第六章：基因组学

浙江大学　陈飘飘

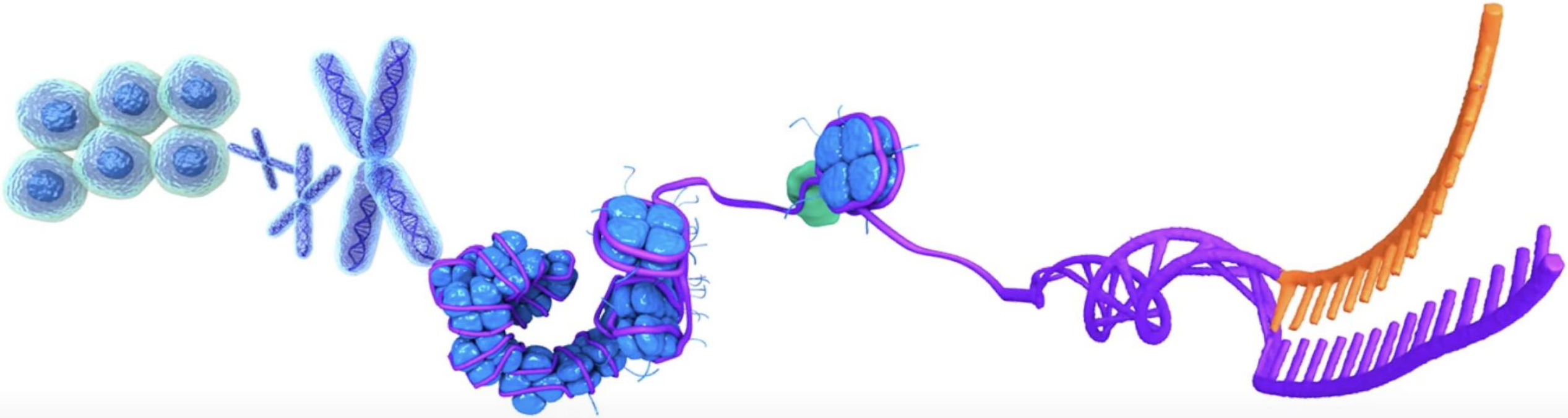
本章内容

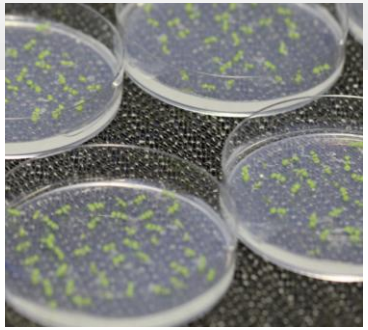
章节结构

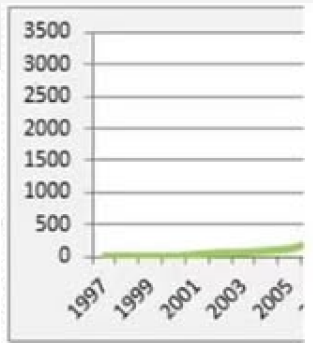
- **第一节：基因组学概述**
- **第二节：序列变异检测**

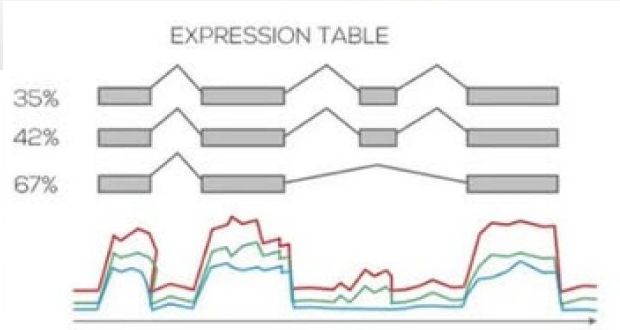
Omics Data Molecular Determinants of a Phenotype

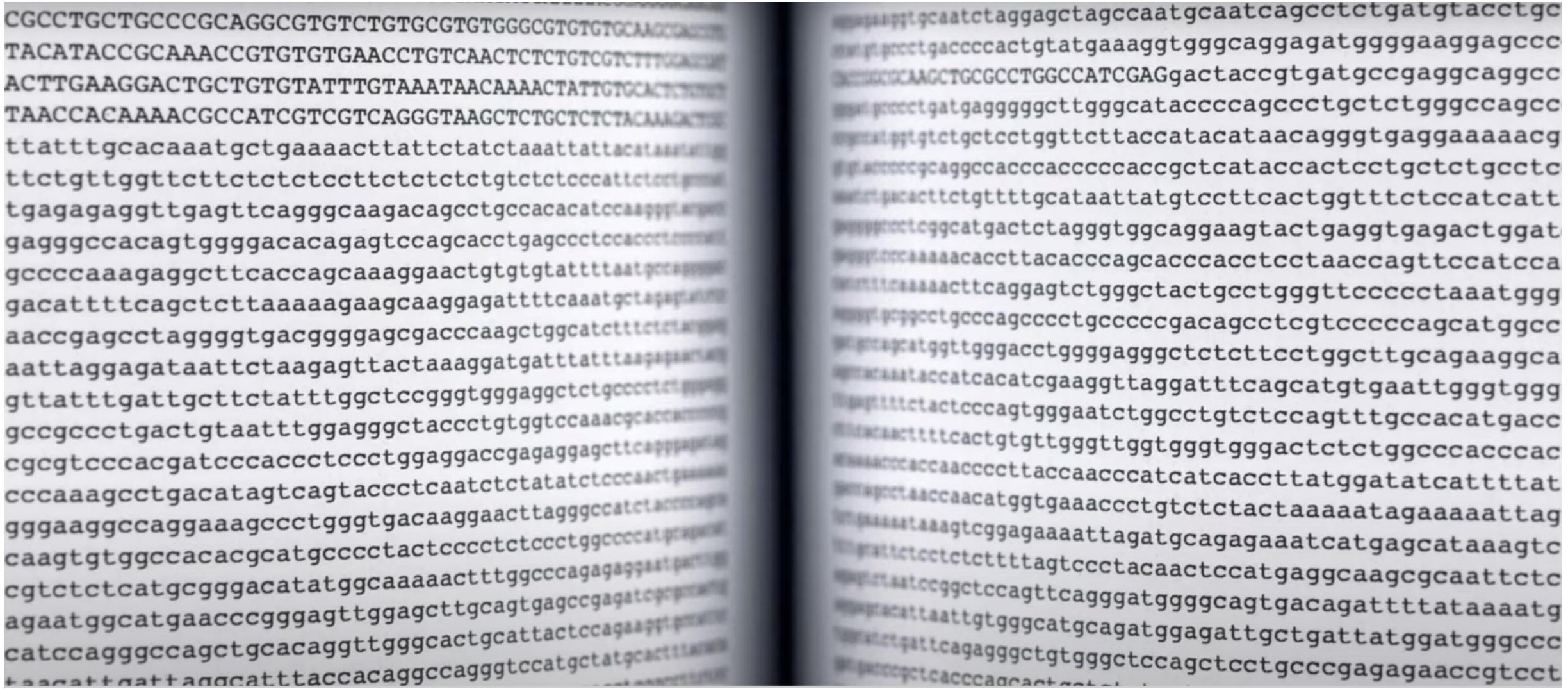
Phenomics **Genomics** **Epigenomics** **Transcriptomics**











Fundamentals of genomics: code and analysis of genomic data

第一节：基因组学概述

基因组学概述

- 基因组学（Genomics），简单来说就是研究基因组（Genome）的科学

- 在基因组学中所研究的问题主要分为三类：
  - 如何获得基因组序列
  - 如何解读/解码基因组
  - 如何重写/编写新的基因组

如何获得基因组序列——DNA测序

测序技术概述

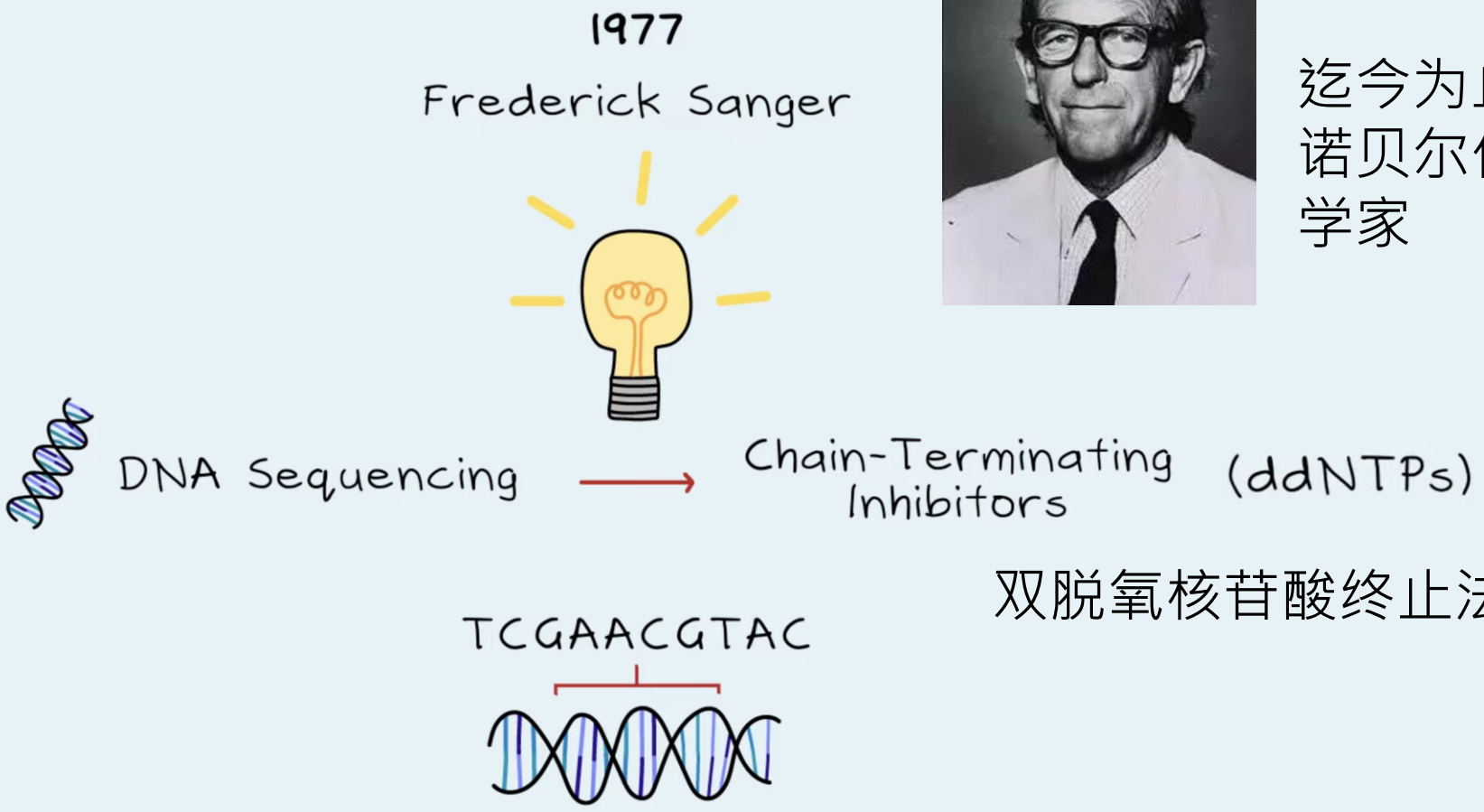
- **一代测序（代表：Sanger 测序）**
  - 优点：测序读长可达1000bp, 准确性高达99.999%
  - 缺点：测序成本高，通量低

- **二代测序（代表：Illumina公司的 HiSeq 技术）**
  - 优点：高准确性、高通量、 高灵敏度、低成本
  - 缺点：读长短 （小于500bp）

- **三代测序（代表：PacBio公司的SMRT、Oxford Nanopore Technologies的纳米孔单分子测序技术）**
  - Nanopore最长的读长可达3万个碱基。是目前最长的测序技术。同时可以测到DNA上的甲基化修饰，测序的速度也很快
  - 缺点：准确性相对低、通量相对低、成本相对高

一代测序

Sanger 测序



迄今为止**唯一**两获诺贝尔化学奖的科学家

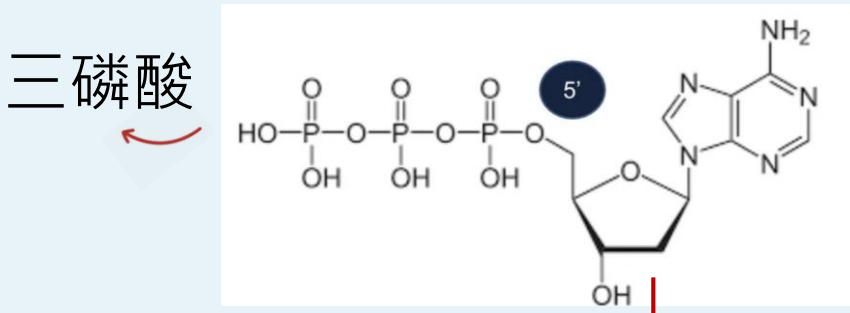
双脱氧核苷酸终止法

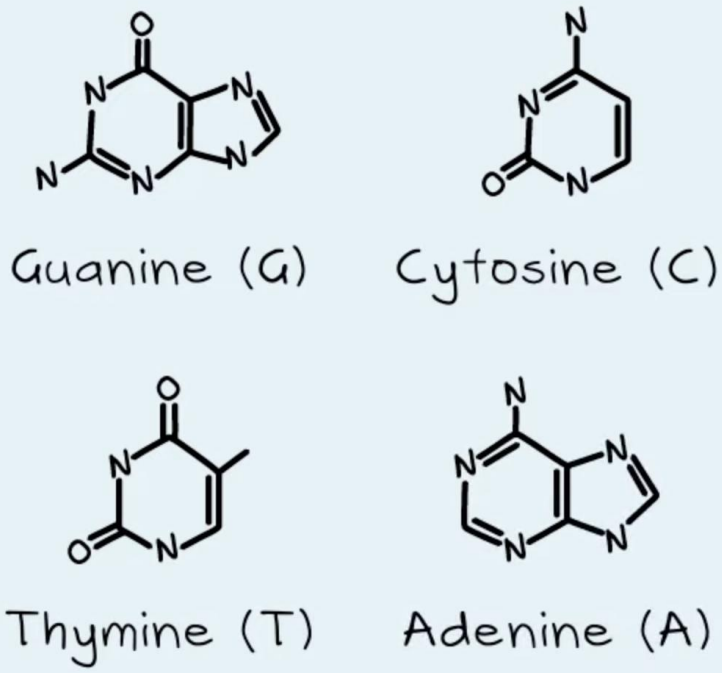
dNTP 结构

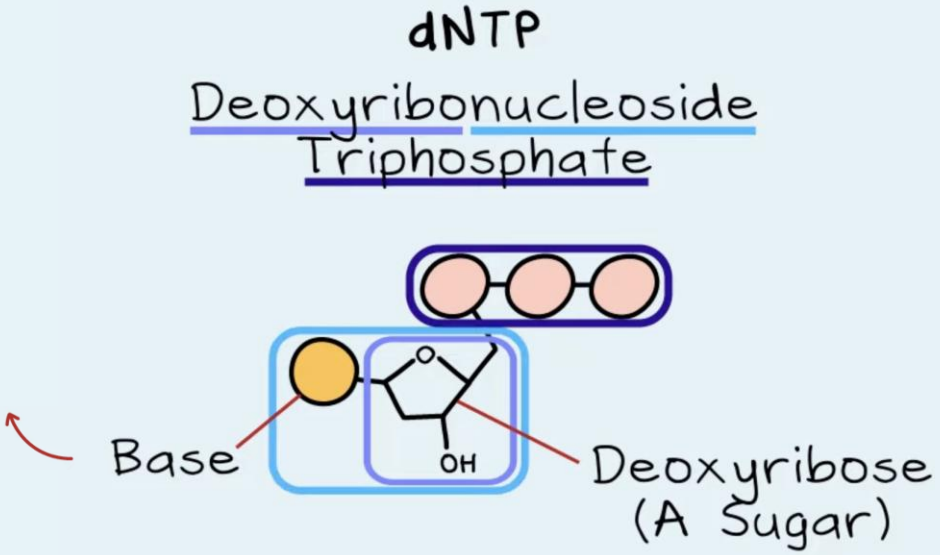
dNTP（脱氧核苷三磷酸）

Deoxyribonucleotide Triphosphate

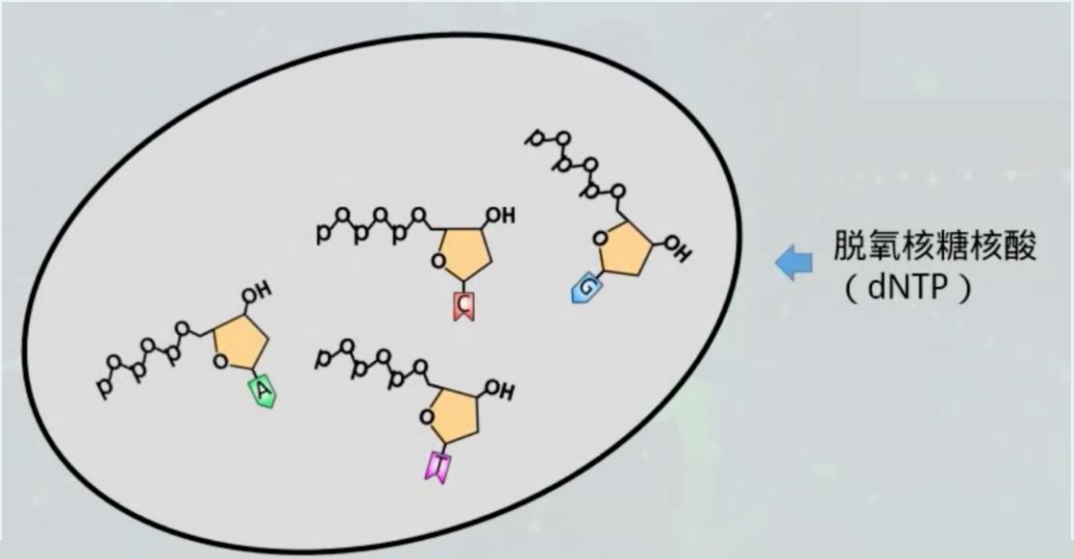
DNA 合成所需的“原材料”

