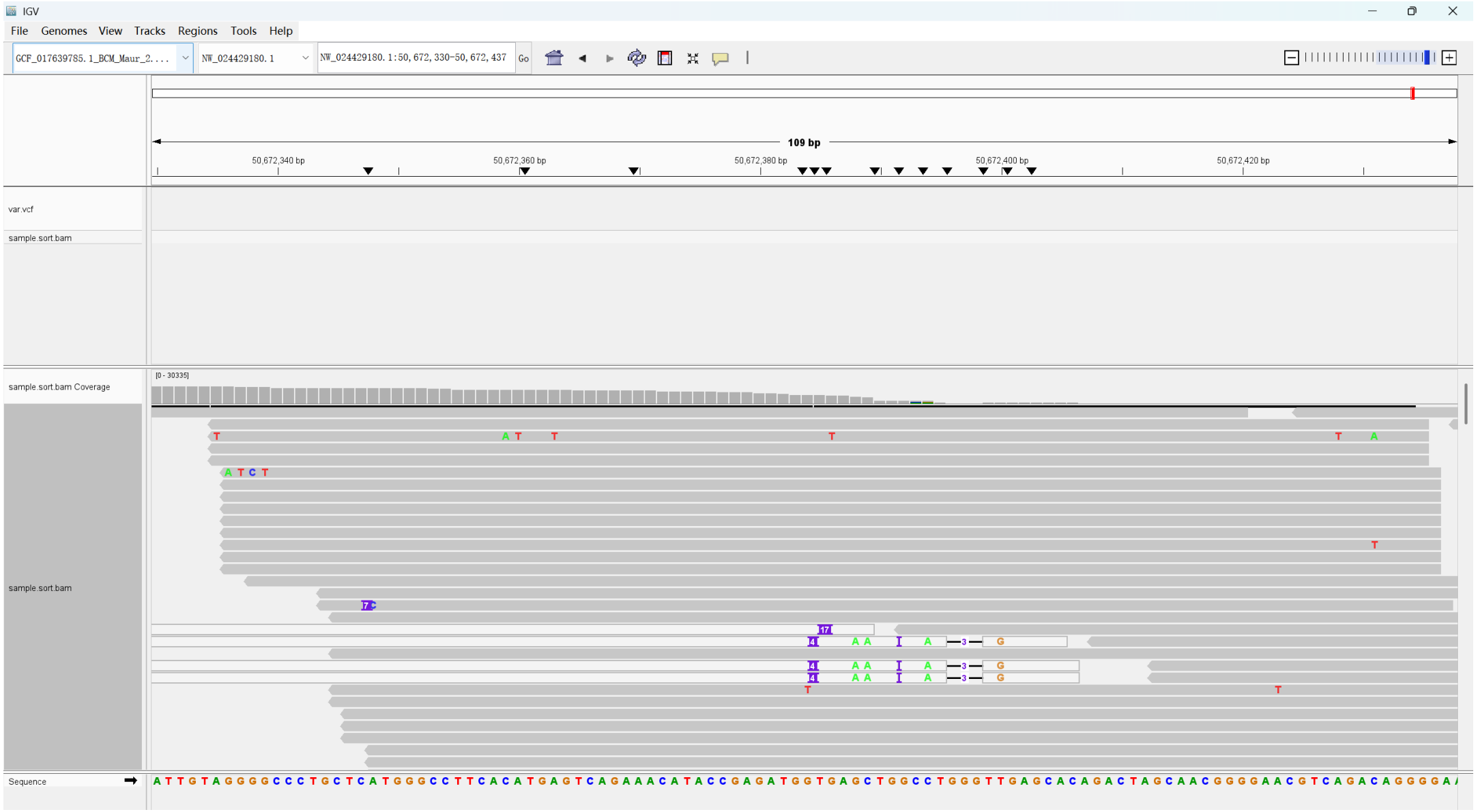
以#开头的部分为注释部分，主体部分每一行代表一个variant的信息。
主体部分10列分别代表的意义：
CHROM： 参考序列名称
POS： variant所在的left-most位置(1-base position)（发生变异的位置的第一个碱基所在的位置）
ID： variant的ID。同时对应着dbSNP数据库中的ID，若没有，则默认使用 ‘.’
REF： 参考序列的Allele，（等位碱基，即参考序列该位置的碱基类型及碱基数量）
ALT： variant的Allele，若有多个，则使用逗号分隔，（变异所支持的碱基类型及碱基数量）这里的碱基类型和碱基数量，对于SNP来说是单个碱基类型的编号，而对于Indel来说是指碱基个数的添加或缺失，以及碱基类型的变化
QUAL： variants的质量。Phred格式的数值，代表着此位点是纯合的概率，此值越大，则概率越低，代表着次位点是variants的可能性越大。（表示变异碱基的可能性）
FILTER： 次位点是否要被过滤掉。如果是PASS，则表示此位点可以考虑为variant。
INFO： variant的相关信息
FORMAT： variants的格式，例如GT:AD:DP:GQ:PL
SAMPLES： 各个Sample的值，由BAM文件中的@RG下的SM标签所决定，这些值对应着第9列的各个格式，不同格式的值用冒号分开，每一个sample对应着1列；多个samples则对应着多列，这种情况下列的数多余10列。

