一、绪论

基因组/Genome：生物体所有遗传物质的总和。 基因组学/Genomics：研究基因组的科学。对生物体所有基因进行集体表征、定量研究及不同基 因组比较研究的一门交叉生物学学科。主要研究基因组的结构、功能、进化、定位、合成/编辑等，以及它们对生物体的影响。

基因组学由技术驱动，特别是测序技术。第一代Sanger，DNA micorarrays，第二代SGS，第三代单分子测序

获得基因组

如何获得基因组？

1. 人类基因组计划HGP
   - 全基因组鸟枪法
   - 层级鸟枪法（Hierachical shotgun）（先拆成短BAC序列）
2. 技术
   - DNA测序
   - 组装
     - 算法
     - 质量控制
   - 三代测序

HGP从1990-2000完成草图，2003完成。

HGP的策略

全基因组鸟枪法或者层级鸟枪法

HGP的技术

测序：测定基因组DNA分子碱基顺序 测序仪技术：测序仪这一机器的设计与生产，配套设备的研发和使用。

DNA测序技术

1. 第一代测序，Sanger法测序
2. 第二代测序（HTS，高通量测序，NGS，下一代测序），测序技术只有边合成边测序（SBS）
3. 第三代测序，纳米孔测序等。

在一个测序结果中，负链的0号位是代表了负链的3‘端，而正链的0号位代表了5’端

DNA组装技术

组装算法

最新技术

1. **Scaffolding** ：

- **Scaffolding** 在基因组学中是一个重要概念，指的是一种用于基因组组装的技术。在基因组测序的过程中，科学家通常会得到大量较短的DNA序列片段，这些片段被称为“contigs”。由于这些contigs只覆盖了基因组的一小部分，因此需要进一步的工作来将它们正确地组装成完整的基因组。
- Scaffolding的目的就是将这些contigs按照它们在基因组中的正确位置和顺序排列起来。这通常涉及到使用额外的信息来确定contigs之间的相对位置和距离，比如通过跳跃克隆（mate-pair sequencing）或光学测图等技术。通过这些技术，研究人员可以确定两个contigs之间是否相邻，以及它们之间的距离，从而把它们连接成更长的序列，即“scaffolds”。

2. **PacBio Contigs** ：

- **PacBio** 是指Pacific Biosciences公司的一种测序技术，这种技术以其高质量的长读长（long reads）著称。在这种技术中，单个的DNA分子被顺序地测序，产生的长序列被称为contigs。
- **Contigs** 是指通过测序获得的较长的、无间隙的DNA序列片段。这些片段是基因组组装过程的基本单元。

3. **Optical Mapping：** 光学测图（Optical Mapping）是一种用于构建基因组高分辨率限制性酶切图谱的技术。它通过对单个染色的DNA分子（称为“光学图谱”）进行分析，来确定未知DNA中限制性酶切位点的位置。这些DNA片段的组合为每个序列提供了独特的“指纹”或“条形码”。

   实验步骤如下

   1. **第一步：DNA条形码：** 首先，通过裂解细胞来释放基因组DNA。这些DNA分子被解开后，放置在带有微流控通道的光学映射表面上，允许DNA通过这些通道流动。随后，通过限制酶对这些分子进行条形码标记，以便通过光学映射技术进行基因组定位。
   2. **第二步：模板切割：** 向固定好的DNA分子中添加DNase I，以随机切割DNA。然后洗涤以去除DNase I。每个模板发生切割的数量取决于DNase I的浓度和孵育时间。
   3. **第三步：间隙形成：** 添加T7外切酶，利用DNA分子中的切口在5'至3'方向扩大间隙。需要小心控制T7外切酶的量，以避免产生过多的双链断裂。
   4. **第四步：荧光染料的结合：** 使用DNA聚合酶将荧光染料标记的核苷酸（FdNTPs）引入每个DNA分子上的多个缺口位点。每个循环中，反应混合物包含一种FdNTP类型，并允许多次添加该核苷酸类型。随后进行洗涤，去除未结合的FdNTPs，为成像和下一个FdNTP添加循环做准备。
   5. **第五步：成像：** 利用荧光显微镜计算缺口区域中结合的荧光标记核苷酸数量。
   6. **第六步：光漂白：** 用于激发荧光染料的激光同时用于破坏荧光信号，重置荧光计数器，为下一个循环做准备。这一步是光学测序的特殊环节，因为它不会在核苷酸结合后去除荧光标记，这使得测序更为经济，但也带来了一些挑战。
   7. **第七步：重复步骤4-6：** 重复进行步骤4至6，每次步骤4使用含有不同荧光染料标记的核苷酸（FdNTP）的反应混合物，直到测序所需区域。

optical mapping技术可以用于scaffolding

Assembly和scaffolding的区别

1. 基因组**Assembly**：

- 组装是指将测序得到的大量短序列（reads）通过重叠区域拼接成较长的序列片段的过程。
- 这些较长的序列片段被称为contigs（连续区）。
- 组装的主要挑战是如何准确地将这些短序列正确地拼接起来，尤其是在基因组中存在大量重复序列的情况下。
- 组装是基因组测序中的第一步，旨在尽可能地还原出基因组的连续区域(contigs)。

2. **Scaffolding**：

- Scaffolding是在组装步骤之后进行的，目的是将已经组装好的contigs按照它们在基因组中的正确顺序和相对位置连接起来。
- 在这个过程中，通常会利用额外的信息（例如mate-pair信息、光学映射数据或其他分子标记）来确定不同contigs之间的距离和顺序。
- Scaffolding的结果是更长的序列片段，这些片段被称为scaffolds（脚手架）。它们提供了比contigs更加接近完整基因组结构的视图。
- Scaffolding的挑战在于解决contigs之间的未知区域，并正确地组织这些contigs。

Hi-C Read Pairs：

- Hi-C是一种用于研究染色体在细胞核内如何折叠和组织的技术。
- 在Hi-C实验中，通过特定的化学方法将空间上接近的DNA片段连接在一起，然后进行测序。
- 测序得到的“read pairs”（读段对）反映了原本在三维空间中接近的两个基因组区域。这些信息对于了解染色体的空间结构非常重要。

Generate Contact Maps（生成接触图）：

- 接触图是一种视觉表示方法，用于展示基因组内各个区域之间的物理接触频率。
- 在Hi-C实验中，接触图是通过分析read pairs之间的相互作用来构建的。
- 这些图表显示了染色体内部不同区域在空间上的接近程度，有助于理解基因表达调控和染色体结构。

Hi-C技术提供了染色体三维结构的信息，这对于基因组的组装和scaffolding非常有价值。特别是在scaffolding阶段，Hi-C数据可以帮助确定不同contigs在空间中的相对位置，从而更准确地组织它们。

```
# contact map 示意图
import numpy as np
import matplotlib.pyplot as plt

# 创建一个随机矩阵来模拟染色体接触频率
data = np.random.rand(20, 20)

'''在实际的染色体接触图中，横坐标和纵坐标通常代表基因组的特定区域，这些区域可以是特定的染色体位置、基因区域，或者更具体地，是基因组上的碱基对（base pairs）的位置。因此，横坐标（以及纵坐标）确实可以表示为基因组上的碱基对序号，即显示某个特定区域是从第几个碱基开始的。

例如，在一个详细的染色体接触图中，横坐标可能标记为1,000,000, 2,000,000, 3,000,000等，表示每个点对应基因组上的一百万个碱基位置。这种表示方法允许研究者准确地识别出基因组中相互作用频率高的区域，从而深入理解基因组结构和功能。'''

# 使用matplotlib绘制接触图
```

```
plt.imshow(data, cmap='hot', interpolation='nearest')
plt.title('Hypothetical Chromosome Contact Map')
plt.xlabel('Genomic Region A')
plt.ylabel('Genomic Region B')
plt.colorbar(label='Contact Frequency')
plt.show()
```

解码基因组

如何识别基因组的信息呢？

1. 基因注释
2. 功能元件
3. RNA组解读
4. 表观遗传组解读
5. 3D结构组解读
6. 比较基因组
7. 单细胞基因组
8. 精准医疗

获得了一大堆ATCG后，哪里是基因呢？

寻找基因

中心法则，从基因组开始，转录出RNA，然后剪接成mRNA（剪接位点的GT做donor，AG做acceptor会被删掉） 隐马模型能帮我们吗？隐状态链本身有状态转移概率，而某一状态的表现又有发射概率。据此可以通过表现推测状态。例如，它是内含子吗等等。

基因注释、功能元件

Gene Annotation（基因注释）是生物信息学中的一个关键过程，它涉及识别基因组序列中的各种功能元素，并提供关于这些元素功能的信息。这个过程包括几个主要方面：

1. **基因的定位**：确定基因组序列中特定基因的位置。这包括识别编码蛋白质的基因（即开放阅读框架或ORFs）以及非编码RNA基因。
2. **功能预测**：对于已经定位的基因，基因注释还包括预测其可能的功能。这可能基于同源性（与已知功能的基因相似），基因表达模式，或者通过各种生物信息学工具的分析。
3. **结构注释**：确定基因的内部结构，如外显子、内含子、启动子、增强子等区域。

4. **实验验证**：虽然基因注释往往依赖于计算方法，但实验验证也是一个重要环节，如通过实验来验证预测的基因表达模式或功能。

功能元件的确定

可以使用ChIP-seq方法

1. **细胞制备和染色质免疫沉淀（ChIP）**

- **细胞固定**：使用甲醛或其他交联剂处理细胞，将蛋白质与DNA暂时固定在一起。
- **染色质裂解**：使用超声波或酶处理将染色质裂解成较小的片段。
- **免疫沉淀**：使用特异性抗体结合目标蛋白（如特定的转录因子或组蛋白修饰，例如，你感兴趣p53结合在哪里）。这些抗体会特异性地结合到带有目标蛋白的DNA片段上。
- **沉淀和洗涤**：通过沉淀复合物并清洗去除非特异性结合的片段。

2. **DNA的提取和净化**

- **逆交联**：加热样品，以逆转甲醛交联，释放DNA。
- **DNA提取**：使用各种方法提取和净化与目标蛋白质相结合的DNA片段。

3. **测序库准备和测序（ChIP-seq）**

- **测序库准备**：将提取的DNA进行修饰，加上测序接头，并进行扩增。
- **高通量测序**：使用NGS（如Illumina平台）进行高通量测序。

4. **数据分析**

- **质量控制**：检查原始测序数据的质量。
- **比对**：将测序读段比对到参考基因组。
- **峰值检测**：使用特定的算法识别富集区域，这些区域表明了目标蛋白质在基因组中的结合位点。
- **功能注释和分析**：对检测到的结合位点进行注释，分析它们在基因调控、基因表达等方面的潜在作用

ChIP-seq信号的组成：

1. **读段富集区域**：特定区域读段数量的增加表明该区域与实验中目标蛋白质有较强的相互作用。例如，转录因子结合位点或活跃的染色质区域通常会显示出较高的读段富集。

2. **峰值（Peaks）**：在ChIP-seq信号中，读段富集的区域被称为peaks。这些峰值通过专门的算法检测，代表了蛋白质-DNA相互作用的热点。

ChIP-seq信号的解释：

- **定位功能性元素**：通过识别峰值位置，研究人员可以确定转录因子的结合位点、组蛋白修饰的位置等功能性元素。
- **定量分析**：ChIP-seq信号的强度可以用来定量评估目标蛋白质与DNA结合的强度。强信号通常意味着较高的结合强度或更多的蛋白质存在。
- **比较研究**：通过比较不同样品或不同条件下的ChIP-seq数据，可以揭示蛋白质结合模式的变化，从而理解基因调控网络的动态变化。

Reads mapping

测序获得了reads，可是不知道这些reads究竟在基因组的哪里。于是需要建立mapping。

Reads mapping：指的是将来自高通量测序（如Illumina测序、PacBio测序或Nanopore测序）的短序列（reads）与一个参考基因组或已知的DNA/RNA序列进行比对的过程。这个过程的目的是确定这些短序列在参考基因组中的具体位置，从而可以进行后续的分析，如变异检测、表达量分析等。

一种策略如下。

1. **建立索引**：对参考基因组进行预处理，创建一个索引（或哈希表）。这个索引包含了基因组中所有可能的短序列（如长度为k的k-mer）及其在基因组中的位置。例如，如果k=3，那么索引将包括所有3个碱基长的序列（如AGC）及其在基因组中出现的位置。
2. **查询过程**：当有新的测序读段需要比对时，系统会首先将其分解成多个短序列，并在索引中查找这些短序列的位置。这样就可以快速定位读段可能匹配的基因组区域。
3. **精确比对**：一旦确定了读段可能的位置，就会进行更详细的比对过程，以确认读段在基因组中的确切位置和最佳比对。

"k-mer"是生物信息学中的一个术语，指的是任意长度为k的DNA、RNA或蛋白质序列片段。在DNA和RNA序列分析中，k-mer通常指的是由A（腺嘌呤）、T（胸腺嘧啶，或在RNA中为U，尿嘧啶）、C（胞嘧啶）和G（鸟嘌呤）组成的长度为k的核苷酸序列。例如，如果k等于3，那么一个3-mer可能是"ATG"、"CAG"、"TTA"等任何可能的三个核苷

酸组合。k值的选择会影响分析的灵敏度和特异性。一般来说，k值较小会增加重叠的可能性，但可能降低特异性；而k值较大则相反。

常用的mapping工具有BIAST，HISAT2等等

Peak Calling

Peak Calling 是生物信息学中用于分析高通量测序数据（如ChIP-seq、ATAC-seq等）的一个重要过程。它专门用于识别测序读段（reads）在基因组上的富集区域，这些区域被称为“峰”（peaks）。在ChIP-seq数据分析中，这些峰通常代表了蛋白质（如转录因子或组蛋白）与DNA的结合位点。

Peak Calling的过程：

1. **数据准备**：首先，需要将测序得到的短读段通过比对（mapping）工具定位到参考基因组上。
2. **峰检测**：然后，使用特定的算法分析这些比对结果，以识别读段富集的区域。这些富集区域就是潜在的峰。
3. **统计分析**：对检测到的峰进行统计分析，以确定哪些峰是显著的，即不太可能是随机事件。
4. **峰注释**：最后，对这些峰进行注释，比如确定它们位于基因的哪个区域（如启动子、增强子等），或者它们可能与哪些生物学功能相关。

MACS（Model-based Analysis of ChIP-Seq）是一种流行的用于ChIP-Seq数据分析的Peak Calling算法。它能够有效地鉴定蛋白质-DNA结合位点（例如转录因子结合位点或组蛋白修饰位点）。以下是MACS进行Peak Calling的基本步骤：

1. 数据预处理

- MACS开始于已经通过比对工具（如BWA或Bowtie）处理过的ChIP-Seq数据，通常这些数据以BAM或SAM格式提供。
- MACS会首先读取ChIP和对照（如Input DNA）的测序数据。

2. 建立模型和峰值检测

- MACS通过估计ChIP和Input DNA之间的偏差来建立模型。它考虑了实验中的各种偏差，如局部偏差和不同的DNA片段长度。
- MACS使用滑动窗口扫描整个基因组，寻找ChIP样本相对于对照样本的富集区域。
- 它计算每个位置的富集得分，并使用动态泊松分布模型来评估这些得分的统计显著性。

3. 筛选峰值

- 根据设定的阈值（例如p值或q值），MACS会筛选出统计显著的峰值。
- 用户可以根据实验的具体需求调整这些参数，以优化峰值检测的灵敏度和特异性。

4. 峰值调整和精化

- MACS会对检测到的峰值进行调整和精化，以更准确地定义峰的边界。
- 它还可以对峰值进行后处理，比如去除ChIP样本中可能的背景噪音。

5. 输出结果

- 最终，MACS输出包含鉴定峰值的信息，如峰值的位置、长度、富集程度等。
- 输出结果通常以BED格式或其他常见的基因组数据格式提供，便于后续的分析和可视化。

MACS的主要优势在于它的模型可以准确地捕捉到ChIP-Seq实验中的特异性峰值。

在ChIP-Seq实验中，对照样本（如Input DNA）是非常重要的，它提供了一个基准，用于与ChIP样本进行比较。这里的对照样本通常指的是未经免疫沉淀的总染色质样本，也就是实验开始前原始的染色质状态。

对照样本（Input DNA）的作用：

1. **背景噪音评估**：对照样本显示了在没有免疫沉淀的情况下，DNA片段在基因组中的分布情况。这有助于识别和校正ChIP样本中的背景噪音。
2. **鉴别特异性结合**：通过比较ChIP样本和对照样本，可以区分出由于特异性蛋白质-DNA相互作用引起的读段富集区域和非特异性背景信号。
3. **数据标准化**：使用对照样本进行数据标准化，可以校正实验和测序过程中的偏差，如测序深度差异、DNA可及性差异等。

ChIP-Seq实验设计中的不同类型对照：

1. **Input DNA**：最常见的对照，直接从同一样本中提取的总DNA。
2. **Mock IP**：使用非特异性抗体（如抗体与实验目标蛋白无关）进行免疫沉淀的样本，有助于识别特异性抗体可能引起的假阳性信号。
3. **No Antibody Control**：没有添加抗体的免疫沉淀实验，用于评估免疫沉淀过程中的非特异性结合。

RNA组学

RNA-seq是RNA组学的重要手段。 使用RNA测序（RNA-Seq）技术来定量基因和其异构体（isoforms）表达水平。这种方法能够精确地测量不同基因及其各种转录变体在不同样品中的表达量。 步骤如下

1. 样本准备和**RNA**提取

- 从细胞或组织样本中提取总RNA，例如可以特别关注富含polyA尾部的mRNA（polyA+ RNA）。

2. **RNA**富集和片段化

- 使用寡聚dT珠子或其他方法富集mRNA，这是因为mRNA通常包含polyA尾部。
- 将富集的mRNA随机断裂成较短的片段（Random Fragmentation）。

3. 测序库构建

- 在RNA片段的两端连接测序接头（Linker Ligation）。
- 将链接了接头的RNA片段转化为cDNA，并进行必要的PCR扩增。

4. 高通量测序

- 将准备好的测序库在高通量测序平台上进行测序，产生大量的短读段（reads）。

5. 读段比对（**Align Reads**）

- 使用生物信息学工具将测序读段比对到参考基因组或转录本数据库。
- 这一步骤考虑到了剪接事件，意味着读段可能跨越基因组上的内含子。

6. 转录本组装（**Assemble Transcripts**）

- 根据剪接比对的结果，组装完整的转录本。
- 这一步骤涉及到识别不同的剪接变体和转录本的重建。

7. 表达量分析

- 根据读段在各个基因或转录本上的覆盖情况，定量基因和转录本的表达水平。
- 对表达水平进行比较，识别在不同样品之间表达量更丰富（more abundant）或更少的（less abundant）基因或转录本。

8. 数据解释和后续研究

- 基于RNA-seq数据分析结果进行生物学解释。
- 可能涉及功能富集分析、差异表达基因的进一步验证、转录因子结合位点预测等。

转录本的图示表达，一般分为三种区域，深浅着色区域，和细线区域。细线区表示内含子，而深浅着色区域则分别表示UTR和CDS，一般浅色区域表示UTR，深色区域是CDS，即起始于起始密码子，终止于终止密码子的一段区域。

RNA结构组学

研究RNA分子内部以及RNA分子间的相互作用和结构。 PARIS（Psoralen Analysis of RNA Interactions and Structures）步骤如下

1. 交联：

- 使用荧光素（Psoralen）和紫外线照射对RNA分子进行交联处理。荧光素是一种能与RNA双链区域形成共价键的化学物质，这种交联可以稳定RNA分子间和分子内的相互作用。