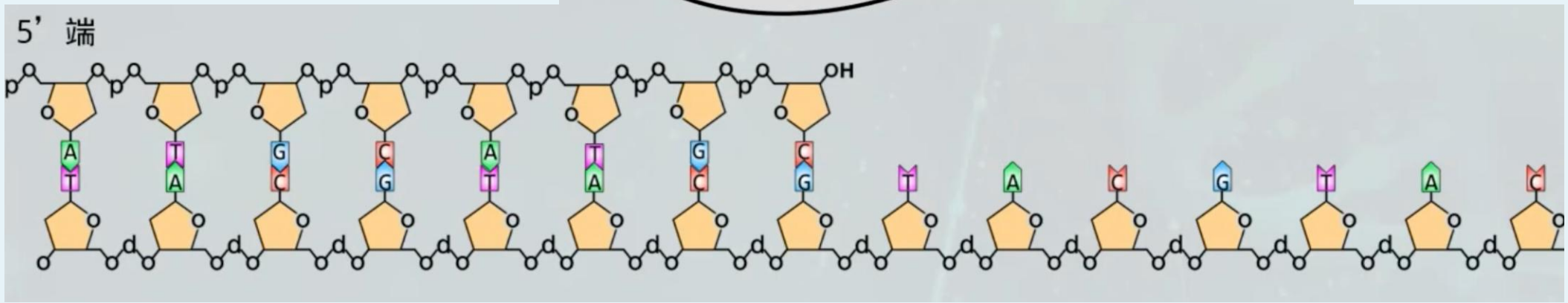






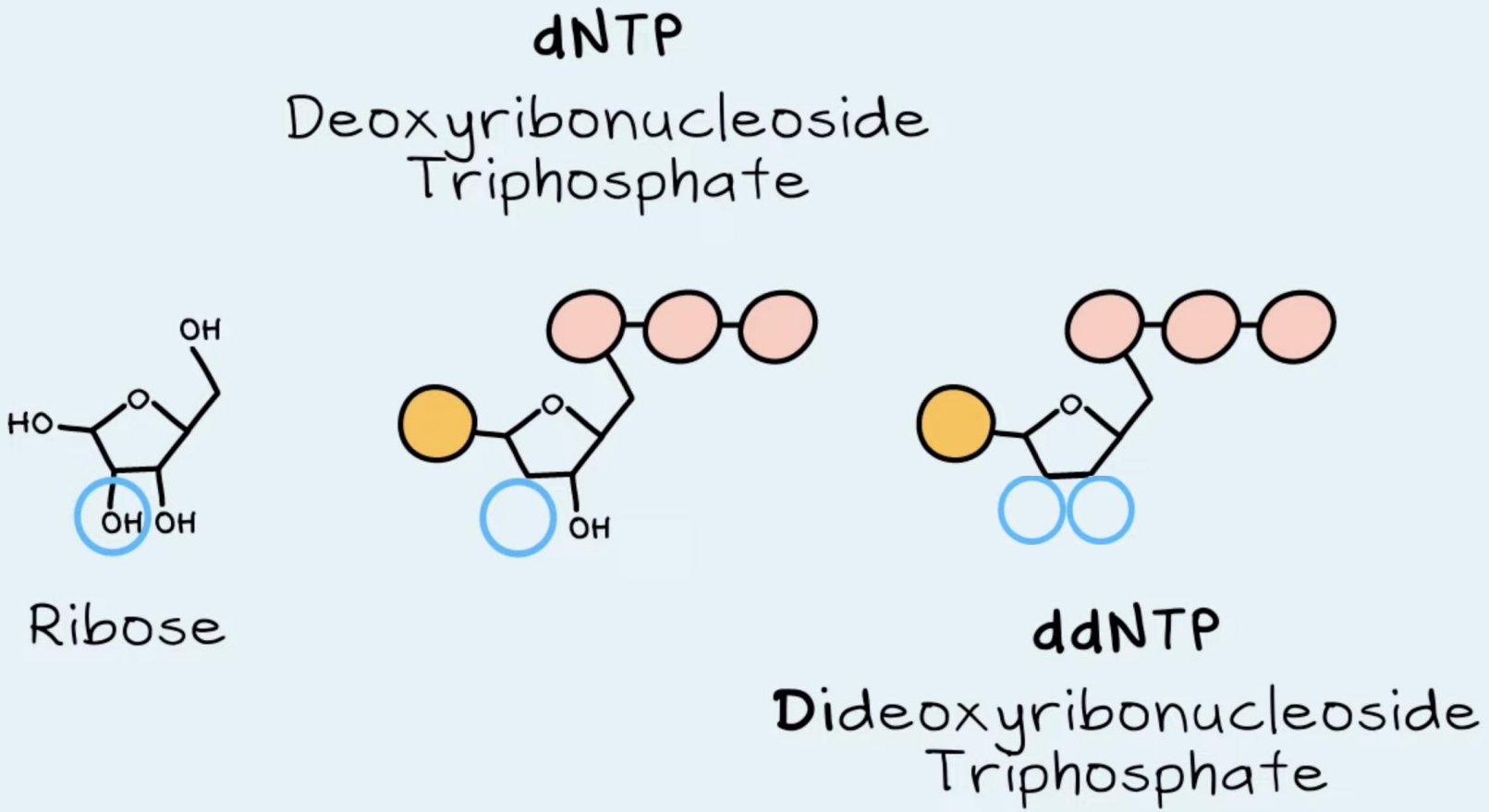
DNA链的延伸



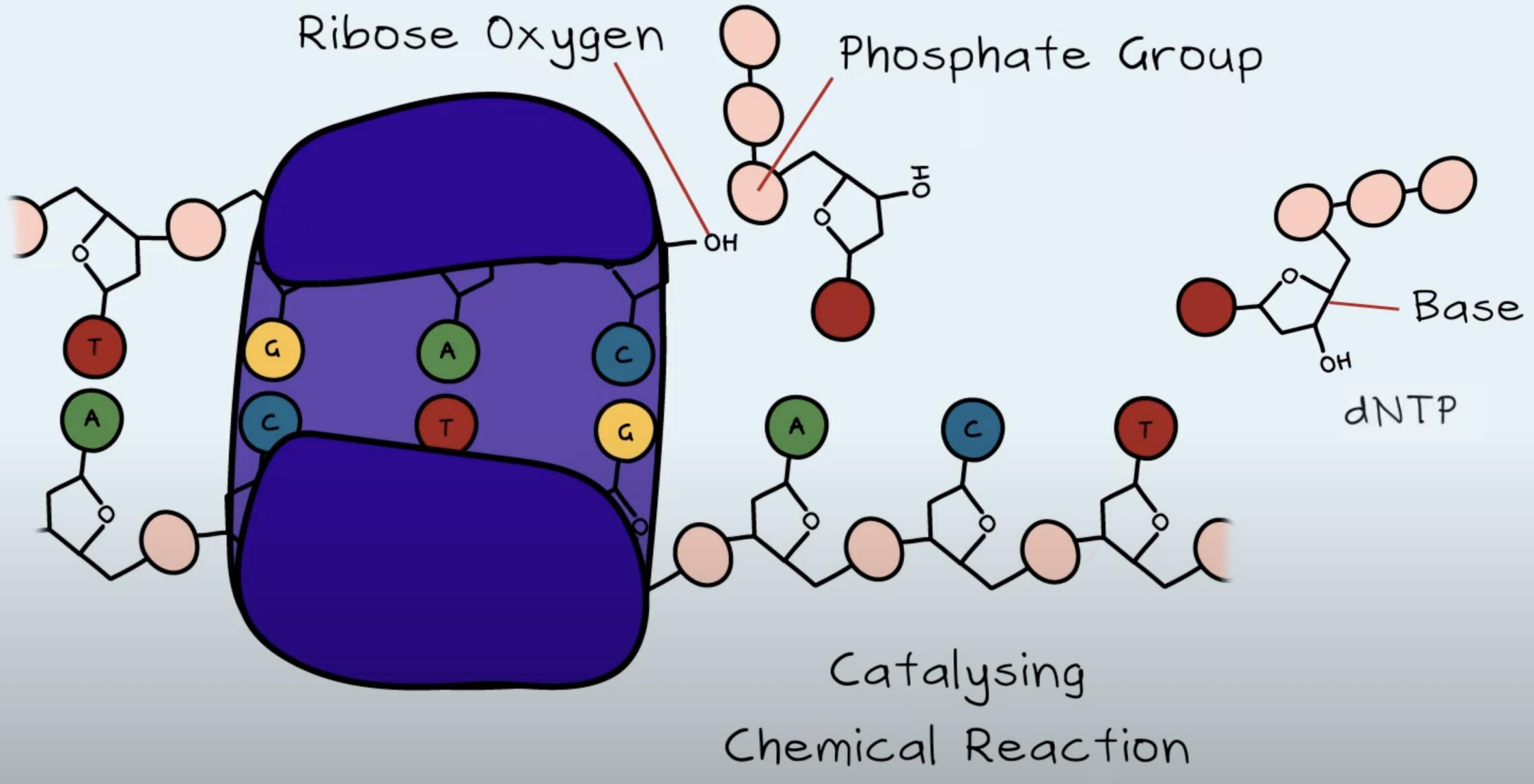


ddNTP 结构

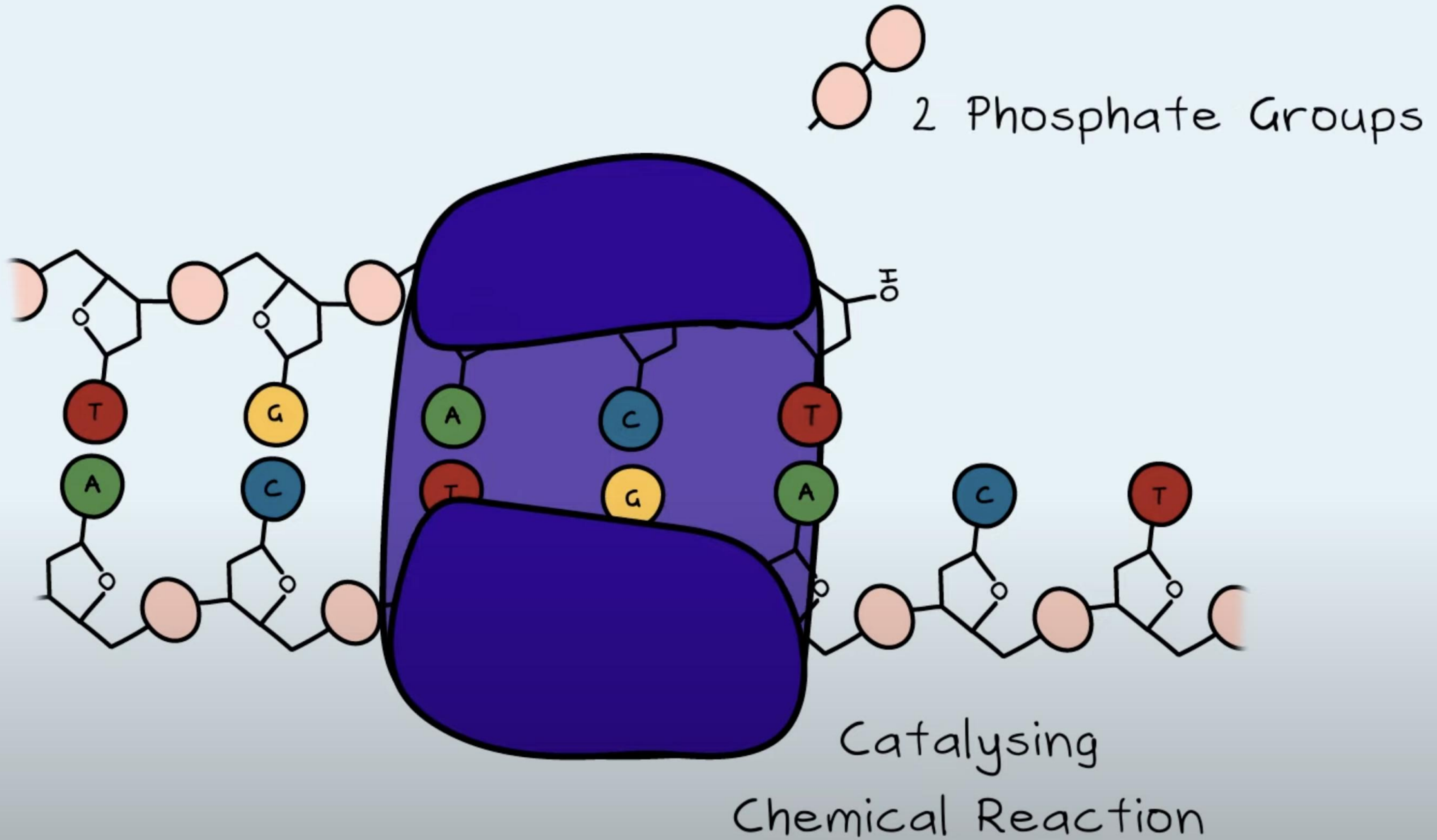
ddNTP（双脱氧核苷三磷酸）



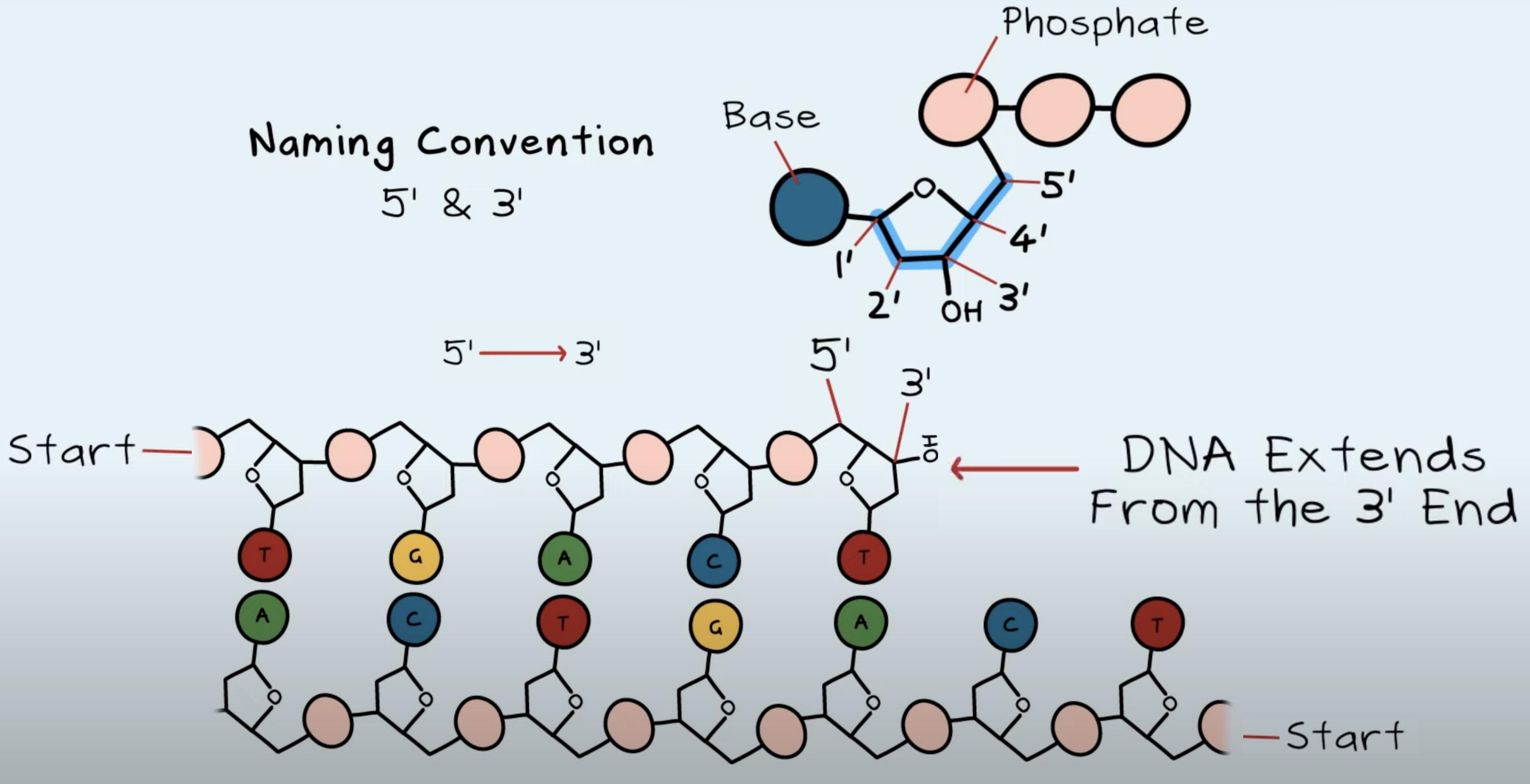
DNA 聚合反应机制


Ribose Oxygen
Phosphate Group
OH
HO
Base
OH
dNTP
T
A
G
C
A
T
C
G
A
C
T
Catalysing
Chemical Reaction

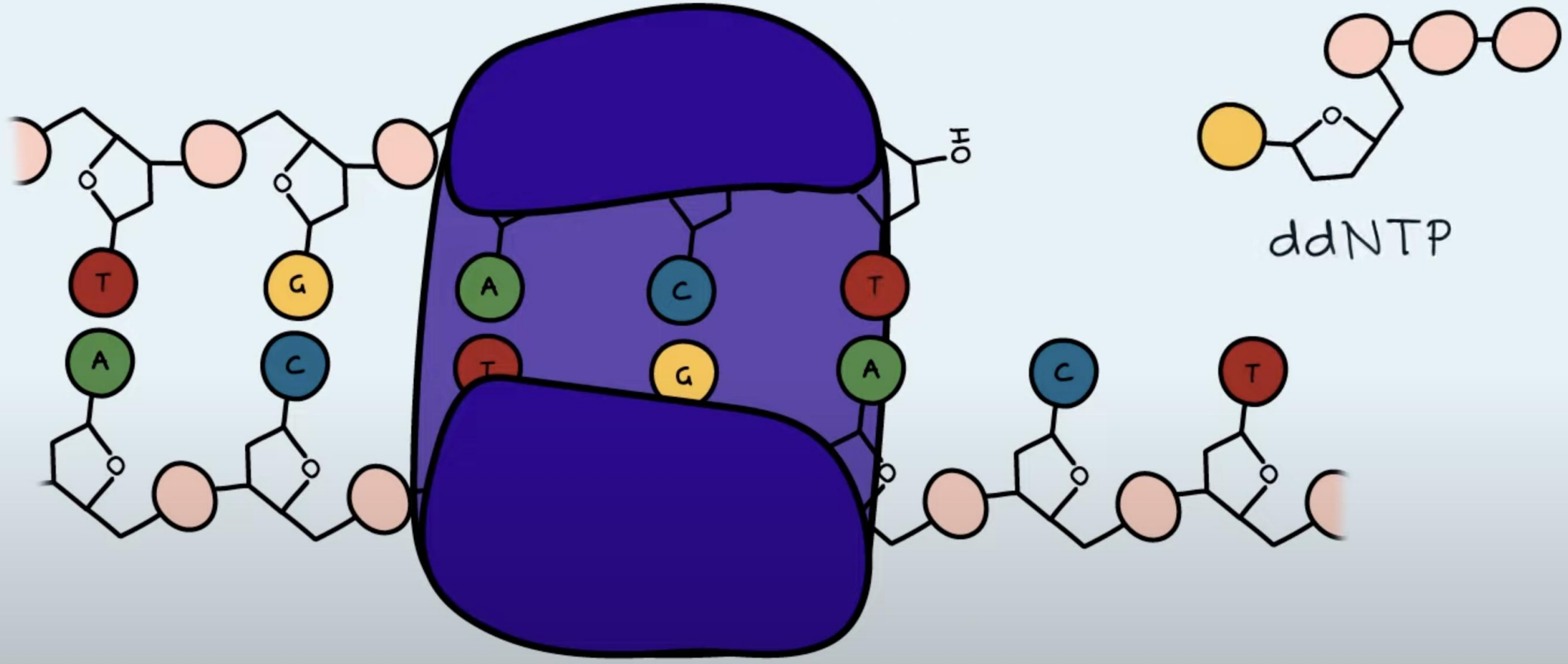
DNA 聚合反应机制


2 Phosphate Groups
T
G
A
C
A
C
T
G
A
C
T
Catalysing
Chemical Reaction

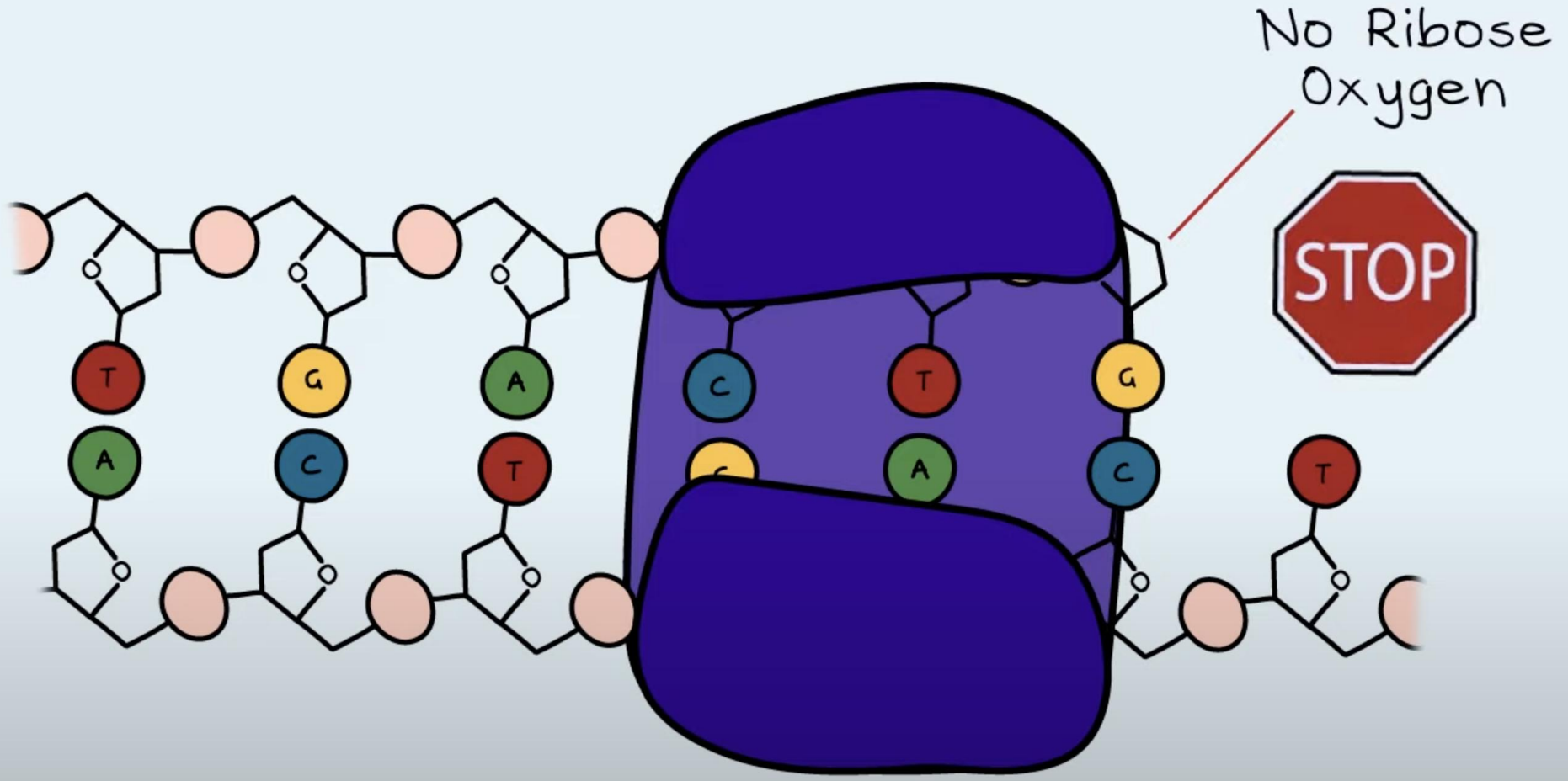
链方向性


Naming Convention
5' & 3'
Phosphate
Base
5'
4'
1'
2'
OH
3'
5' → 3'
5'
3'
OH
Start
DNA Extends
From the 3' End
T
G
A
C
T
A
C
T
G
A
C
T
Start

链终止原理



链终止原理

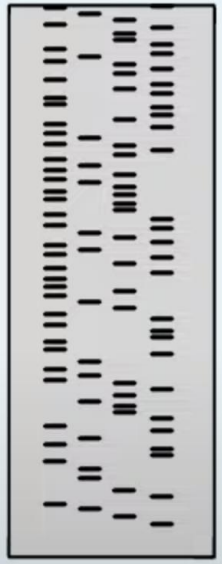


Sanger 测序的手动时代和自动化时代

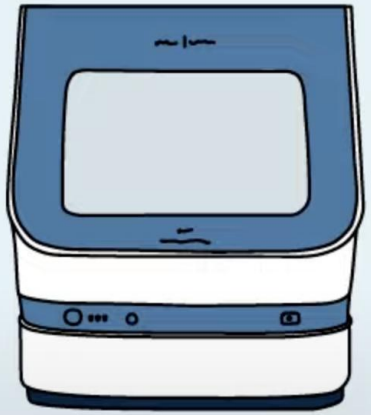
Original

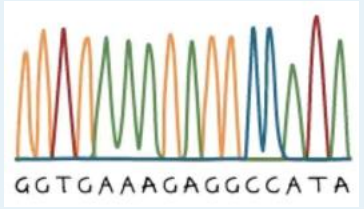
Manual

Radioactive Dyes

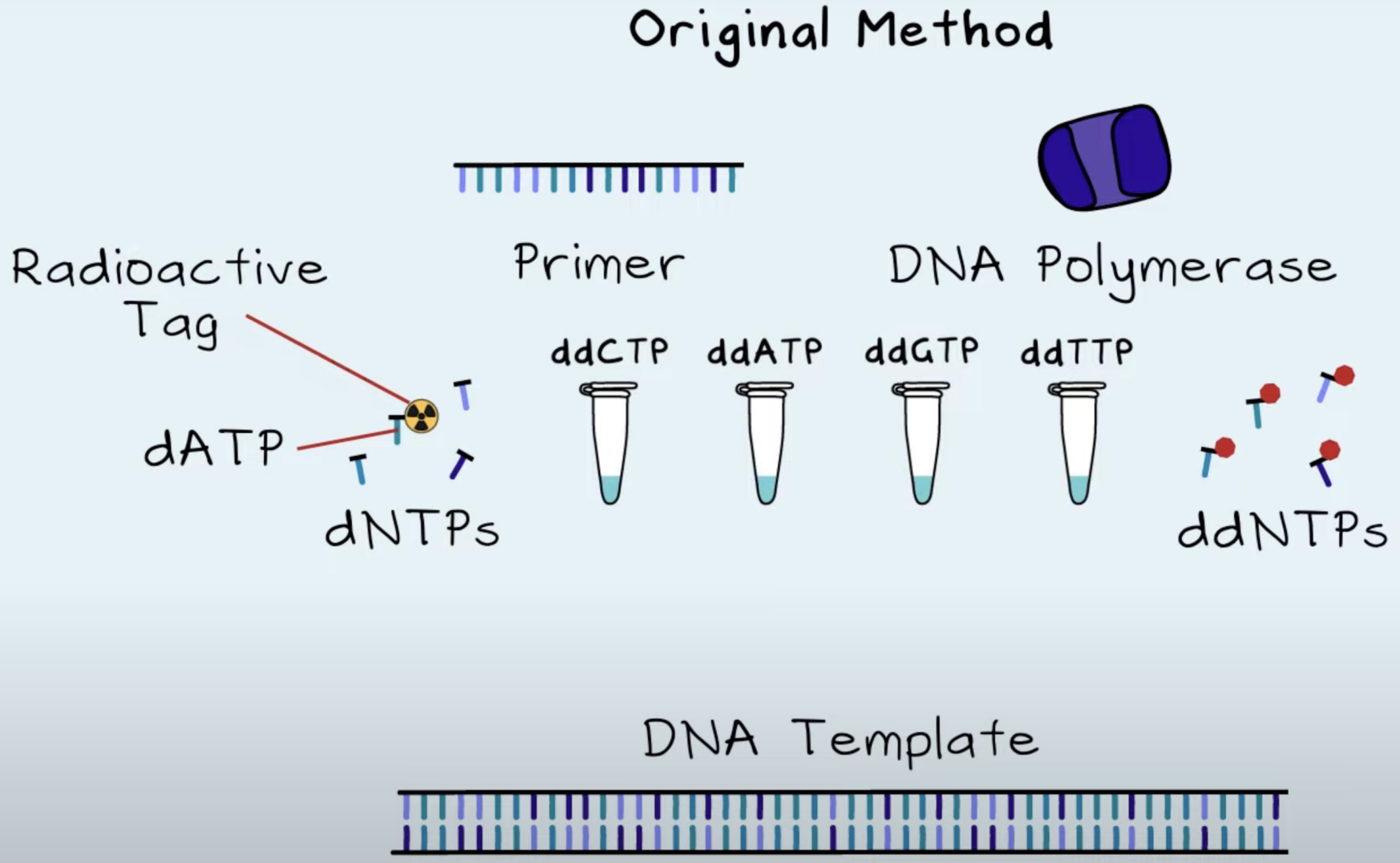


Today

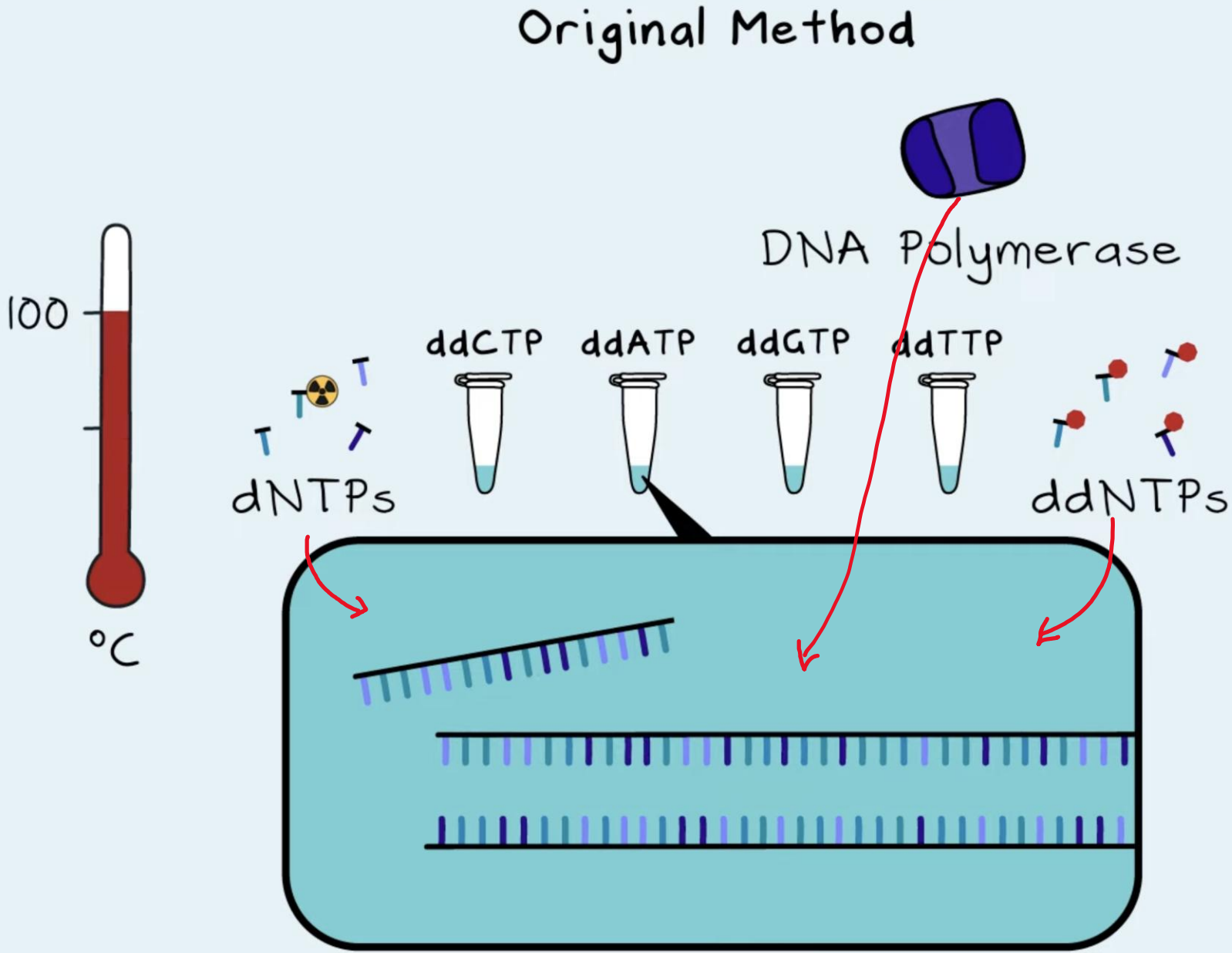




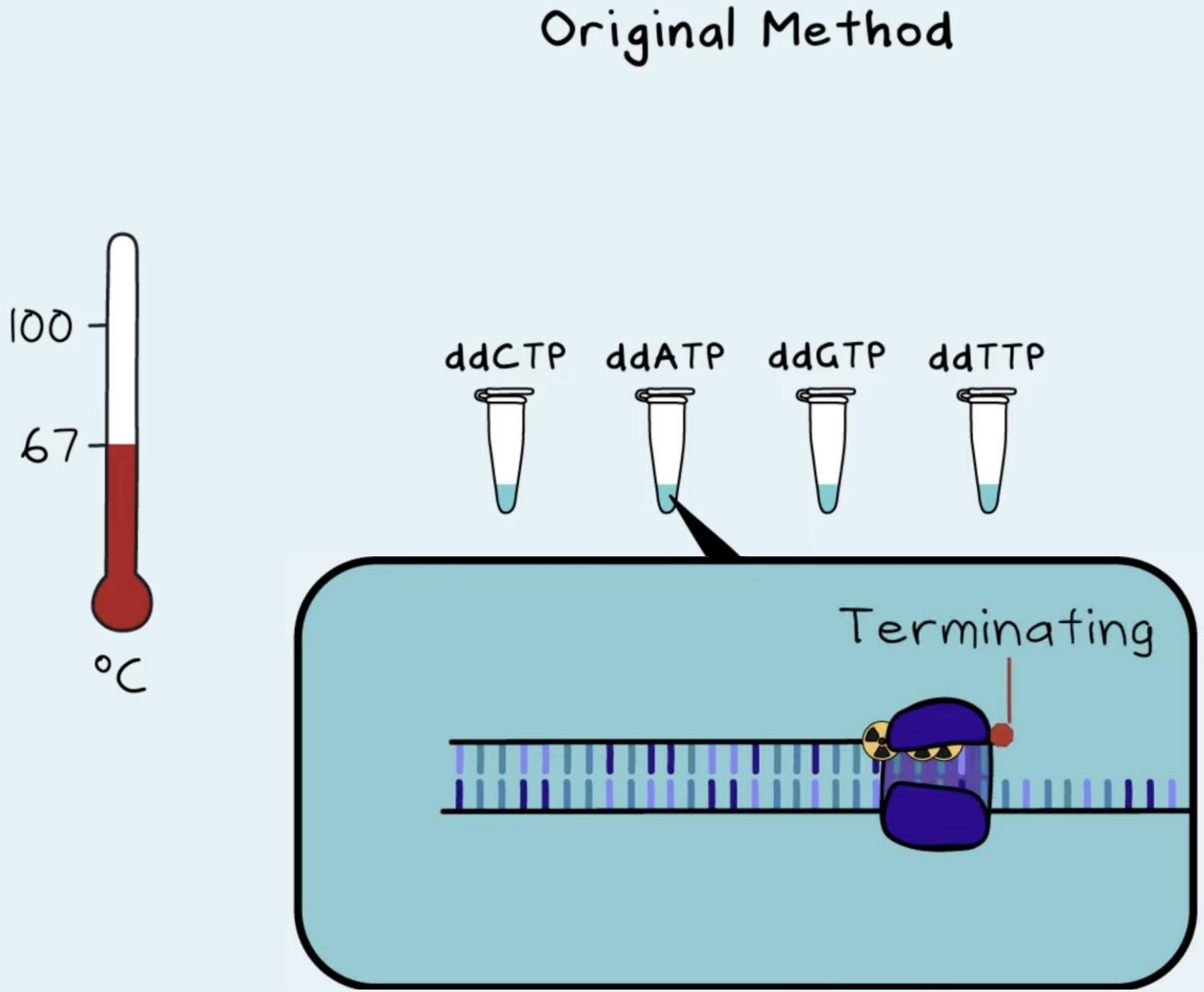
Sanger Sequencing 反应过程



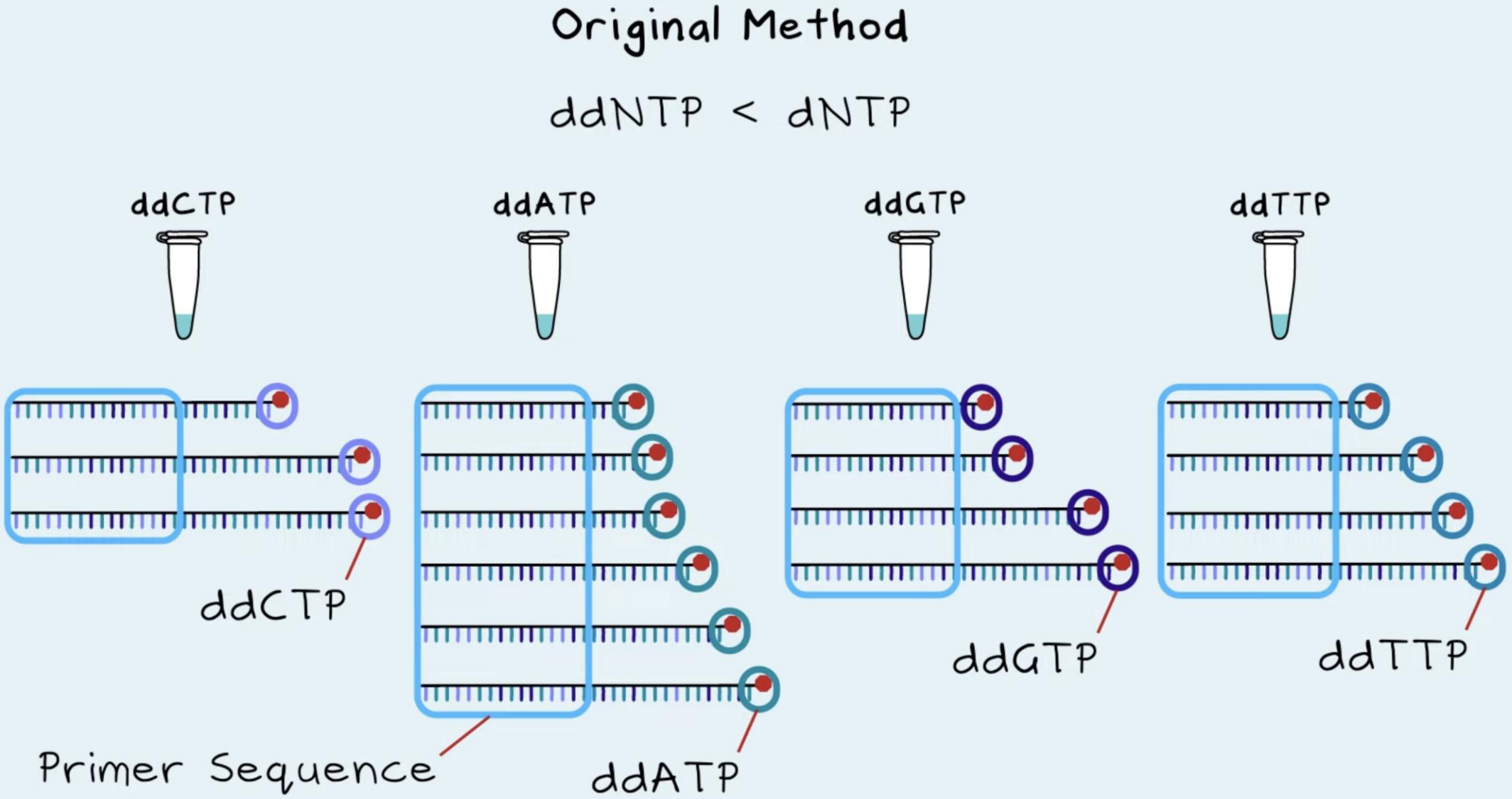
Sanger Sequencing 反应过程



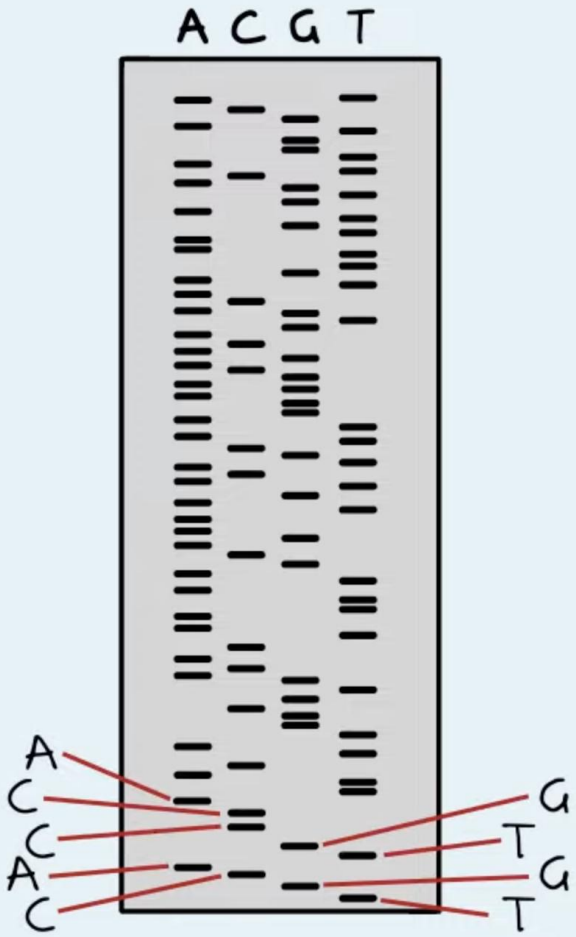
Sanger Sequencing 反应过程



终止反应结果


Original Method
ddNTP < dNTP
ddCTP
ddATP
ddGTP
ddTTP
ddCTP
Primer Sequence
ddATP
ddGTP
ddTTP