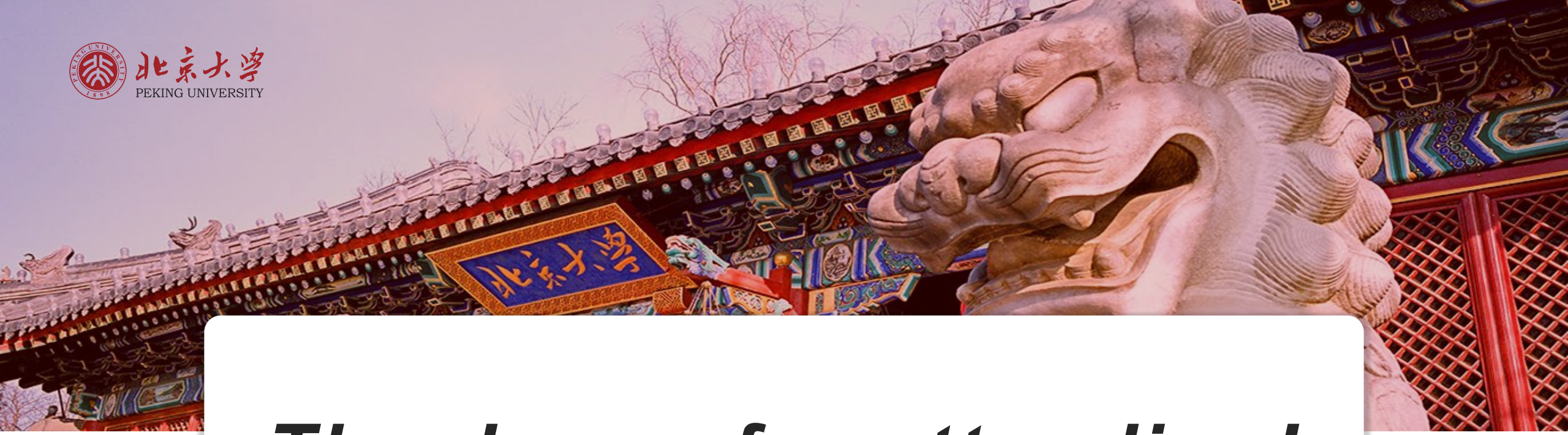
Visualization

samtools tview --reference /lustre/home/2100012177/group/Linux_final/ref/GCF_017639785.1_BCM_Maur_2.0_genomic.fna sample.sort.bam

点表示正链比对，逗号表示负链比对。
AGTCN代表正链上与参考序列不同的碱基，agtcn代表负链上与参考序列不同的碱基。
不同的颜色代表不同的比对质量值：白色>=30，黄色20-29，绿色 10-19，蓝色0-9
在tview模式里按下?问号，获得帮助。
NW_024429180.1:675915 5M3I93M

Visualization





Thank you for attending!

SNP Identification Using Next-Generation Sequencing (NGS)

冯婉滢 刘鹤 钱喜静

2025/06/11