2. **RNA提取和片段化：**

- 从处理过的样品中提取RNA，并将其随机断裂成较短的片段。

3. **测序库构建和高通量测序：**

- 利用特定的方法从断裂的RNA片段中富集含有荧光素交联位点的RNA，并构建用于高通量测序的库。
- 然后在测序平台上进行高通量测序。

4. **数据分析：**

- 分析测序数据，识别RNA分子间和分子内的相互作用区域。
- 通过计算方法和生物信息学工具，可以揭示RNA分子的二级和三级结构。

3rd Generation RNA结构组学技术的特点：

- **更高的分辨率：**与传统方法相比，第三代技术如PARIS能够提供更高分辨率的结构信息，揭示更复杂和更精细的RNA结构。
- **更全面的视角：**不仅能够研究RNA分子内部的结构，还能探索RNA分子间的相互作用，这对于理解RNA在细胞内的功能非常重要。
- **技术整合：**这些方法通常结合了多种技术和学科的优势，如化学、分子生物学、生物信息学等。

Epi-Genome

表观遗传组学，染色质状态研究等等

3D/4D 基因组学

Hi-C技术检测基因组如何互作从而确定三维结构。

比较基因组学

比较基因的演化进程。综合分析多个个体的基因组。可以看看功能元件的演化特征之类的。例如，可以利用功能元件的保守性鉴定功能原件。

1. 选择目标基因和物种

- 确定你感兴趣的基因或基因组区域。
- 选择几个与你的研究物种相关的其他物种，理想情况下应该在进化上有一定的距离，但又足够近，以确保基因的同源性。

2. 获取和比对序列

- 从不同物种中获取你感兴趣的基因区域的序列。
- 使用序列比对工具（如BLAST、ClustalW、MAFFT等）对这些序列进行多序列比对，以识别保守区域。

3. 分析保守性

- 利用比对结果，寻找在多个物种中高度保守的序列区域。这些区域可能代表了重要的功能性元素，如转录因子结合位点。
- 可以使用专门的保守性分析工具（如PhyloP、GERP等）来量化保守性。

4. 预测转录因子结合位点

- 将保守区域与已知的转录因子结合模体（TF binding motifs）进行比较，这可以通过数据库（如JASPAR、TRANSFAC）或预测工具（如MEME Suite）完成。
- 寻找模体与保守区域的匹配情况，以预测潜在的转录因子结合位点。

单细胞基因组学

单细胞基因组学是指对单个细胞的基因组（全套DNA）进行分析的技术。这项技术使科学家能够在最精细的水平上探究遗传信息，揭示细胞间的微小差异，这些差异在用传统

的、整体组织水平的基因组学方法时可能无法检测到。

空间转录组学是指结合组织的空间信息与转录组数据（全套RNA）的分析技术。它使研究人员能够不仅了解哪些基因在特定细胞中表达，还能知道这些细胞在组织中的确切位置。

临床应用

1. GWAS GWAS（全基因组关联研究）是一种研究方法，用来识别基因组中与特定疾病、性状或生物学特征相关联的遗传变异。这种方法通过比较不同个体（例如疾病患者和健康对照组）的基因组序列，寻找频率在两组之间显著不同的遗传标记，通常使用的遗传标记是SNP。GWAS不依赖于关于疾病或性状的先验生物学假设，因此能够揭示之前未知的遗传风险因素。通过检测群体中成千上万个SNP的频率分布，GWAS能够识别出与特定性状或疾病相关联的遗传区域。这些关联并不直接表明因果，但可以为进一步研究遗传机制和病理过程提供线索。
2. SNP SNP（单核苷酸多态性）是指在人类或其他物种的基因组中，某个特定位置的单个核苷酸存在两种或更多的变体。人类基因组中有大约0.1%的差异，而这些差异中有90%是由SNP引起的。在群体中，如果某个特定位置的核苷酸变异的频率达到或超过1%，则该变异被称为SNP。SNP可以位于编码区、调控区或基因间区域。虽然许多SNP对个体的表型没有直接影响，一些特定的SNP与疾病易感性、药物反应性以及其他生物学性质有关。

还可以鉴定突变。

基因组学的深度学习

特征提取（使用卷积），人工神经网络方法。

编写基因组

如何人工构建基因组？从解读基因组到书写基因组，构建人工生命。使用合成生物学

基因组碱基编辑器

Genome Editing（基因组编辑）是一种现代生物技术，用于精确修改生物体基因组中的特定DNA序列。这种技术的核心是利用特定的酶系统，如CRISPR-Cas9，来在基因组的

特定位点引入切口，然后利用细胞自身的修复机制来引入所需的修改。基因组编辑技术在基础生物学研究、疾病模型制备、基因治疗、农业生物技术等多个领域展现出巨大潜力。

Base Editors（BEs） Base Editors（BEs）是一类特殊的基因组编辑工具，它们是对传统CRISPR-Cas9系统的扩展和改进。BEs能够在不引入DNA双链断裂的情况下，直接对单个DNA碱基进行精确的化学修改。这种技术主要用于实现碱基转换，比如将腺嘌呤（A）转换为鸟嘌呤（G），或将胞嘧啶（C）转换为胸腺嘧啶（T）。

BEs的工作原理

- **融合构造**：BEs通常由两部分组成，一部分是类似于CRISPR-Cas9的DNA结合域，用于将编辑器引导至特定的基因组位点；另一部分是碱基修饰酶，如脱氨酶。
- **碱基修改**：BEs通过与目标DNA结合并转换特定的碱基，实现点突变而不引起DNA链断裂，从而减少了因DNA修复引起的非特异性插入或缺失。

基因组合成

基因组合成（Genome Synthesis）

基因组合成是指使用化学方法从头合成整个基因组或基因组的大片段。这包括设计和构建新的基因序列，甚至是整个微生物的基因组，以研究基因的功能或创造具有新功能的生物体。基因组合成在合成生物学中扮演着重要角色。

应用：

- **创造新生物**：设计新的微生物，用于药物生产、生物能源、环境修复等领域。
- **基因功能研究**：通过合成改变的基因组来研究特定基因或基因网络的功能。
- **生物技术和医学**：开发新的治疗方法或生产有用化合物的微生物。

功能基因组学的研究使人们知道如何合成基因组，合成基因组学的发展让人们了解更多的功能基因。

二、HGP

在疾病遗传中，假设某基因D会导致遗传病，那么，与疾病基因靠的足够近的DNA标记往往会与疾病基因一起遗传。虽然常见的DNA标记在大的群体中会各不相同，但是在所研究的群体中由于遗传，一部分DNA标记（例如子代的一条染色体上的DNA标记与母亲

相同，另一条与父亲相同）往往相同，能够使得研究者从子代的DNA标记追踪DNA是从哪里来的。

DNA标记

常见的DNA标记有

1. RFLP(限制性片段长度多态性) **RFLP（限制片段长度多态性）** 是一种遗传标记，用于在DNA序列中检测多态性（即序列变异）。这种技术基于一个事实：限制性内切酶（一种酶）可以特异性地识别并切割DNA上的特定序列。如果这些识别位点因遗传变异（如插入、缺失或点突变）而改变，限制酶的切割模式也会随之改变，从而产生长度不同的DNA片段。这些不同长度的DNA片段就是所谓的“限制片段长度多态性”。
2. STR(简单串联重复多态性) 即某一个小片段的重复次数不同

CF基因定位（以前的基因定位技术）

CF基因定位是指确定引起囊性纤维化（Cystic Fibrosis，简称CF）的基因在人类基因组中的确切位置的过程。

在确定CF基因位置的早期研究中，科学家们寻找与囊性纤维化表型密切相关的遗传标记。如果某个标记经常与疾病一起出现（即在受影响的家族成员中共遗传），这表明该标记与疾病基因在染色体上靠得比较近，从而揭示了基因的大致位置。

1. 遗传连锁分析

- **初步研究：** 囊性纤维化是一种常见的遗传性疾病，表现为呼吸道和消化系统的问题。科学家通过对受影响家庭的遗传研究，确定了这是一种常染色体隐性遗传疾病。
- **寻找遗传标记：** 1980年代，研究人员开始寻找与CF表型相关的遗传标记。他们使用了限制片段长度多态性（RFLP）分析来识别可能与CF相关的遗传标记。通过分析受影响家庭的DNA样本，研究人员寻找与CF遗传表型一起出现的遗传标记。

2. 确定染色体位置

- **利用连锁映射：** 通过遗传连锁映射，科学家们确定了一个与CF表型紧密相关的遗传标记。这个标记位于第7号染色体的长臂上。

- **重组分析：** 进一步的重组分析帮助研究人员更精确地确定了CF基因的位置。他们分析了在CF患者家族中出现的重组事件，这些事件表示了基因与特定标记之间的物理距离。

3. 染色体步移（chromosome walking）

- **技术介绍：** 染色体步移是一种逐步沿着染色体特定区域识别和克隆DNA片段的方法。科学家们使用这种方法来逐步逼近CF基因的确切位置。
- **逐步克隆：** 通过从已知标记开始，逐步克隆邻近的DNA片段，研究人员逐渐缩小了CF基因的搜索范围。

4. 基因克隆和鉴定

- **克隆成功：** 经过持续的努力，1989年，研究人员最终成功克隆了囊性纤维化跨膜调节器（CFTR）基因。这是通过分析多个从步移中获得的克隆，最终确定了包含CFTR基因的克隆。
- **基因鉴定：** 对CFTR基因的进一步分析揭示了这个基因编码一个跨膜蛋白，该蛋白在盐分和水的细胞运输中发挥作用。在CF患者中，由于CFTR基因的突变，这个蛋白的功能受损，导致囊性纤维化的症状。

染色体步移（Chromosome Walking）主要是一种用于发现并定位染色体上特定区域的任意基因的方法，而在初始阶段并不需要知道这个基因具体是什么或者它的确切功能。在囊性纤维化（CF）基因定位的案例中，这个技术被用来沿着第7号染色体逐步移动，以找到与CF疾病表型相关联的未知基因。

染色体步移的目标

- **定位基因：** 步移的目的是确定某个区域内是否存在基因，以及这些基因的大致位置。
- **探索未知：** 在步移的过程中，研究人员可能并不知道他们最终会找到什么样的基因，他们只知道这个区域内可能存在与疾病相关的基因。

后续研究

- **功能鉴定：** 一旦通过步移找到了一个候选基因，就需要进行后续的实验来确定这个基因的功能，以及它是否真的与特定的疾病或表型相关。
- **突变分析：** 对于CF，一旦CFTR基因被确定，研究人员随后对其进行了深入研究，包括寻找导致囊性纤维化的特定突变。

（染色体walking已经是过时的技术，有了全基因组测序后根本用不上费时费力walking，直接测序后从计算机层面寻找可能的基因）

全基因组测序WGS

遗传图，物理图，序列图，基因图，在全基因组测序后全都被测序结果替代了。

全基因组测序（WGS）与传统图谱的关系

- **遗传图：** WGS数据可以用来分析基因之间的连锁关系和重组频率，从而构建出遗传图。
- **物理图：** 通过WGS，可以直接获得基因或标记在染色体上的确切物理位置，从而创建物理图。
- **序列图：** WGS本质上就是提供了完整的基因组序列，即最详尽的序列图。
- **基因图：** WGS数据可以用来识别和定位基因组中的所有基因，创建出详细的基因图。

HGP

基因组学发展史即从HGP讨论伊始，直至其完成及所有后续计划的历史，HGP于1990启动，2003完成。

HGP的科学依据：“遗传信息储藏在 DNA 序列中”（生命是序列的）这一理念

HGP的目标

HGP想要获得四张图，遗传图，物理图，转录图，DNA全序列图

连锁图（Linkage map）

表示基因或 DNA 标记在染色体上的相对位置与遗传距离的图谱 完成于1998年，含8325个STR标记，平均密度为0.36cM或2.8个/cM 遗传图是路标。

物理图

以STS位点（Sequence-Tagged Sites，一小段特异的 DNA 序列）为图标， Mb或 Kb为图距，表示基因组的物理大小或标记距离的图谱。STS，是已知序列的特异性单拷贝DNA片段，100~300 bp 完成于 1998 年 10 月，5.2 万个 STS，平均距离为 60Kb

1. 建立BAC克隆文库

- **操作：** 人类基因组的DNA被随机打断成较大的片段，然后这些片段被克隆到BAC载体中，形成一个大规模的BAC克隆文库。每个BAC克隆包含基因组中的一个独特片段。
- **目的：** 这个文库覆盖了整个人类基因组，每个BAC克隆代表了基因组的一小部分。

2. 使用STS筛选重叠克隆

- **操作：** 序列标签位点（STS）是基因组中独一无二的小DNA序列，可以作为“路标”来识别特定的基因组区域。使用STS两端的特异性引物，可以通过PCR方法筛选出包含这些STS的BAC克隆。
- **重叠判断：** 如果同一对引物可以从多个BAC克隆中扩增出相应的STS片段，这表明这些BAC克隆可能是重叠的，即它们包含了基因组中相互重叠的区域，同一段STS。

3. 通过消化和电泳判断重叠

- **操作：** 使用特定的限制酶对每个BAC克隆进行消化，然后通过凝胶电泳分析消化产物。
- **重叠分析：** 如果两个或多个BAC克隆在电泳图谱中显示出一些相同位置的条带，这意味着它们包含相同的DNA序列，从而进一步证实了这些BAC克隆之间的重叠。

转录图

所有基因的转录本（transcript）序列（一个基因完整的 cDNA 序列和不完整的 EST）的总和。 mRNA 逆转录得到 cDNA 或 EST（Expressed Sequence Tags，表达序列标签），EST是不完整的cDNA，只是cDNA的片段。因此组装EST可以得到完整的cDNA

HGP的技术路线

结合“重叠克隆（overlapped clones）”和“霰弹法（shotgun sequencing）”的双重策略，即逐个克隆霰弹法（clone-by-clone shotgun） 逐个克隆霰弹法：

1. 将初步定位的 BAC 逐个用霰弹法测序，以末端重叠组装，补“克隆内小洞 gap” 将所有相关克隆的一致序列按末端重叠组装成 Contig（序列重叠群）
2. 将 Contig 定位到物理图与遗传图上

3. 用 Contig 两侧序列设计 PCR 再在 BAC 文库中筛选新的克隆来补“克隆间大洞”

HGP的模式生物

大肠杆菌（Escherichia coli） 酿酒酵母（Scharomyces cerevisiae） 秀丽线虫（Caenorhabditis elegans） 拟南芥（Arabidopsis thaliana） 果蝇（Drosophila melanogaster） 河豚鱼（Takifugu rubripes ） 小鼠（Mus musculus） 是因为它们本身具有的重要科学和医学意义，以及生命世界中的代表性。多年的遗传学和其它生物及医学基础，特别是遗传图和物理图。前 6 种比人类小得多，便于发展和改进技术和策略（从易到难）。比较基因组学，染色体、基因等同源性，基因分布、排列顺序。运作上便于评估进展和成本，报告阶段性成果

HGP的完成

人类基因组草图完成于2000. 覆盖全基因组90%以上，平均准确率99%。 测序深度至少5X，即所有reads的总长度约为基因组估计长度的至少5倍。

Hierarchical Shotgun

- **过程描述：** Hierarchical shotgun（分层shotgun）方法与clone-by-clone策略类似，但在某些方面更加综合和高效。在这种策略中，基因组也被切割成大片段并克隆到BAC中，但是接下来的步骤不仅仅是对每个克隆进行测序。相反，这些BAC克隆首先与其它BAC克隆建立重叠关系，组织好所有BAC克隆的相互关系。然后再对每个BAC进行shotgun测序，测序后得到的contig对应一个BAC，根据之前组织好的信息可以直接组装成更大的染色体。

人类基因组的精细图完成于2003年，覆盖全基因组99%，准确率99.99%，测序深度10X

HGP完成后的后续

国际HapMap计划

- **目的：** 国际HapMap计划旨在建立人类基因组的单核苷酸多态性（SNP）的全球参考图谱。
- **关键点：** HapMap计划通过分析不同人群的基因组，识别了成千上万的SNPs，这些SNPs是人类基因组多样性的重要标记。

- **应用：** 这个计划帮助科学家理解遗传变异如何影响人类健康和疾病，促进了个体化医学和药物研发的发展。

ENCODE（DNA百科全书计划）

- **目的：** ENCODE项目的目标是识别和解释人类基因组中所有功能性元素，包括基因、调控区域和其他非编码序列。即提供基因组注释
- **关键点：** 该项目通过一系列的实验技术，如转录组测序、染色体共沉淀等，揭示了大量先前被认为是“垃圾DNA”的非编码区域的功能。
- **应用：** ENCODE项目的成果对于理解基因如何被调控、基因表达的复杂性以及非编码DNA在疾病中的作用至关重要。

国际千人基因组计划（G1K）

- **目的：** 国际千人基因组计划旨在建立一个详细的人类基因组变异图谱，通过测序来自不同人群的大约1000个人的基因组。
- **关键点：** 该计划提供了关于人类基因组多样性和遗传变异的丰富信息，包括SNPs、拷贝数变异（CNVs）和结构变异。
- **应用：** 这个项目的数据对于研究人类遗传病、人类进化以及开发个性化医疗策略具有重要意义。

国际癌症基因组计划

- **目的：** 国际癌症基因组计划旨在通过全面分析不同癌症类型的基因组，来理解癌症的遗传基础。
- **关键点：** 该计划涉及对成千上万个癌症样本的全基因组测序，以识别与癌症相关的遗传变异和信号通路。
- **应用：** 这些研究成果对于发现新的癌症治疗靶点、开发个性化癌症治疗方案以及理解癌症的分子机制至关重要。

HGP的意义和影响

创造了一种新的文化：合作。HGP 精神：共需、共有、共为、共享 催生了一门新的学科：组学 提供了一个新的技术：测序 HGP 的运行过程就是测序技术发展的过程 测序技术此后的发展也归功于 HGP、基因组学及其他“组学”的推动。测序技术使生命变成了数据。生命科学的数据化也汇入当今世界的大数据潮流。促进了 Bioinformatics 的发展

三、DNA测序

本节中我们考虑最细枝末节的DNA测序方法，最技术的最底层的方法，例如给你1000个bp，你该如何检测。而不是整个基因组的测序策略。基因组测序显然依赖于DNA测序。可以把DNA测序理解成最底层的一个封装函数。之后基因组测序只需要调用它。

前直读法，即酶切测序，使用不同的酶进行酶切得到小片段序列，推导完整序列

直读法

直读法，依赖于三个技术

1. 分子克隆
2. PAGE，能分辨一个碱基的差别
3. 放射自显影（现在可以直接荧光染色）

化学法

Maxam-Gilbert测序的基本步骤

1. **标记DNA：** 首先，DNA片段的一个末端被标记（通常是放射性标记）。
2. **化学修饰：** DNA样本分别暴露在一系列不同的化学试剂中，每种试剂都特异性地修饰DNA上的一种特定的核苷酸（比如A、G、C或T）。
3. **断裂DNA：** 经过化学修饰的DNA在修饰位点附近被切割（断裂）。
4. **凝胶电泳分离：** 切割后的DNA片段通过凝胶电泳进行分离。由于DNA片段的大小不同，它们在凝胶中移动的速度也不同，从而根据长度进行分离。
5. **读取序列：** 最后，通过分析凝胶上的放射性信号，可以确定DNA序列。不同长度的片段对应于从标记末端到不同修饰位点的距离，从而可以推断出DNA序列。

链终止法

Sanger方法，掺入ddNTP

Sanger测序法的基本步骤

1. **DNA合成：** 首先，单链DNA模板与一个特定的引物、DNA聚合酶、四种正常的脱氧核苷酸（dNTPs）以及少量的链终止脱氧核苷酸（ddNTPs）混合。这些ddNTPs是标记了不同荧光标签的。
2. **链终止：** 当ddNTP被嵌入到DNA链中时，它会阻止链的进一步延伸，因为ddNTP缺少3'端的羟基，这是链延伸所必需的。

3. **产生长度不同的DNA片段：** 这种方法会产生一系列长度不同的DNA片段，每个片段都以特定的核苷酸（A、T、C或G）结尾。
4. **电泳分离：** 这些不同长度的DNA片段通过毛细管电泳进行分离。
5. **检测和读取序列：** 当片段通过荧光检测器时，不同的荧光标签发出信号，从而确定每个片段的终止核苷酸。通过这些信号的顺序，可以推断出DNA模板的序列。