电泳结果与序列读取

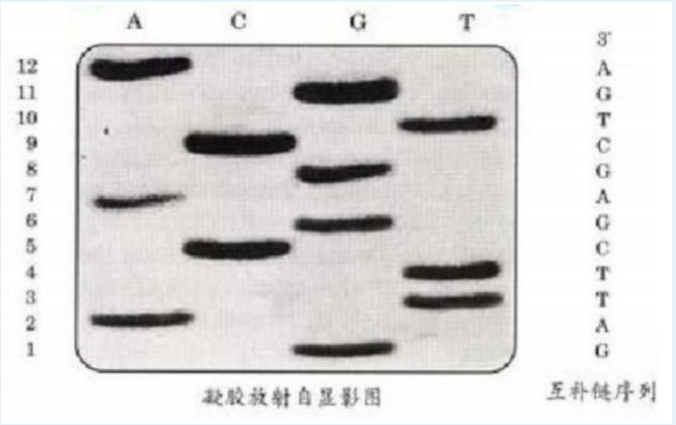


聚丙烯酰胺凝胶

Base Calling

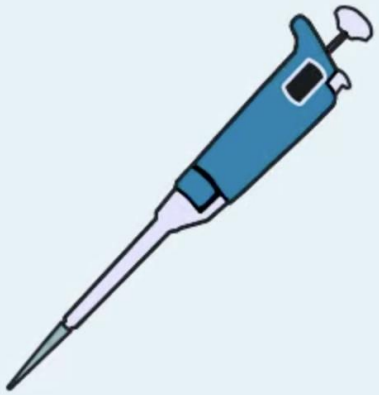
5' to 3'

TGCATGCCA



早期 Sanger 测序的局限性

Labour Intensive



4 Days

200 Nucleotides

Few Samples

自动化时代的开启：第一台商业化测序仪（1987）

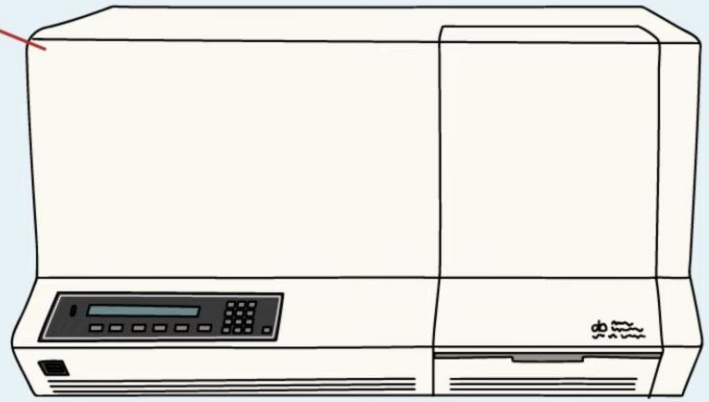
Applied Biosystems

First Commercial Sequencer

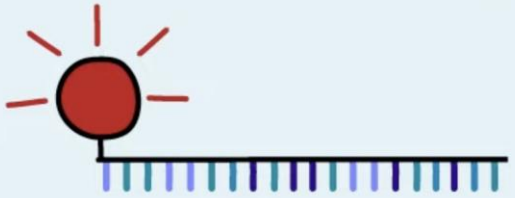
1987

Radioactive Dyes

AB370A



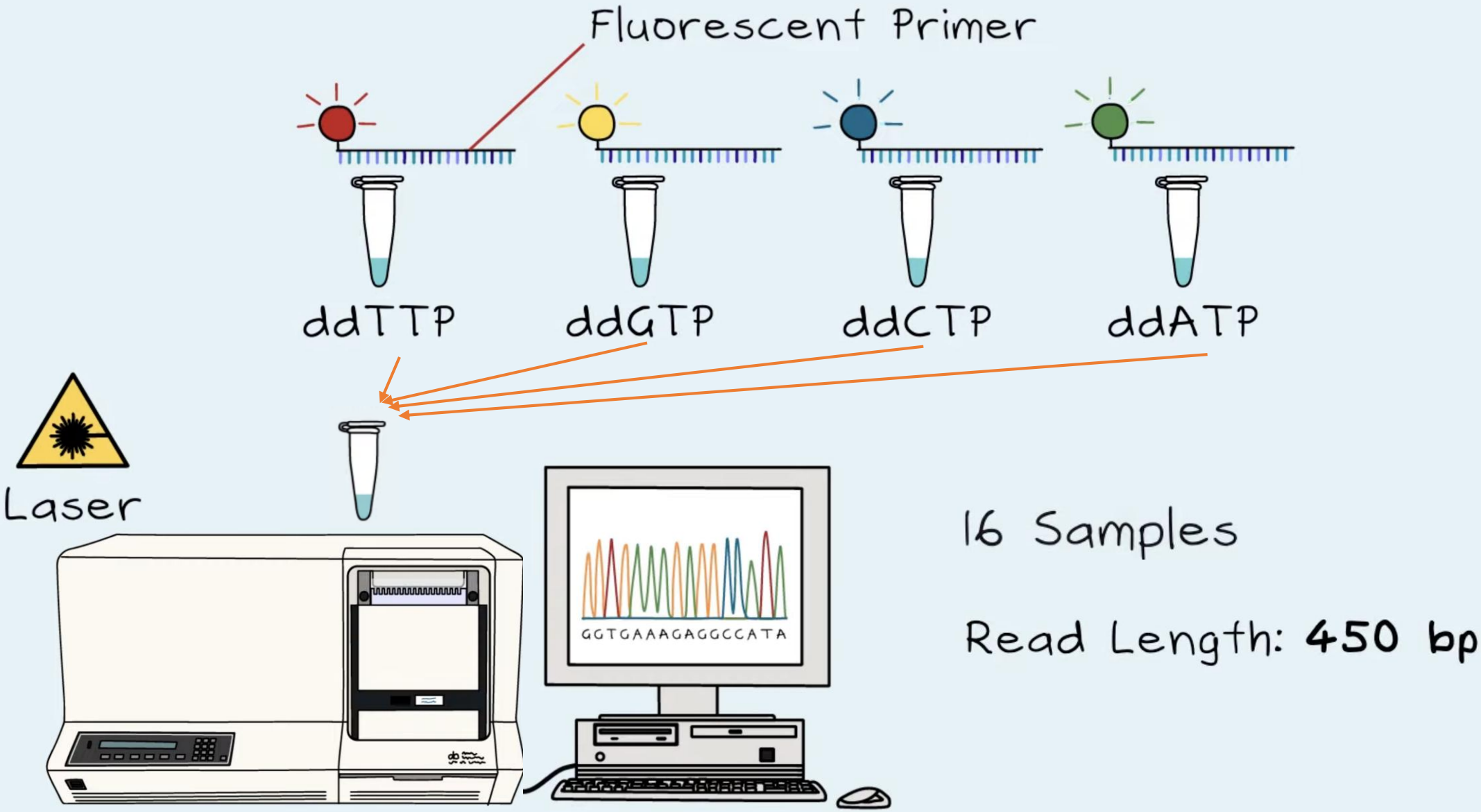
Fluorescent Dyes



- Safer
- No X-Ray Film

Saves 3 Days

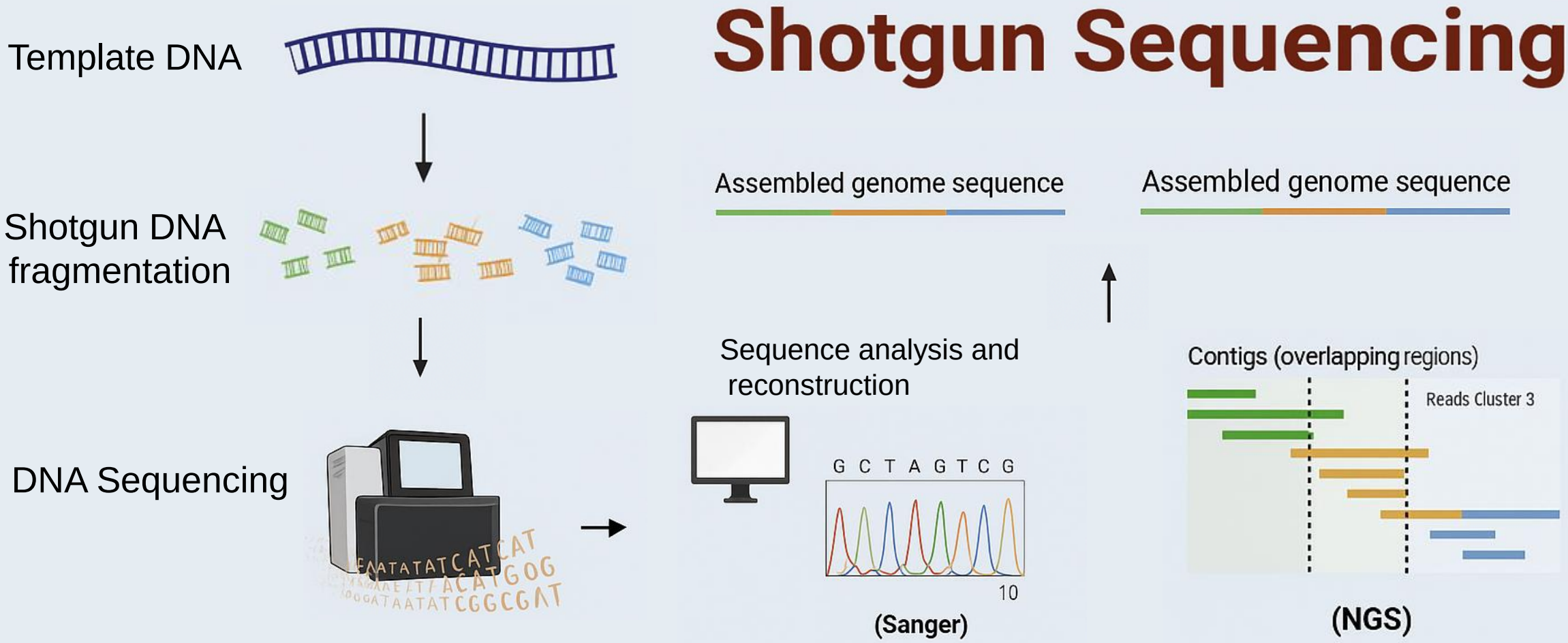
自动化 Sanger 测序



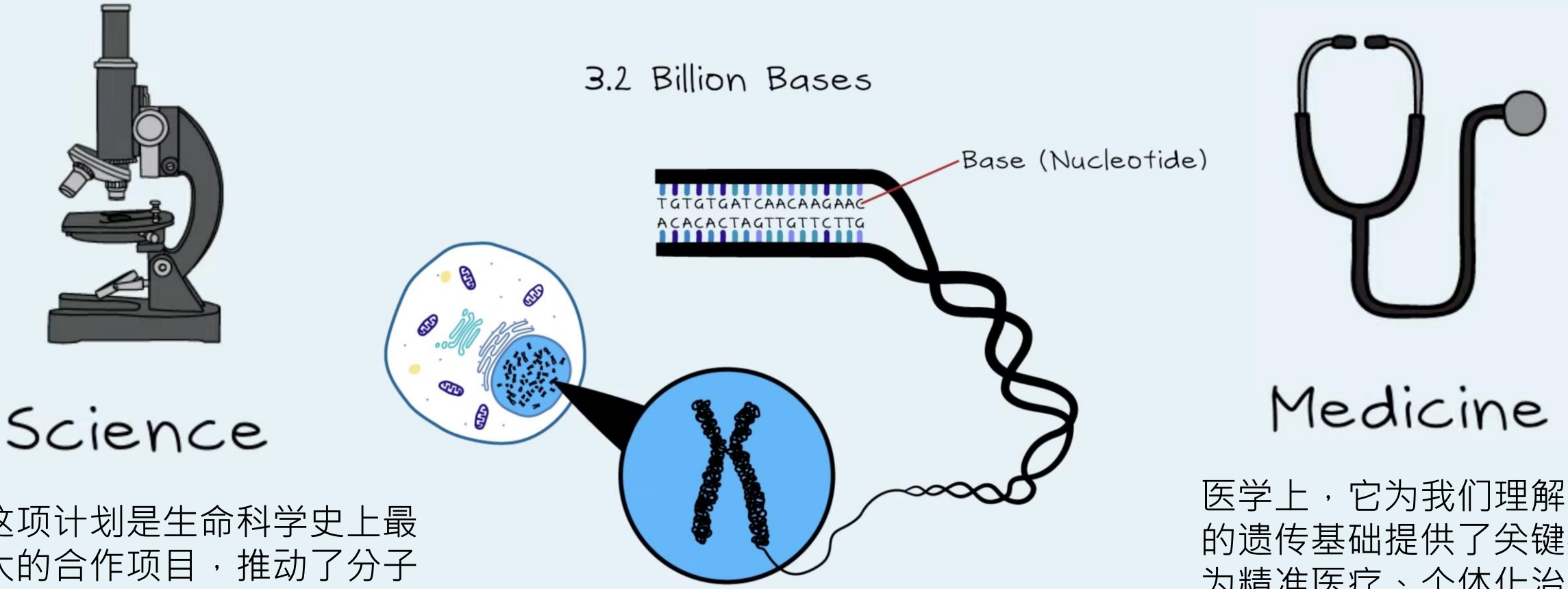
16 Samples

Read Length: **450 bp**

鸟枪法（shotgun sequencing)



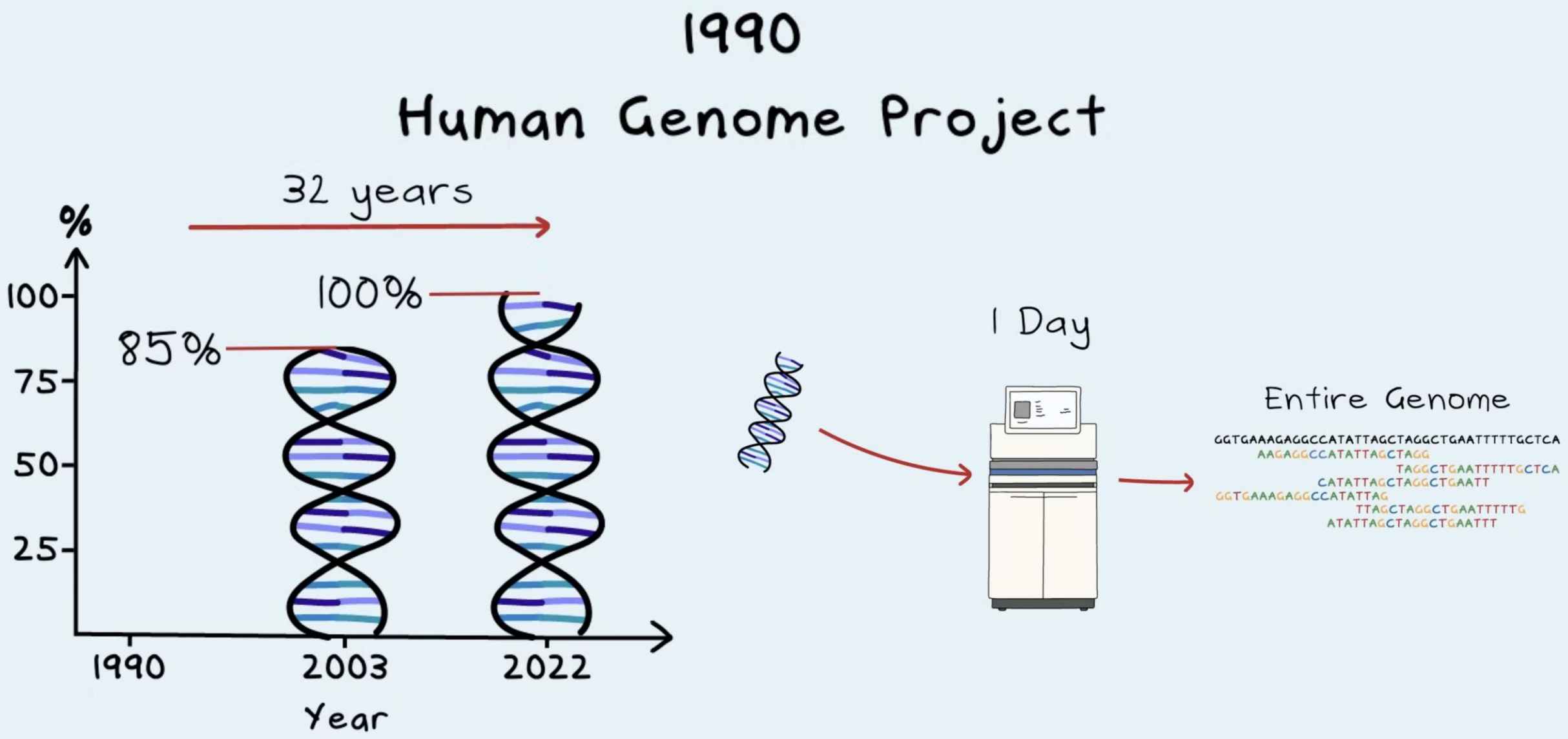
人类基因组计划（Human Genome Project）



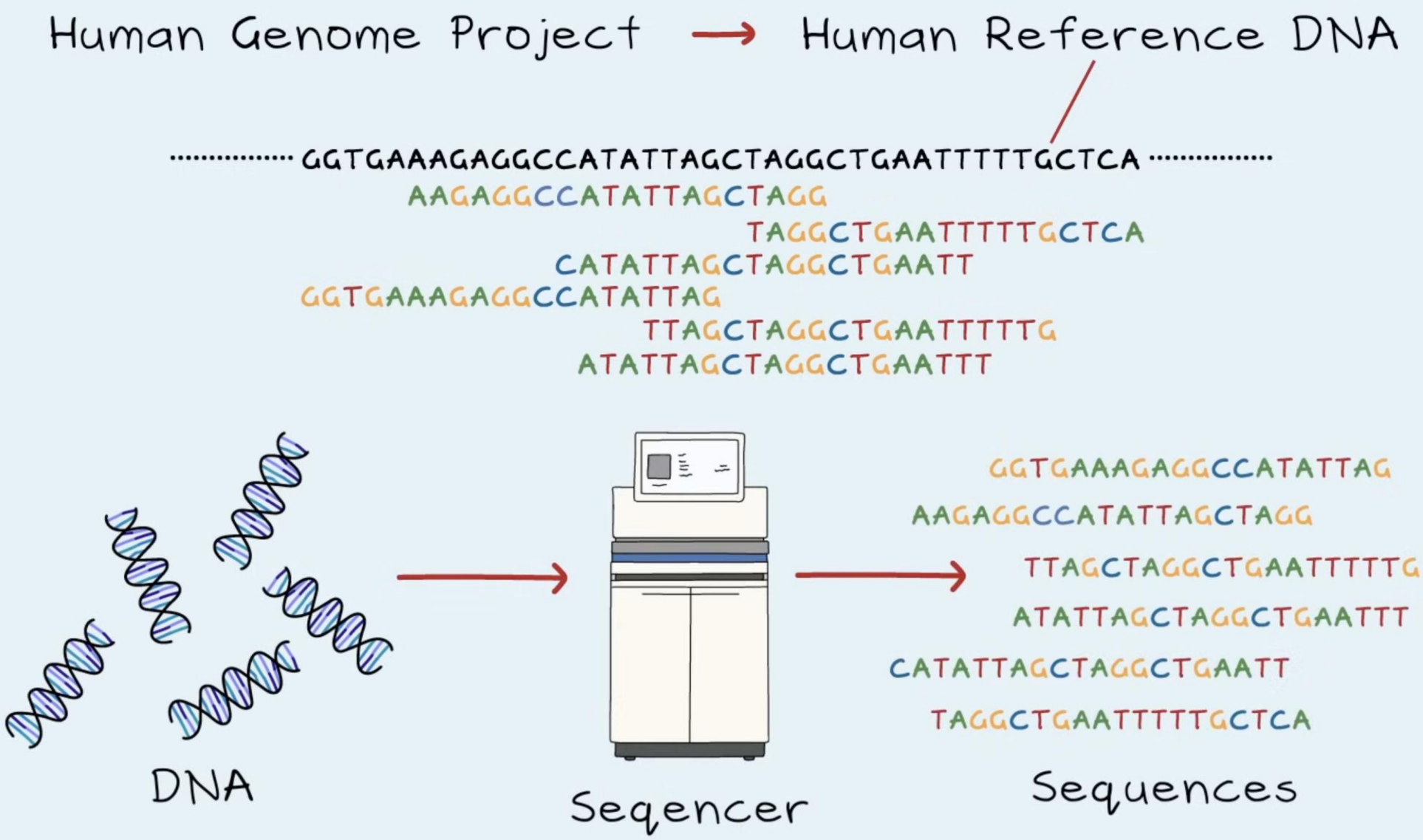
这项计划是生命科学史上最大的合作项目，推动了分子生物学、遗传学、以及生物信息学的迅速发展。

医学上，它为我们理解疾病的遗传基础提供了关键数据，为精准医疗、个体化治疗奠定了基础。

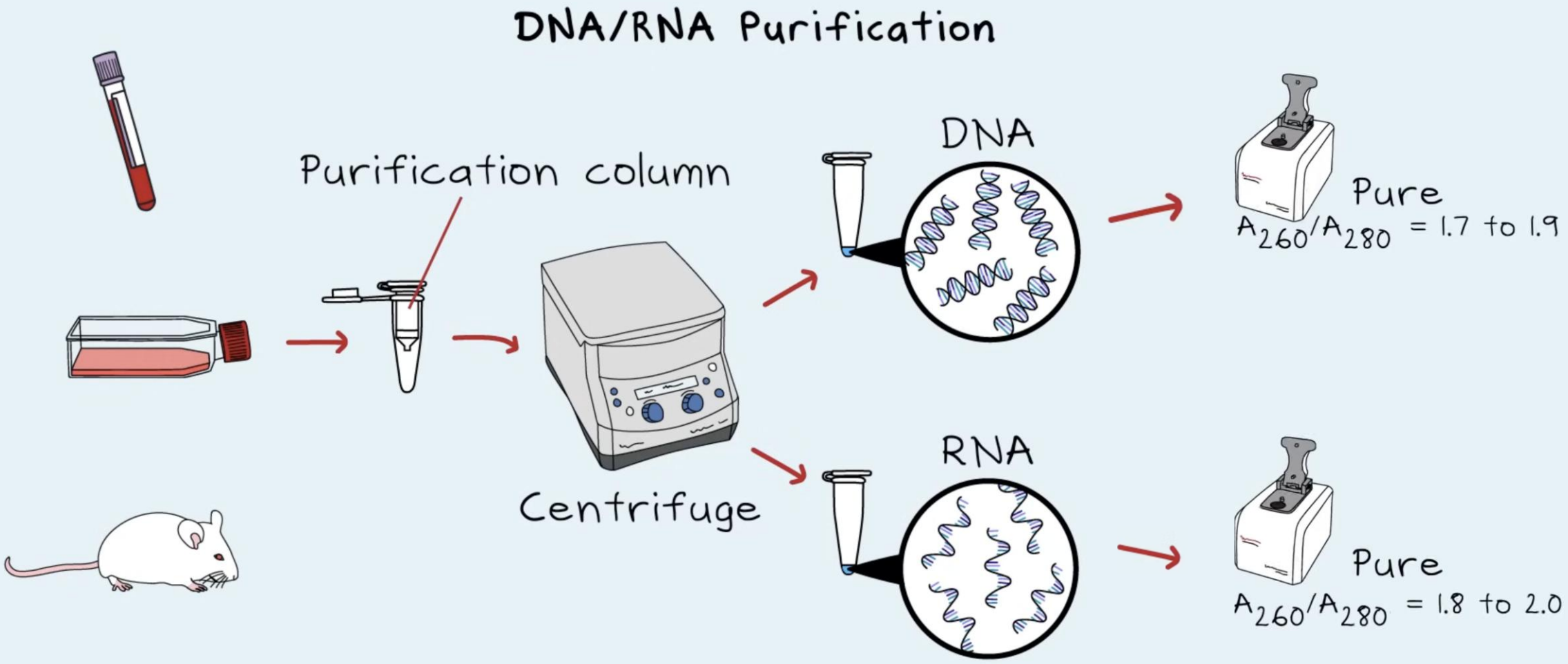
人类基因组计划（Human Genome Project）



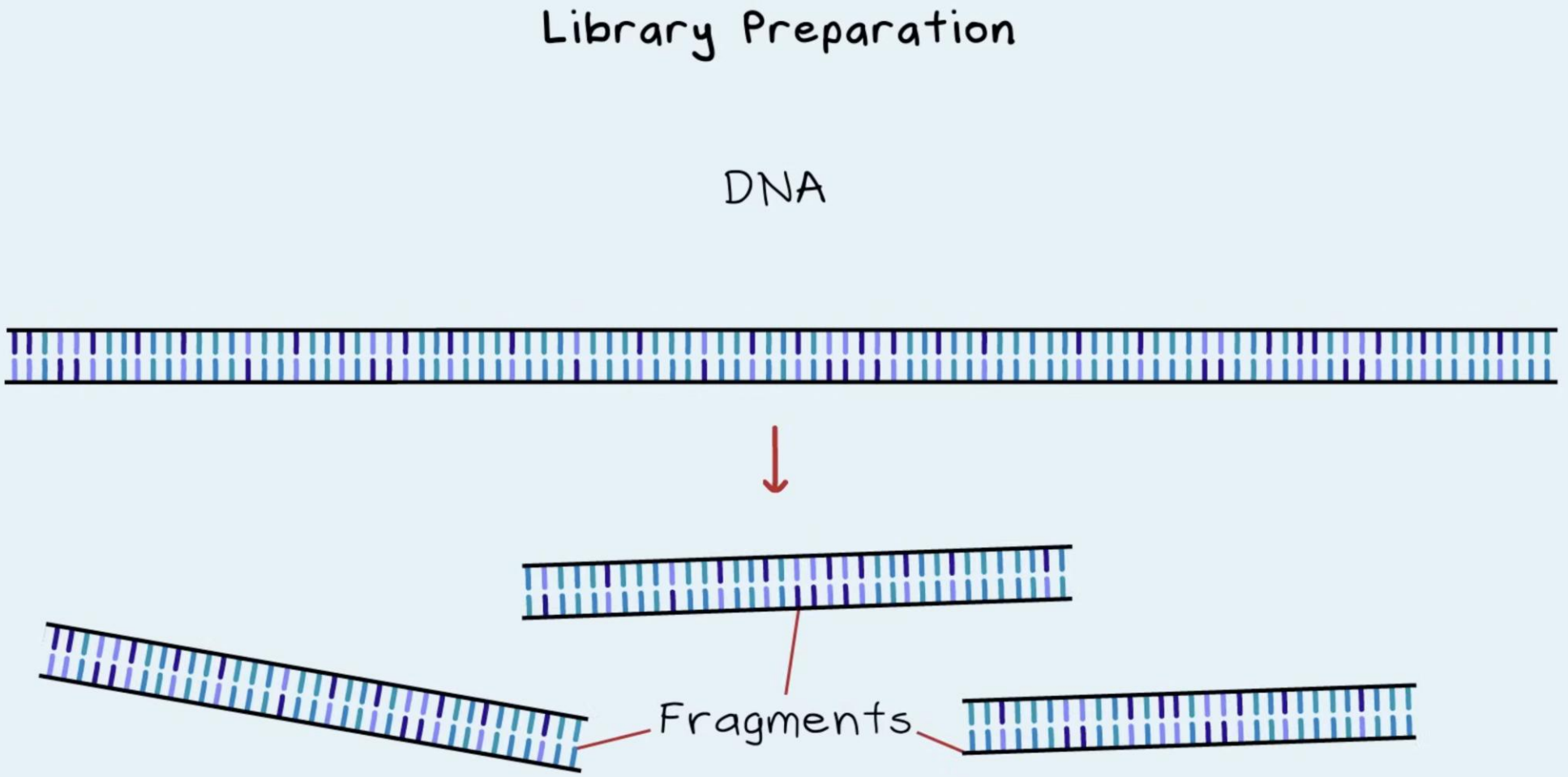
二代测序（Next Generation Sequencing, NGS）



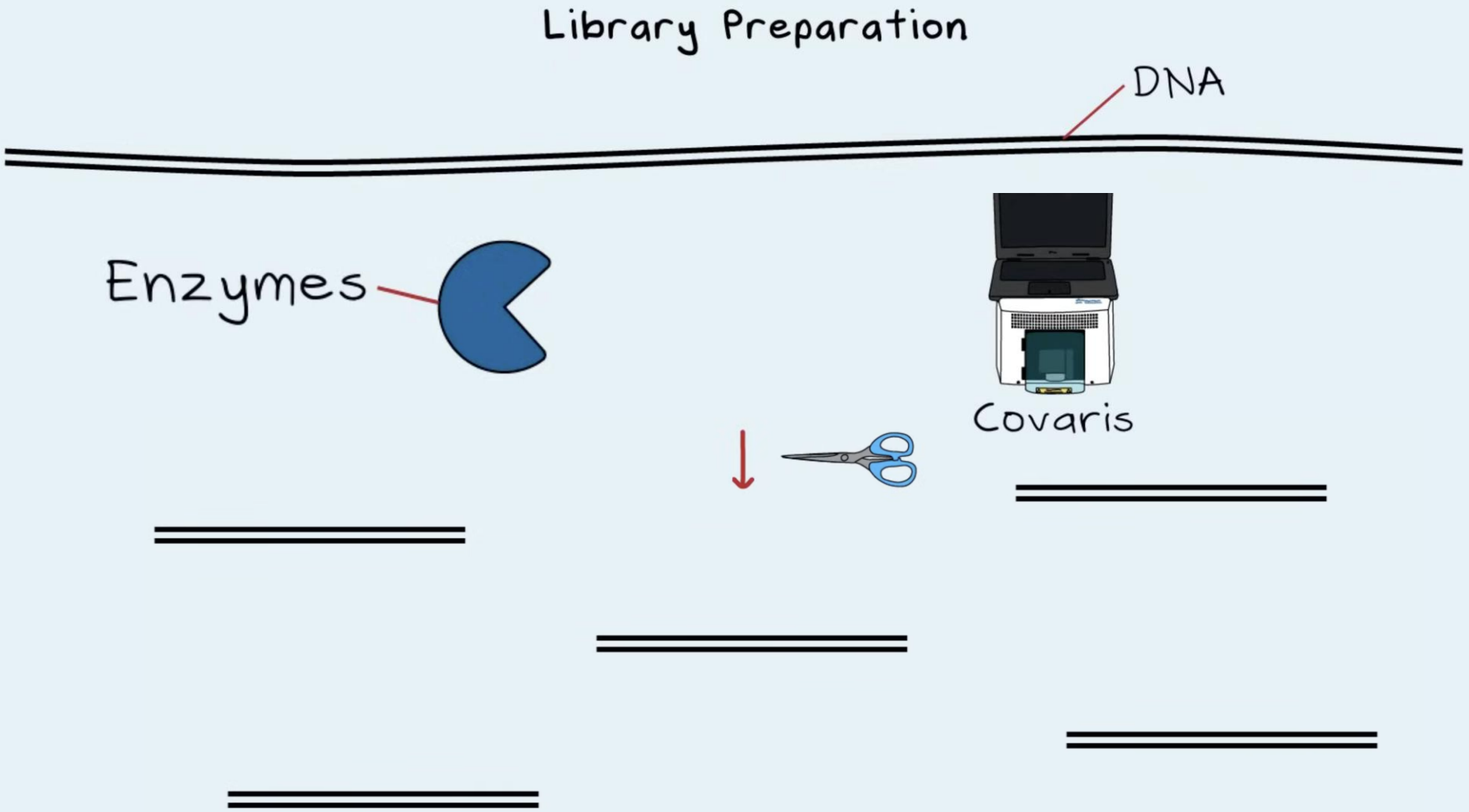
DNA/RNA 纯化——高质量测序样品的准备



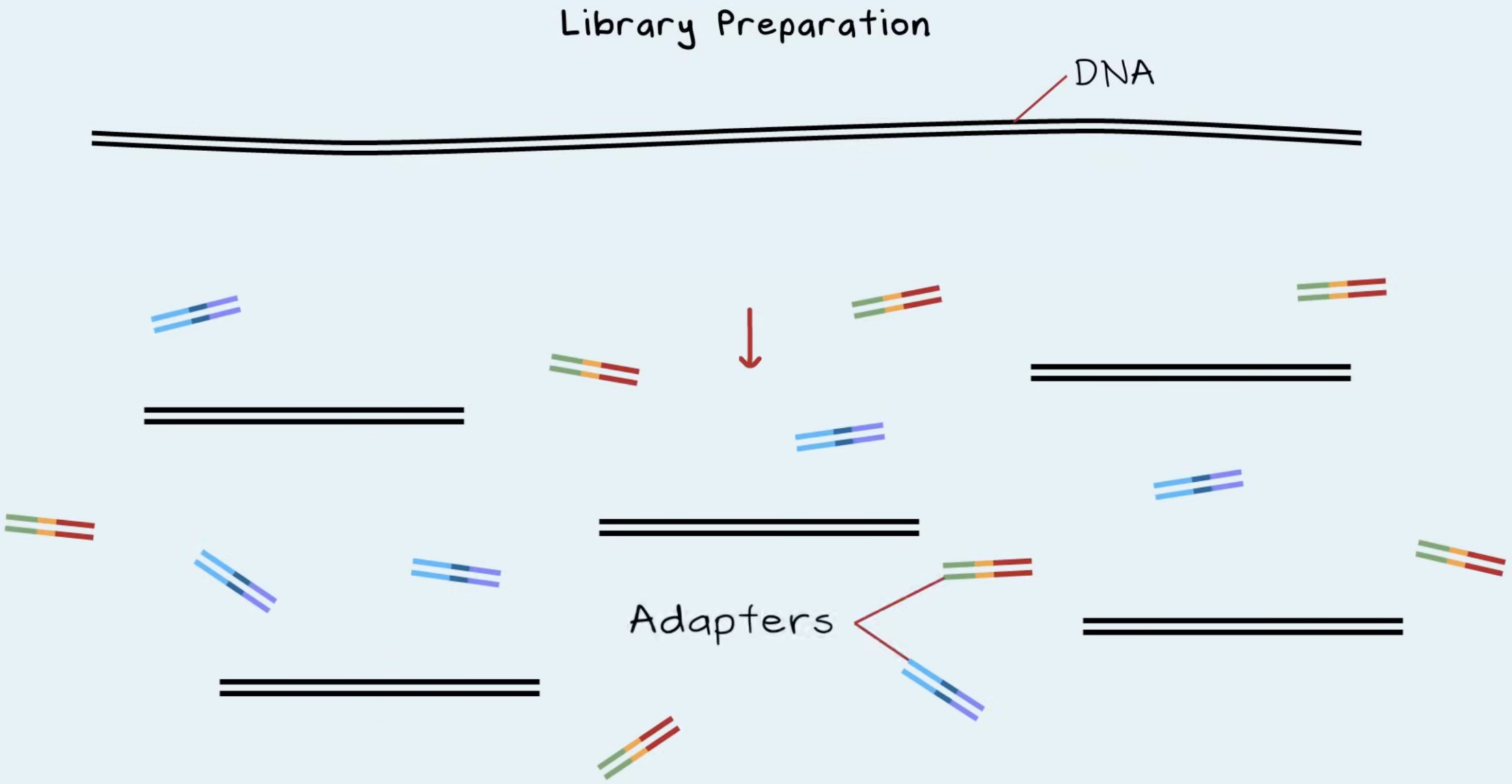
文库构建（DNA Fragmentation）



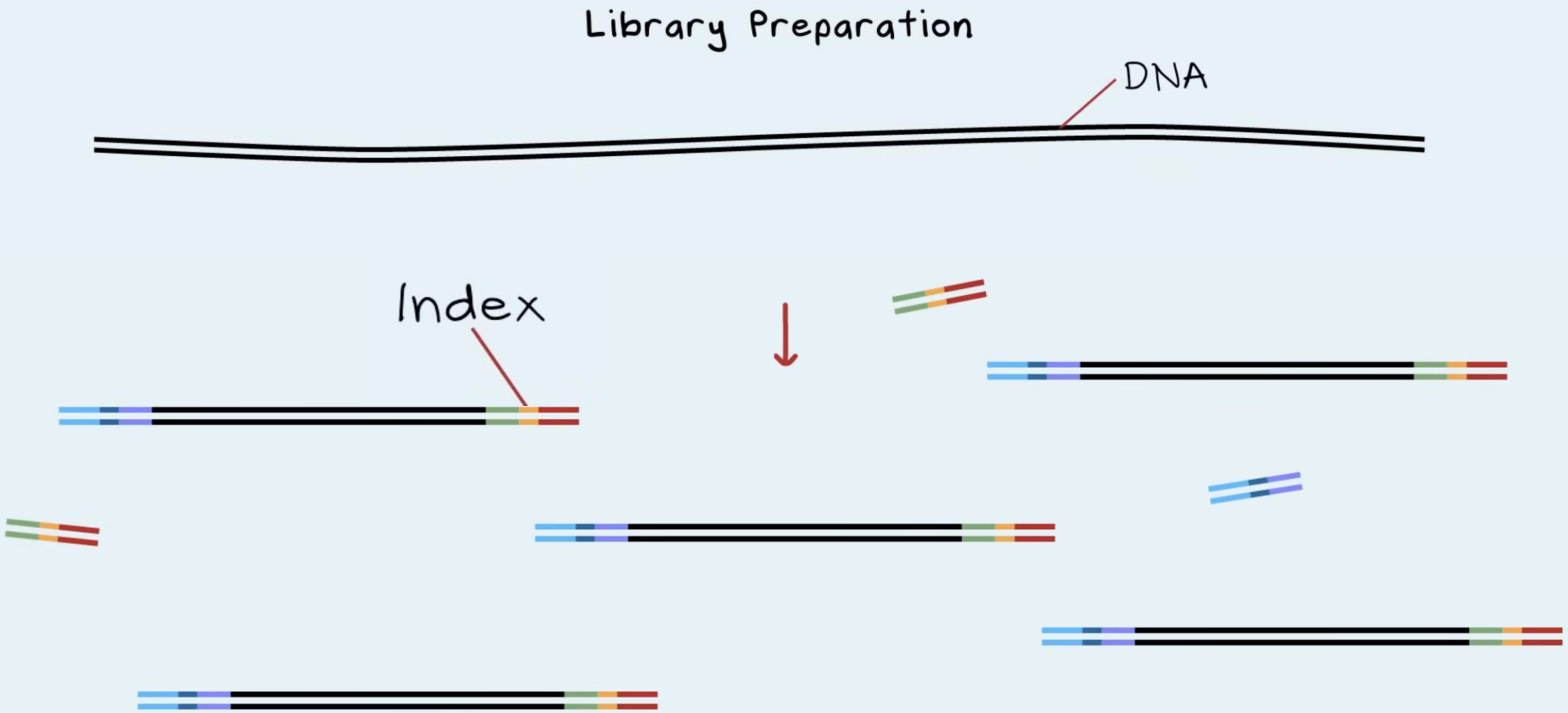
文库构建（DNA Fragmentation）



文库构建（Adapter Ligation）



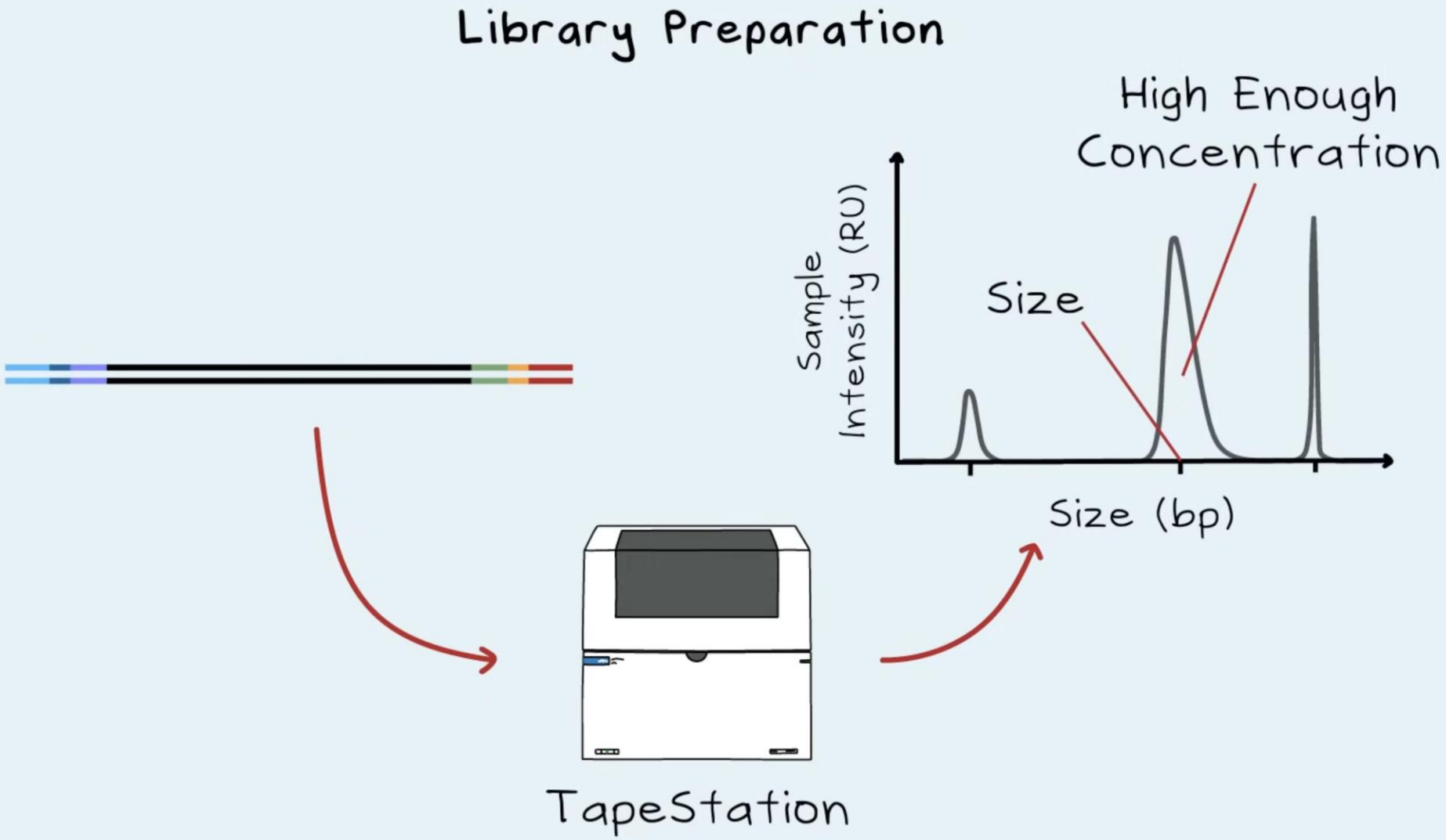
文库构建（Adapter Ligation）



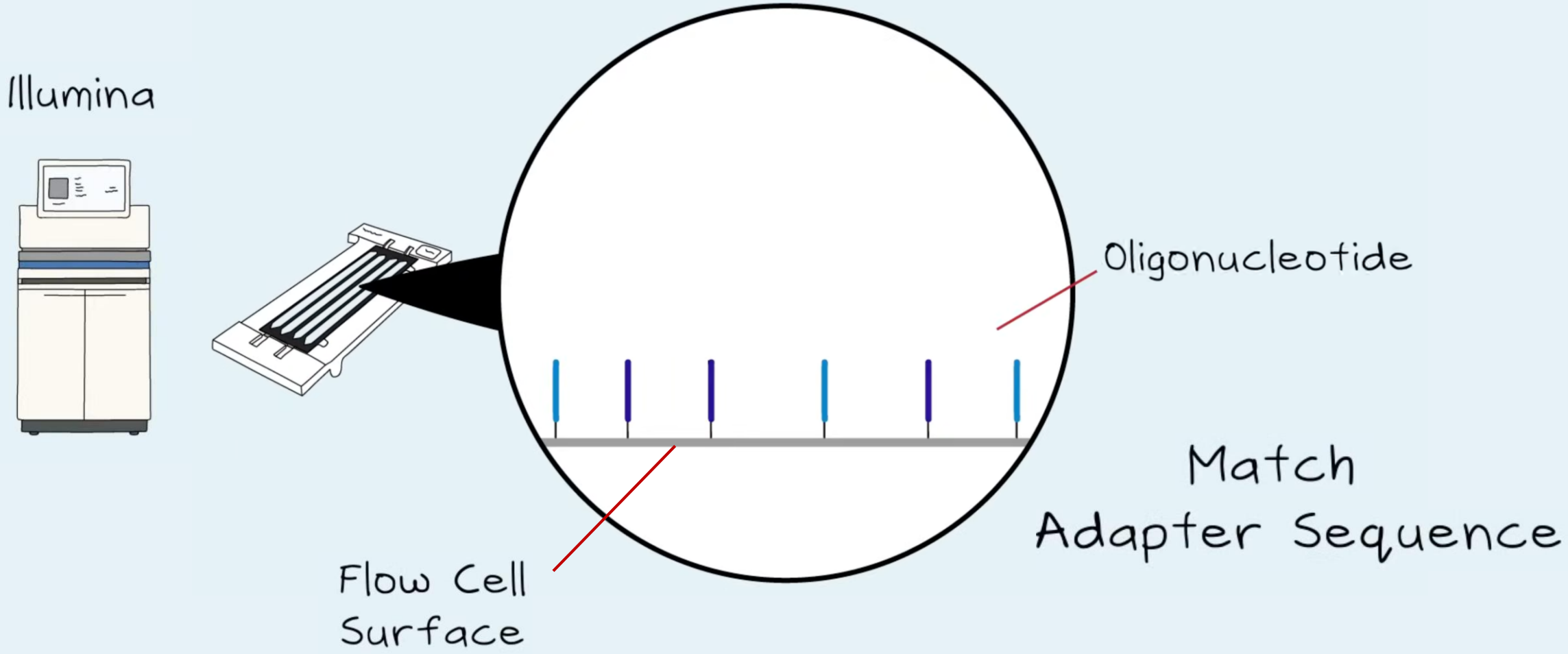
文库构建（PCR Amplification）

PCR

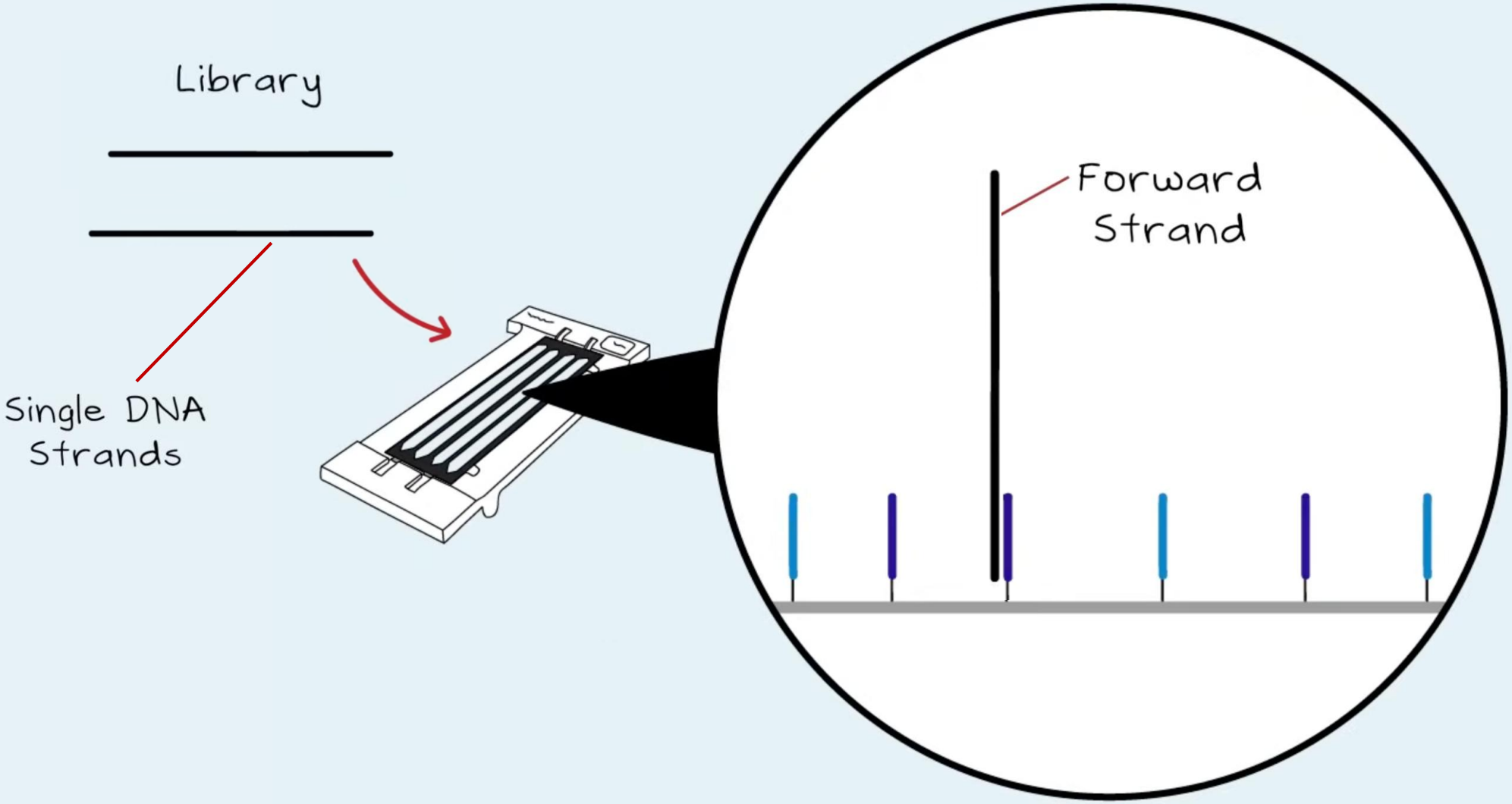
文库质控


Library Preparation
High Enough
Concentration
Sample
Intensity (RU)
Size
Size (bp)
TapeStation