Sanger法测序获得的是模板链的互补链的序列。

自动化测序

Sanger法可以自动化后让机器来做反应。

规模化

毛细管电泳技术一次能电泳96个样本，比普通凝胶电泳多太多。

基本原理

- **电动力：** 在毛细管电泳中，样品中的带电粒子在电场作用下根据它们的电荷和大小进行迁移。通常，带负电的分子会向正极移动，而带正电的分子则向负极移动。
- **分离机制：** 分子的迁移速度取决于它们的电荷密度（电荷与体积的比率），大小和形状。不同的分子因此在毛细管中以不同的速度移动，从而实现分离。

毛细管电泳的组成

- **毛细管：** 一根细长的、通常是石英制的管子，内径通常在几十到几百微米之间。
- **电极和电源：** 毛细管的两端分别连接到高压电源的两个电极。
- **检测器：** 在毛细管的出口处通常有检测器，如紫外线吸收光谱仪或荧光检测器，用于检测和分析分离后的样品。

```
毛细管如何上样?
1. 静电注入 (Electrokinetic Injection)

* **原理：** 利用电场力来吸引带电的样品进入毛细管。在这种方法中，毛细管的一端浸入含有样品的溶液中，然后在毛细管两端施加电压，使样品因电动力进入毛细管。
* **特点：** 这种方法简单快捷，但可能受样品的电荷和电泳缓冲液的pH值影响，导致上样量的不一致。

2. 水力注入 (Hydrodynamic Injection)

* **原理：** 通过在毛细管一端施加正压或负压，利用压力差来驱动样品进入毛细管。例如，将毛细管
```

的一端浸入样品中，然后提升这一端的高度，利用重力或外部压力将样品引入毛细管。
* **特点：** 这种方法的上样量更为一致，但可能不适用于极易受压力影响的样品。

3．真空注入

* **原理：** 在毛细管的出口端施加真空，吸引样品进入毛细管。
* **特点：** 类似于水力注入，但使用真空代替正压。

4．微注射器注入

* **原理：** 使用微注射器直接将样品注入毛细管。
* **特点：** 这种方法可以精确控制上样量，但操作相对复杂，且可能对毛细管造成损伤。

高通量大规模并行测序

之前的限制在于每个样本做一个反应，占用一个泳道。

SBS

边合成边测序。依赖于DNA合成的末端可逆的荧光修饰。 每次合成，都上一个被修饰的碱基，于是一轮反应只能上一个碱基，然后把荧光修饰去掉，进行下一轮反应。每次上一个碱基会发出荧光信号，读取荧光信号就知道每一轮上了什么碱基。

illumina测序的基本步骤：

1. **DNA片段的准备：** 首先，从样本中提取的DNA被切割成较短的片段，并加上两端的接头序列。
2. **固定到流动池：** 这些处理过的DNA片段随后被固定到覆盖有寡核苷酸的流动池（flow cell）表面。
3. **桥接扩增：** 在PCR过程中，固定的DNA片段被热解离成单链，然后与流动池表面的另一组互补寡核苷酸杂交，形成一个桥形结构。DNA聚合酶扩增这些单链，生成双链DNA。
4. **循环重复：** 通过反复的热解离和扩增步骤，每个单一的DNA片段在流动池表面形成成千上万的密集克隆簇。（这些cluster在flow cell的表面在物理上是聚集的，把flow cell想象成一块板子，一个cluster就是上面的一个小圆点。）
5. **线性化和单链化：** 一旦扩增完成，双链DNA片段被线性化。
6. **洗脱互补链：** 然后，采用化学或酶学方法去除或洗脱非模板的互补链。这可以通过多种方式实现，例如使用碱性溶液来破坏氢键，使双链DNA解离成单链。
7. **固定模板链：** 洗脱后，保留在流动池表面的是单链DNA模板，准备进行后续的测序反应。

接下来进行测序

1. 测序引物的杂交

- **操作：** 专门设计的测序引物与流动池表面上每个克隆簇的单链DNA模板进行杂交。

2. 测序反应

- **操作：** 在测序反应过程中，加入四种标记了不同荧光标签的脱氧核苷酸（dNTPs），每种dNTP都带有可被逆转录酶识别的可切除修饰。
- **同步化：** 通过使用特殊的化学修饰，确保每个扩增循环中只加入一个碱基。每次加入一个碱基后，流动池被照相，以记录荧光信号。

3. 图像采集和信号分析

- **操作：** 每次加入dNTP后，流动池表面被照相，荧光信号被记录。不同颜色的荧光对应于不同的碱基（A、T、C、G）。
- **去除荧光标记：** 加入的荧光标记和阻止剂随后被化学去除，使得下一个dNTP可以加入。

4. 迭代过程

- **操作：** 这个过程（加入dNTP、记录信号、去除荧光标记和阻止剂）在每个扩增循环中重复进行，直到达到所需的测序长度。

5. 数据处理和序列组装

- **操作：** 测序完成后，荧光图像被转换成序列数据（即原始的碱基序列）。这些数据随后通过生物信息学软件进行处理和分析，以组装完整的序列。

Reads：最终，每一个cluster生成的序列就是一个“Read”。

不过一个cluster不一定读完，而是只读一部分，大约150bp，因此也有双端测序技术。

华大的SBS测序技术

华大DNBSEQ测序 讲解DNBseq

1. DNA单链环化

- **过程：** 将带有接头序列的双链DNA（dsDNA）高温变性成单链DNA（ssDNA）。使用环化引物与ssDNA的两端进行互补配对，并在连接酶的催化下将ssDNA的首尾相连接，形成单链环状DNA（sscDNA）。

2. DNB制备

- **滚环扩增：** 以单链环状DNA为模板，在DNA聚合酶作用下进行滚环扩增（RCA），将单链环状DNA扩增成100-1000拷贝，形成的扩增产物称为DNA纳米球（DNB）。
- **优点：** 这种扩增技术有效地避免了PCR扩增过程中错误积累的问题，提高了测序的准确性。

3. 规则阵列载片

- **制备：** 利用半导体精密加工工艺，在载片表面形成结合位点阵列，实现DNB的规则排列吸附。确保活性位点间距整齐一致，每个位点只固定一个DNB。

4. DNB加载

- **过程：** DNB在酸性条件下带负电，通过正负电荷相互作用，被加载到载片上带正电荷的活化位点。尽量避免多个DNB结合到同一个位点，提高DNB的有效利用率。

5. 测序过程

- **引物杂交和荧光探针聚合：** 在DNA聚合酶催化下，将测序引物锚定分子和荧光探针在DNB上进行聚合。
- **荧光信号采集：** 洗脱掉未结合的探针后，激发荧光信号并通过高分辨率成像系统采集、读取和识别，从而获取待测碱基的序列信息。
- **信号处理：** 加入再生洗脱试剂，去除荧光基团，进入下一个循环的检测。

6. DNA二链合成和双端测序

- **操作：** 在完成一链测序后，加入具有链置换功能的DNA聚合酶进行DNA二链合成反应。形成大量单链DNA作为二链测序的模板。
- **优点：** 二链拷贝数更多，获得更强的荧光信号，提高测序准确性。

7. 数据处理和算法优化

- **图像处理：** 利用Sub-pixel Registration算法，实现亚像素级别的图像配准精度，提高碱基识别的准确度。
- **GPU加速：** 实现高精度算法下的快速数据处理，实现实时化的图像处理和碱基识别。

8. 质量得分和识别准确率

- **质量评估：** 利用训练好的数据模型，根据每个碱基的信号特征输出预估的错误率，采用phred-33质量得分标准进行质量评分。
- **提高准确率：** 通过这些技术和方法的优化，华大智造的DNBseq技术能够在保持高通量的同时，大大提高测序的准确性和效率。

为了更清楚地阐述双端测序的过程，假设我们已经从生物体中提取了一段待测的DNA序列，我们将其标记为A。在使用测序平台前，我们在A的两端分别添加了接头序列，分别标记为X和Y。因此，经过接头添加后的直链DNA序列可以表示为XAY。接下来，通过环化处理，X和Y相连形成一个环状结构。

由于X和Y是人工添加的，我们可以选用与X、Y互补的特定引物来进行DNA复制。假设沿着YAX方向进行滚环扩增（Rolling Circle Amplification, RCA），扩增后的产物将会是一系列重复的Y'A'X'序列，形成的DNA纳米球（DNB）包含多个Y'A'X' Y'A'X' Y'A'X' Y'...的重复序列。

首先进行第一链测序。测序引物与DNB上所有Y'互补结合，随后引入带荧光标记的ddNTP以及DNA聚合酶进行测序。第一链测序通常只覆盖大约50个碱基对，即大约从A'的起点出发合成了约50bp的新链，这条在测序过程中合成的新链被称为正链。由于这一过程并没有完全覆盖整个A'序列，所以A'并未被完全测序。

接下来，我们将停止使用荧光标记的ddNTP，转而添加普通的dNTP，并使用具有链置换功能的DNA聚合酶。这时，每个A'对应的未完成的正链将继续延伸。当第一个A'的正链延伸到下一个A'区域时，由于该区域已经有一条链存在，具有链置换功能的DNA聚合酶会置换掉原有链，以使新的正链能继续延伸。被置换的链原本是第二个A'的正链，现在则变成了一端未互补的游离链。当所有A'对应的正链都经历了类似的置换过程后，我们便得到了一系列的游离链，即第二链。

最后，我们加入第二链的测序引物。这些测序引物将特异性地结合到所有游离端的特定接头上，并开始进行第二链测序。第二链测序实际上是从每个A'的末端开始，反向进行的测序过程。

华大公司不同的机器对应不同的发光选择，有的机器是不同的碱基发不同的光，有的机器是4个碱基用二进制编码发光。

Barcode测序

- **定义：** Barcode测序，也称为多重索引测序或带标签的测序，是一种方法，其中不同样本的DNA片段被添加上独特的序列标签（即“条形码”或“索引”）。这些条形码使得多个样本可以在同一测序反应中被混合和测序，而后可以根据这些独特的序列

标签来区分和重建原始样本。（序列全都变成电子文本了，在字符串里查特定的barcode就知道这条链是哪个样本。）
- **应用：** Barcode测序广泛应用于需要同时处理大量样本的研究，如群体遗传学、生态遗传学、癌症研究和微生物组研究。
- **优势：** 这种方法显著提高了测序的通量和成本效率，因为它允许在单次测序运行中同时分析多个样本。

Illumina测序仪技术与DNBseq测序仪技术有哪些不同？

1. 测序技术基础

- **Illumina测序：** 基于桥式PCR（Bridge PCR）和测序-终止反应（Sequencing by Synthesis, SBS）。在固定表面上进行桥式PCR以形成簇（clusters），然后使用荧光标记的dNTPs进行测序。
- **DNBseq测序：** 使用滚环扩增（Rolling Circle Amplification, RCA）技术生成DNA纳米球（DNBs），每个DNB包含数百至数千个重复的DNA序列。然后通过一系列特定步骤进行测序。

2. DNA扩增方法

- **Illumina：** 使用桥式PCR在流动池表面上形成高密度的簇，每个簇包含许多相同的DNA序列的拷贝。
- **DNBseq：** 采用RCA技术在体外生成DNBs，每个DNB包含许多重复的DNA序列的拷贝。

3. 测序过程

- **Illumina：** 测序过程包括逐步加入荧光标记的dNTPs，然后进行成像和信号分析，以确定每个位置的碱基。
- **DNBseq：** 同样利用逐步加入荧光标记的dNTPs的方法，但在特定步骤（如二链生成）中使用具有链置换功能的DNA聚合酶。

4. 测序准确性和效率

- **Illumina：** 提供高准确性和高重复性的测序结果，适用于各种规模的基因组测序。
- **DNBseq：** 通过RCA避免了PCR扩增过程中的错误累积，提高了测序准确性。使用链置换聚合酶进行二链测序，提高了测序效率。

基因组信息学

插入片段（Insert） - 用于测序的DNA部分。 读段（Read） - 被测序的插入片段的其中一部分。 单端测序（Single End） - 一种测序程序，其中插入片段只从一端进行测序。 双端测序（Paired End） - 一种测序程序，其中插入片段从两端进行测序。

从两端测序得到两个读段，中间可能没有完全被读到，也就是insert有一段 inner distance 没被读到。

正链，一般表示合成方向，也有其它意思，例如DNA聚合时正链是正在延伸的链，负链是模板链。 碱基质量，表示这个碱基测序的准确度。碱基质量为10表示准确度1个9，90%，50表示准确度5个9，99.999%准确。

四、DNA测序的更高层、DNA组装

对于任何一个未知的基因组，首先需要进行基因组概貌评估。

基因组概貌评估

在进行完整的基因组测序和注释之前，了解基因组的大致特征（如大小和复杂度）对于规划测序项目和数据分析策略非常重要。这有助于确定所需的测序深度和所采用的组装算法。 评估内容：

- 基因组的大小
- 复杂度、重复序列和 GC 含量
- 测序深度的预测

确定基因组的大小，例如，可以直接称重。流式细胞仪等。

流式细胞仪方法：

1. **样本准备**：

- **细胞固定**：首先需要收集待测样本的细胞，并通过固定剂（如甲醛）进行固定，以保持细胞结构的完整性。
- **染色质释放**：使用适当的方法（如机械破碎或化学处理）破坏细胞膜，释放染色质。

2. **DNA染色**：

- 将固定的细胞用含有DNA特异性荧光染料的染色液处理。常用的染料包括丙啶碘（PI）或荧光素等。
- 这些染料可以结合到DNA的双螺旋结构中，荧光强度与DNA含量成正比。

3. 流式细胞仪分析：

- 将染色后的样本通过流式细胞仪。流式细胞仪可以测量通过激光束时发出的荧光强度。
- 每个细胞的荧光强度被记录下来，这反映了每个细胞的DNA含量。

4. 使用标准样本校准：

- 同时，需要使用一个或多个已知基因组大小的标准样本（如某种标准的植物或动物细胞）进行校准。
- 通过比较待测样本和标准样本的荧光强度，可以估计出待测样本的基因组大小。

5. 数据分析：

- 收集的数据通常以荧光强度的分布图（如直方图）呈现。
- 通过分析这些直方图，可以估算出样本的平均DNA含量，并进一步推算出基因组大小。

6. 结果解释：

- 结果通常表示为相对于标准样本的基因组大小倍数。
- 这个估计值可以用于比较不同物种或不同条件下的样本。

基因组评估方法

基因组速览

genome survey是指对未知的基因组DNA进行低深度WGS测序。然后进行估算大小，复杂度等。

K-mer分析

1. K-mer计数

- **提取K-mer**：首先，从测序数据中提取所有可能的K-mer。这意味着将每个读段（read）分解为长度为K的所有可能的连续子序列。
- **计数频率**：计算每个独特K-mer在所有reads中出现的次数，以建立一个K-mer频率分布图。

2. 构建K-mer频率分布图

- 在这个分布图中，横轴表示每个K-mer的出现次数（频率），纵轴表示具有该频率的K-mer的数量。
- 对于一个理想的基因组，K-mer分布应呈现为一个清晰的波峰，即大多数K-mer有着相似的出现频率。

3. 估算基因组大小

- **计算总K-mer数**：计算所有K-mer的总数m，不是独特的。
- **确定峰值**：在K-mer频率分布图中，找到代表基因组平均覆盖深度的峰值。
- **计算基因组大小**： 假设基因组平均覆盖深度为d，一共有m个K-mer（每个reads，设长为l，会提供l-k+1个），那么基因组大小约为m/d，基因组越大，k-mer越小，该公式越准。

基因组测序

由于之前已经发展了最底层的DNA测序，因此可以考虑更高层次的基因组如何测序。显然最基础的想法是讲基因组分解成为DNA测序能够使用的长度。

clone-by-clone shotgun

构建重叠clone群

BAC Contig Map

BAC（细菌人工染色体）是一种用于复制大段DNA（一般是100-300千碱基对）的载体。BAC contig map是指通过将大量BAC克隆排列成顺序的方式来创建一个覆盖整个基因组的连续DNA片段（contigs）的地图。

1. **BAC克隆**：通过将较大片段的DNA插入BAC载体中并转入细菌中进行复制，从而制备BAC克隆。每个BAC克隆含有基因组的一个独特片段。
2. **建立Contigs**：通过比较不同BAC克隆中DNA片段的重叠区域，可以将这些BAC克隆按照它们在基因组中的位置排列起来，形成较长的连续DNA序列，即contigs。

3. **创建BAC Map**：继续这个过程，可以将整个基因组的BAC克隆按顺序排列，从而创建出一个BAC contig map，它代表了整个基因组的大致框架。

克隆指纹图谱（Clone Fingerprinting） 克隆指纹图谱是一种用于分析和比较DNA克隆（如BAC克隆）中的DNA片段的技术。它通过产生每个克隆独特的指纹图谱来识别和区分不同的DNA克隆。这些指纹图谱基于DNA片段的大小、数量和组成。

实现过程

1. **DNA切割**：使用限制性内切酶对BAC克隆中的DNA进行切割，产生一系列不同大小的DNA片段。
2. **电泳分离**：通过凝胶电泳对这些片段进行分离。电泳的结果是基于片段大小的一系列条带，这些条带的模式对于每个克隆是独特的。
3. **图谱生成**：记录每个克隆的条带模式，即它的“指纹”。

使用指纹图谱建立Contigs

1. **比较指纹**：通过比较不同克隆的指纹图谱，可以识别出它们之间的重叠区域。具有相似指纹模式的克隆可能具有重叠的DNA序列。
2. **建立重叠关系**：通过识别这些重叠区域，可以将克隆按顺序排列，形成连续的DNA片段，即contigs。
3. **创建地图**：通过这种方式，可以构建一个覆盖整个基因组的BAC contig map，即使每个克隆没有被完整地测序。

Clone selection

既然已经有了contig map，我就可以从中挑选数目最少的能覆盖整个基因组的一组BAC clone。（让人回忆起有限覆盖定理）

BAC克隆shotgun测序

交付DNA测序即可

shotgun测序结果组装

计算机方法进行组装，获得的BAC序列的一部分。（因为shotgun测序不一定全测全了）

BAC克隆组装成基因组草图

根据contig map以及其它算法进行组装。

基因组组装算法

后缀前缀相同，那有可能重叠。 重叠越长信息越长。 短重复会导致组装困难。（该问题可以通过加长reads长度解决）

三代测序技术

第三代测序可以测的reads非常长，可以实现单分子测序。

PacBio测序

PacBio测序，也称为单分子实时测序（SMRT），使用了一种独特的实时测序方法。

原理

1. **SMRT Cell**：使用一种被称为SMRT Cell的装置，其中包含数以百万计的微小井点，每个井点能够容纳一个单分子的DNA聚合酶。
2. **DNA模板准备**：准备闭环的DNA模板（通常是通过连接测序接头形成的闭环结构），并将它定位在井点中的DNA聚合酶上。（要测序的DNA链被制备成了闭环的DNA模板。）
3. 测序：

- 当DNA聚合酶开始复制闭环DNA模板时，它会逐个加入互补的核苷酸。
- 每种核苷酸被设计为在加入时释放荧光信号。这些信号在特定的波长下发光，并通过下方的探测器捕捉。
- 荧光信号随着DNA链的合成而实时捕捉，每个荧光标记的颜色代表了不同的核苷酸（A、T、C、G）。

1. 优势：

- PacBio平台能够产生非常长的读段（可达数十千到数百千碱基），使其在基因组组装、揭示结构变异等方面非常有用。
- 它还能够直接检测DNA的甲基化状态。
- 循环测序：由于DNA模板是闭环的，DNA聚合酶可以在模板上多次循环，重复地进行读取。这种循环读取可以提高测序的准确性。