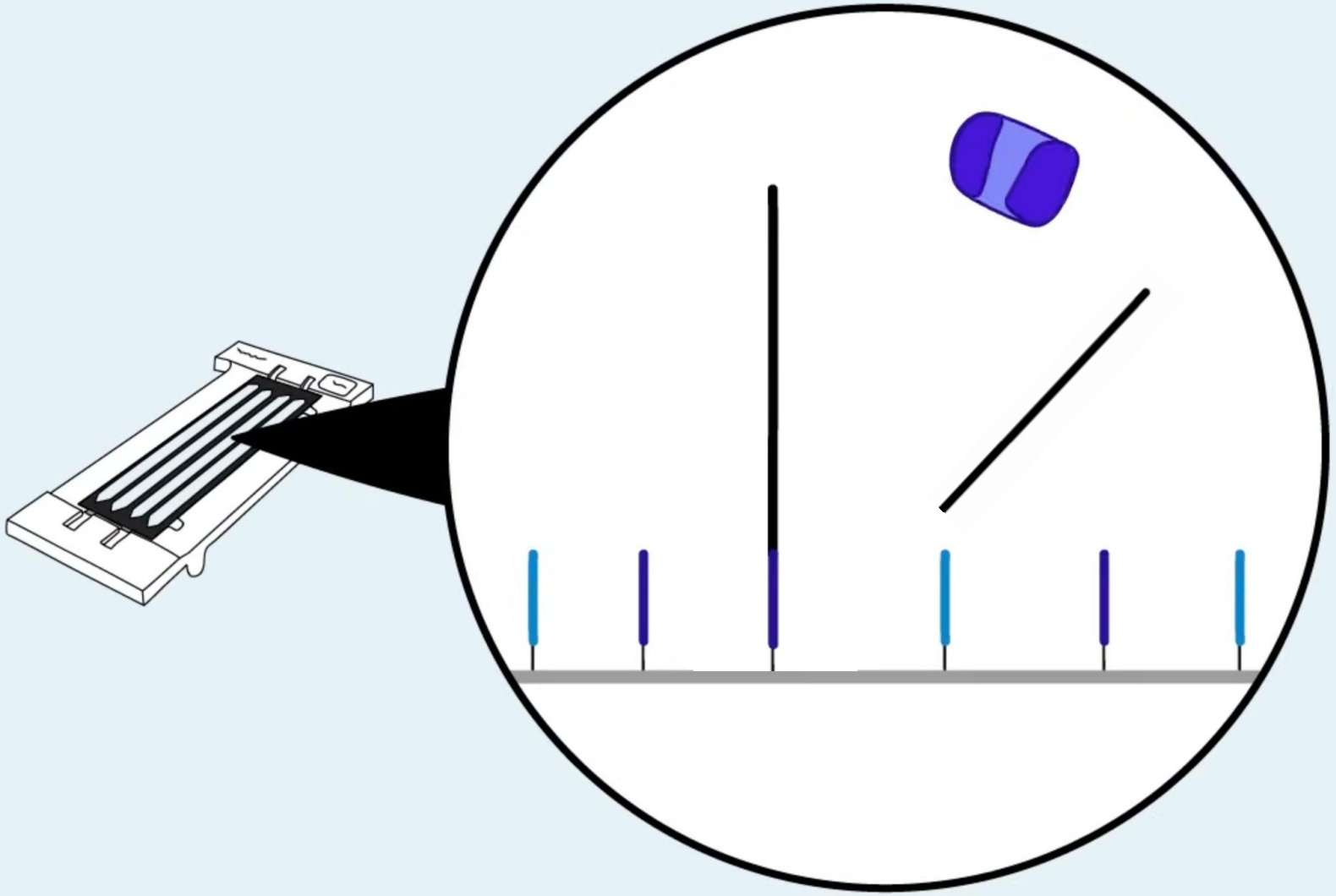
二代测序上机（Flow Cell Loading）



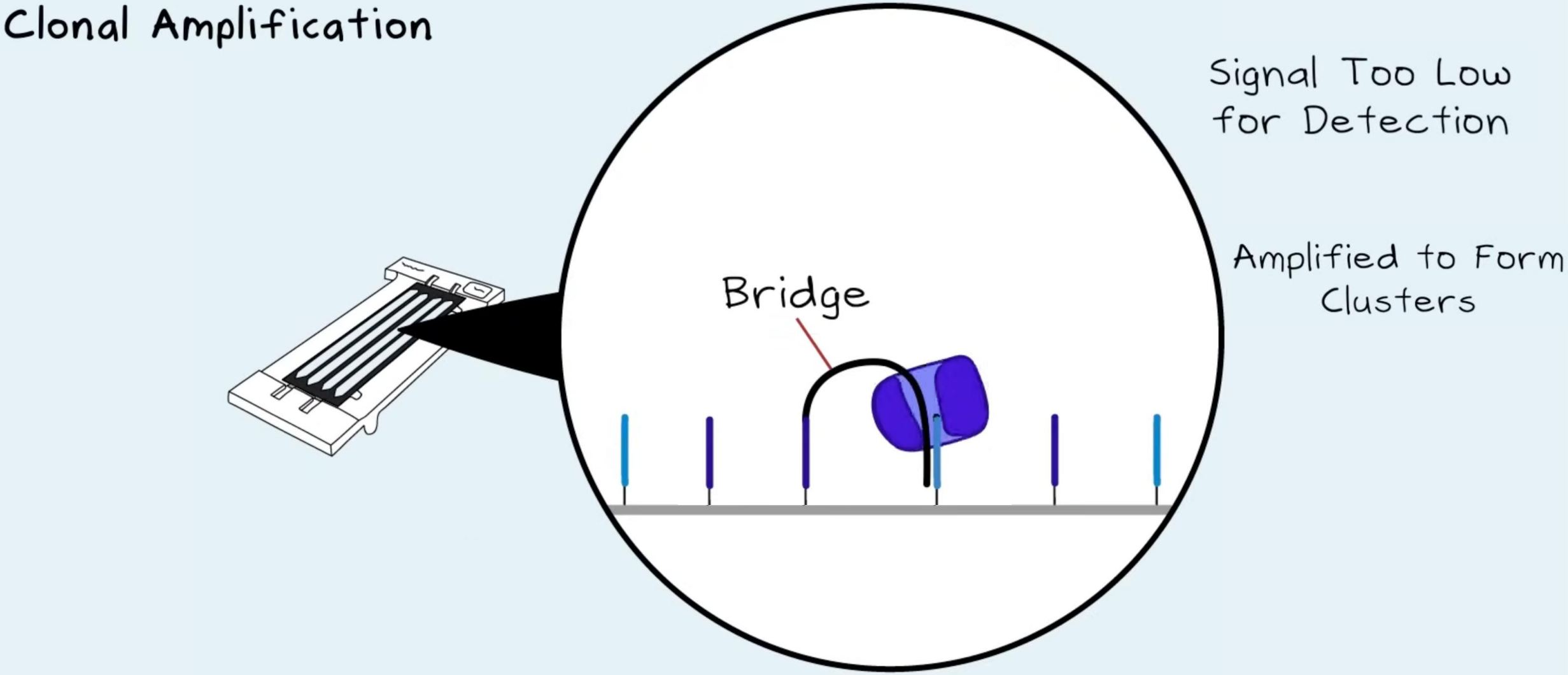
Library Hybridization on the Flow Cell



PCR Amplification

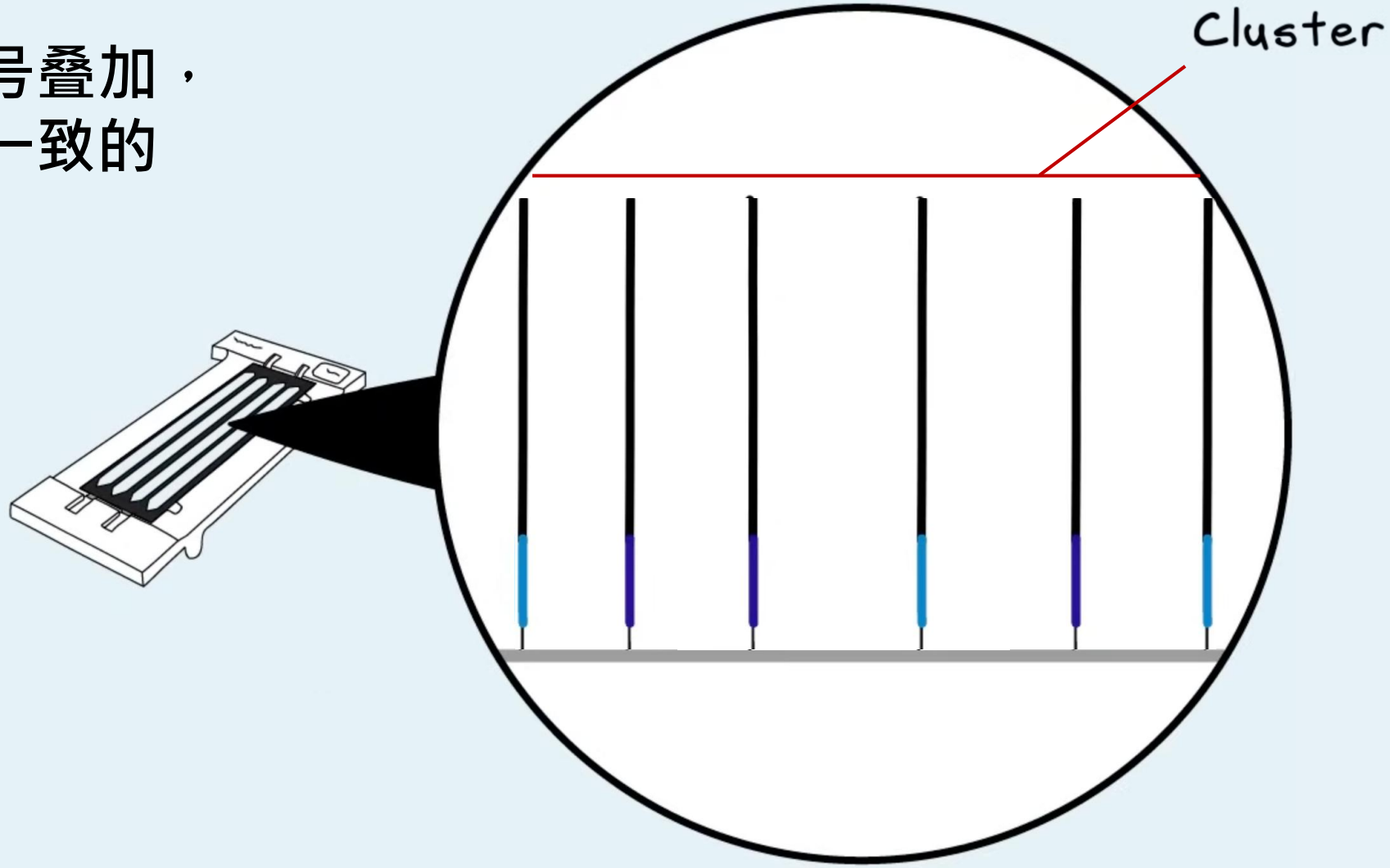


桥式扩增 (Bridge Amplification) 产生DNA簇（Cluster）



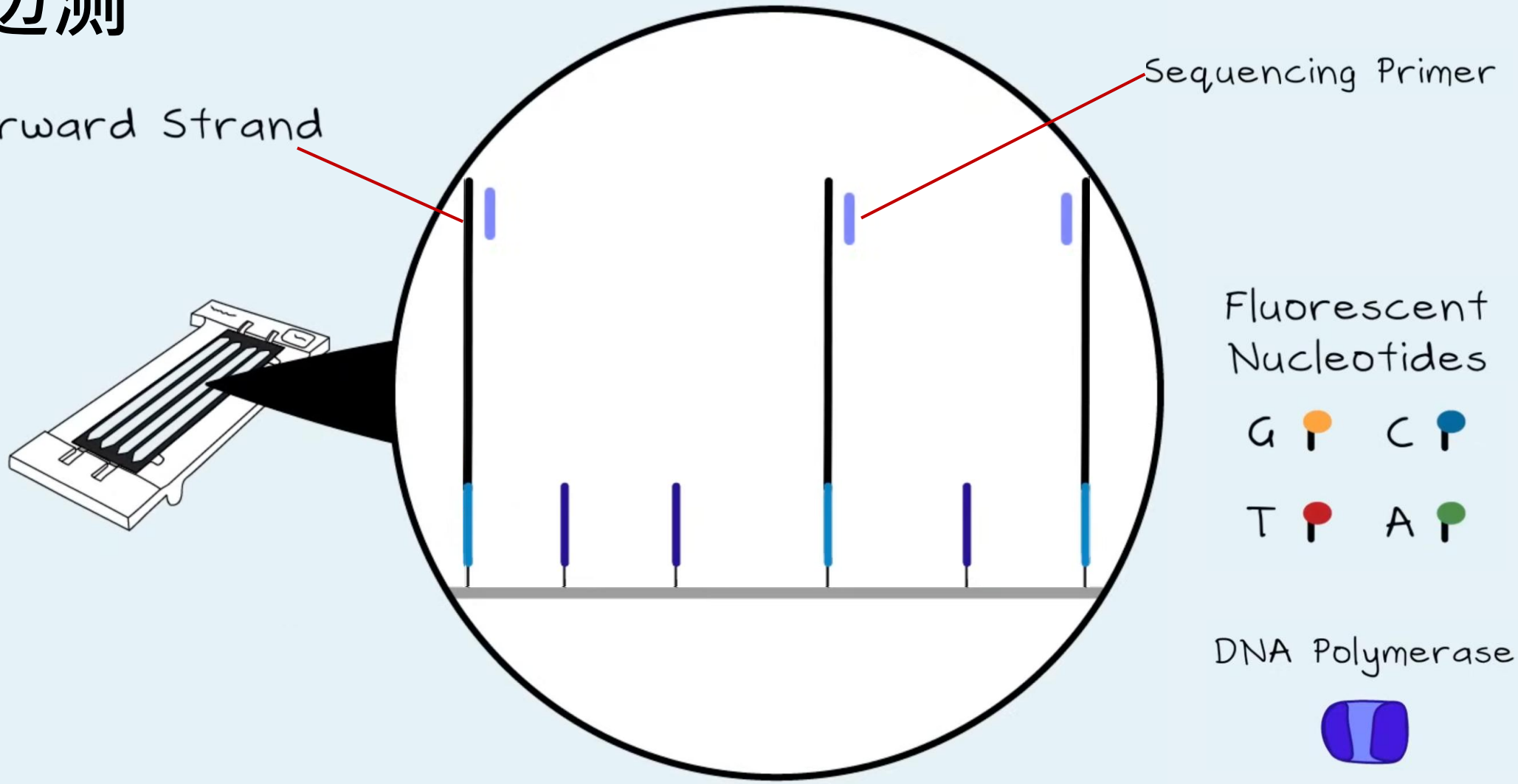
桥式扩增 (Bridge Amplification) 产生DNA簇（Cluster）