Nanopore测序

Nanopore测序技术由Oxford Nanopore Technologies开发，它使用纳米孔道直接读取单分子DNA或RNA序列。

原理

1. **纳米孔**：使用含有纳米级孔道的膜。每个孔道都有一个电导探测器。
2. **DNA通过孔道**：在电压作用下，单链DNA或RNA分子通过这些孔道。当分子通过孔道时，会改变孔道的电流。
3. **电流变化检测**：探测器记录通过孔道的DNA或RNA分子引起的电流变化。每种核苷酸（A、T、C、G）通过孔道时产生的电流变化是独特的。
4. **数据解析**：通过分析这些电流变化的模式，可以确定通过孔道的核苷酸序列。
5. **优势**：

- Nanopore技术能够产生极长的读段，理论上没有上限，常见的读段长度为数千到数十万碱基。
- 它同样可以直接从RNA样本进行测序，无需转录成cDNA。
- 这项技术还可以用于检测DNA和RNA的修饰，如甲基化。

Irys系统的工作原理

1. **DNA样本准备**：

- 从细胞中提取高分子量的DNA，并在特殊的条件下保持其为超长的单分子状态。

2. **利用内切酶处理**：

- 使用特定的内切酶对DNA进行处理。这些酶可以识别DNA上特定的序列并在这些位置进行切割。

3. **DNA标记和修复**：

- 在切割点附近引入荧光标记。这通常是通过在酶切后的DNA上添加带有荧光标记的短DNA序列来实现的。
- 修复切割的DNA，使其重新变成连续的长链，但这次含有荧光标记的位置。

4. **线性化DNA分子**：

- 将标记的DNA分子线性化，以便于分析。这通常是通过在特制的表面上展开或拉直DNA分子来实现的。

5. **荧光成像扫描**：

- 通过高分辨率的显微镜系统扫描线性化的DNA分子。
- 荧光标记在成像过程中被激发，并产生荧光信号。

6. **图像分析和基因组映射**：

- 通过分析荧光标记的模式，可以确定DNA分子上特定位置的序列特征。
- 这些数据被用来构建基因组的物理地图，尤其是关于大规模结构变异的信息。

组装质量评估

一个组装，长contigs越多越好。

N50的定义

- N50值是指，在所有contigs或scaffolds的长度从长到短排列时，长度累积到达或超过基因组总长度一半时的最小contig或scaffold的长度。
- 换句话说，当把所有contigs或scaffolds按长度排序并累积它们的长度，N50是长度累积到总基因组长度一半时的那个contig或scaffold的长度。

假设一个基因组组装项目产生了五个contigs，长度分别为8kb, 7kb, 6kb, 5kb, 4kb。基因组总长度为30kb。按照长度排序后，累积长度达到15kb（一半的总长度）是在第三个contig（6kb）。因此，这个组装的N50值是6kb。

总结

如何对一个未知物种的基因组测序？

1. 基因组概貌评估 例如利用流式细胞仪进行基因组大小估计或者K-mer分析估计基因组大小
2. 基因组测序 根据基因组大小等特定决定测序策略，例如使用Clone-by-clone shotgun。首先建立多个BAC文库。构建重叠clone群，然后进行Clone selection，从contig map中挑选能覆盖整个基因组的最少BAC clone，然后BAC克隆shotgun测序，可以使用illumina测序手段或者使用第三代测序。最后将Shotgun测序结果组装，BAC克隆组装成基因组草图。
3. 进行组装质量评估 BAC克隆建立了BAC contigs，根据这些contigs进行N50评估。

五、基因分析

获得了序列之后呢？开始解码吧！

基因组注释

从序列到生物功能。三个级别

1. 核酸级别：基因在哪，重复在哪，变异在哪
2. 蛋白级别：这个序列的蛋白有什么功能？结构域什么样子？
3. 过程级别：它怎么表达的，怎么相互作用的。

基因组注释(Genome annotation)是利用生物信息学方法和工具，对基因组所有基因的生物学功能进行高通量注释，是当前功能基因组学研究的一个热点。

- 结构注释(Structural annotation) : 基因位置及其结构等
- 功能注释(Functional annotation) : 基因功能及其调控等 目的：识别基因组序列中存在的基因和其他多种功能元件(包括编码基因、非编码RNA、转座子等重复序列、调控元件等)，并推测其生物学功能(如 ENCODE)。 意义：基因组注释是生物学研究的基础，一个基因组的价值取决于该基因组注释的质量，基因组注释建立了从未知功能的基因组序列到该物种生物学研究的桥梁

注释方法有：计算预测，结合实验数据的方法

计算预测

基因预测

基因预测的方法，湿实验手段，直接做实验看看效果。 干实验手段。通过计算机对DNA序列进行特征查找。又可分为同源搜索，结构特征搜索。

结构特征搜索

通过结构特征，比如六框翻译法直接预测基因。

Kozak序列是一种在真核生物中发现的特定核苷酸序列，它位于信使RNA（mRNA）上，紧邻起始密码子（通常是AUG）。这个序列对于识别起始密码子和开始蛋白质的合成至关重要。它由Marilyn Kozak发现，因此以她的名字命名。

Kozak序列的典型模式是 GCC(A/G)CCAUGG，其中AUG是蛋白质合成的起始密码子，而其前后的核苷酸则起到辅助作用，增强核糖体的结合效率和识别准确性。这个序列不是绝对固定的，但通常至少包含起始密码子的前三个和后一个位置的特定核苷酸。kozak序列规则描述了起始密码子前的碱基偏好，因此可以用来预测起始密码子。（终止密码子则遇到了就终止，因此不太需要预测）

同源基因搜索

可以看直系同源基因，例如如果一个序列在小鼠中发现了相似的序列，并且已知小鼠这个序列的基因是肌红蛋白什么的，就知道这个序列也应该类似肌红蛋白。

同源性(Homology):进化过程中源于同一祖先的不同分支，用来描述物种之间的进化关系，所以在同源性的表达 中只能用“ 有” 或者“ 无” ，属于定性描述。 相似性(Similarity):指所序列之间相似位点占整个序列的比例，属于定量描述。 一致性(Identity):是对序列间相同位点占整个序列的比例，相对精确度更高的一个描述，属于定量描述。

非编码RNA预测

例如tRNA的预测 由于tRNA的结构非常特殊，可以直接从碱基序列上预测这里是否会是tRNA，例如是不是形成茎环结构等。

共变异模型预测

一个tRNA的茎环结构中相配对的茎总是倾向于一起变异。因此有时光看序列看不出什么同源性，但是如果加上看序列是否形成类似的茎环结构，则可以推断同源性。

重复序列注释方法

与已知的比对再进行分类，或者从头预测。

为什么要预测重复序列？ 测序困难：由于重复序列的高度相似性，它们在DNA测序中容易导致读序错误和组装问题，尤其是在使用短读段的测序技术时。 数据解释：重复区域的存在使得从测序数据中准确重建基因组结构变得更加困难。这些区域往往是基因组中重排和变异的热点。 组装算法的复杂度：在基因组组装过程中，重复序列的存在大大增加了算法的复杂度和计算需求。

隐马模型

建立足够的状态也可以预测donor，acceptor等。

结合实验数据的方法

small RNA-seq 等。"Integrative Gene Isoform Assembler"（IGIA）是一个用于重建和量化转录本异构体（即不同的基因表达形式或亚型）的计算方法。

基因功能分析

普遍使用BLAST方法对预测出来的基因进行功能注释。预测蛋白结构域也可以用同源搜索的方法，在数据库如Interpro，CDD中找答案。

GO（基因本体）

Ontology 是特定领域信息组织的一种形式，是领域知识规范的抽象和描述，是表达、共享、重用知识的方法。 基因本体(Gene Ontology，简称GO)是—种系统地对物种基因及其产物属性进行注释的方法和过程。基因本体知识库是世界上最大的基因功 能信息资源。这些知识既是人类可读的，又是机器可读的，是生物医学研究中大规模分子生物学和遗传学实验的计算分析的基础。

Term（术语）

- **定义：** GO术语是GO体系中的一个基本单元，每个术语代表一个特定的生物学属性。
- **举例：** 一个术语可以是“细胞凋亡”（cell apoptosis），它描述了一个特定的生物学过程。
- **属性：** 每个术语都有一个唯一的ID、名称、定义、以及与其他术语的关系。

Ontology（本体）

- **定义：** 在GO中，ontology指的是一组有组织的术语集合，用于描述基因和蛋白质的特定方面。GO有三个主要的ontology：生物过程（Biological Process），分子功能（Molecular Function），和细胞组分（Cellular Component）。
- **生物过程：** 描述了生物学活动的过程，如“光合作用”或“细胞分裂”。
- **分子功能：** 指的是基因产品（如蛋白质）的分子级活动，如“酶活性”或“信号受体”。
- **细胞组分：** 涉及基因产品在细胞内的特定位置或结构，如“线粒体”或“细胞膜”。
- **目的：** 通过将这些术语系统地组织起来，GO本体能够提供一个框架，用以描述基因产物的功能和特性。

Entry（条目）

- **定义**：在GO的上下文中，entry通常指的是与特定GO术语相关联的具体的基因或蛋白质。
- **功能**：每个entry通常会被注释为具有特定的生物过程、分子功能和/或细胞组分，这些注释基于实验结果或其他类型的证据。
- **重要性**：这种注释使得研究人员能够在大规模基因组分析中理解特定基因或蛋白质的功能。

Entry

- **cytochrome（细胞色素）**：这里的"cytochrome"是一个基因或蛋白质的例子，它在GO系统中被注释或分类。在GO的上下文中，它是一个entry，因为它是一个具体的生物学实体，被描述或分类为具有特定的生物学功能、过程和组件。

Term

- **Oxidoreductase activity（氧化还原酶活性）**：这是一个特定的GO term，它属于"Molecular Function" ontology。
- **Oxidative phosphorylation（氧化磷酸化）**：这也是一个GO term，属于"Biological Process" ontology。
- **Mitochondrial inner membrane（线粒体内膜）**：这个term属于"Cellular Component" ontology。

代谢通路

KEGG是代谢通路的权威数据库 reference pathway:根据已有的知识绘制的的具有一般参考意义的代谢图。通路图中的小框都是白色，在KEGG中名字以map开头，比map00010。 species-specific pathway:物种特有代谢通路图。绿色小框为该物种特有的基因或酶。名字为特定物种种属英文缩写，比如人的糖酵解通路图hsa00010。 以ko/ec/rn开头的Reference pathway:ko通路中的节点只代表基因；ec通路中 的节点只代表相关的酶；rn通路中的节点只表示该点参与的某个反应、反应物及反应类型。底色以蓝色表示。

基因功能富集分析

基因功能富集分析可以从成千上万的基因中筛选出生物学上重要的信号，使研究人员能够集中关注那些最有可能影响研究现象的基因和途径。

1. **GO富集分析**：

- **GO的基础：** GO是一个大型的生物信息学数据库，它为基因和蛋白质提供了一套统一的分类体系，用于描述它们的生物过程（Biological Process），分子功能（Molecular Function）和细胞组分（Cellular Component）。
- **富集分析的目的：** GO富集分析的目的是识别在特定生物条件下（例如，疾病状态、不同发育阶段、治疗响应等）显著表达或调控的基因集合中，GO术语是否比随机预期更频繁地出现。
- **分析过程：** 这通常涉及统计测试（如卡方检验或Fisher精确检验），比较特定基因集与整个基因组的GO术语分布，从而识别特定的生物过程、分子功能或细胞组分在该基因集中是否过度表示（即“富集”）。

2. **KEGG富集分析：**

- **KEGG的基础：** KEGG是一个数据库资源，提供了关于基因和蛋白质功能以及它们在细胞内网络中的相互作用的信息。它包括代谢途径、细胞信号通路、疾病机制等数据。
- **富集分析的目的：** KEGG富集分析用于确定在特定基因列表中，是否有特定的代谢途径或信号通路比随机预期更为常见。
- **分析过程：** 与GO富集分析类似，这通常涉及统计方法来比较特定基因集与整个基因组在KEGG途径中的分布，从而鉴定出显著富集的途径。

六、RNA

RNA生成与加工

加工部分分为

1. 5' 加帽
2. RNA剪接
3. 3' polyA化
4. RNA修饰
5. mRNA降解

基因表达调控：指使用一系列机制来增加或减少基因产物。 转录调控 (Transcriptional regulation) ：指通过改变转录效率从而调控 RNA 转录本表达水平的过程。转录调控可以控制基因的时空动态表达。原核与真核生物具有不同的转录调控机制。 顺式作用元件 (cis-regulatory elements ， CRE) ：位于基因旁侧序列中能影响基因表达的 DNA. 序列，可影响基因转录活性。 反式作用元件 (trans-regulatory elements ， TRE) ：转录模

板上游基因编码的一类蛋白调节因子，又称转录因子 (Transcription Factors) ，通过与特异的顺式作用元件相互作用反式激活基因转录，分为两类：1.通用转录因子：所有mRNA 转录启动共有。 2.特异转录因子：个别基因转录所必需，决定基因的时空特异性表达。

转录组

转录组：在某一特定条件下，所能转录出的所有 RNA 的总和，包括信使 RNA (mRNA) 、核糖体 RNA (rRNA) 、转运RNA (tRNA) 及非编码 RNA ；狭义上指细胞所能转录出的所有信使 RNA (mRNA) 。 转录组可以被视为蛋白质组的前体，即由基因组表达的整组蛋白质。 转录组学：研究在单个细胞，或特定类型细胞、组织、器官或发育阶段的细胞群内所产生的各类 RNA 分子的类型和数量。

转录组与相对稳定的基因组相比，是高度动态和复杂的。

转录组反映了在任何给定时间正在积极表达的基因——基因表达。 理解转录组对于解释基因组的功能元件、揭示细胞和组织的分子成分，以及理解发育和疾病至关重要。 转录组学的主要目标包括：

- 目录所有的转录物，包括信使RNA、非编码RNA和小RNA；
- 确定基因的转录结构，包括它们的起始位点、5'和3'端、剪接模式和其他转录后修饰；
- 在发育过程中以及在不同条件下，量化每个转录物表达水平的变化。

RNA-seq一般步骤

1. **RNA的片段化（Fragmentation of RNA）：**

- 这一步骤的目的是将长的RNA分子断裂成较短的片段。这对于后续的cDNA合成和测序是必要的，尤其是在使用短读长测序技术时。片段化可以通过物理（如超声波）或化学方法（如加热或用酶处理）实现。

2. **cDNA合成（cDNA Synthesis）：**

- 这个步骤涉及使用逆转录酶将RNA转录为互补的DNA（cDNA）。这是必要的，因为大多数测序平台是针对DNA而非RNA设计的。通常会使用随机引物来初始化逆转录，以确保整个RNA片段都能被转录。

3. **接头连接（Adaptor Ligation）：**

- 在cDNA合成后，需要将测序接头（adaptors）连接到cDNA的两端。这些接头对于后续的PCR扩增和测序仪的识别是必需的，它们包含必要的引物结合位点和识别序列。

4. **PCR扩增（PCR Amplification）：**

- 此步骤使用聚合酶链反应（PCR）放大cDNA，以获得足够的材料进行测序。PCR扩增也有助于引入额外的序列元素，如索引或条形码，用于样本的识别和多重测序。

5. 大小选择（**Size Selection**）：

- 在PCR扩增之后，通常会进行大小选择，以去除过短或过长的片段，确保测序库的一致性。这一步可以使用凝胶电泳或其他方法来实现。

6. 测序（**Sequencing**）：

- 最后一步是实际的测序过程。根据使用的技术（如Illumina、PacBio或Oxford Nanopore），测序可以生成短读长或长读长的序列数据。这些数据随后通过生物信息学分析来确定基因表达水平、变异、剪接事件等。

不同测序方法比较

1. 短读长**RNA-seq（Short-read RNA-seq）：**

- **RNA片段化：** 通常需要将RNA分子断裂成较短的片段。
- **cDNA合成：** 使用逆转录酶将RNA转录为cDNA。
- **接头连接：** 将测序接头连接到cDNA的两端。
- **PCR扩增：** 放大cDNA，增加样品量，同时可能引入条形码或索引。
- **大小选择：** 去除不合适长度的片段，保证测序库的质量。
- **测序：** 使用高通量测序技术（如Illumina平台）生成短读长的序列数据。

2. 长读长**RNA-seq（Long-read RNA-seq）：**

- **RNA片段化：** 对于某些长读长平台（如PacBio），片段化可能不是必需的，因为它们能处理较长的RNA分子。
- **cDNA合成：** 长读长测序也需要逆转录步骤，但过程可能略有不同，以适应长读长数据的特点。
- **接头连接：** 与短读长类似，但接头设计可能有所不同。
- **无PCR扩增或特定的PCR扩增：** 某些长读长平台（如Oxford Nanopore）可以直接测序未经PCR扩增的样品。
- **大小选择：** 不是必需的，取决于具体平台和实验设计。

- **测序：** 使用长读长测序技术（如PacBio或Oxford Nanopore）。

3. **直接RNA测序（Direct RNA-seq）：**

- **无需RNA片段化：** 直接RNA测序技术通常不要求RNA片段化。
- **无需cDNA合成：** 这种技术直接对RNA分子进行测序，跳过逆转录步骤。
- **接头连接：** 需要接头连接
- **无需PCR扩增：** 直接RNA测序避免了PCR扩增，以减少潜在的偏见和错误。
- **测序：** 直接使用专门的平台（如Oxford Nanopore）进行RNA分子的测序。

基因表达模式热图 横坐标表示不同的组织样本，例如，第一列取自普通组织，第二列取自癌细胞 纵坐标表示不同的基因，例如，第一行表示A基因，第二行表示B基因。可以从图上直观的看到哪些基因在哪些组织里高表达，或者低表达

GTEx, Genotype-Tissue Expression 一个组织，收集了健康人的基因在各个组织的表达差异以及个体差异。

高通量测序技术与RNA生成和加工

如何知道RNA在体内的生成加工过程呢，可以利用高通量测序技术。

很多基因拥有多个潜在的PolyA位点，这意味着相同的前体RNA（pre-mRNA）可以在不同的位点被剪切和聚腺苷酸化，从而产生具有不同3'端的成熟mRNA。这种现象被称为“替代性聚腺苷酸化”（alternative polyadenylation），是基因表达调控的一个重要方面。

使用PolyA site RNA-seq技术可以分析在不同组织或者不同实验条件下对一个mRNA使用的特定的polyA位点

1. **mRNA稳定性：**

- 不同的polyA位点可能导致3' UTR长度的变化，影响mRNA分解的速率。一般来说，较长的3' UTR可能含有更多的降解信号，使mRNA更不稳定。

2. **转录后调控元件：**

- 3' UTR通常包含多种调控元件，如microRNA结合位点、RNA结合蛋白位点等。不同的polyA位点可能导致这些调控元件的增加或丢失，从而影响mRNA的翻译效率和稳定性。

3. **mRNA的翻译调控：**

- 不同长度和序列的3' UTR可能影响翻译启动效率，从而影响蛋白质的表达水平。

4. 细胞定位：

- 某些3' UTR序列可影响mRNA在细胞内的定位，不同的polyA位点可能导致这些序列的变化，影响蛋白质在细胞内的分布。

5. 可变剪接：

- polyA位点的选择与可变剪接过程密切相关，这可能影响到其他剪接位点的使用，从而改变最终产生的蛋白质的性质。

检测新生RNA（nascent RNAs）的技术

GRO-seq（全核苷体标记法）

GRO-seq（Global Run-On Sequencing）是一种用于研究活跃转录基因的技术。它主要用于测定正在进行的转录活动，尤其是新生RNA。

1. 原理：GRO-seq的核心是核苷体标记试验（nuclear run-on assay）。首先，细胞被裂解，然后使用带标记的核苷酸在原位进行短暂的转录延伸反应，从而标记新生的RNA。
2. 步骤：

- 细胞裂解后，核内的RNA聚合酶被固定在其活跃的转录位点。
- 在体外添加带标记的核苷酸（BrU）进行短暂的转录延伸，从而标记新生RNA。
- 提取RNA，并通过高通量测序分析新生RNA。

mNET-seq（核RNA聚合酶活动测定法）

mNET-seq（Mammalian Native Elongating Transcript sequencing）是一种用于精确测量哺乳动物细胞中RNA聚合酶活性的技术。

1. 原理：这种技术专注于RNA聚合酶活跃的转录延伸复合体，可以捕获和测序正在被转录的RNA。
2. 步骤：

- 从细胞中提取正在延伸的RNA聚合酶复合体。
- 使用一种特殊的方法提取与聚合酶复合体关联的RNA。
- 通过高通量测序来分析这些RNA。

SLAM-seq（标记代谢物组成分析）

SLAM-seq（Thiol(SH)-Linked Alkylation for the Metabolic sequencing of RNA）是一种快速和敏感的转录活动测量技术。

1. **原理**：SLAM-seq基于新合成RNA的代谢标记，通过将核苷酸类似物（如4-thiouridine）纳入新生RNA。
2. **步骤**：

- 细胞在存在标记的核苷酸类似物的条件下培养，细胞在培养基中合成新RNA。
- 使用化学方法选择性地修饰这些标记的RNA。
- 通过高通量测序来分析这些RNA。

Ribosome profiling，也被称为Ribo-seq，是一种革命性的技术，用于定量分析细胞内正在被核糖体翻译的RNA分子。

基本原理

Ribosome profiling的基本原理是利用测序技术来识别和量化核糖体保护的RNA片段（RPFs）。这些片段是在翻译过程中由核糖体遮蔽的mRNA区域，因此不会被核酸酶降解。

主要步骤

1. **细胞裂解**：首先，细胞被裂解，释放出正在进行翻译的核糖体-RNA复合体。
2. **核酸酶处理**：接下来，使用核酸酶部分降解未受核糖体保护的RNA。这一步骤留下了核糖体保护的RNA片段，通常长度约为28-30核苷酸。
3. **核糖体RNA片段的分离和纯化**：保护的RNA片段随后被分离和纯化。通常使用密度梯度离心来实现这一步。
4. **构建测序文库**：纯化的RNA片段被用于构建适用于高通量测序的文库。
5. **高通量测序**：对文库进行测序，生成大量关于核糖体保护片段的数据。
6. **数据分析**：最后，通过生物信息学方法分析测序数据，识别被翻译的RNA，并量化它们的丰度。

一些常识问题

1. **平均每个基因的外显子数目**：

- 在人类基因组中，平均每个基因大约有8-9个外显子。但这个数值在不同物种和不同基因中有很大的变化。

2. **外显子的平均长度**（按5'UTR、CDS或3'UTR划分）：

- 5'UTR（5'非翻译区）的平均长度一般在100-200个核苷酸之间。
- CDS（编码序列）的外显子平均长度在人类中大约为145个核苷酸。
- 3'UTR（3'非翻译区）的平均长度通常更长，可以达到几百到几千个核苷酸。

3. **内含子的平均长度及其最小和最大长度**：

- 在人类基因组中，内含子的平均长度大约是3,363个核苷酸。
- 内含子的最小长度可以短至几十个核苷酸（通常至少20个核苷酸），而最大长度可以达到上百千个核苷酸（如DMD基因中的某些内含子超过200,000个核苷酸）。

4. **不同物种中这些特征的差异**：

- 不同物种之间，这些基因组特征的大小差异很大。例如，在酵母中，外显子数量少，长度较短，而内含子则非常稀少。在哺乳动物中，外显子数量更多，长度也有所增加，内含子则普遍存在且长度较长。
- 植物和某些其他生物（如一些昆虫）的基因组结构也有其特有的特征，如内含子长度和外显子数量可能与人类和其他哺乳动物有所不同。
- 这些差异反映了基因组复杂性和进化历程的不同。

七、基因组映射

如何进行基因组映射，也就是说，你获得了很多reads，你怎么知道这些reads在基因组的哪里？

可以使用种子建立索引。或者使用BWT方法

Burrows-Wheeler Transform (BWT) 在 read mapping 中的应用是基于其构建高效索引的能力。这种方法特别适合于处理大规模基因组数据，如高通量测序（HTS）产生的数据。下面是BWT在read mapping中的应用步骤：

1. **构建BWT索引**：

- 首先，对参考基因组应用BWT，创建一个压缩且易于搜索的基因组表示。
- 这通过对基因组所有可能的循环移位进行排序，然后提取排序后方阵的最后一列来完成。
- 这一过程产生的BWT字符串，连同其他辅助数据结构（如Suffix Array和Rank表），构成了FM-index。

2. **使用FM-index进行快速搜索**：

- FM-index允许我们在BWT变换的基因组中高效地搜索特定的序列（如reads）。

- 它利用了BWT的特性，即重复序列在变换后的字符串中聚集在一起，从而使得搜索更加高效。

3. **Read Mapping**：

- 将测序得到的reads逐一与FM-index进行比对。
- 这个过程通常从read的末尾开始，逐步向前搜索，以在BWT字符串中找到匹配的位置。
- 由于FM-index支持从后向前的搜索，这使得即使在面对大规模数据时也能快速准确地找到reads在参考基因组中的位置。

4. 定位和分析：

- 一旦read在基因组中的位置被确定，就可以进行进一步的分析，如变异检测、等位基因特异性表达分析等。

BWT应用例子 参考基因组序列

假设参考基因组序列为 `AGCTAGC`。

构建BWT索引

1. 构造循环移位（添加一个特殊字符"$"来标记序列的开始和结束）：

- AGCTAGC$
- GCTAGC$A
- CTAGC$AG
- TAGC$AGC
- AGC$AGCT
- GC$AGCTA
- C$AGCTAG
- $AGCTAGC

2. 按字典顺序排序：

- $AGCTAGC
- AGC$AGCT
- AGCTAGC$
- C$AGCTAG
- CTAGC$AG
- GCTAGC$A
- GC$AGCTA

- TAGC$AGC

3. 提取最后一列 （BWT结果）：

- C
- T
- $
- G
- G
- A
- A
- C

4. 构建**FM-index** ： 这个索引包括BWT的结果和辅助数据结构。

Read Mapping

假设我们的read是 `CTAG`。

1. 使用**FM-index**进行搜索 ：

- 从read的最后一个字符 `G` 开始。
- 在BWT结果中找到所有的 `G`。在此例中，它们位于第4和第5行。
- 接着查找 `A`，它应位于与 `G` 相邻的前一列。在此例中，第6和第7行的 `A` 与 `G` 相邻。
- 接下来，查找 `T`，它应位于与 `AG` 相邻的前一列。只有第2行满足条件。
- 最后，查找 `C`，它应位于与 `TAG` 相邻的前一列。在此例中，第1行的 `C` 符合条件。

2. 确定位置 ：

- 通过FM-index，我们确定read `CTAG` 在参考基因组中对应的位置。在此例中，它位于序列的第2到第5位置。

常用的DNA映射工具有bowtie2，minimap2 常用的RNA映射工具有 HISAT2，minimap2

Bowtie2

- 原理 ：Bowtie2 是一种广泛使用的DNA序列比对工具，它基于Burrows-Wheeler Transform (BWT)。它首先构建参考基因组的BWT索引，然后使用这个索引来有效地比对短DNA序列（reads）。
- 特点 ：
- 高效的内存使用，使得大基因组（如人类基因组）的索引可以容纳在普通电脑的内存中。

- 支持模糊匹配和间隔，允许reads中有一定程度的不匹配或插入/缺失。
- 快速，适合大量数据集。

Minimap2

- **原理**：Minimap2 是一个通用的序列比对工具，适用于DNA和RNA序列。它使用一种叫做“minimizer”的技术来索引参考基因组和reads。
- **特点**：
- 支持长reads（如来自PacBio或Oxford Nanopore的reads）和短reads的比对。
- 快速处理大数据集，特别是在长reads的映射上表现出色。
- 支持结构变异的检测和全基因组比对。

HISAT2

- **原理**：HISAT2 是一个高效的RNA-seq reads比对工具，它也是基于BWT和FM-index。不同于Bowtie2，HISAT2 使用了一种分层图FM索引（Hierarchical Graph FM Index），这使得它在处理剪接位点时更加高效。
- **特点**：
- 特别适用于RNA-seq数据，能够有效处理剪接现象。
- 较低的内存占用和快速的比对速度。
- 支持多种剪接位点和变异类型的检测。

Minimap2（针对RNA

对于RNA应用）

- **原理**：在RNA映射方面，Minimap2同样使用了minimizer技术。它对于RNA-seq数据的处理特别优化，能够有效处理剪接事件。
- **特点**：
- 能够处理长reads RNA-seq数据，例如来自Oxford Nanopore技术的数据。
- 支持剪接位点的检测，这对于RNA-seq数据至关重要。
- 快速高效，适合大规模数据集和高通量数据处理。

DNA map 工具

而BLAST则几乎是所有映射工具的前身，使用种子建立索引。

早期read mapping工具算法是基于hash表的，类似于BLAST。读段短 早期算法的问题是 假设任务是找到32bp read的最多2个错配位点。 背包计算（back-of-the-envelope calculation）指出，为了找到最多两个错配位点的11-mer种子，必须进行大量计算。 人

类基因组大约有3亿碱基对（3×10^9），且有双链。 平均来说，每个11-mer在基因组中出现约1432次（2×3×10^9/4^11），导致有太多候选位置，这使得搜索变得缓慢。 因此，要么搜索过程变得缓慢，要么为了提速而牺牲准确性。

第二代短基因组Read映射

1. **Read**长度 ： 25-51bp

- 这代表着与第一代相比，第二代技术可以处理稍长的reads。

2. 代表性工具 ：

- **Bowtie** （Langmead et al, 2009）
- **BWA** （Li & Durbin, 2009）
- **SOAP2** （R. Li et al, 2009）
- 这些工具标志着映射技术的进步，特别是在处理更长的reads上。

3. **Bowtie**和**BWA**的算法 ：

- 关键原理 ： 如果β是α的子串，可以快速（在常数时间O(1)内）测试ββ是否是α的子串，不论ββ在α中出现的频率如何。
- **“魔法”** ： 这种方法的效率不受ββ在α中出现次数的影响。
- 回溯机制 ： 对α进行编辑，使编辑后的α与β完美匹配。
- 这些方法的处理速度比第一代快数十倍。

4. 第二代技术的问题 ：

- 回溯算法在处理错配数量时是指数级的。
- 更长的reads意味着允许更多的错配，从而导致处理速度指数级减慢

第三代短基因组Read映射

1. **Read**长度 ： 76-300bp

- 第三代技术能处理更长的reads，相比前两代显著增长。

2. 代表性工具 ：

- **Bowtie 2** （Langmead & Salzberg, 2012）
- **BWA-MEM** （Li, 2013）
- 这些工具代表了read映射技术的进一步发展。

3. 算法 ：

- **种子与双向BWT**：使用双向Burrows-Wheeler Transform（BWT）来创建映射的种子点。
- **扩展**：使用SSE优化的Smith-Waterman算法进行扩展。
- 有点类似于BLAST算法（与第一代技术相似），但更加高效和适应于长reads。

4. **针对25-51bp Reads的问题**：

- 对于较短的25-51bp reads，这些第三代工具可能会遇到与第一代类似的问题，如在处理速度和准确性方面的平衡。

RNA-seq map 工具

第一代：TopHat 和 TopHat2

- **TopHat (Trapnell et al, 2009) 和 TopHat2 (Kim et al, 2013)**：
- **识别潜在外显子**：这些工具首先识别出RNA-seq数据中可能的外显子区域。
- **映射未映射的reads到潜在剪接位点**：对于那些没有直接映射到已知外显子的reads，TopHat尝试将它们映射到潜在的剪接位点（即外显子之间的连接处）。
- **对于25-51bp reads的必要性**：对于较短的reads（如25-51bp），这种识别剪接位点的方法尤为重要，因为短reads可能不足以跨越整个剪接位点。

第二代：QPALMA, STAR 和 HISAT

- **QPALMA (De Bona et al, 2008), STAR (Dobin et al, 2012) 和 HISAT (Kim et al, 2015)**：
- **使用分割reads识别剪接位点**：这些工具通过识别分割reads（即部分位于一个外显子，部分位于另一个外显子的reads）来直接识别剪接位点。
- **基于全文索引**：它们基于全文索引（如后缀数组、FM-index等）来实现这一功能，这使得对reads的搜索和映射更加高效。
- **对于≥76bp reads更简单快速，但对约30bp reads不适用**：这些工具对处理较长的reads（如76bp及以上）更加有效，因为较长的reads更有可能跨越整个剪接位点。然而，对于大约30bp长度的reads，这些工具可能不如专门为短reads设计的工具

长read map 工具

第一代：BWA-SW

- **BWA-SW** （Li & Durbin, 2010）
- 这是一种早期的长读长映射工具，现已不再常用（deprecated）。
- 主要用于处理比短读长的测序数据，但其性能和准确度随着测序技术的发展已不再是最佳选择。

第二代：

- **BLASR** （Chaisson et al, 2012）和 **BWA-MEM** （Li, 2013）
- 这些工具在某种程度上类似于BLAST算法，用于处理长读长测序数据。
- **BLASR** 特别使用了“chaining”技术，这是一种将相似或相关序列片段连接起来的方法，以提高映射的准确性。
- **BWA-MEM** 是一种广泛使用的工具，适用于中到长读长的测序数据，以其高效的性能和准确度而知名。

2.5代：

- **GraphMap** （Sović et al, 2016）
- 该工具以其对复杂序列和重复区域的高灵敏度而闻名，但在处理现代高质量数据时，其速度可能慢于第二代工具。
- **GraphMap** 特别适用于处理较差质量的数据，但随着测序技术的进步，这种需求减少了。
- **NGMLR** （Sedlazeck et al, 2018）
- 这是一个较新的工具，其性能和准确度与第二代工具相当。
- **NGMLR** 专为处理结构变异和复杂基因组区域而设计，因此在某些应用中可能更为合适。

第三代长读长映射

1. **Read长度** ： 1千碱基对（kb）到250兆碱基对（Mbp）

- 这代表了相比于前两代，第三代技术能处理极其长的reads，这是一项重要的技术进步。

2. 代表性工具 ： **minimap2** （Li, 2018）

- Minimap2是目前处理长reads映射的主流工具之一，以其高效率和灵活性著称。

3. 算法 ： 种子-链-扩展（seed-chain-extend）

- 这是由基因组比对器和BLASR等工具所采用的一种常见方法。

  种子（**Seed**） ：

- 用于找到精确的k-mer匹配点（也称为锚点）。
- 这些锚点作为基本单位帮助识别reads中的潜在映射区域。

链（Chain）：

- 将这些锚点中的一组线性对齐的锚点识别为一条“链”。
- 这一步骤是为了确定一组锚点是否共线，即它们是否按照它们在read和参考序列中的顺序排列。

扩展（Extend）：

- 从链的末端开始填补缺口并向外扩展。
- 这一步是为了完成映射，通过扩展锚点链来填补和精确地映射整个read。

想象一个坐标系，横轴是参考基因组(单位bp)，纵轴是查询序列(bp)。例如说，reference的100-150bp与纵轴的2-52bp exact匹配。那么图上就会显示一条斜率为正的短线段，这就是锚点。如果有很多斜率为正的短线段恰好组成一条直线，即它们的截距都一致，那么这些短线段可以组成长线段，也就是长的匹配。

比对算法

动态规划算法

"生物序列的近似匹配" 是生物信息学中一种用于比对 DNA、RNA 或蛋白质序列的计算技术，它允许这些序列之间存在差异。

"编辑距离"（Edit distance）是计算机科学和生物信息学中常用的一个概念，用于量化两个字符串之间的差异程度。具体来说，它衡量将一个字符串转换为另一个字符串所需的最少编辑操作次数。这些编辑操作通常包括插入、删除和替换字符。

当我们提到“由最优编辑器从左到右引入的编辑”，我们是在设想一个过程，其中编辑器（一个理论上的工具或算法）以最有效的方式逐步修改字符串。在这个过程中，编辑器从字符串的左侧开始，按顺序对字符进行操作，直到将原始字符串（x）转换为目标字符串（y）。

“编辑转录（Edit transcript）”是记录这一过程的一种方式。它概括了编辑器如何将字符串 x 转换为字符串 y 的具体步骤。编辑转录不仅显示了进行了哪些操作（插入、删除、

替换），还显示了这些操作发生的具体位置。这对于理解两个字符串之间的相似性和差异性非常有用，特别是在生物序列分析和文本处理领域。 例如 x:GC G TATGCGGCTA - ACGC TTGCAC y:GC - TATGCGGCTA T ACGC ------ MM D MMMMMMMMMM I MMMM DDDDDD （最优的编辑，指的就是不要直接替换，而是看看能不能插入或删除） 编辑距离为8

编辑距离就是 edit transcript中非匹配步骤（I，D，R）的长度。

D[i,j] 表示x字符串的前i长度的前缀与y字符串前j长度的前缀之间的编辑距离。 D[i,j] 可能写成递推的形式

1. D[i,j] = D[i-1,j-1]，如果x中的第i个字符和y中第j个字符匹配（M），D[i,j] = D[i-1,j-1]+1，如果最好的方式是需要把x中的第i个字符替换成y中第j个字符（R）。也就是对 D[i-1,j-1] 的transcript 加上一个M或R
2. D[i,j] = D[i-1,j]+1，如果x中第i个字符不与y中第j个字符匹配，但是最好的方式是在x字符串中插入y中的第j个字符。也就是对 D[i-1,j]的 transcript 加上一个I
3. D[i,j] = D[i,j-1]+1，如果x中第i个字符不与y中第j个字符匹配，但是最好的方式是在x字符串中删除第i个字符。也就是对 D[i,j-1]的 transcript 加上一个D

令D[i,0] = i，D[0,j] = j。 立刻有 D[i,j] = min{D[i-1,j]+1,D[i,j-1]+1,D[i-1,j-1]+delta(x[i-1],y[j-1])}（这个取min的过程就是最优编辑的过程） delta(a,b)是Kronecker 符号，相等时为0，不等时为1.