让荧光信号叠加，
形成强而一致的
亮点



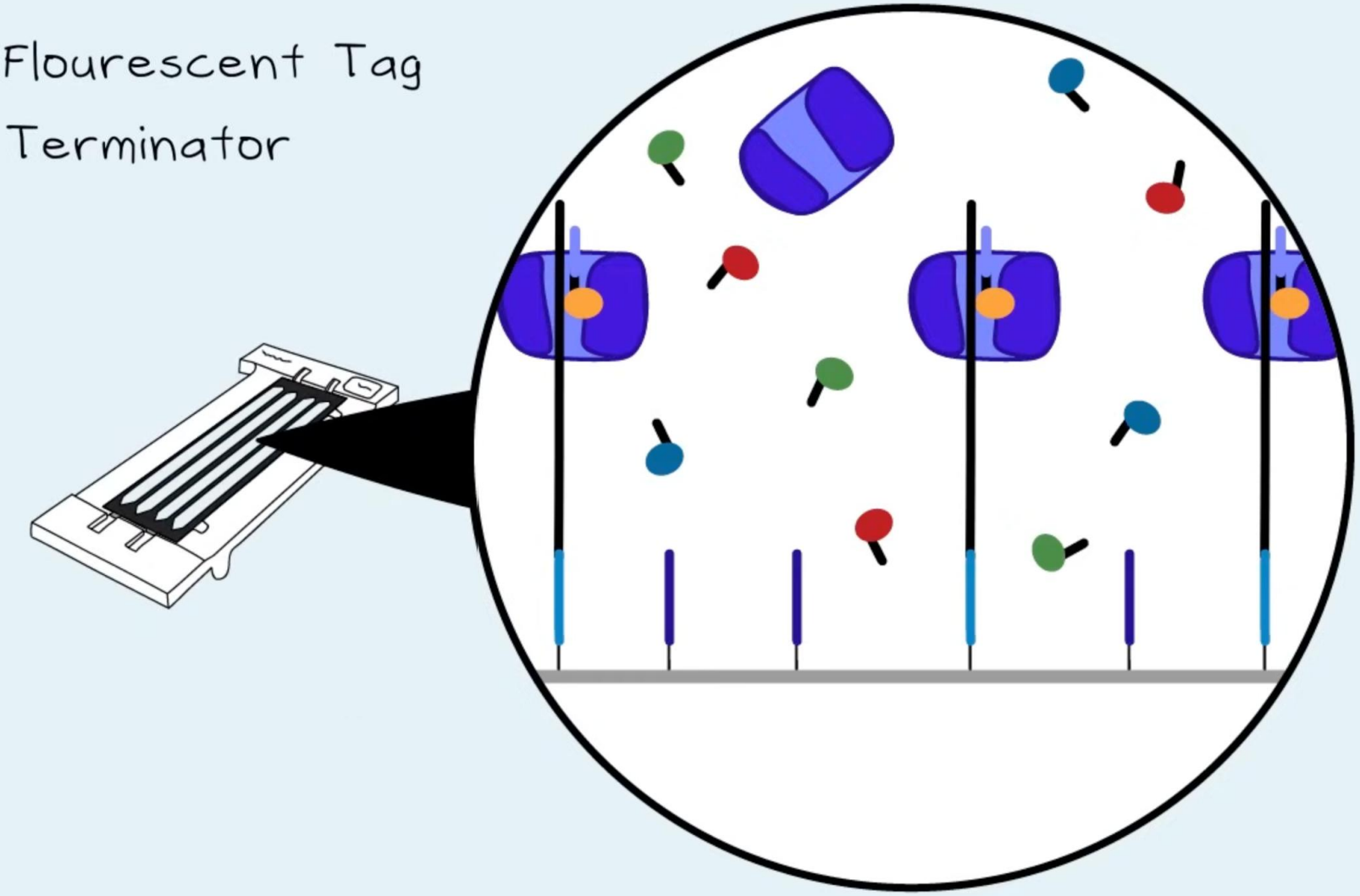
Sequencing by Synthesis — Initiation

边合成边测序

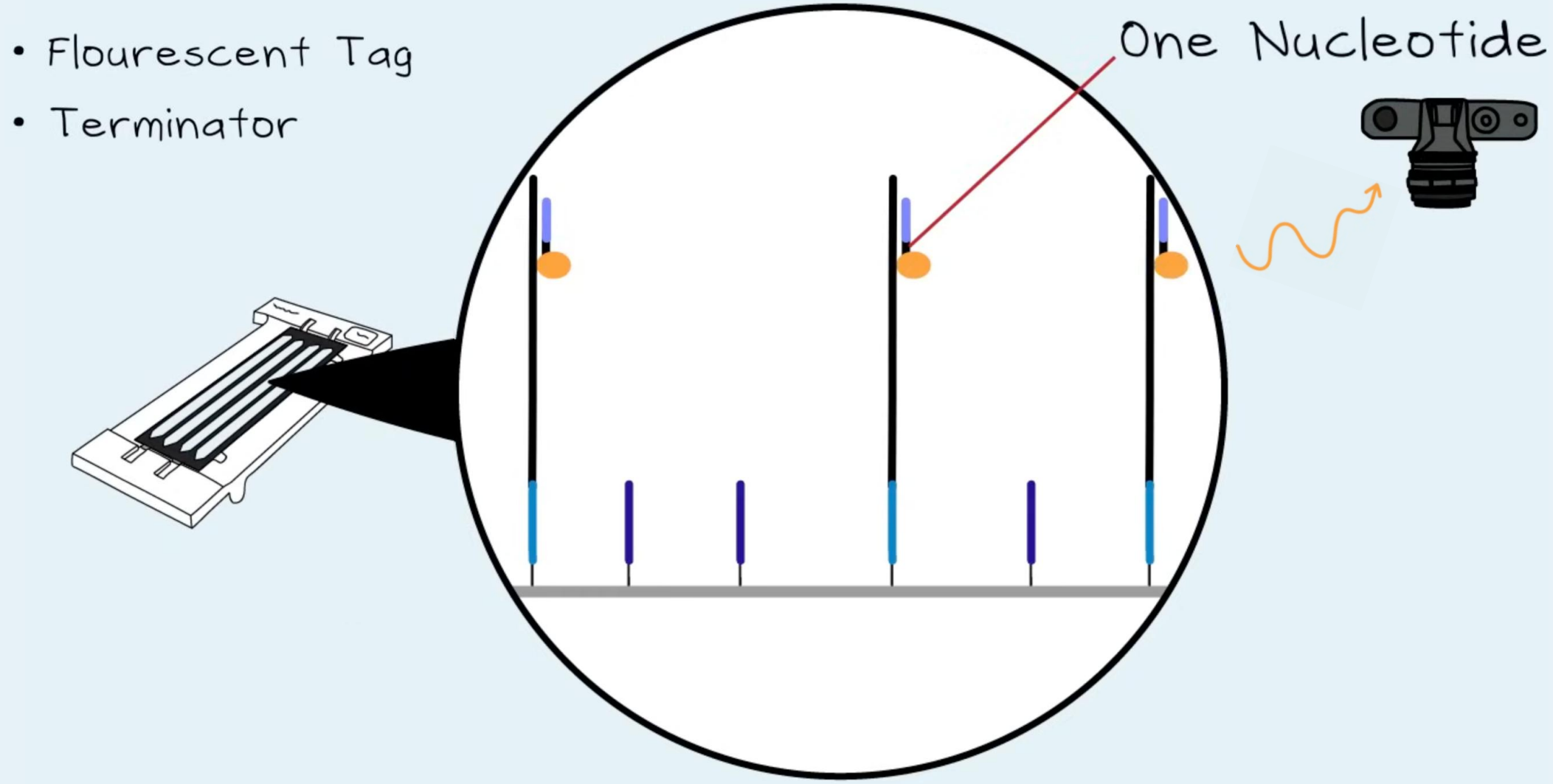


Sequencing by Synthesis — Fluorescent Terminator Incorporation

- Flourescent Tag
- Terminator

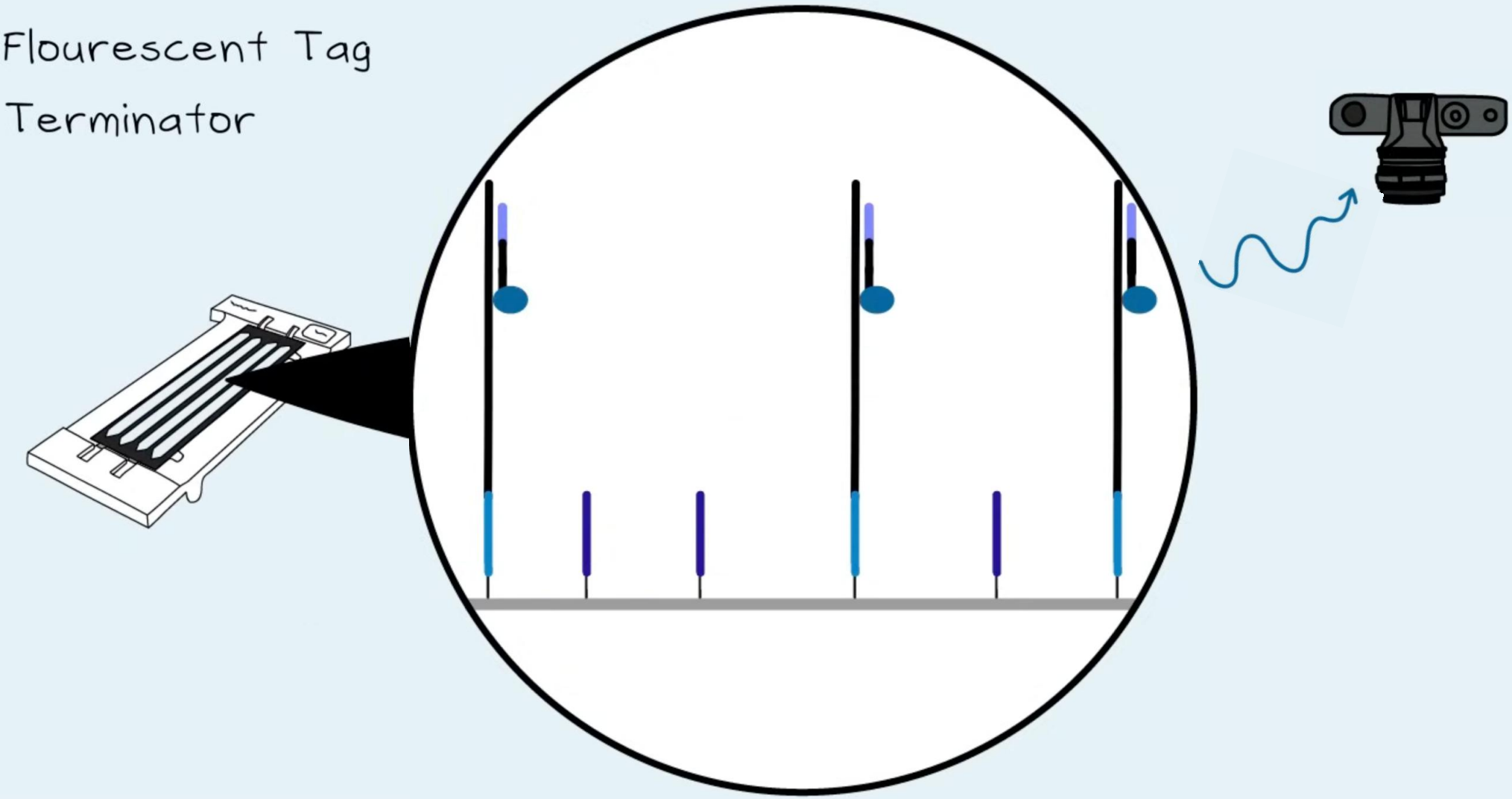


Sequencing by Synthesis — Fluorescent Terminator Incorporation

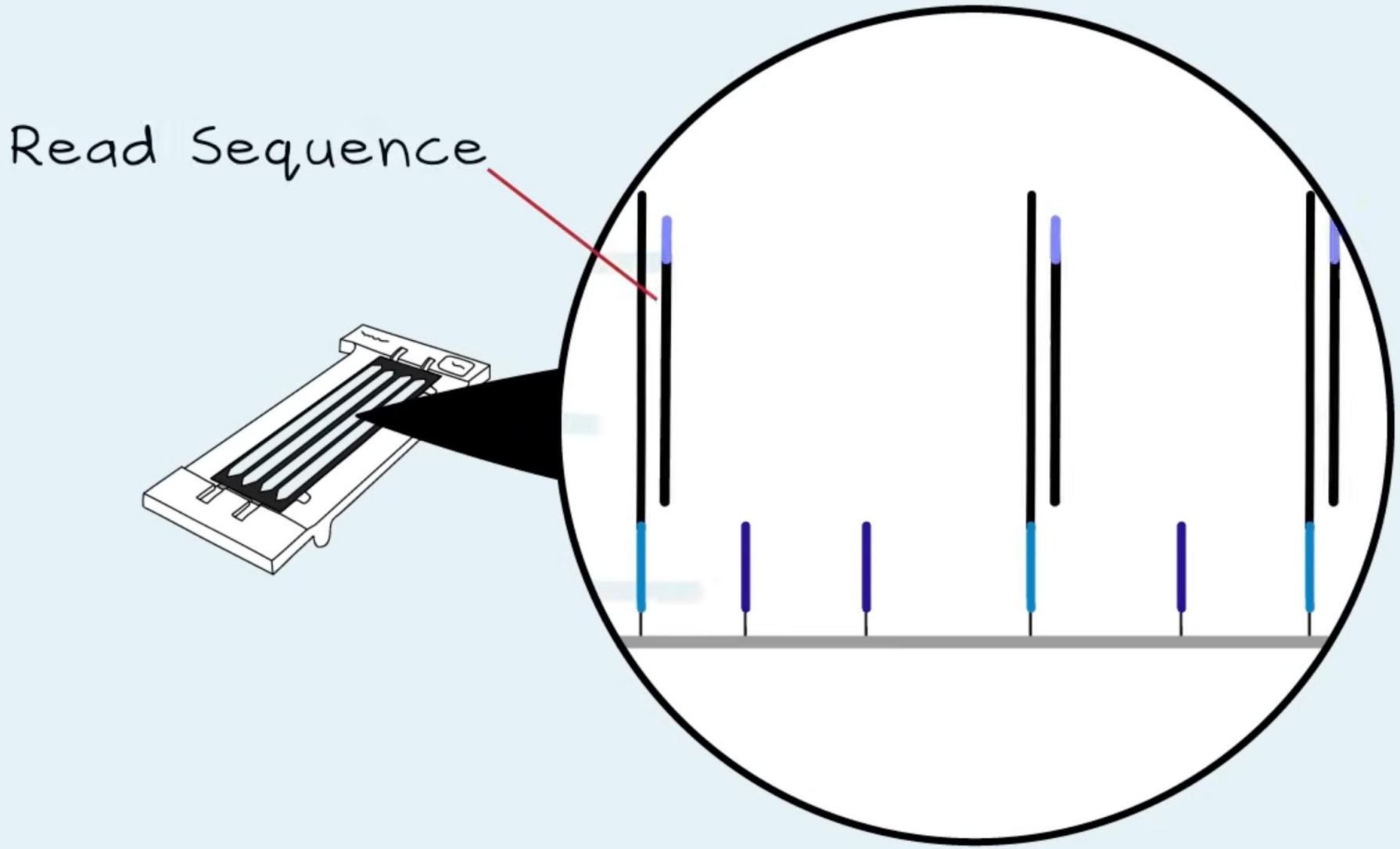


Sequencing by Synthesis — Cleavage and Next Cycle

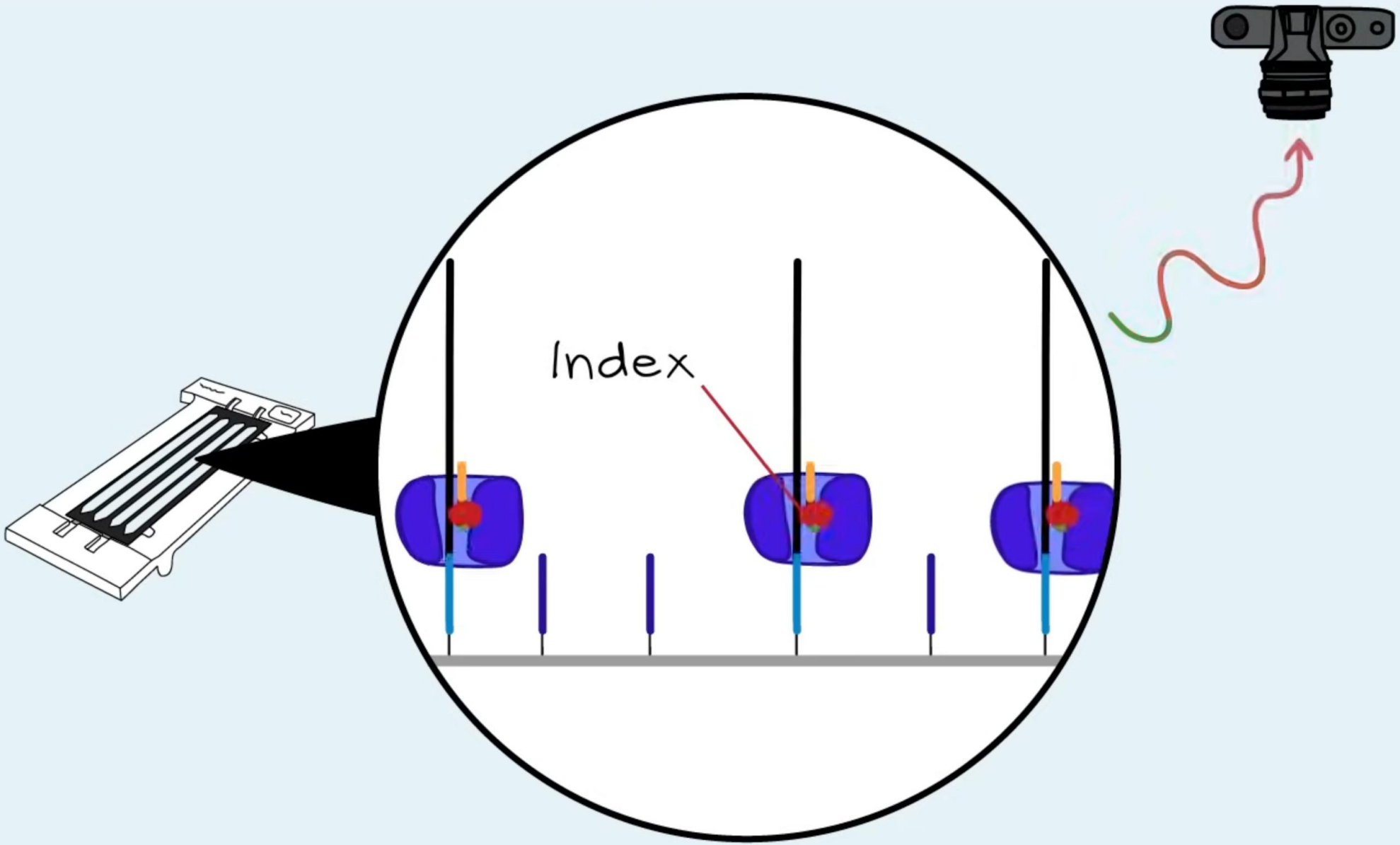
- Flourescent Tag
- Terminator



Sequencing by Synthesis — Read Sequence



Sequencing by Synthesis — Read Index



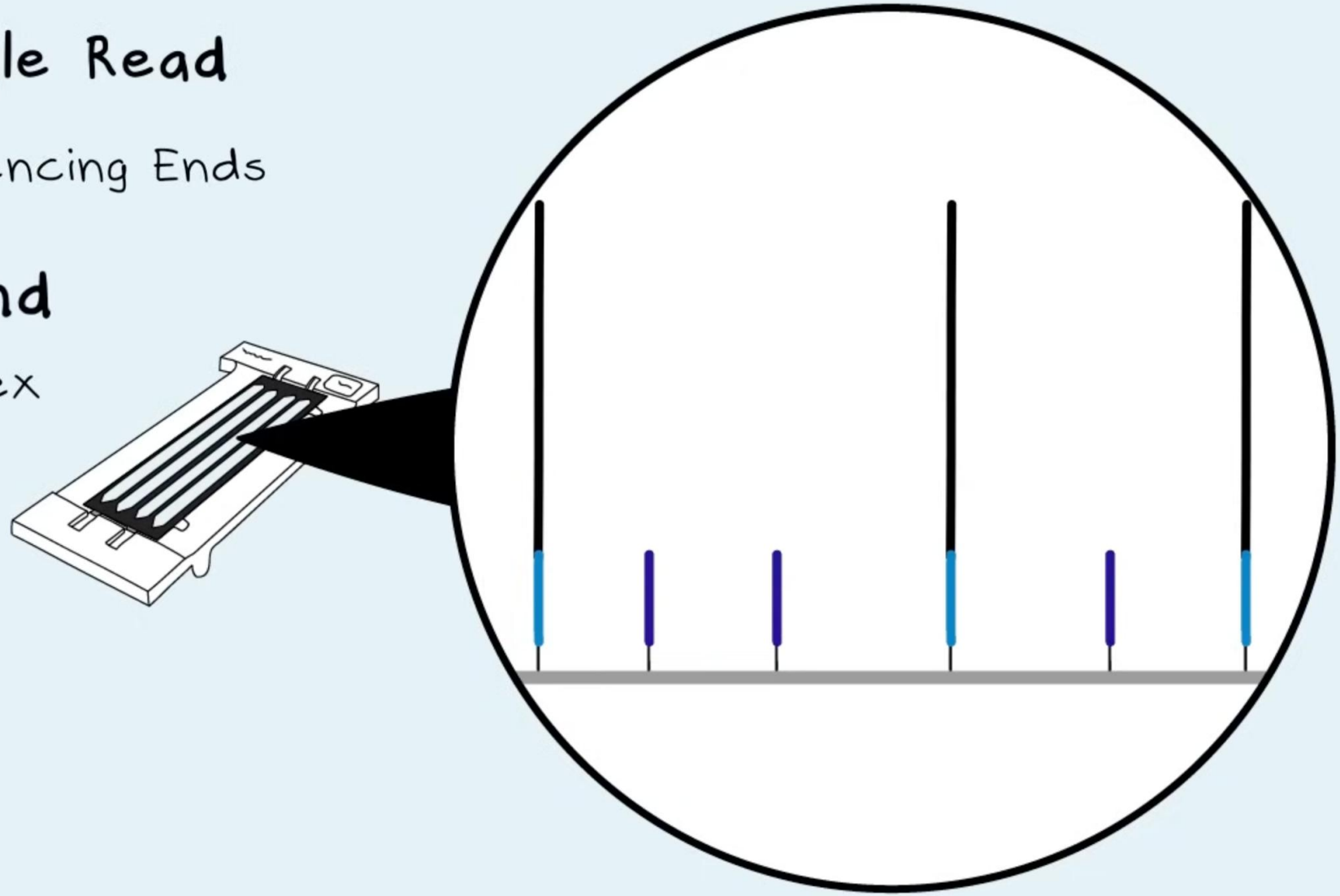
Single-end vs. Paired-end Sequencing

Single Read

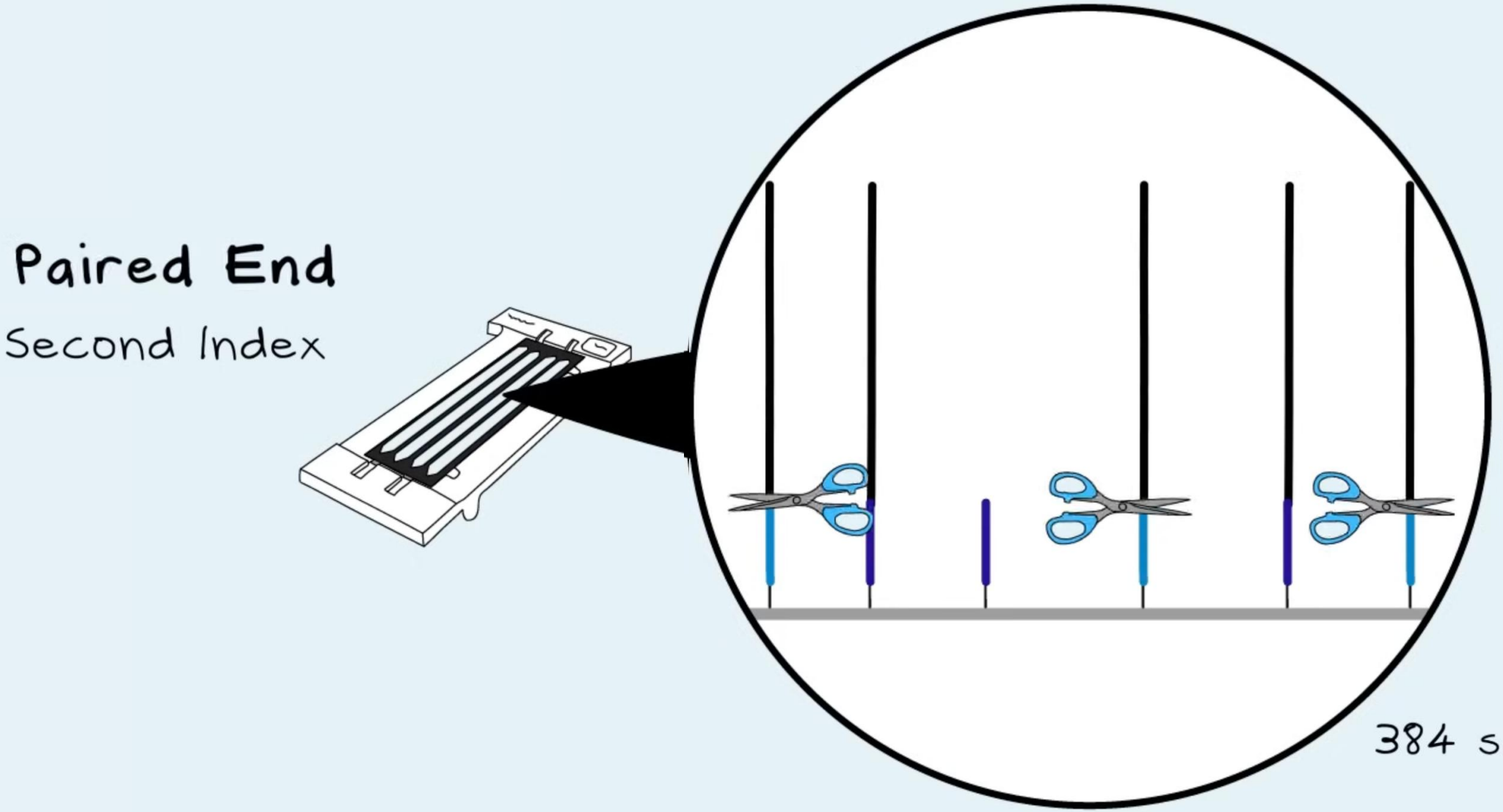
Sequencing Ends

Paired End

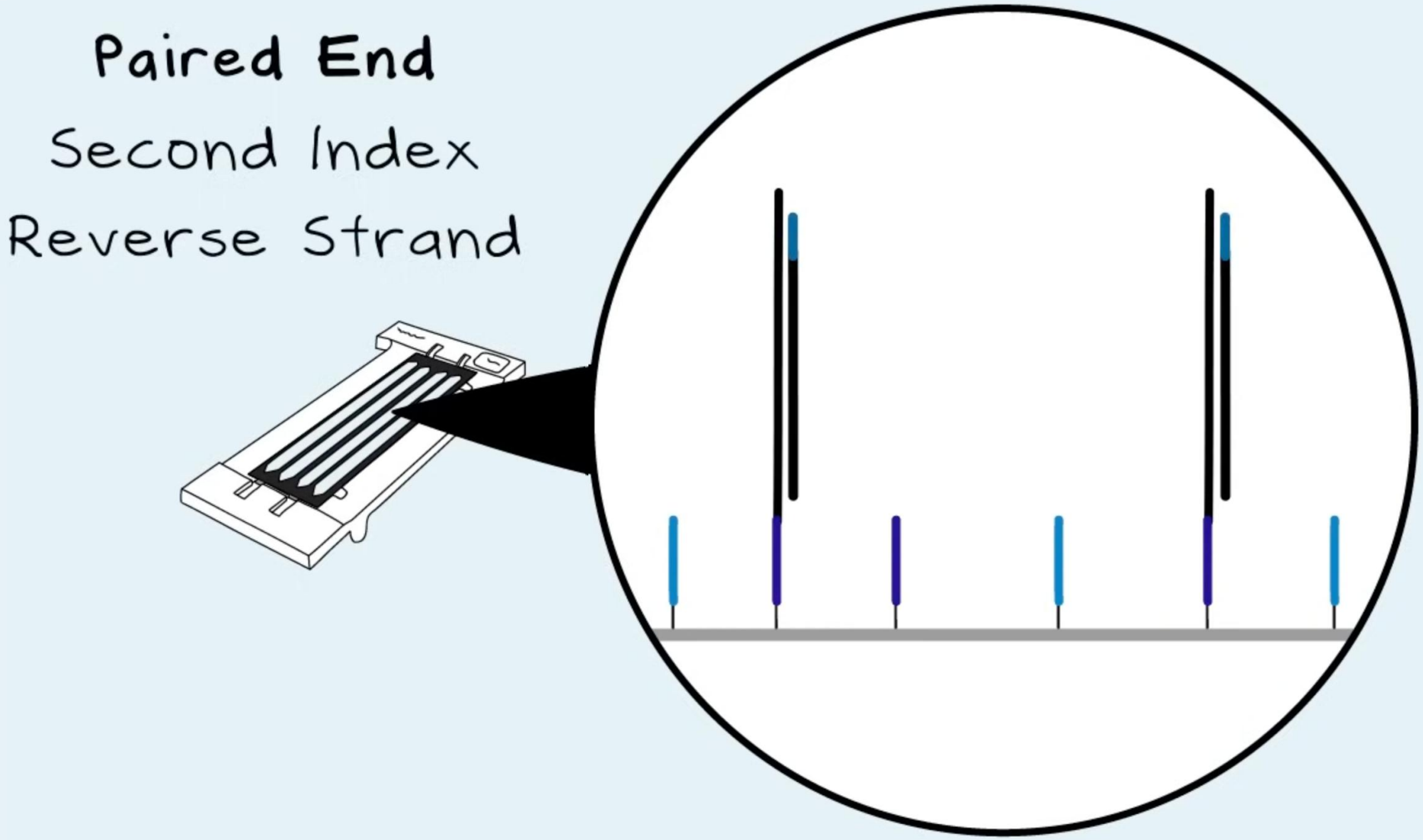
Second Index



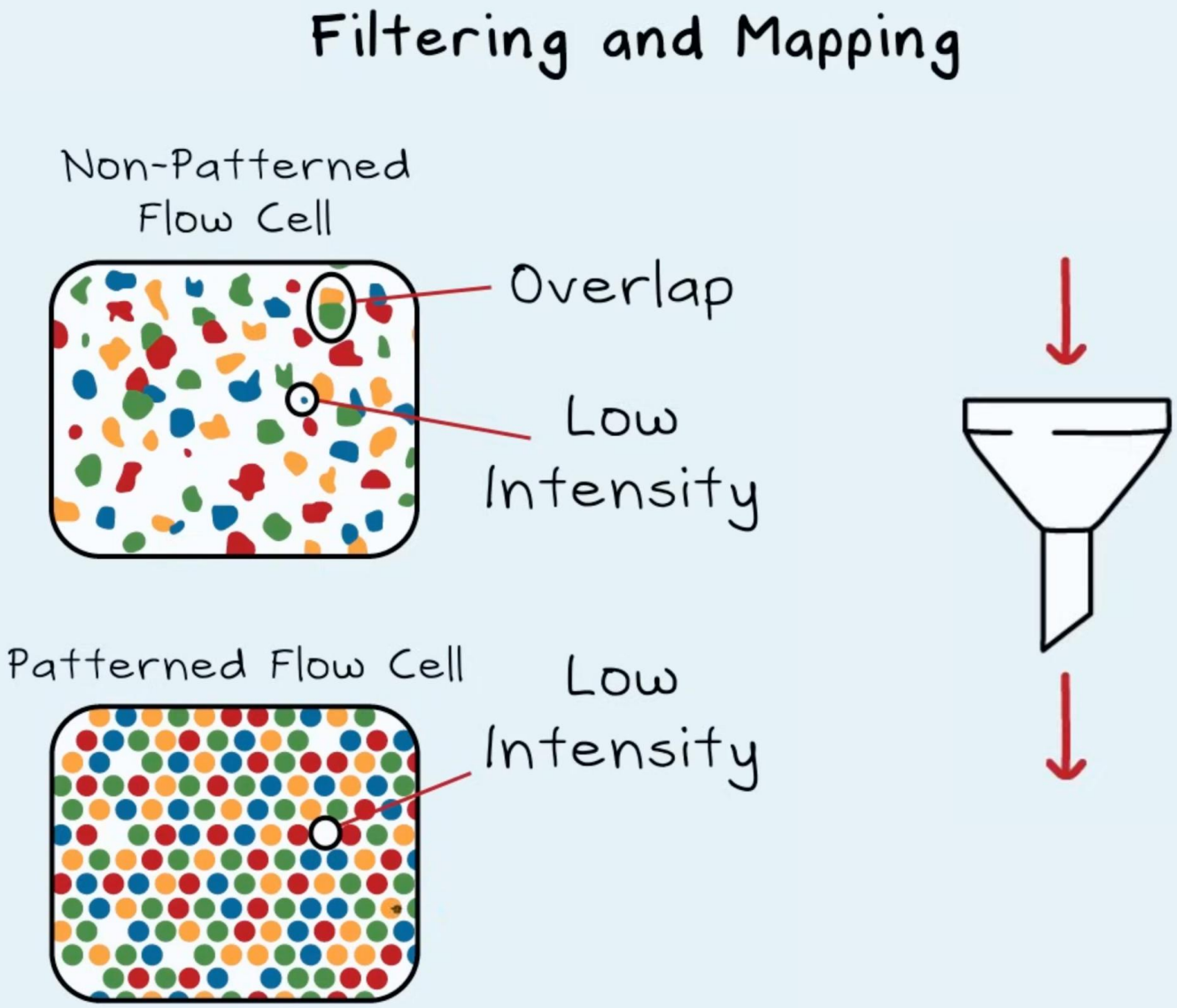
Paired-End Sequencing



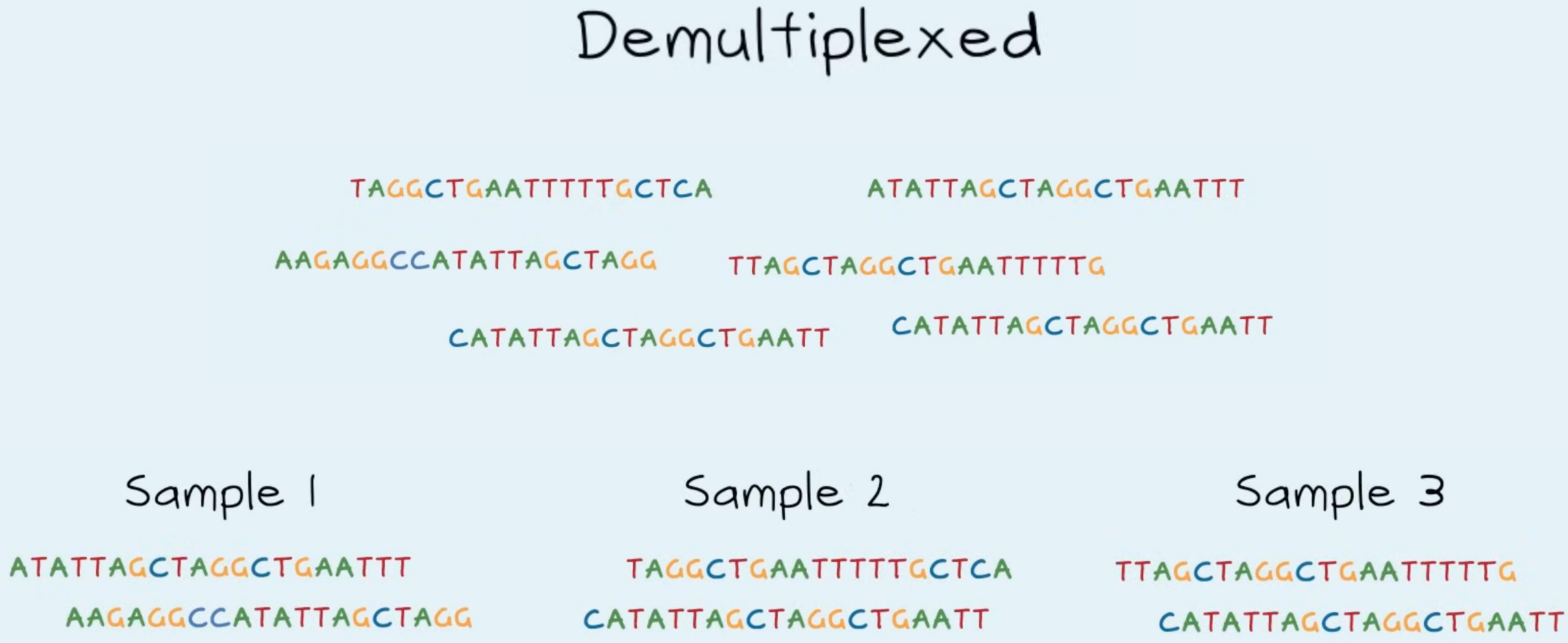
Paired-End Sequencing



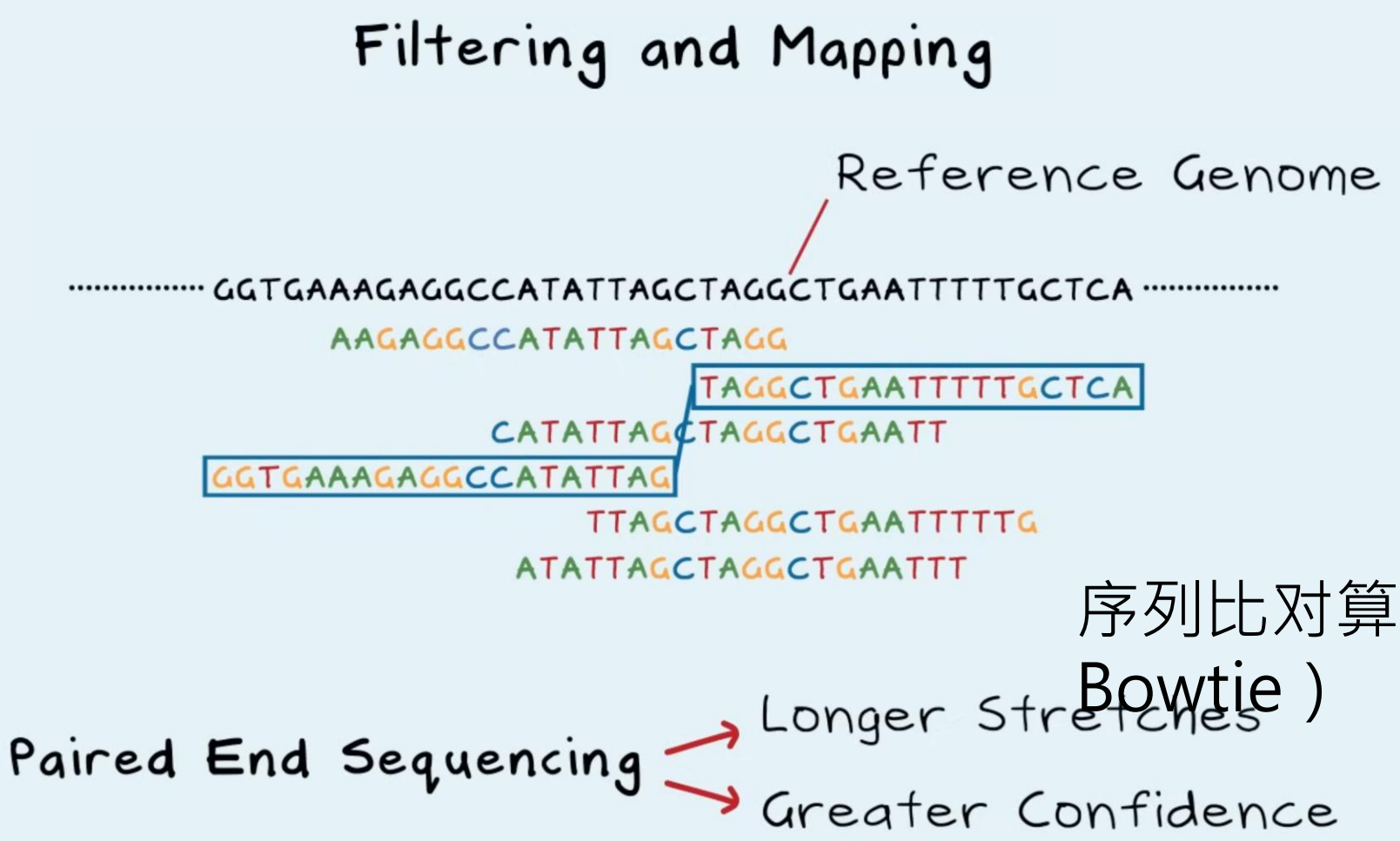
信号过滤与比对：从原始信号到可靠序列



Demultiplexing （样本拆分：将reads分配到各自样本）

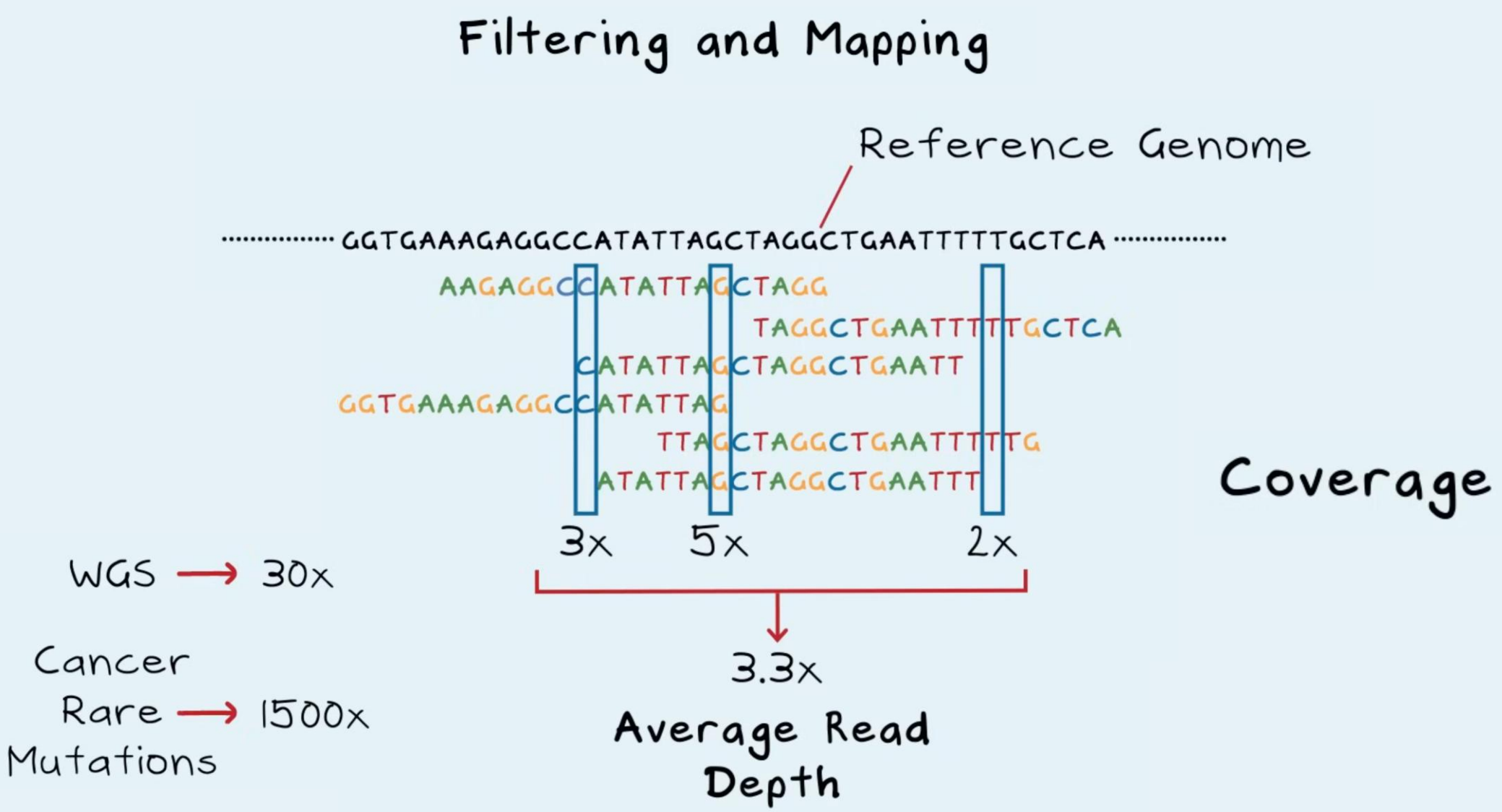


Mapping — Aligning Reads to the Reference Genome



序列比对算法（如BWA、Bowtie）

Coverage and Read Depth



测序数据格式

FASTQ 数据格式

高通量测序得到的原始图像数据文件，经过碱基识别（**Base Calling**）分析转化为原始测序序列（**Sequenced Reads**），我们称之为 **Raw Data** 或 **Raw Reads**，结果以 **FASTQ** 文件格式存储，其中包含**测序序列（Reads）**的序列信息以及其对应的测序质量信息。测序样品中真实数据随机机截取结果如下图：