利用以上内容可以进行动态规划比对两个序列，很容易用python编程。

```
定义序列 x 和 y
x = "GCGTATGCACGC"
y = "GCTATGCCACGC"

# 初始化动态规划矩阵 D
D = [[0 for j in range(len(y) + 1)] for i in range(len(x) + 1)]

# 填充矩阵的第一行和第一列
for i in range(1, len(x) + 1):
    D[i][0] = i
for j in range(1, len(y) + 1):
    D[0][j] = j

# 动态规划填充其余的矩阵
for i in range(1, len(x) + 1):
    for j in range(1, len(y) + 1):
        delt = 1 if x[i-1] != y[j-1] else 0
        D[i][j] = min(D[i-1][j-1] + delt, D[i-1][j] + 1, D[i][j-1] + 1)

# 返回填充后的动态规划矩阵
D
```

```
[[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
 [1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11],
 [2, 1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
 [3, 2, 1, 1, 2, 3, 3, 4, 5, 6, 7, 8, 9],
 [4, 3, 2, 1, 2, 2, 3, 4, 5, 6, 7, 8, 9],
 [5, 4, 3, 2, 1, 2, 3, 4, 5, 5, 6, 7, 8],
 [6, 5, 4, 3, 2, 1, 2, 3, 4, 5, 6, 7, 8],
 [7, 6, 5, 4, 3, 2, 1, 2, 3, 4, 5, 6, 7],
 [8, 7, 6, 5, 4, 3, 2, 1, 2, 3, 4, 5, 6],
 [9, 8, 7, 6, 5, 4, 3, 2, 2, 2, 3, 4, 5],
 [10, 9, 8, 7, 6, 5, 4, 3, 2, 3, 2, 3, 4],
 [11, 10, 9, 8, 7, 6, 5, 4, 3, 3, 3, 2, 3],
 [12, 11, 10, 9, 8, 7, 6, 5, 4, 4, 3, 3, 2]]
```

填完矩阵后看右下角就知道总的D。然后回溯，寻找哪里做了改动。回溯的方法很简单，到的地方大于等于来时的地方，并且看一眼这个位置上两个元素是否相等。相等则来自左上，否则应该来自-1的地方。 这个算法本质上是把所有的情况看了看，然后找最小的。

但问题是，认为每次修改的花费相同这事合理吗？在人中transition(A-G,C-T)的概率比transversion大，这种修改的花费应该比transversion小。而对应到空的删除要花费很大。

因此把每个对位修改加上权重。transition权重2，transversion权重4，碱基对gap权重8可以修改动态规划的步骤。

```python
import numpy as np

def scoring_matrix(a, b):
    """
    Define the scoring matrix for alignment.
    Transition (A <-> G, C <-> T) has a weight of 2.
    Transversion (other base substitutions) has a weight of 4.
    Gap has a weight of 8.
    """
    if a == '-' or b == '-':
        return 8  # Gap penalty
    elif (a == 'A' and b == 'G') or (a == 'G' and b == 'A') or \
         (a == 'C' and b == 'T') or (a == 'T' and b == 'C'):
        return 2  # Transition penalty
    elif a == b:
        return 0  # No penalty for a match
    else:
        return 4  # Transversion penalty

def globalAlignment(x, y):
    """ Calculate global alignment value of sequences x and y using dynamic
programming with modified weights. """
    D = np.zeros((len(x)+1, len(y)+1), dtype=int)
```

```
    # Initialize the first row and column of the matrix
    for j in range(1, len(y)+1):
        D[0, j] = D[0, j-1] + scoring_matrix('-', y[j-1])
    for i in range(1, len(x)+1):
        D[i, 0] = D[i-1, 0] + scoring_matrix(x[i-1], '-')

    # Fill the rest of the matrix
    for i in range(1, len(x)+1):
        for j in range(1, len(y)+1):
            diag = D[i-1, j-1] + scoring_matrix(x[i-1], y[j-1])  # Diagonal
            vert = D[i-1, j] + scoring_matrix(x[i-1], '-')       # Vertical
            hori = D[i, j-1] + scoring_matrix('-', y[j-1])       # Horizontal
            D[i, j] = min(diag, vert, hori)

    return D, D[len(x), len(y)]

# Define sequences
x = "TACGTCAGC"
y = "TATGTCATGC"

# Calculate global alignment with modified weights
alignment_matrix, global_alignment_value = globalAlignment(x, y)
```

	ε	T	A	T	G	T	C	A	T	G	C
ε	0	8	16	24	32	40	48	56	64	72	80
T	8	0	8	16	24	32	40	48	56	64	72
A	16	8	0	8	16	24	32	40	48	56	64
C	24	16	8	2	10	18	24	32	40	48	56
G	32	24	16	10	2	10	18	26	34	40	48
T	40	32	24	16	10	2	10	18	26	34	42
C	48	40	32	24	18	10	2	10	18	26	34
A	56	48	40	32	26	18	10	2	10	18	26
G	64	56	48	40	32	26	18	10	6	10	18
C	72	64	56	48	40	34	26	18	12	10	10

填矩阵复杂度O(mn)，而回溯为O(m+n)

氨基酸替换的权重矩阵则更为复杂。

局部比对

给定字符串x和y，x的子串与y的子串的最优全局比对值为x和y的局部比对。 局部比对的思路是，相似性给高分，不相似给低分。考虑所有结束为i的x的子串和结束为j的y的子串，之间的相似性。迭代时取最大值。即为Smith-waterman算法。

动态规划总结

是很强大，可是花费时间太久O(mn)

Index-assisted approximate matching

索引辅助的近似匹配

核心思想

1. **使用索引进行精确匹配子问题：** 在大规模序列比对任务中，通过构建索引（例如哈希表或后缀树），可以快速找到短序列（如读序列）在较长序列（如基因组）中的精确匹配位置。
2. **动态规划处理近似匹配：** 在找到精确匹配的基础上，使用动态规划来处理序列周围的近似匹配问题，即在精确匹配周边进行局部序列比对。

步骤

1. **分区 P：** 将序列 P（例如read序列或查询序列）分为多个部分（p1, p2, p3, ... pk），这些部分可以单独进行精确匹配搜索。
2. **索引查找精确匹配：** 使用索引来快速定位这些小片段在较长序列 T（例如基因组或数据库中的序列）中的精确匹配位置。
3. **在精确匹配附近使用动态规划：** 一旦找到精确匹配的位置（称为hit），在其周围的区域内使用动态规划进行局部序列比对。这样可以处理插入、删除和替换等类型的序列变异。
4. **整合结果：** 最后，将这些局部比对的结果整合起来，形成整个序列 P 的比对结果。

在比对矩阵中，当两个序列在某个区域内的字符大量一致时，这些匹配会在矩阵中形成一条对角线。这种对角线延伸的匹配通常表明序列在这一区域有较高程度的相似性。使用索引技术（如哈希表或后缀数组），可以高效地识别这些对角线匹配延伸。

FM index

则使用BWT算法压缩形成索引，索引更快更好。

八、RNA组学

RNA-bingding protein在多个调控通路中有重要作用。

RNA定位

CeFra-seq（Cell Fractionation followed by Sequencing）和APEX-Seq都是用于研究RNA在细胞内分布和局部化的技术。

CeFra-seq的步骤：

1. **细胞分离**：首先将细胞进行分离，通常是通过物理或化学方法将细胞核和细胞质等不同部分分开。
2. **RNA提取**：分别从每个细胞分离出的部分中提取RNA。
3. **RNA测序**：使用高通量测序技术对提取的RNA进行测序，如RNA-seq。
4. **数据分析**：分析测序数据，确定不同RNA在细胞不同部位的丰度和分布。

CeFra-seq通过这些步骤可以揭示RNA在细胞内的具体分布，有助于理解RNA在不同细胞组分中的功能。

APEX-Seq的步骤：

APEX-Seq是一种基于APEX（Ascorbate Peroxidase）生物酶标记的技术，用于在活细胞中精确地定位RNA分布。

1. **构建表达APEX融合蛋白的细胞系**：首先在细胞中引入编码APEX融合蛋白的基因。这种融合蛋白通常是将APEX酶与特定的细胞器定位信号或蛋白质结合，使其特异性地定位于细胞的特定区域。
2. **处理细胞以诱导生物标记**：将含有APEX融合蛋白的细胞处理以产生活性氧种，这些活性氧种能够修饰附近的生物分子，包括RNA。
3. **修饰RNA**：利用APEX酶的过氧化活性，将生物素（一种小分子）特异性地添加到APEX融合蛋白附近的RNA上。
4. **RNA提取和富集**：破坏细胞并提取总RNA，然后使用生物素亲和素纯化技术从中富集被生物素标记的RNA。
5. **RNA测序**：对富集的RNA进行高通量测序。
6. **数据分析**：分析测序数据来识别在细胞特定区域局部化的RNA种类。

RBP-RNA 接触组

Bind-seq是一种用于确定RNA结合蛋白（RBP）的RNA结合特异性的实验方法。这种技术能够帮助研究者了解特定的RBP是如何识别和结合到RNA上的

Bind-seq的步骤：

1. **合成随机RNA库**：首先合成一个包含大量随机序列的RNA库。这些RNA序列是多样化的，可以广泛代表不同的RNA结构和序列。
2. **RNA与RBP的结合**：将这个RNA库与特定的RBP混合，允许它们相互作用。在这一步骤中，RBP会结合到它们偏好的RNA序列上。
3. **分离未结合和已结合的RNA**：使用免疫沉淀或其他方法分离已经与RBP结合的RNA和未结合的RNA。
4. **测序和分析**：对已结合的RNA进行测序，并分析数据以识别RBP的结合序列和模式。通过比较结合和未结合的RNA序列，可以推断出RBP的结合特异性。

RIP-seq是一种结合免疫沉淀技术和深度测序的实验方法，用于研究RNA结合蛋白（RBP）与其目标RNA之间的相互作用。

RIP-seq的步骤：

1. **细胞裂解**：首先从细胞或组织中提取全细胞裂解液，释放出RNA和蛋白质。
2. **免疫沉淀**：使用针对目标RBP的特异性抗体进行免疫沉淀。这个步骤会捕获RBP以及与之结合的RNA。
3. **RNA提取**：从沉淀的复合物中提取RNA。此时提取的RNA是那些与目标RBP直接相互作用的。
4. **RNA测序**：对提取的RNA进行高通量测序（深度测序）。这通常涉及到RNA的转录成cDNA，然后进行测序。
5. **数据分析**：分析测序数据以识别RBP结合的RNA种类和位置。这通常包括比对测序读段到参考基因组，以及识别富集区域

与ChIP类似。

CLIP-seq（Cross-Linking ImmunoPrecipitation coupled with deep sequencing），包括其一个变种HITS-CLIP，是用于研究RNA结合蛋白（RBP）与RNA之间相互作用的先进技术。

CLIP-seq的基本步骤：

1. **紫外线交联**：细胞或组织首先暴露于紫外线下，使得RNA结合蛋白与其结合的RNA分子之间形成共价键。这一步骤固定了RBP与RNA之间的相互作用，确保它们在后续的实验处理中不会被打乱。
2. **细胞裂解和RNA降解**：将细胞裂解，释放RBP和RNA，然后用核酸酶部分降解RNA，留下与RBP直接接触的RNA片段。
3. **免疫沉淀**：使用针对目标RBP的抗体进行免疫沉淀，以分离出RBP-RNA复合物。
4. **RNA提取和制备**：从RBP-RNA复合物中提取RNA，并进行适当的处理，如链接适配器，以备测序使用。
5. **深度测序**：对提取的RNA进行高通量测序。
6. **数据分析**：通过生物信息学工具分析测序数据，鉴定RBP的结合位点。

RNA结构组学

理解RNA的二级结构对于促进RNA功能研究非常重要。RNA的二级结构是指RNA链上核苷酸之间形成的局部折叠和配对，这些结构决定了RNA的三维形状，从而影响其功能和与其他分子的相互作用。

RNA结构预测

首先预测二级结构。它们可能形成茎环，茎环中的茎就像是左括号和右括号。

Vienna格式

Vienna格式是一种广泛使用的RNA二级结构表示格式。它通常包含两行：第一行是RNA的核苷酸序列，第二行是对应的结构表示，使用点和括号来表示配对关系。

- **点（.）** 表示未配对的碱基。
- **括号** （(和)）表示碱基配对。左括号 (表示一个配对的开始，右括号)表示配对的结束。配对的括号数量和位置表示RNA链中相互作用的碱基。

BPSeq格式

BPSeq格式是另一种RNA结构文件格式，通常包含三列数据：

1. **位置**：碱基在RNA序列中的位置（通常是从1开始的整数）。
2. **碱基**：相应位置的核苷酸（如A、U、C、G）。
3. **配对位置**：如果碱基在结构中形成配对，则显示配对碱基的位置；如果未配对，则显示为0或空。

基础预测方法，最小自由能折叠。 最小自由能（Minimal Free-Energy, MFE）折叠是一种常用于预测RNA分子功能构象（即其天然状态）的计算方法。这种方法的基本假设是，RNA分子会折叠成为自由能最低的配置。这个假设基于这样一个观察：在生物体内，RNA分子倾向于采取能量最低、最稳定的三维结构。

也有动态规划算法来完成。 时间复杂度O(n^3)，空间复杂度O(n^2)

概率方法在RNA折叠研究中的应用代表了计算RNA结构的一种范式转变。传统上，RNA结构的预测主要集中在寻找单一的、最小自由能（MFE）的二级结构。然而，实际上，一个RNA分子在细胞内并不总是仅采取单一的最稳定结构，而是可能存在于多种不同的构象中，形成一个结构集合（ensemble）。因此，概率方法被引入以更全面地描述RNA分子的结构多样性和动态性。

1. 从单一结构到结构集合 ：传统的MFE方法聚焦于预测单一的、最稳定的RNA二级结构。而概率方法转向考虑RNA分子可能采取的所有可行结构的集合，即结构集合。
2. 集合方法 ：这些方法计算RNA分子在给定序列下所有可能结构的分布，而不是仅仅预测单一的最可能结构。这种方法认为RNA分子在一定条件下可能同时存在于多种不同的构象中。

PARS（第一代）

PARS（Parallel Analysis of RNA Structure）是一种早期的全局RNA结构map技术。PARS利用特定的核酸酶（如RNase V1和S1 nuclease），它们分别切割双链和单链RNA。通过测序切割产物，PARS可以区分RNA分子中的双链和单链区域，从而推断RNA的二级结构。

DMS-seq（第二代）

DMS-seq使用二甲基亚砜（DMS）这种化学试剂，它可以修饰未配对的腺嘌呤（A）和胞嘧啶（C）碱基。通过测序处理过的RNA，并分析DMS修饰的位置，DMS-seq可以识别RNA分子中未配对的区域，从而揭示RNA结构。

RNA G-quadruplexes（rG4s）

RNA G-quadruplexes（rG4s）是RNA分子中特殊的结构，由富含鸟嘌呤的序列形成的四链体结构组成。这些结构在调控基因表达和细胞过程中扮演重要角色，并且与多种疾病相关。

icSHAPE（第二代）

icSHAPE（in vivo Click-Selective 2'-Hydroxyl Acylation and Profiling Experiment）是一种先进的技术，用于在活细胞中测量RNA结构。icSHAPE利用核糖核酸的2'羟基选择性酰化反应，然后进行测序，以确定哪些核苷酸是结构上访问的或受限的。这种方法可以提供更精细的RNA结构图谱。

PARIS（第三代）

PARIS（Psoralen Analysis of RNA Interactions and Structures）是一种基于接近连接的第三代RNA结构测绘技术。它使用荧光素（一种化学试剂）在活细胞中交联相互接近的RNA分子。然后，通过利用高通量测序技术，PARIS可以鉴定在三维空间中互相接近的RNA区域，从而揭示复杂的RNA-RNA相互作用和高级结构。

总结起来方法是

1. 酶切法
2. 化学试剂修饰法
3. 交联法

深度学习方法

九、比较基因组学

不同的物种之间如何不同的？演化是不可不品的一环。

演化：即一个随着时间过去而发生的变化。

分子演化是进化生物学的一个分支，它研究DNA和蛋白质序列层面上的进化变化。我们可以将其视为时间推移下基因序列的变化过程。这个过程中，生物体的遗传物质（如DNA、RNA）在一代代中发生着序列上的改变。这些变化反映了生命的历史和进化的动态。

在分子演化的研究中，有几个主要话题值得关注：

1. **序列变化的速率和模式**：这涉及到研究遗传物质变化的速度以及这些变化遵循的模式。比如，有些基因区域的变化速率可能会非常快，而其他区域则相对稳定。
2. **适应性和中性变化的相对重要性**：在进化过程中，一些基因变化是为了适应环境（适应性变化），而另一些则可能是随机发生、对生物的生存和繁殖无显著影响的（中性变化）。
3. **中性演化与自然选择**：中性演化是指那些不受自然选择影响的遗传变化，而自然选择则是指有利于生存和繁殖的遗传特征在种群中逐渐增多的过程。

4. **新基因的起源**：研究新基因是如何在进化过程中产生的，这些新基因可能来源于基因复制、基因重组等机制。
5. **物种形成的遗传基础**：探讨在物种形成（物种分化）过程中，基因序列是如何发生变化的。
6. **基因组和表型变化上的进化力量**：这包括了研究影响整个基因组结构和表型特征

演化是生命最重要的特点，演化学说是现代生命科学的灵魂。 基因组的比较分析是对一个物种的多个个体基因组（群体内）或多个相 似或相差很大的物种基因组（跨群体）的综合分析。

比较基因组学可以做基因注释。

在比较基因组学中，科学家们集中研究以下几个特征：

1. **DNA序列**：比较不同物种中相似的DNA序列，以发现进化上保存的区域。
2. **基因**：研究不同生物体中基因的存在、缺失或变异，这些变化可能揭示了生物的进化历史。
3. **基因顺序**：比较基因在基因组中的排列顺序。在一些情况下，基因的物理顺序在进化过程中得到保留，这可以提供进化关系的线索。
4. **调控序列**：这些是控制基因表达的DNA序列。比较调控序列有助于理解不同物种中基因表达调控的进化。
5. **其他基因组结构标志**：如重复序列、转座元件等，这些元素的存在和分布也能提供关于进化的信息。

比较基因组学的核心原则是，如果两个生物体之间的某些特征在进化过程中被保留下来，那么这些特征很可能在它们的DNA中被编码。例如，如果两个远缘物种共享某个特定的基因序列，这可能表明这个序列在它们的共同祖先中就存在，且由于其重要性而被保留下来。

基因组比较分析

1. **功能元素大多保守**：在进化过程中，对生物体功能至关重要的DNA区域（如编码蛋白质的基因、重要的调控序列）往往在不同物种间高度保守。这意味着这些区域在漫长的进化历史中变化较小，因为任何显著的变异可能会对生物体产生负面影响。
2. **非功能区域大多发散**：与功能区域相比，非功能区域（如一些非编码DNA）在进化过程中更可能发生变化。这些区域的变化对生物体的生存和繁衍影响较小，因此在进化过程中更容易积累变异。

3. **功能区域因而更加突出**：由于功能区域的保守性和非功能区域的差异性，比较基因组学可以通过识别保守区域来突显出功能性DNA。

- **蛋白质、RNA、基序各自不同地进化**：这些不同的生物分子和序列元素在进化过程中受到不同的约束。例如，蛋白质的结构和功能可能限制了某些氨基酸的变化，而RNA分子的三维结构可能影响其核苷酸的变异。
- **通过各自独特的进化模式来发现它们**：每种类型的元素都有其独特的进化“签名”。通过比较不同物种的基因组，科学家可以识别出这些模式，从而推断出特定区域的功能。

比较基因组学的应用

1. **通过比较基因组学揭示功能元素**：

- 例如，外显子（exons）在不同物种（如小鼠、鸡、鱼）中普遍保守，意味着这些基因序列在进化过程中发生的变化较少，这通常是因为它们执行着对生物体至关重要的功能。
- 许多其他元素，如调控序列，也展现出了强烈的保守性。这表明它们在不同物种的进化过程中保持了较为稳定的结构和功能。

2. **确定每个区域的具体功能**：

- 不同类型的功能元素会因其特定功能而受到不同的选择压力，从而在进化过程中表现出不同的变化模式。这些模式包括基因序列的突变、插入和删除等。
- 通过分析这些变化模式，科学家可以推断出基因组的特定区域的功能。例如，一个高度保守的序列可能是一个重要的蛋白编码区或关键的调控元件。

3. **发展针对每种功能的进化特征签名**：

- 进化特征签名是指在不同物种间保持相对稳定的基因组特征，这些特征通常与特定的生物学功能相关联。
- 通过识别和分析这些特征签名，科学家可以更准确地预测基因或基因组区域的功能。例如，如果一个特定的DNA序列在多个物种中都高度保守，且这些物种共有某个生物学特征，那么这个序列很可能与该特征相关。

不同功能元件的演化特征

编码蛋白质的基因由于简并性，可以很好的度过自然选择，只要这个改变不是太离谱。

1. **蛋白质编码基因**：

- **密码子替换频率（Codon Substitution Frequencies）**：观察不同密码子在进化过程中如何变化。由于自然选择和遗传密码的冗余性，某些密码子的替换可能更常见。
- **阅读框架的保守性（Reading Frame Conservation）**：重要的蛋白质编码基因通常会保持稳定的阅读框架，以防止产生非功能性或有害的蛋白质。

2. **RNA结构**：

- **代偿性变化（Compensatory Changes）**：在RNA分子中，结构的完整性对功能至关重要。如果一个位置的突变破坏了结构，则另一个位置的突变可能会补偿这一缺陷。
- **静默的G-U替换（Silent G-U Substitutions）**：在RNA结构中，鸟嘌呤-尿嘧啶配对可以在不改变结构的情况下替换成其他配对，这种变化通常不影响RNA的功能。

3. **微RNA（microRNAs）**：

- **保守性轮廓的形状（Shape of Conservation Profile）**：微RNA的某些区域（如种子序列）可能显示出高度的保守性。
- **结构特征**：如环、配对等，这些特征对微RNA的功能至关重要。
- **与3'UTR基序的关系（Relationship with 3’UTR Motifs）**：微RNA通常通过与靶基因的3'非翻译区（3'UTR）内的特定基序结合来调控基因表达。