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCCATTTGTTCAACTCACAGTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC6
```

- 第一行： 以 @ 开头，后接 Illumina 测序识别符（Sequence Identifier）和描述文字
- 第二行： 为碱基序列（Sequence Line），由 A、T、C、G、N 构成。
- 第三行： 以 + 开头
- 第四行： 为对应碱基序列的 测序质量值（Quality Score），用 ASCII 字符编码 表示每个碱基的置信度。

测序质量（Quality Score）

Illumina 测序中，每个碱基的质量值用 **Phred Q 值** 表示，计算公式为：

$$Q = -10 \log_{10}(e)$$

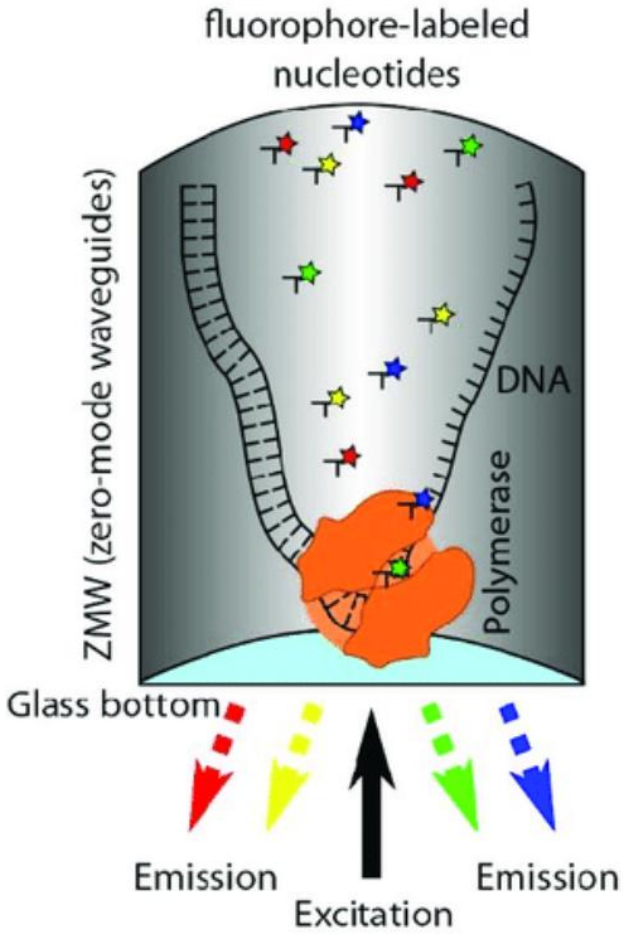
其中，e 是该碱基测错的概率

常见 Q 值及对应错误率：

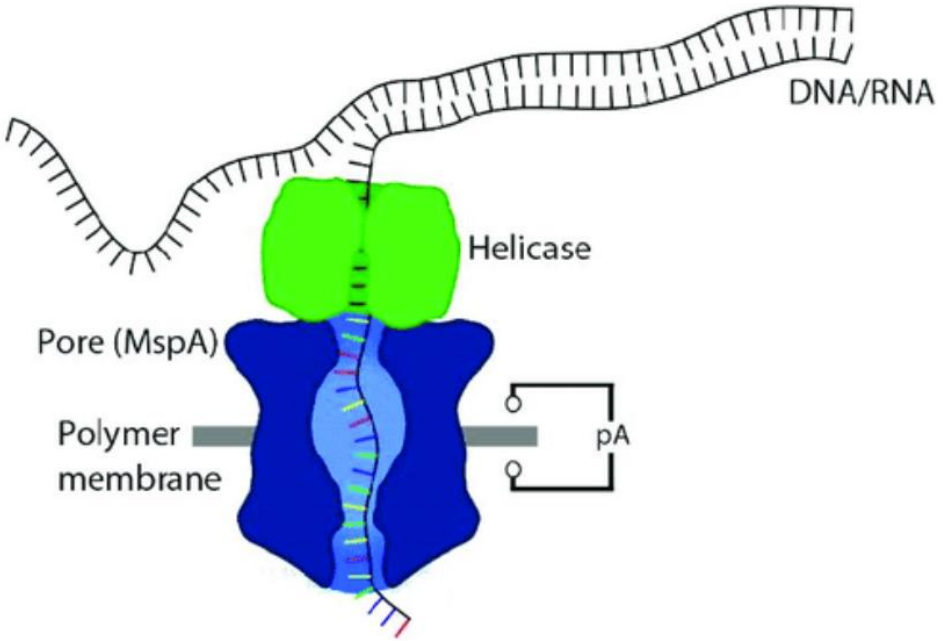
Q值	错误率	准确率
Q40	1 / 10,000	99.99%
Q30	1 / 1,000	99.9%
Q20	1 / 100	99%

三代测序

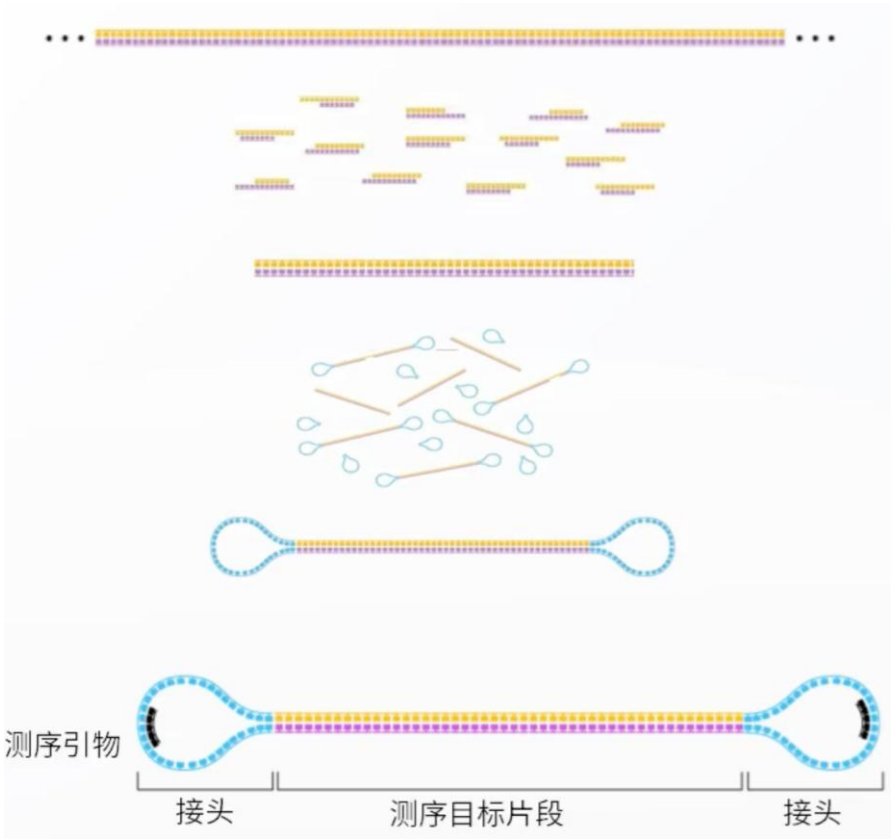
PacBio 三代测序



Nanopore 三代测序



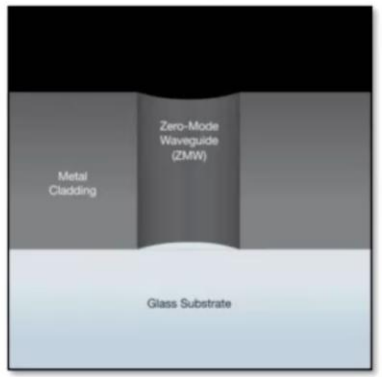
PacBio sequencing



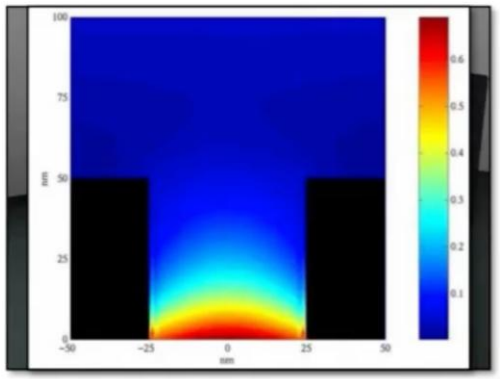
- ◆ 投入一定量的DNA或RNA
- ◆ 打断
- ◆ 末端修复
- ◆ 加接头
- ◆ 纯化，去除多余的接头
- ◆ 在文库中添加测序引物和聚合酶

完整文库结构

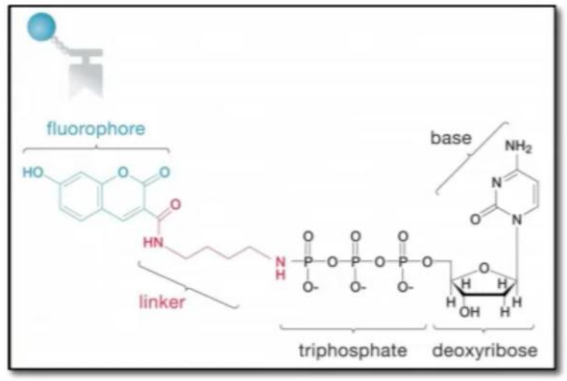
PacBio sequencing



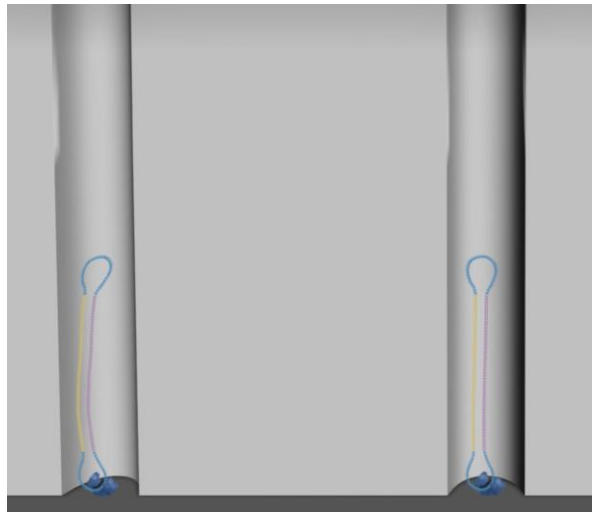
直径100nm纳米小孔
简称ZMW（Zero-Mode Waveguide）

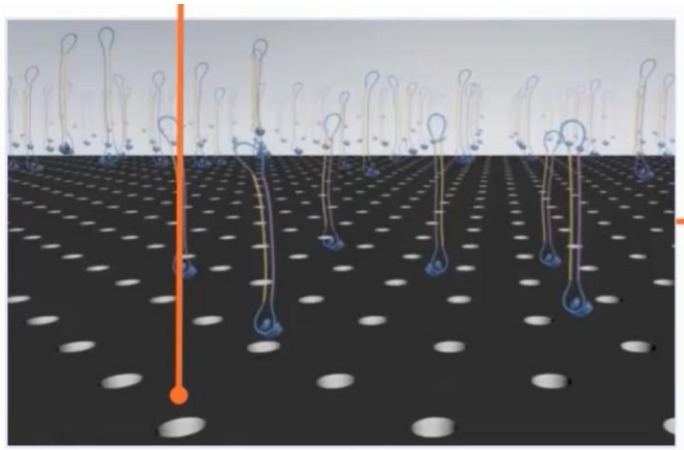


微弱光照亮ZMW底部

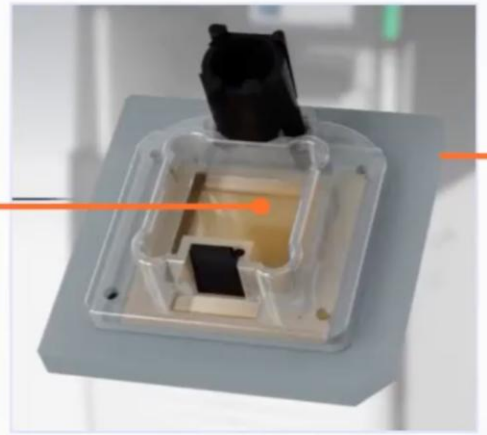


4种核苷酸 (A, T, G, C) 被标记有不同颜色的荧光基团





1个cell上800万个纳米小孔



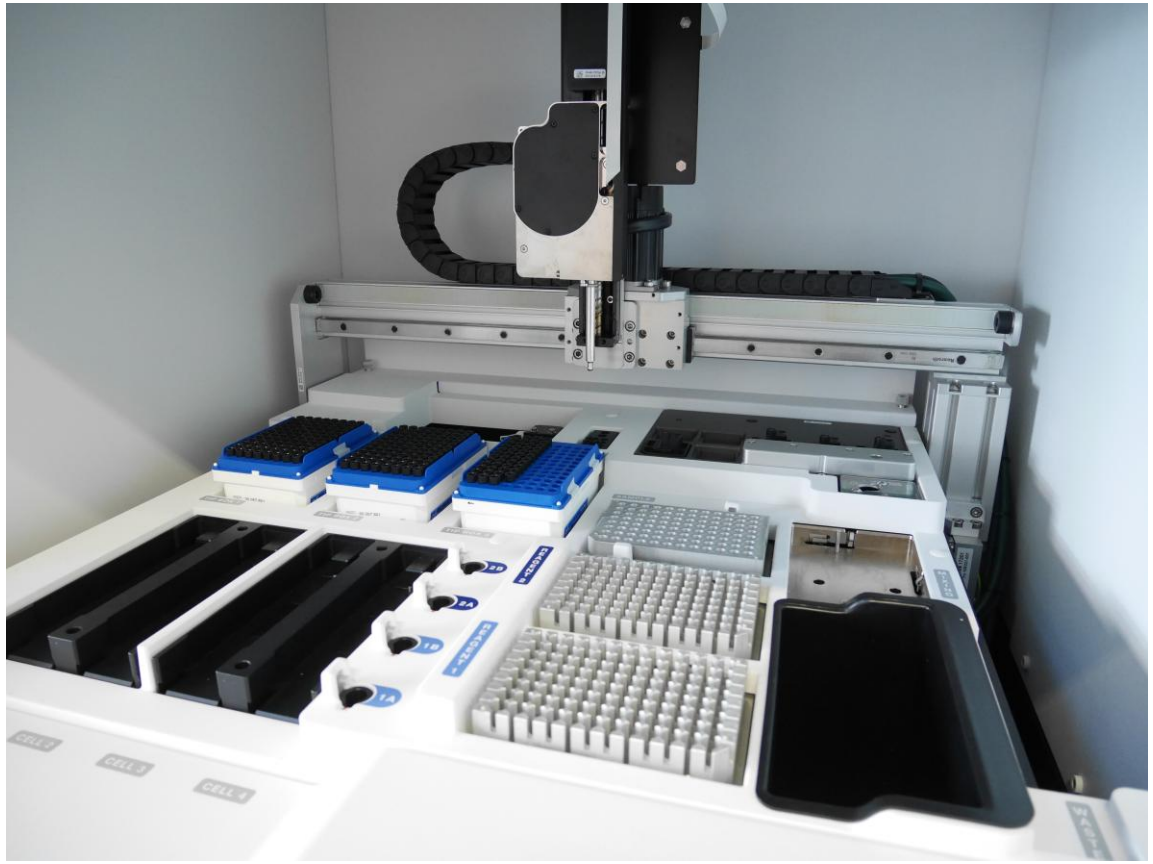
SMRT Cell测序芯片

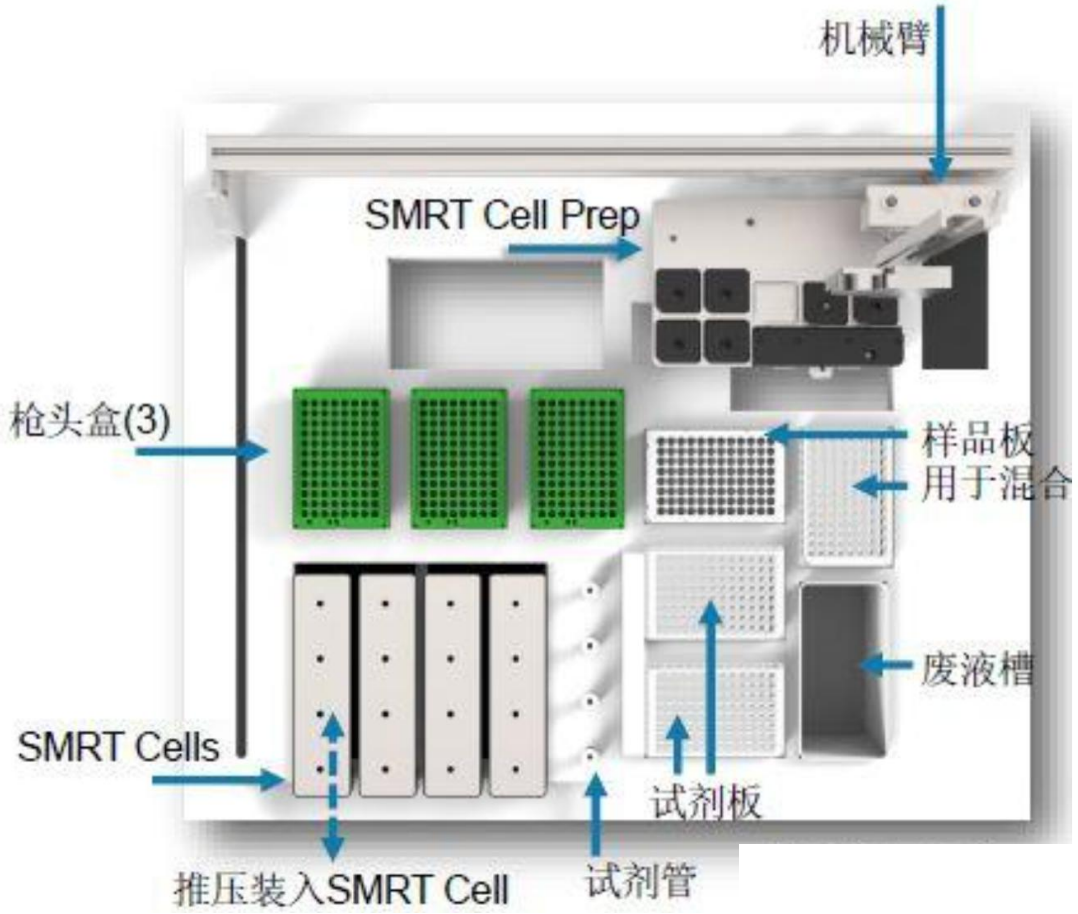
Pacbio测序仪

过程比喻：像一个高速摄像机在井底（ZMW）拍摄一个正在砌墙的工人（聚合酶），每拿起一块砖（dNTP），砖头闪一下光，记录下他拿的砖的颜色（序列）

PacBio 测序上机流程

将建好的文库和测序试剂、耗材放入测序仪，操作测序仪开始运行。仪器会显示结束时间倒计时。





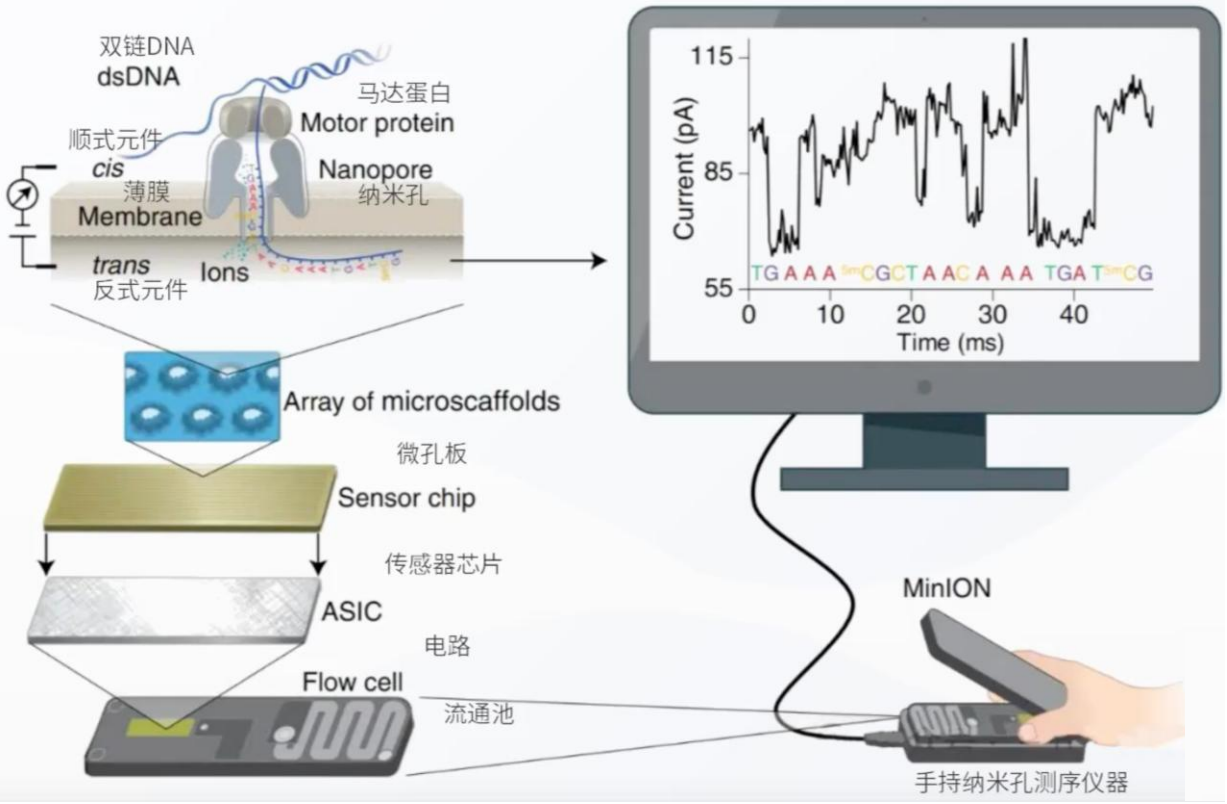
PacBio 两种测序模式

模式	全称	主要特点	典型应用
CCS (Circular Consensus Sequencing)	环状一致性测序	- 将目标DNA片段做成**环状模板（SMRTbell)** - DNA聚合酶可多次读取同一分子，从而得到多个重复reads - 多次读取后取**共识序列（Consensus Read)**，大大提高准确率（>Q30，错误率 < 0.1%)	高精度reads； 适用于**全基因组组装**、**转录组分析**、**突变检测（SNV/indel)**
CLR (Continuous Long Read Sequencing)	连续长读长测序	- 聚合酶从模板一端开始连续读取，不循环 - 可获得**超长reads（>20–50 kb)**，但错误率较高(约10–15%)	适用于**复杂基因组装（de novo assembly)**、**结构变异检测（SV)**、**重复序列分析。**

Nanopore sequencing

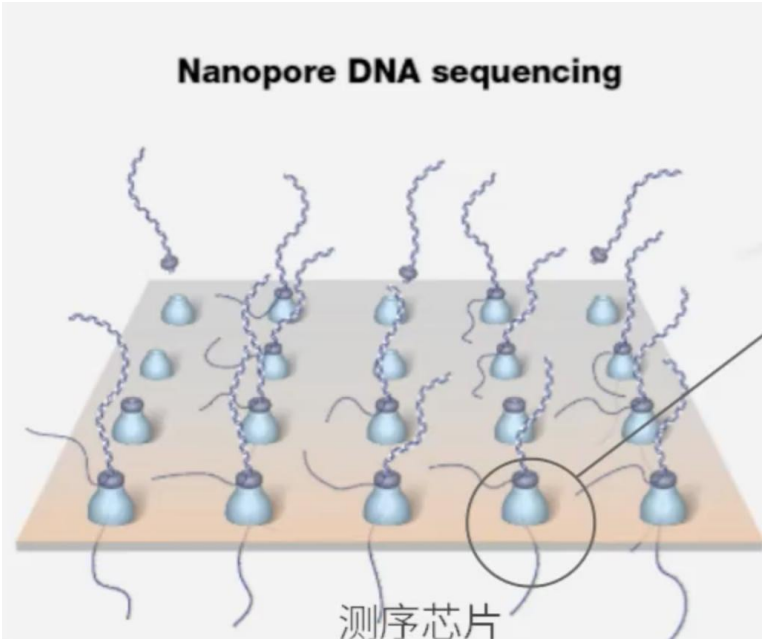
牛津纳米孔技术（Oxford nanopore technologies, ONT) 是第一家提供纳米孔测序仪的公司，第一个原型MinION于2014年发布。

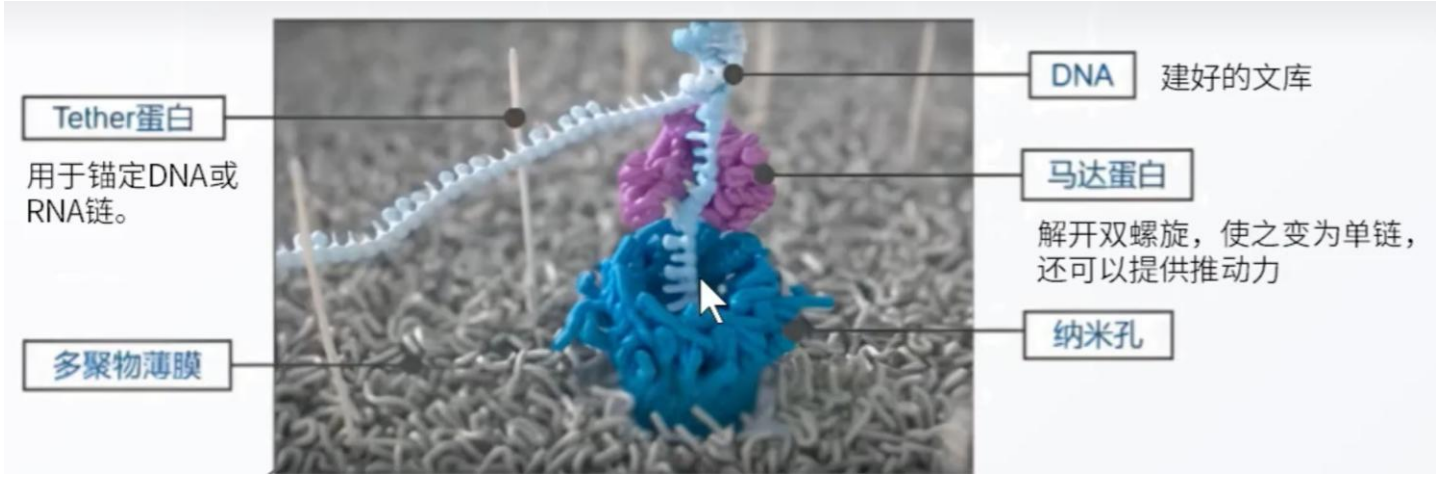
核心技术的原理是：核酸分子通过动力蛋白引导通过特殊的纳米孔，核酸通过时会引发电阻膜上电流的微小变化，基于ATCG每个碱基的带电性质不同产生不同的电流信号，从而推导确定碱基序列。



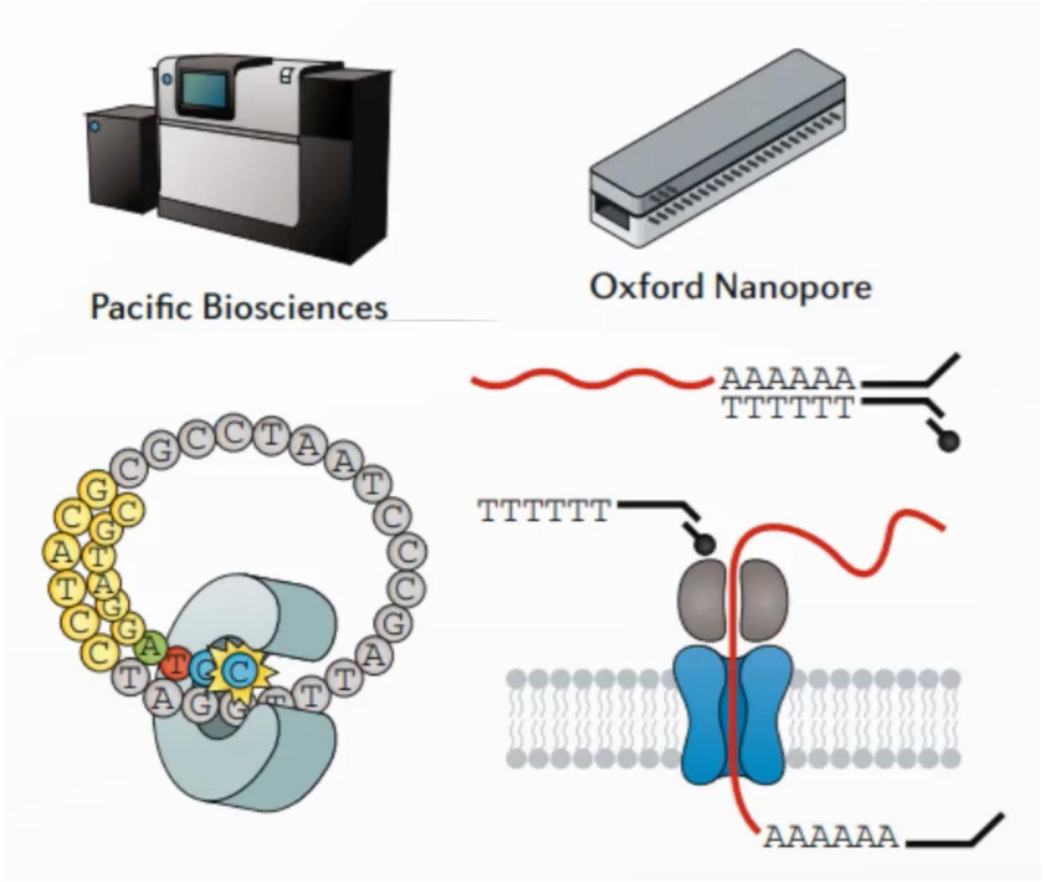
比喻：就像一根绳子（DNA）被拉着（马达蛋白牵引）穿过一个非常窄的门（纳米孔），门上有个传感器，通过“听”绳子不同粗细部分（碱基）通过时发出的“声音”（电流变化）来判断绳子的组成（序列）

Nanopore sequencing





三代测序技术特点



边合成边测序，长链分子，少拼接或不拼接

- ✓ 长读长：从几千到几十万个碱基对。
- ✓ 无需PCR扩增：避免了大量扩增过程中模板信息的丢失，提高了测序的准确性。
- ✓ 可以直接测RNA的序列，直接检测甲基化。
- × 高错误率：长读序列技术在读取过程中容易出现插入、删除和替代错误，错误率15%-40%。
- × 数据处理难度大：由于产生的读长通常较长，数据处理和分析的复杂性也相应增加。

测序技术比较

- 一代、二代、三代PacBio测序技术的共同点在于基于在DNA复制中对掺入的A/C/G/T引入标记信号，通过不同的方法读出DNA序列：
  - 一代测序技术通过按片段大小依次读出末端终止碱基；
  - 二代测序技术并行、循环可逆地边合成边读取信号（掺入终止碱基/读取信号/去除终止基团和信号）；
  - 三代测序技术并行、以单分子实时读取掺入的碱基信号。
- 它们的主要区别在于：
  - 一代测序技术通量最低、读长稍长于二代测序技术；
  - 二代测序技术读长最短，但通量最大；
  - 三代测序技术读长最长，但通量低于二代测序技术。

NGS 数据获取

NCBI https://www.ncbi.nlm.nih.gov/**sra**

EMBL-EBI https://www.ebi.ac.uk/

DDBJ DNA Data Bank of Japan https://www.ddbj.nig.ac.jp/

国家基因组科学数据中心 National Genomics Data Center https://ngdc.cncb.ac.cn/

常用工具

常用工具及其作用

需求	工具
基因鉴定	GENSCAN、AUGUSTUS、BRAKER2