4. **调控基序（Regulatory Motifs）**：

- **保留共识序列的突变（Mutations Preserve Consensus）**：在调控基序中，即使发生突变，通常也会保持特定的核苷酸共识序列，这表明这些序列对于基因的调控功能至关重要。
- **增加的分支长度得分（Increased Branch Length Score）**：这表示在进化树上，具有特定调控基序的物种分支可能显示出更长的分支长度，暗示这些基序在进化中有显著的变化。
- **基因组范围内的保守性（Genome-wide Conservation）**：关键的调控基序在不同物种的基因组中通常显示出高度的保守性，这表明它们在生物体的调控网络中扮演着重要角色。

基因共线性（Gene Colinearity）是指在不同物种之间，相似或相同功能的基因在基因组上保持着相似的排列顺序。这种现象在进化生物学和比较基因组学中非常重要，因为它揭示了物种之间共同的进化起源和遗传物质的保守性。

在基因共线性的背景下，进行基因组范围内的比对（Genome-wide alignments）可以揭示所谓的“同源片段”（orthologous segments）。这些同源片段是指在不同物种中可以找

到的具有共同祖先的基因序列，通常它们在功能上是一致的。这种比对涵盖了整个基因组，帮助科学家识别和比较不同物种间的功能元素。

基因的选择与演化Ka/Ks

1. **非同义替换率 (Ka)**：

- 非同义替换是指那些改变了氨基酸序列的核苷酸替换。也就是说，这种替换会改变由DNA序列编码的蛋白质的氨基酸。
- Ka 是指在一个非同义位点（即一个可以发生非同义替换的位点）上发生非同义替换的比率。它反映了蛋白质编码区域的进化速度，通常与功能或结构上的变化相关。

2. **同义替换率 (Ks)**：

- 同义替换是指不改变编码氨基酸的核苷酸替换，即尽管DNA序列发生了变化，但由于遗传密码的冗余性，编码的氨基酸保持不变。
- Ks 是指在一个同义位点（即一个可以发生同义替换的位点）上发生同义替换的比率。由于这些替换不影响蛋白质的氨基酸序列，因此它们通常被认为是中性的，不受自然选择的影响。

选择压力

- 如果 **Ka/Ks > 1**，这表明非同义替换的频率高于同义替换，可能意味着这个基因区域受到了正向选择（适应性进化）。
- 如果 **Ka/Ks < 1**，则表示同义替换的频率高于非同义替换，暗示这个区域可能受到了纯化选择（保守性进化），维持了蛋白质的功能稳定性。
- 如果 **Ka/Ks ≈ 1**，则可能表示这个基因区域的进化既非受正向选择也非受纯化选择，可能是中性演化的结果。

纯化选择（Purifying Selection），又称为负向选择（Negative Selection），是自然选择的一种形式，其主要特征是淘汰那些有害的或非有利的基因变异。在纯化选择的作用下，不利于生物体生存和繁殖的突变被消除，从而维护了基因的稳定性和生物体的适应性。

这个概念可以通过以下几个要点进行理解：

1. **淘汰有害突变**：纯化选择主要针对那些对生物体有负面影响的基因突变。这些突变可能会损害重要的生物学功能，如蛋白质的结构和活性，或者影响生物体的生理机能。

2. **保持基因的功能稳定性**：通过淘汰有害突变，纯化选择有助于保持基因序列的完整性和功能稳定性。它对于保持物种的生存和繁衍至关重要，因为这有助于防止有害的遗传变异在种群中累积。
3. **与正向选择的对比**：与正向选择（或适应性选择）相对，纯化选择不是促进有利基因变异的积累，而是防止有害变异的积累。正向选择促进那些提高生物适应性的突变的传播，而纯化选择则抑制那些减少生物适应性的突变。
4. **遗传多样性的影响**：虽然纯化选择有助于去除有害的突变，但过强的纯化选择也可能限制基因的遗传多样性。这是因为它可能会导致有益变异的丢失，特别是在强烈的选择压力下。

如何计算Ka/Ks（初等计算法）

1. **计算位点（Counting Sites）**：

- **同义位点（S, Synonymous Sites）**：这些是序列中可以发生同义替换（不改变编码的氨基酸）的位点。计算同义位点是估计Ks的基础。
- **非同义位点（N, Nonsynonymous Sites）**：这些是序列中可以发生非同义替换（改变编码的氨基酸）的位点。非同义位点的计算对于估计Ka至关重要。

2. **计算替换（Counting Substitutions）**：

- **同义替换（Sd, Synonymous Substitutions）**：指在同义位点上发生的实际替换次数，这些替换不改变编码的氨基酸。
- **非同义替换（Nd, Nonsynonymous Substitutions）**：指在非同义位点上发生的实际替换次数，这些替换会改变编码的氨基酸。
- 计算替换的目的是确定在给定的演化时间内，每种类型的位点发生了多少次替换。

3. **修正多重替换（Correcting for Multiple Substitutions）**：

- 由于一些替换可能在同一个位点上多次发生，观测到的替换次数往往低于实际发生的替换次数。因此，需要对这种“多重替换”进行修正。
- 使用数学模型和方法（如Jukes-Cantor模型或Kimura的两参数模型）来估算实际替换次数，并根据这些估算值修正Ka和Ks的计算。

同义位点和非同义位点：UCX编码Ser（X表示任意核酸），因此第三个位置是同义位点，而改变第二个和第一个都会改变Ser，因此是非同义位点。UUU，UUC编码Phe，UUA，UUG编码leu，CUX编码leu，因此第三个位点既可以是同义位点又可以是非同义位点，因此计算时各算0.5个位点。而当关注的序列是UUA时，第一个位点也既可以是同义位点，又可以是非同义位点，所以此时各算0.5个位点。关注UUU时，第一个位点就是非同义位点。

问题是，一个突变可能有多种演化路径，例如GGC->AGA 。路径可能是GGC->AGC->AGA，可能是GGC->GGA->AGA 而且突变可能变来变去，中间不一定只有一步，所以真实替换率大于观察值。

有很多概率模型来计算Ka/Ks

系统发育树

基本term

1. 根（**Root**）：

- 系统发育树的“根”是指所有树中生物的共同祖先。这个祖先是推测出的，通常位于系统发育树的底部或起始处。
- 树的根表示了所有生物在该树中的最远共同起源，通常由一个进入该点的线表示，表明它来源于一个更大的类群（clade）。

2. 节点（**Nodes**）：

- 每个节点代表一个假设的共同祖先，这个祖先物种分化（speciated）产生了两个或更多的后代类群（taxa）。
- 节点是系统发育树上的关键点，标志着进化分支点，每个节点都代表了一个物种分化事件。

3. 外群（**Outgroup**）：

- 在系统发育树中，外群是与其他所有物种（内群）相比最为远亲的物种或群体。
- 外群的作用是提供一个参照点，帮助确定树的根部位置和理解内群物种之间的进化关系。通过比较内群和外群，科学家可以推断内群的共同特征是从共同祖先继承来的还是独立进化而来。

4. 类群（**Clades**）：

- 一个类群（单系类群）包含了一个共同祖先及其所有的后代。这意味着一个类群由一个节点及其所连接的所有分支组成。
- 类群是系统发育学中一个重要的概念，因为它代表了一个完整的进化支系，包括所有由特定祖先演化而来的物种。

系统发育树：亦称系统发生树、分子演化树、分子进化树 依据分子序列分化程度，重构演化关系，预测 分子序列功能 建树方法： 距离法（Distance Matrix）：UPGAM、

Neighbor-Joining（邻接法） 最大简约法（Maximum Parsimony）：替换数最小的拓扑结构，作为最优树 最大似然法（Maximum Likelihood）：需要替换模型，概率最大的拓扑结构作为最优树

以下是使用UPGMA构建树的基本步骤：（有点像Huffman编码树的构建）

1. **从比对序列中派生距离矩阵：**

- 首先，需要从已比对的序列中生成一个距离矩阵。这个矩阵表示了序列间的距离或差异，通常基于序列间的相似性或差异性（如核苷酸或氨基酸替换的数量）来计算。

2. **将每个叶子/序列视为一个簇（Cluster）：**

- 初始时，每个序列都被视为一个单独的簇或分支。

3. **重复以下步骤，直到只剩下两个簇：**

- **步骤1：通过最短距离聚类一对叶子（物种）：**在每一轮中，选择距离矩阵中距离最短的一对簇（或序列），然后将它们合并成一个新的簇。这个步骤基于假设距离最近的序列在进化上最为相关。
- **步骤2：重新计算新簇与其他叶子的新平均距离，并生成新的距离矩阵：**合并两个簇后，需要重新计算这个新簇与树中其他簇之间的平均距离，并更新距离矩阵。新的距离是原来两个簇到其他簇距离的算术平均。

构建树的一般步骤

Step 1: 选择共同基因/蛋白质 在这一阶段，研究者会选择一个或多个对所研究的一系列生物体普遍存在的基因或蛋白质。这个基因或蛋白质通常具有高度保守和足够变异的特点，以便反映物种间的进化关系。例如，在细菌和古菌中，16S rRNA基因是一个常用的选择，因为它的序列在不同物种间既保持了足够的保守性以进行有效比对，又包含足够的变异信息来揭示物种间的亲缘关系。

Step 2: 获取选定生物体的分子序列 接下来，从所选生物体中提取并收集对应于选定基因或蛋白质的分子序列数据。这些序列可以从公共数据库（如NCBI GenBank）获取，或者通过实验方法（如PCR扩增和测序技术）直接从生物样本中获得。

Step 3: 运行多序列比对 得到所有选定生物体的分子序列后，使用专门的生物信息学软件（如MEGA、Clustal Omega、Muscle或MAFFT等）来进行多序列比对。这个过程旨在排列各序列，使得同源区域尽可能地对齐，从而确定每个位置上的变异模式。

Step 4: 生成系统发育树（ cladogram） 基于比对结果，利用系统发育分析方法构建系统发育树（也称为 cladogram）。根据不同的分析策略和假设，可以选择不同的建树算法，包括但不限于：

- 距离法（Distance-based methods），如邻接法（Neighbor-Joining, NJ）
- 最大简约法（Maximum Parsimony, MP）
- 最大似然法（Maximum Likelihood, ML）
- 贝叶斯推断法（Bayesian Inference）

常用软件有MEGA，PAML，iTOL等

基因比对

动态规划多序列比对的时间复杂度更是要命。k个序列比对则为O(n^k)

LAGAN（Local Alignment of Genomic Sequences）是一种生物信息学工具，用于在基因组尺度上进行局部序列比对。以下是LAGAN算法的四个主要步骤：

1. **Find Local Alignments**：在这个阶段，LAGAN首先通过应用快速、灵敏的局部比对方法来识别两个或多个基因组序列中的相似片段。这些局部比对通常聚焦于保守区域和重复序列，它们可能代表了进化上的同源性或者功能相关性。局部比对允许序列即使在存在插入、缺失和点突变等变异的情况下也能匹配。
2. **Chain Local Alignments**：找到一系列局部比对后，LAGAN会将高度相关的局部比对链接起来形成“链”，这些链跨越整个基因组长度并尽可能保持连续性。这一过程有助于捕捉更大规模的结构和功能单位，比如基因簇或调控区域，并且可以克服由于序列重组而造成的局部比对之间的断开问题。
3. **Refine Alignments**：LAGAN在链接局部比对的基础上进一步优化比对结果。它可能会使用更复杂的模型和算法调整比对区块的位置和得分，以增强比对的整体一致性。这一步骤确保了生成的全局比对更加准确地反映物种间的遗传相似性和差异。
4. **Restricted Dynamic Programming (DP)**：虽然标准的全局动态规划（DP）方法对于长序列比对过于耗时和计算密集，但LAGAN采用了受限的动态规划技术。这种方法仅在第一步中找到的局部比对区域内执行精确的动态规划比对，从而有效地提高了算法效率。这种策略既能保留全局比对的优势，又能处理大规模基因组数据。

对多个基因组的比对。

目标：寻找对应区域

方法论：

1. **通过包含的基因锚定基因组片段：** 首先，基于每个基因组中所包含的基因来定位和识别基因组内的关键区域。由于基因是功能单元且在不同物种间相对保守，因此可以作为比对的基准点。
2. **解决每一对物种间的基因对应关系：** 对于任意两个物种，要确定它们之间的基因是否同源（即共享共同祖先），以及是否存在多对一或多对多的关系。这一步骤需要分析基因间的氨基酸相似性，并考虑基因的位置信息。
3. **构建核苷酸水平的比对：** 在明确基因对应关系的基础上，进一步细化到核酸序列层面进行精确比对，以便揭示详细的序列变异和结构变化。

面临的挑战：

- **并非所有区域都存在一对一的对应关系：** 因为进化过程中会发生基因的分歧、复制和丢失现象，使得某些区域可能存在复杂的同源关系，无法简单地一一对应。
- **基因组重排与基因演化：** 基因在染色体上的排列顺序可能发生改变，如倒位、易位等，这些都会增加基因组比对的复杂性。

解决方案：

- **BUS算法（Best Unambiguous Subgroups）：** 这是一种用于解决基因和区域对应关系的方法，它利用完全二分图连通性原理整合了蛋白质相似性和基因顺序信息。该算法有助于区分具有明确一对一关系的基因群组和那些经历快速变化的区域或蛋白家族。
- **正确解析基因对应关系的结果：** 应用上述方法后，通常能够成功解析出超过90%基因的一对一对应关系，同时还能识别出发生快速演变和多样性的基因区域及蛋白质家族。

UCSC的算法

1. **Multiple Progressive Alignment (MULTIZ)**：MULTIZ是一种用于构建多序列比对的方法，特别是在处理多个相关基因组或蛋白质序列时。这种方法采用了逐步递增的方式，首先将最相似的序列进行比对，然后逐步加入其他序列，并根据之前已经比对好的信息调整新加入序列的位置。在基因组层面，MULTIZ可以用来比对多个物种间的同源区域，从而揭示它们之间的进化关系和保守结构。
2. **Conservation Modeling (Phylo-HMM)**：在基因组学中，通过构建一个名为“phylo-HMM”（Phylogenetic Hidden Markov Model）的模型来量化基因组序列的进化保守性。Phylo-HMM结合了隐马尔可夫模型（HMM）和系统发育树（phylogeny），能够模拟DNA或蛋白质序列随时间演化的过程，同时考虑不同物种间共享的保守区

域和发生变异的部位。这种模型能更精确地识别那些在多个物种中高度保守的功能重要区域。

3. **Conservation Score: phastCons** : phastCons是一个基于概率的工具，它利用上述的phylo-HMM框架来估计每个碱基或者氨基酸残基在整个基因组或者蛋白质序列集合中的保守程度。phastCons计算出的分数代表了一个位置上的保守性评分，得分越高表示该位置在多个物种间的保守性越强，这往往意味着这个位置可能包含有重要的功能信息或调控元件。最终生成的phastCons得分矩阵可用于可视化基因组的保守区域，并辅助科学家们研究基因功能、调控区域和其他重要的生物学问题。

比较基因组学研究

同源外显子比较保守。 使用保守区域寻找TF结合位点。识别重要的突变。

一些常识问题

要找到染色体22上SNPs最多的外显子，你需要访问一个包含基因组变异数据的数据库，比如dbSNP，并结合基因结构注释信息进行分析。这可以通过利用公共资源如UCSC Genome Browser、Ensembl或通过生物信息学工具及计算平台（如Galaxy）进行复杂的数据挖掘和计算来实现。用户可以根据SNP密度或者直接计数特定外显子内的SNP数量来找出SNPs最多的外显子。

十、表观基因组学

什么是表观基因组学？

一个人的基因组明明都一样，为什么会有不同的细胞？

定义：在核苷酸序列不发生改变的情况下，研究基因组上的化学修饰和空间结构变化如何影响基因功能 和表达调控的一门学科。 研究内容： 化学修饰变化：DNA、RNA、 蛋白质 空间结构变化：核小体（ nucleosome）、染色质（ chromatin）、基因组 表观特点：DNA序列不变、可逆、 可遗传、动态变化

DNA甲基化

甲基作为一个化学基团（-CH_3），它能够结合在DNA上某些特定部位，这个甲基和DNA结合过程叫甲基化。相反，甲基从DNA上 脱落的过程就叫做去甲基化。 DNA甲基化能引起染色质结构、 DNA构象、DNA稳定性及DNA与蛋白质相互作用方式的改变，从而控制基因表达。

DNA甲基化是指DNA碱基上特定位置的碳被添加甲基的过程。其主要形式有：N5-mC（胞嘧啶）、N6-mA（腺嘌呤）、N7-mG（鸟嘌呤）

CG甲基化：在DNA甲基转移酶（DNA methyltransferase）的作用下， 在基因组CpG二核苷酸的胞嘧啶5'碳位共价键结合一个甲基基团。

非CG甲基化是指在DNA分子中，胞嘧啶（C）的甲基化并不发生在经典的CpG二核苷酸对中，而是发生在CHG或CHH序列上下文中。这里的“H”代表A、T或C。

人类5mC甲基化的特点 除了胚胎和脑组织，其它组织的5mC甲基化通常发生在CpG二核苷酸上 人类基因组中大约含有3千万个CpG二核苷酸 CpG岛（一段CpG含量较高的区域）甲基化水平较低，非CpG岛甲基化水平较高 人类一生中5mC的含量动态变化 受精过程中会发生5mC重编程

CpG岛通常位于基因启动子区域附近，富含非甲基化的CpG二核苷酸对。以下是一个简化的CpG岛的例子（实际序列会更长且复杂）：
ATCGCCGGACGTCGAGCCGCCGCGTACGCGTGCCTAGCGTAGC... 在这个例子中，“CG”就是 CpG 二核苷酸对的实例，它们在序列中相对密集出现。一个真正的CpG岛不仅需要包含高密度的CpG位点，还需要满足一定的长度（一般超过200bp）、GC含量较高（通常大于60%）以及与基因转录起始点紧密关联等特点。

甲基化的建立与维持由一个从头甲基化酶和一个维持甲基化酶。维持甲基化酶能将对面还没有甲基化的C也甲基化（例如半保留复制中能让对面迅速甲基化）

甲基化也可以被去除主动或者被动。

大多数情况下启动子区域甲基化程度与基因表达量成负相关。

组蛋白修饰

组蛋白修饰（Histone Modifications）是组蛋白翻译后发生的共价修饰，包括甲基化、磷酸化、乙酰化、泛素化、类泛素化等。 有很多种组蛋白修饰，对每个不同的组蛋白的不同的修饰。

组蛋白与DNA是染色质的稳定成分， 非组蛋白与RNA的含量则随细胞生理状态不同而变化（染色体中组蛋 白以外的蛋白质成分称非组蛋白， 绝大部分非组蛋白呈酸性） N-末

端氨基酸残基可发生多种共价修饰 组蛋白修饰的作用：

- 改变与DNA及其他核蛋白的相互作用
- 改变染色质结构和活性
- 调控基因表达

组蛋白甲基化：组蛋白甲基转移酶将一至三个甲基基团从S-腺苷-L-蛋氨酸转移到组蛋白的赖氨酸或精氨酸残基上的过程。

组蛋白乙酰化：通过酶催化使乙酰辅酶A的乙酰基（COCH3）转移至组蛋白上的过程。功能：与许多细胞过程的调控密切相关，包括 染色质动力学、基因转录、基因沉默、细胞周期、凋亡、分化、DNA复制、DNA修复、入核 、神经元抑制。

组蛋白修饰命名规则 [组蛋白名称][氨基酸单字符简称][氨基酸位置][修饰种类] H3K4Me=组蛋白H3从N端开始起计第4个赖氨酸的甲基化

正向交叉对话（Positive cross-talk）： 当一种组蛋白修饰促进或增加另一种组蛋白修饰的可能性时，这种关系通常用箭头表示，即“→”。例如，某个特定位置上的组蛋白乙酰化可能有助于吸引或激活某些酶，这些酶又能进一步催化同一或邻近残基的甲基化。这样，一个修饰就“积极地”影响了另一个修饰的发生，共同影响染色质的开放程度和转录活性。