重复DNA序列的鉴定	RepeatMasker
全基因组的多重比对	PHAST、CACTUS
保守DNA元件的鉴定	phastCons, PhyloP
二代测序读段与参考基因组的比对	Bowtie、**BWA**、STAR
比对文件的存储和解析	**SAMtools**
ChIP-seq的peak鉴定	MACS2、PeakSeq
基因表达的定量	StringTie、FeatureCounts、Salmon
差异表达的统计性检验	edgeR、DESseq2
可变剪接事件的鉴定和定量	rMATS
转录因子基序（Motif）的发现	MEME、Homer
染色体的突变鉴定	**GATK**、VAAST
染色质状态的鉴定	ChromHMM
基因调控网络的解析	PECA、ANANSE

基因组学数据可视化

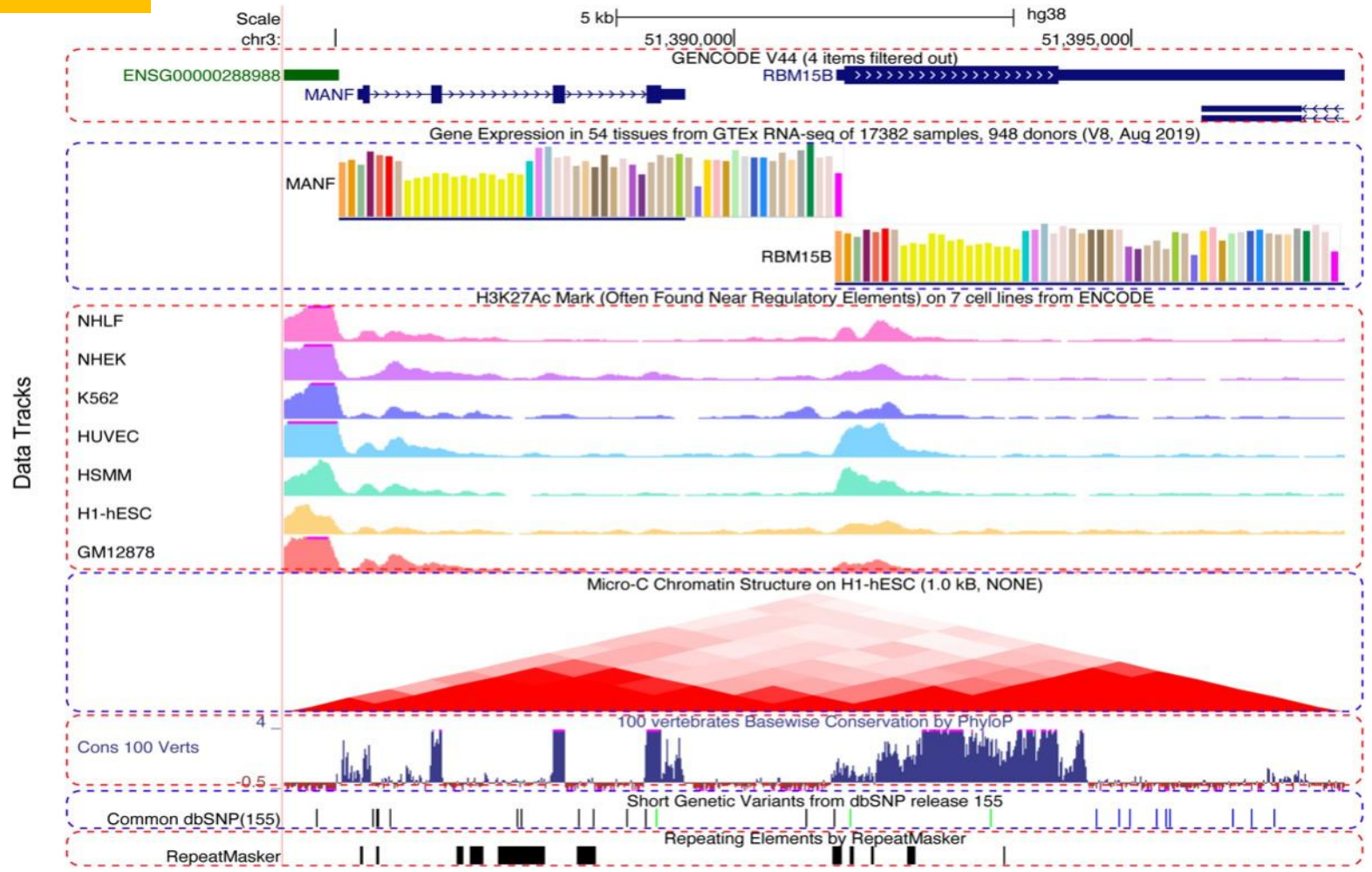
基因组学数据可视化

- 常见的基于网络的集成基因组浏览器包括： **UCSC** 和 Ensembl Genome Browser、NCBI Genome Data Viewer 等，提供了涵盖多个物种的基因组数据和工具

- 本地基因组浏览器包括 **IGV** (Integrative Genomics Viewer)、IGB (Integrated Genome Browser)等。

基因组学数据可视化

https://genome.ucsc.edu/



UCSC Genome Browser数据界面示例

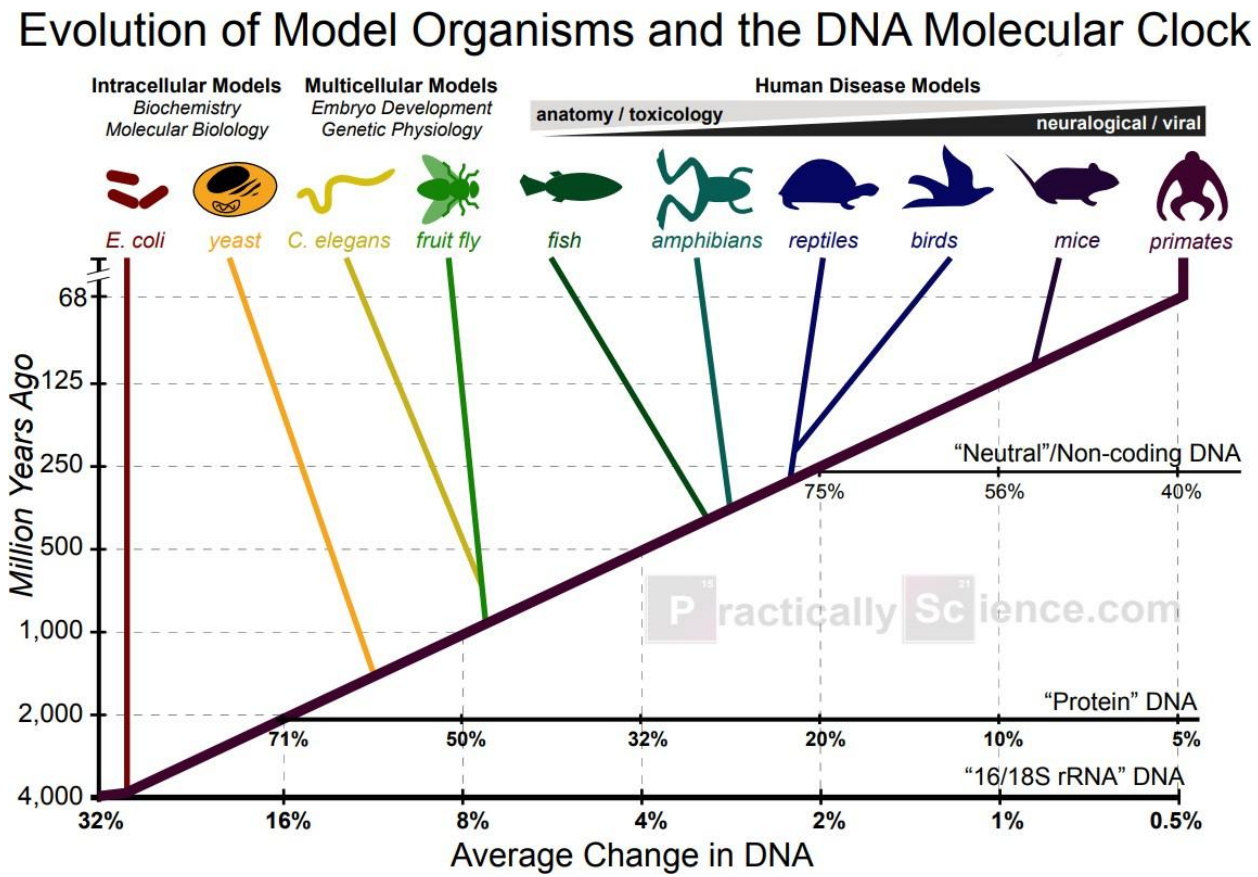
本章内容

章节结构

- **第一节：基因组学概述**
- **第二节：序列变异检测**

分子进化与变异积累

遗传物质的序列并非恒定不变的，而是动态演化的。基因组突变 (genomic mutation) 作为进化论的核心机制之一，对物种的演化历程起着决定性作用。

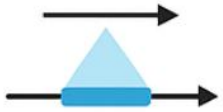


常见的变异类型

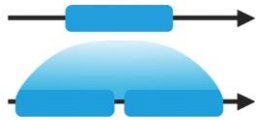
Single Nucleotide Variant

Deletion

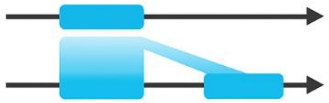
Insertion



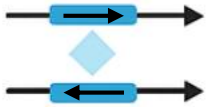
Tandem Duplication



Interspersed Duplication

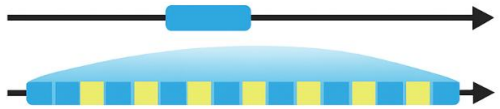


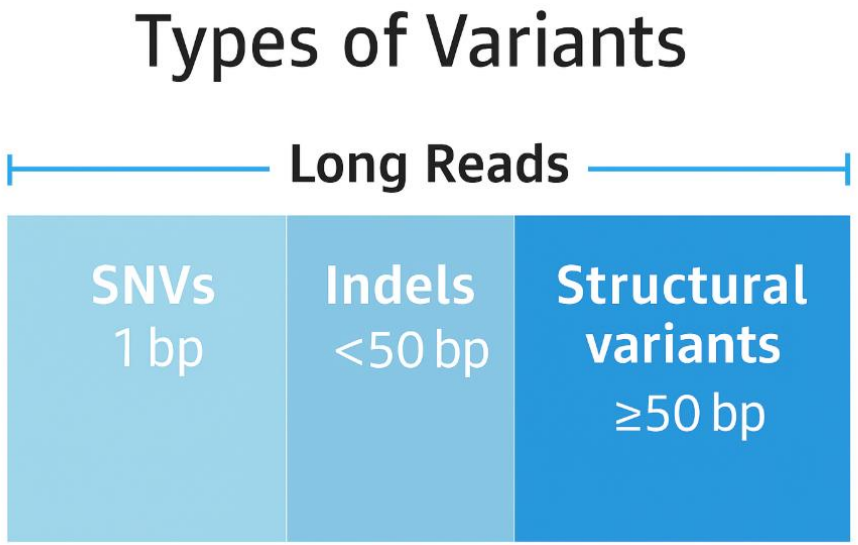
Inversion



Translocation

Copy Number Variant



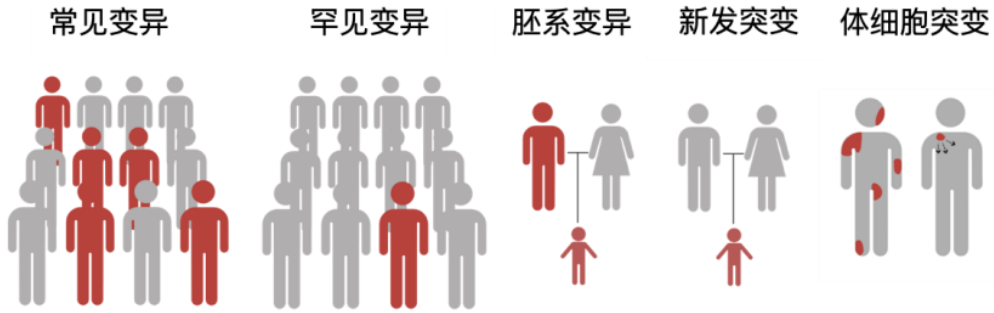


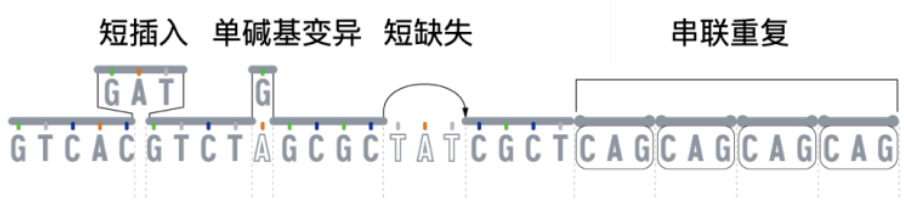
人群遗传差异与变异类型

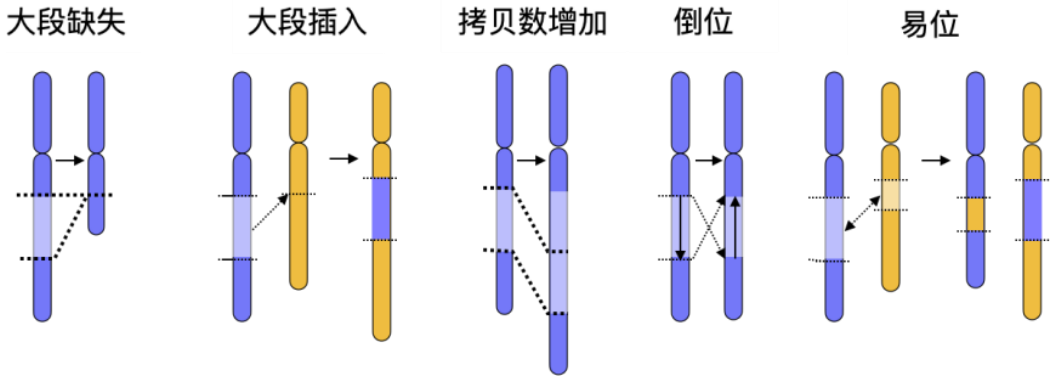
千人基因组计划的深入研究显示人类基因组**个体差异**大约占总基因组总长的**0.4%**，基中0.1%为单碱基差异，约0.3%为其他类型差异。

这些差异导致了我们在人类中观察到的巨大差异性。

各种类型的基因组变异。其中主要包括单碱基变异（single nucleotide variation，**SNV**)、短插入缺失（short insertion/deletion, **InDel**) 以及各种结构变异(structural variation, **SV**)





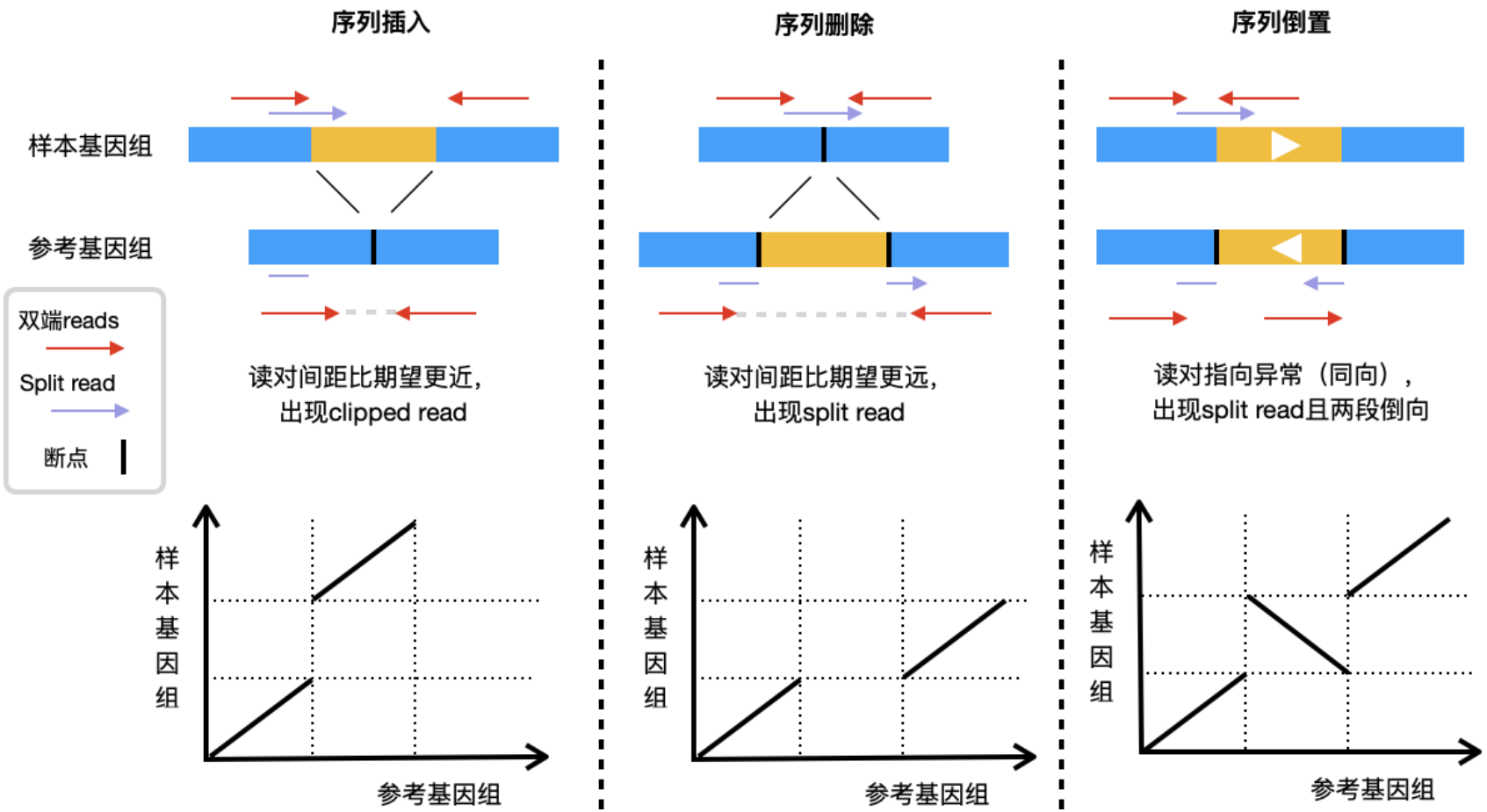


序列变异检测

检测单碱基序列变异的常用工具

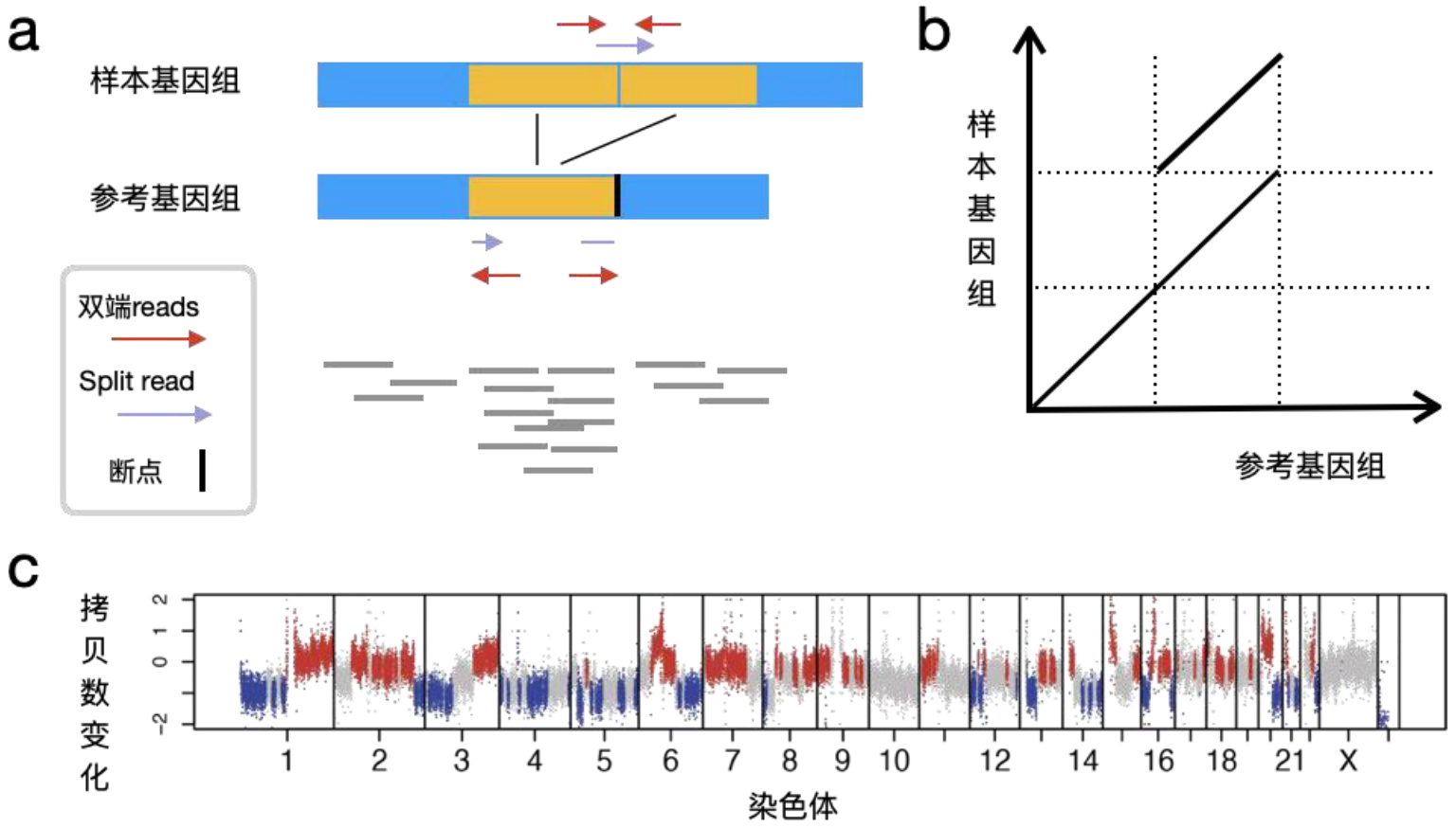
方法分类	全身性变异	癌症体细胞突变	非癌体细胞突变、单细胞突变
启发式算法	/	Varscan	LiRA
统计模型	GATK-HaplotypeCaller, Samtools, FreeBayes, Platypus, Octopus	Scomatic, Mutect2, Octopus, Vardict	Monovar, Sccaller, CAN2,LoFreq, MosaicHunter, Monopogen
机器学习和深度学习模型	DeepVariant, Strelka2	Strelka2	MosaicForecast, DeepMosaic
图模型	Dragen, Pangenie	/	/

结构变异的检测



结构变异断点附近的双端测序数据（上）和用reads拼装好的连续片段（下）比对到参考基因组后呈现的特征

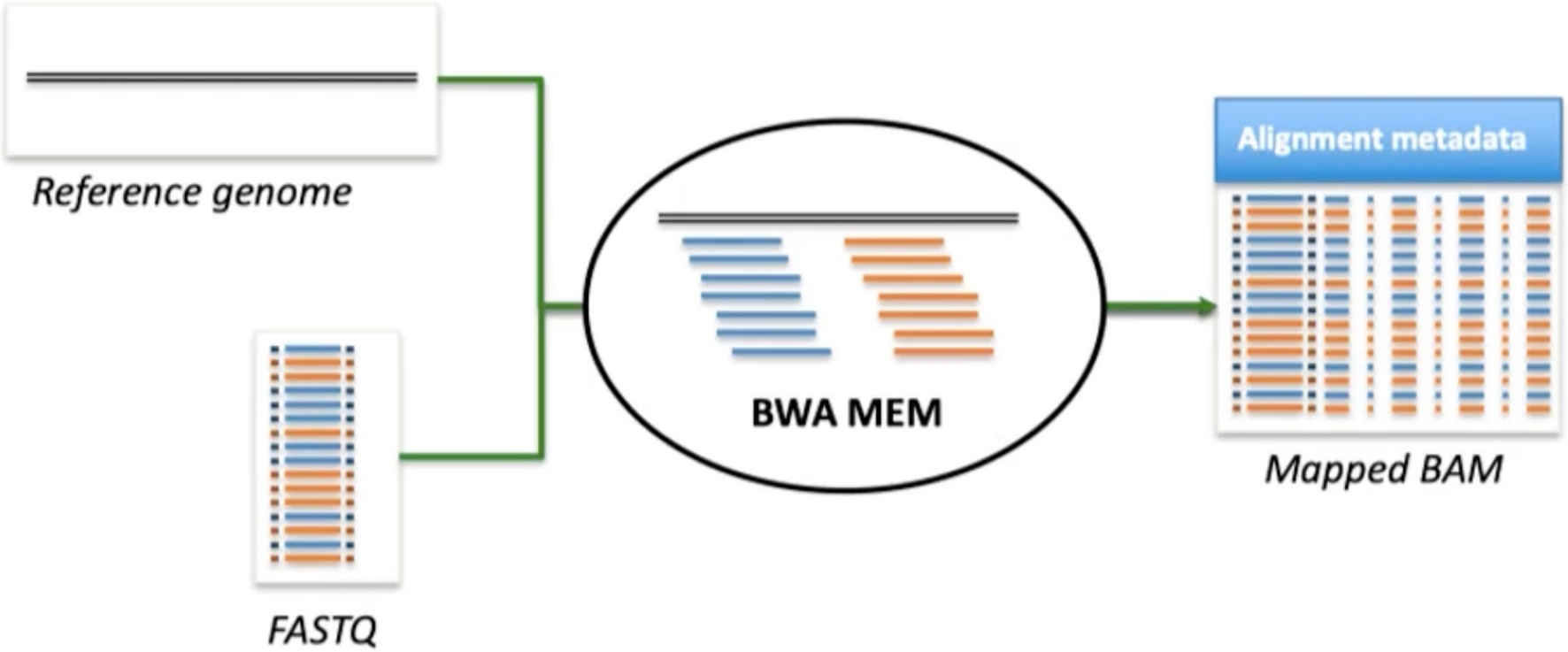
结构变异的检测



拷贝数增加变异引起的测序数据特征变化（a）以及用reads拼装好的连续片段(b)特征。（c）一组真实癌症数据中的拷贝数变异。红色区段为拷贝数增加，蓝色区段为拷贝数减少

BWA

Burrows–Wheeler Aligner (**BWA**) 和 Bowtie 是基于 **Burrows–Wheeler Transform (BWT)** 和 **FM-index** 的高效比对算法，用于将测序得到的短序列（reads）快速比对到参考基因组上。



BWA： http://bio-bwa.sourceforge.net/

Burrows–Wheeler Transform

- **可逆的排列变换（Reversible permutation）**，最初用于**数据压缩。**
- **数据库序列（Database sequence）**：T = acaacg $ ←

Burrows–Wheeler Transform

- 一旦构建出最后一列（**Last column**），其他中间过程都可以丢弃
- BWT矩阵的首列（**First column**), 可以通过对last column排序得到
- 在Last column中，字符会自然聚集在一起，从而使得压缩算法更高效

gc$aaac -> compression -> gc$3ac

T = acaacg

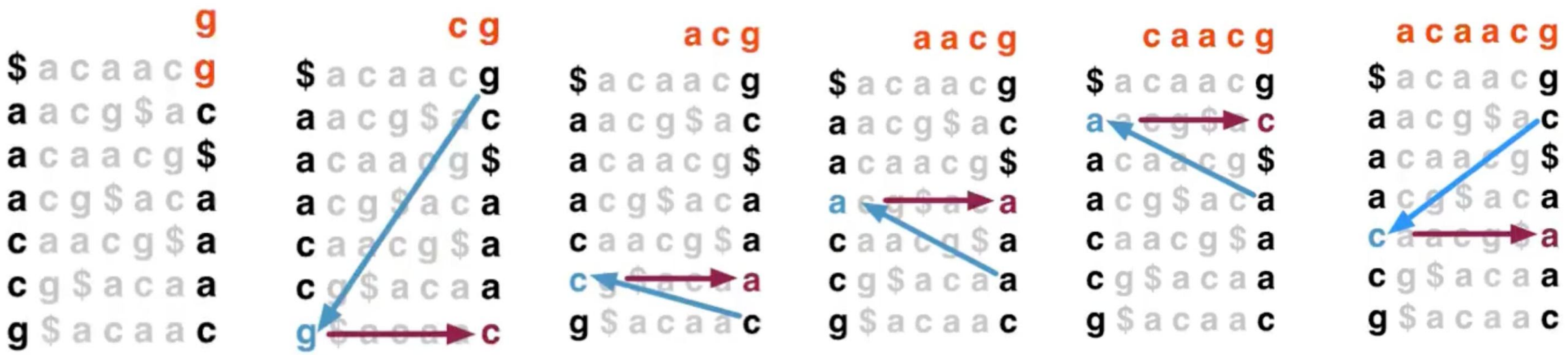
如何利用 BWT(T) 恢复原始序列 T?

—— LF 映射（LF mapping）

Burrows–Wheeler Transform

- LF mapping

- 核心原理
  - F 列和 L 列包含相同的字符集合，只是顺序不同。
  - 同一行中，L列 实际上 F列 前一个字符
  - 每个字符在 L 列中的第 *k* 次出现，对应 F 列中该字符的第 *k* 次出现。



数据下载

UGENE的下载链接：https://ugene.net/download-all.html