负向交叉对话（Negative cross-talk）： 相反，如果一种组蛋白修饰阻止或降低了另一种组蛋白修饰的存在或维持，这种关系则用扁平头表示，即“⊥”或“⊣”。比如，某一残基的甲基化状态可能干扰相关酶对同一或附近残基进行乙酰化的识别和催化过程，导致乙酰化水平降低或者被移除。这种“负向”交互作用可以抑制基因表达或者引起染色质压缩。

表观转录组

RNA上也有表观修饰。 例如hm5C 5羟甲基修饰，m6A 6甲基修饰。 它们能调控转录等

RNA编辑

RNA编辑是一种生物体内发生的转录后修饰过程，它发生在基因转录生成mRNA（信使RNA）之后，通过一系列酶催化反应对mRNA分子中的碱基序列进行精确或随机的修改。这种修改不同于直接由DNA模板复制的原始信息，而是以一种非模板驱动的方式改变了mRNA的遗传编码。

具体来说，RNA编辑的过程可以包括以下几种类型：

1. **碱基替换**：在mRNA链上的某个核苷酸被另一种不同的核苷酸所取代，例如将腺嘌呤（A）变为胞嘧啶（C），或者鸟嘌呤（G）变为尿嘧啶（U）等。这种变化可能会影响对应的氨基酸编码，从而改变最终翻译出的蛋白质结构和功能。
2. **核苷酸插入**：在mRNA的特定位置增加一个或多个额外的核苷酸，这会改变后续密码子读框，可能导致氨基酸序列的改变、提前出现终止密码子（截短蛋白）或是产生新的开放阅读框。
3. **核苷酸缺失**：从mRNA链中移除一个或多个核苷酸，同样可以影响到读框和编码信息，导致氨基酸序列的变化或提前终止翻译。

为什么要有表观修饰

发育过程中一个受精卵能发育成那么多种不同的细胞，很神奇吧，生命。

以下是一些表观遗传调控的主要分子标志：

1. **DNA甲基化**：DNA甲基转移酶（DNMTs）催化在胞嘧啶（C）的第5位碳原子上添加一个甲基基团，形成5-甲基胞嘧啶（5-mC），这通常导致转录沉默。甲基化的模式在发育、细胞分化和疾病中具有重要意义。
2. **组蛋白修饰**：组蛋白是构成染色体结构的重要蛋白质，它们上的化学修饰如乙酰化、甲基化、磷酸化等可以影响染色质的紧密程度，从而影响基因的可及性和活性。例如，组蛋白乙酰化通常与活跃转录相关联，而特定的组蛋白甲基化状态则可能与基因沉默或激活有关。
3. **非编码RNA调控**：包括微小RNA（miRNA）、长非编码RNA（lncRNA）等，它们可以通过与mRNA结合或影响染色质结构间接调控基因表达。
4. **染色质重塑**：由一系列染色质重塑复合体介导的过程，通过改变核小体定位或DNA螺旋结构来重新排列染色质构型，进而影响基因表达。
5. **X染色体失活**：在雌性哺乳动物中，一种典型的表观遗传现象是两个X染色体中的一个会随机失活以保证性别决定基因的剂量补偿。
6. **印记遗传**：是指某些基因根据父母来源的不同，其表达受到不同表观遗传标记控制的现象，只有一条等位基因（通常是父源或母源）是活性的。

DNA甲基化能精准的预测实际年龄。运动、良好生活习惯能降低表观年龄。 生活方式和环境因素，如吸烟、饮食习惯及感染性疾病等，都会使人体面临化学反应的压力。当身体暴露于这些压力源时，会触发一系列生化反应以应对这些挑战。这些反应过程通常会导致表观基因组发生改变。

表观基因组包含了DNA甲基化、组蛋白修饰以及其他调控因子等非遗传性的化学标记，它们不改变DNA序列本身，但能调控基因的表达模式。例如，长期吸烟可能导致某些基因区域过度甲基化，从而抑制了相关基因的正常功能；不良饮食习惯（如高脂肪或高糖摄入）可以影响体内代谢途径，并通过表观遗传机制改变基因表达；而感染病原体也可能引发炎症反应和其他免疫调节机制，间接地导致表观遗传变化。

虽然环境因素引起的表观遗传变化有时可能对健康有害，但表观基因组具有一定的适应性和可塑性，它在一定程度上允许生物体根据外部环境和内部生理状态调整基因表达，这对于维持正常的人类健康至关重要。然而，当表观遗传调控系统出现问题时，比如负责识别和添加（“读”和“写”）表观遗传标记的蛋白质出现功能障碍，则可能导致疾病的发生。许多人类疾病，包括但不限于癌症、神经退行性疾病以及心血管疾病等，都与表观遗传学异常有关。

如何研究表观遗传组学

组蛋白修饰（在基因的哪一个区域）、TF结合位点可以使用ChIP-seq DNA甲基化，可以使用Bisulfite-seq RNA修饰（m6A）可以用RIP-seq、CLIP-seq、GLORI-seq

ChIP-exo

ChIP-exo（染色质免疫沉淀结合exo核酸酶）是一种高分辨率的技术，用于研究转录因子（TF, Transcription Factor）与DNA的相互作用位点。这种技术结合了传统的染色质免疫沉淀（ChIP）方法和exo核酸酶的精确切割特性，以达到单核苷酸分辨率水平地确定TF在基因组上的结合位置。

在ChIP-exo实验中，首先通过抗体特异性地捕获与目标转录因子结合的DNA片段。然后，使用具有5'到3'方向 exonuclease活性的exo核酸酶对捕获的DNA-蛋白质复合物进行处理。Exo核酸酶会从DNA片段的5'端逐个去除核苷酸，直到遇到紧密结合的转录因子所保护的DNA区域为止。由于转录因子通常只覆盖DNA上的一小段区域，因此当核酸酶到达这个保护区域时就会停止降解。

最后，通过高通量测序技术（如深度测序或下一代测序NGS）分析这些被exo核酸酶切割后保留下来的DNA片段，可以精确地确定转录因子在DNA上的结合位点，并且能够分辨出相邻非常接近的结合位点，甚至识别单个核苷酸差异的结合偏好性。

因此，ChIP-exo技术极大地提高了我们对转录因子调控网络的理解，尤其是在识别转录因子如何精细调节基因表达以及它们在基因组上的精确分布方面的认知。

5mC 测序

5mC测序，全称为5-甲基胞嘧啶测序，是一种用于检测DNA分子中胞嘧啶（C）碱基甲基化状态的技术。在基因组DNA中，胞嘧啶可以被甲基化形成5-甲基胞嘧啶（5-methylcytosine, 5mC），这一表观遗传修饰对于调控基因表达、细胞分化以及许多生物学过程至关重要。

Bisulfite sequencing（重亚硫酸盐测序）是研究5mC的经典方法之一，其工作原理如下：

1. **Bisulfite处理**：通过将基因组DNA暴露于强还原性条件下的重亚硫酸盐溶液中，未发生甲基化的胞嘧啶（C）会被脱氨基转变为尿嘧啶（U）。由于甲基化的胞嘧啶（5mC）的甲基基团保护了它不受此化学转化的影响，因此甲基化的胞嘧啶保持不变。
2. **PCR扩增与转化**：经过bisulfite处理后的DNA进行PCR扩增时，原本未甲基化的胞嘧啶（已转化为尿嘧啶）会按照PCR引物的设计和DNA复制规则，在合成新的DNA链过程中变成胸腺嘧啶（T）。而甲基化的胞嘧啶则仍然保持为C。
3. **高通量测序**：得到的PCR产物进一步纯化并构建测序文库，然后利用高通量测序技术（如Illumina平台）进行测序。通过对测序数据的比对分析，可以精确识别出哪些C碱基在原始DNA序列中是甲基化的，哪些是非甲基化的。
4. **结果解读**：最终，通过比较处理前后的序列差异，可以在单碱基分辨率上绘制出全基因组或目标区域的DNA甲基化图谱，从而揭示特定细胞类型或生物体发育阶段的DNA甲基化模式及其可能的功能意义。

5mC甲基化水平的本质是一个定量指标，指的是该特定胞嘧啶位置在不同细胞群体或样本中发生甲基化的概率或者频率。 用于描述一个或多个CpG位点在整个细胞群体中的平均甲基化状态，反映的是表观遗传调控的异质性及其在细胞分化、发育和疾病进展中的潜在作用。

GLORI-seq

GLORI-seq是一种高通量测序技术，它专门用于在全基因组范围内以单碱基分辨率来检测和量化RNA分子中的N6-甲基腺苷（m6A）修饰。这种表观遗传修饰是RNA上最常见的内部修饰之一，对转录后调控具有重要作用。

本质仍是使用化学标记保护被甲基化的碱基。

表观遗传信息学

MACS模型peak calling

MACS（Model-based Analysis of ChIP-Seq）是一种用于分析染色质免疫沉淀测序（ChIP-Seq）数据的生物信息学工具，它通过建立模型来精确地识别转录因子结合位点和其他蛋白质-DNA相互作用区域。在处理ChIP-Seq数据时，MACS考虑了芯片实验中 reads 的偏移特性。

在ChIP-Seq实验中，当抗体与DNA上的目标蛋白（如转录因子）结合并进行免疫沉淀时，由于实验过程中的物理和化学因素，读段（reads）可能会发生一定的偏移。这种偏移可能是由于核小体定位、DNA片段化过程中的偏差等因素导致的。为了准确估计这些偏移，并将reads映射到实际的蛋白质结合峰位置，MACS引入了“shifting size”的概念。

Shifting size 参数是指对原始测序reads的位置进行校正的距离。默认情况下，MACS会基于实验数据自动估算最优的shift size值，以便更准确地确定转录因子或其他蛋白质在基因组上的具体结合位置。这个参数通常根据ChIP实验的目标蛋白以及组织类型或细胞系的特性进行优化，确保最终得到的峰分布图谱能够真实反映蛋白质-DNA结合的实际情况。通过这种模型驱动的方法，MACS能够在全基因组范围内高效且准确地检测出潜在的蛋白质结合峰值（peaks）。

在MACS（Model-based Analysis of ChIP-Seq）模型中，动态泊松分布被用于分析ChIP-Seq实验数据中的序列标签分布情况，以识别和量化潜在的转录因子或组蛋白修饰等蛋白质-DNA相互作用位点。

泊松分布： 在经典的统计学框架下，泊松分布描述的是单位时间内独立随机事件发生的次数的概率分布。对于ChIP-Seq数据分析，可以将基因组上每个碱基位置视为可能发生一个“事件”的单元，在MACS模型中，这些事件就是测序reads映射到该位置的计数。假设这些事件的发生遵循泊松过程，那么在一个特定位置i上的reads计数可以服从泊松分布，其参数λ表示平均期望计数。

λBG： 在MACS中，背景标记密度（λBG）代表了整个基因组范围内序列标签均匀分布时的预期计数率。这意味着如果没有特定的蛋白质-DNA结合发生，理论上读段应该均匀地分布在基因组所有区域。λBG可以通过对全基因组数据进行全局评估来估计。

λlocal： 然而，在实际的ChIP-Seq数据中，由于各种原因（如DNA复制、染色质结构和实验偏差等），序列标签在局部区域内的分布往往不是均匀的，存在一定的偏倚。因

此，MACS引入了一个动态的局部标记密度（λlocal），用来捕获基因组各区域内可能存在的局部偏差特征。

- **有对照样本的情况下**：λlocal 由λBG 和 控制样本（ctl）在不同窗口大小（1k, 5k, 10k 等bp长度的滑动窗口）下的最大计数率决定，即 λlocal = max(λBG, λ1k-ctl, λ5k-ctl, λ10k-ctl)。通过比较实验组与对照组在同一窗口大小下的峰值密度，可以更准确地识别出实验组特有的、显著高于背景噪声的峰区。
- **无对照样本的情况下**：λlocal 的计算则仅依赖于λBG以及实验组自身在5k和10k bp窗口下的计数率，即 λlocal = max(λBG, λ5k, λ10k)。此时需要通过对实验数据的不同区域进行内部比较，寻找显著偏离背景水平的区域。

通过这种方式，MACS能够有效地适应不同区域的生物学差异，并准确地区分真实的蛋白质结合位点与随机噪音，从而提高ChIP-Seq数据分析的精度和可靠性。

在MACS（Model-based Analysis of ChIP-Seq）中，动态性体现在对基因组上不同区域的序列标签密度分布（reads计数）进行灵活且适应性的建模。具体来说：

1. **局部背景模型的动态调整**：

- 在传统的泊松分布模型中，通常会有一个固定的背景发生率（λBG），但在MACS中，针对基因组的不同区域，尤其是具有显著信号或偏差的地方，MACS引入了一个动态的、局部的标记密度估计值——λlocal。
- λlocal是根据每个窗口内的实际数据动态计算得到的，能够反映该区域内序列标签分布的实际特性，包括但不限于实验噪音、生物学过程引起的区域特异性标记偏好以及测序深度不均一等。

2. **窗口大小的动态选择**：

- MACS通过使用不同大小的滑动窗口（如1k, 5k, 10k bp等）来估算λlocal，这意味着它不是依赖于一个全局固定的参数，而是依据数据特征自适应地选取合适的窗口大小以优化峰值识别效果。

3. **有无对照样本情况下的动态处理**：

- 对于有对照样本的数据，λlocal 是基于对照样本和全基因组背景共同确定的，在不同窗口下取最大计数率，从而可以更好地排除非特异结合带来的影响。
- 对于没有对照样本的情况，MACS依然能够利用不同的窗口大小动态地评估并比较实验组数据本身的局部背景，从而寻找潜在的蛋白质-DNA相互作用位点。

有两种基本的方法来生成控制DNA样本：

1. **Input样本**：

- 输入样本是从与ChIP DNA相同条件下进行交联和破碎细胞所提取的DNA。这意味着细胞经过同样的交联步骤以固定蛋白质-DNA复合物，并通过超声波处理使DNA片段化。通过比较ChIP样本与输入样本的测序数据，可以识别出相对于随机分布的DNA片段而言显著富集的区域，这些区域可能是目标蛋白的真实结合位点。

2. **IgG抗体控制样本**：

- IgG（免疫球蛋白G）是一种“模拟”ChIP反应的控制方法，它使用一种针对无关、非核内抗原的抗体来进行ChIP实验。这种抗体不会与任何特定的转录因子或其他核内蛋白质结合，因此从理论上讲，它捕获到的DNA片段应当代表的是非特异性的背景信号。
- 与输入样本相比，IgG控制样本更接近于模拟真实的ChIP实验过程，因为它包括了抗体与DNA碎片之间的非特异性物理吸附等所有可能的实验误差来源。通过对比ChIP样本与IgG控制样本的数据，能够更加准确地排除实验过程中的假阳性信号，从而提高鉴定转录因子或组蛋白修饰真实结合位点的可靠性。

在表观遗传学和基因调控研究中，转录因子（TFs）与DNA的相互作用模式可以分为以下几种类型：

1. 点源（**Point-source**）：

    - 这种类型的TF-DNA结合通常发生在特定且局部化的DNA序列上，生成ChIP-seq实验中高度集中的信号峰。
    - 大多数具有明确序列特异性的转录因子及其协同因子都属于这一类。它们直接识别并结合到DNA上的特定顺式作用元件，如启动子区域（TSS, Transcription Start Site）或增强子区域，并在此处调节基因表达。
    - 同样，某些与启动子或增强子相关的组蛋白修饰标记（例如H3K4me3、H3K27ac等），它们的存在也是以点源模式出现，指示了活跃的转录起始位点或增强子活性。

2. 宽域来源（**Broad-source**）：

    - 这些蛋白质与DNA的相互作用遍布较大的基因组区域，而不是单一的固定位置，导致ChIP-seq信号表现为较宽的峰值或连续的区域。
    - 某些组蛋白修饰标志如H3K9me3（常与异染色质相关）、H3K36me3（与转录延伸相关）等往往表现出这样的分布特征。
    - 另外，一些与转录过程中的延长或抑制相关的染色质蛋白质，例如ZNF217，其结合模式也可能表现为宽域性，影响大片段染色质的状态或功能。

3. 混合来源（**Mixed-source**）：

- 这类转录因子及某些染色质修饰蛋白在不同基因座上的结合模式是混合的，既能在部分位点上像点源那样精确地结合，又能在其他位点形成较大范围的结合域。
- RNA聚合酶II是一个典型的例子，它在启动子区域的结合可能相对精确，而在转录单元内则随着转录过程而产生更广泛但非均匀的分布。
- 类似地，SUZ12（Polycomb复合体的一个组件）等某些染色质修饰蛋白在不同的基因调控环境中有时会显示点源结合，有时则形成较大的结合域。

测量全局ChIP富集程度：

- FRiP (fraction of reads in peaks) 是一个衡量整体ChIP-seq实验富集效果的关键指标。它计算的是所有有效reads中有多少比例落在了peak区域（即显著富集区）内。

FRiP值的应用和特性：

- FRiP值与被识别为显著富集区域的数量呈正相关，并且这种关系通常是线性的。也就是说，当有更多的显著峰被call时，FRiP值往往更高。
- FRiP非常适用于对比在同一细胞系中使用相同抗体得到的不同实验结果，或是在不同细胞系中使用同一抗体的结果，以及使用不同抗体针对同一种转录因子的研究结果，从而评价抗体特异性和实验有效性。
- FRiP值受peak calling算法的影响较大。不同的峰值识别策略可能导致不同的FRiP值，因此在比较不同实验时需要确保采用一致或可比的peak calling方法。

在ChIP-seq数据分析中，重复实验的一致性是评估结果可靠性的重要指标。通过比较不同生物学重复之间的数据，可以区分真实信号与随机噪声。

IDR（Irreproducible Discovery Rate）：

- IDR是一种衡量ChIP-seq实验中检测到的峰值在重复样本间重现性的统计方法。显著的峰值（即具有生物学意义的真实信号）通常在多个独立实验复制中的排序更加一致，而那些低显著性的峰值可能更多地由于随机波动产生，一致性较差。
- IDR通过对各个峰进行分类，将其划分为可重现组和不可重现组来量化这种从真实信号到噪音的转变过程。如果一个峰在多个重复实验中的排名都非常稳定，那么它被归类为可重现的，并且更有可能代表真实的转录因子结合位点或者组蛋白修饰区域。

通过计算IDR，研究人员可以设置阈值以确定哪些峰被认为是高度可信的、能够在不同生物学条件下稳定出现的结果，从而有助于过滤掉假阳性信号，提高最终分析结果的准确性。

Peak calling是ChIP-seq数据分析中的关键步骤，用于识别并量化转录因子或组蛋白修饰在基因组上的显著富集区域（即潜在的结合位点）。这一过程涉及到选择合适的峰值检测算法和归一化方法，并对用户可设置参数进行优化。

选择适当的峰值调用算法与归一化方法：

- 不同类型的蛋白质（如转录因子、组蛋白修饰酶等）可能具有不同的结合特性。因此，在执行peak calling时，需要根据所研究蛋白质的性质选择最适合的算法。例如，某些算法可能更适合识别点状且强峰信号的转录因子，而另一些算法则擅长处理形成宽域分布的组蛋白修饰数据。
- 归一化方法的选择也至关重要，以消除测序深度、样本间差异等因素的影响。例如，基于控制样本的归一化可以校正背景噪音，使得不同实验间的比较更为准确。

用户可设置参数及其影响：

- 现有的peak calling软件提供了许多用户可调整的参数，包括但不限于峰值宽度范围、峰强度阈值、窗口大小、局部背景估计方式等。这些参数的设置将直接影响到最终被调用出的峰的数量和质量，也就是说，相同的原始数据使用不同的参数配置可能会得到迥异的结果。

峰值数量的可比性问题：

- 即使采用相同显著性水平的阈值，如p-value或FDR（False Discovery Rate），也不能保证在不同文库之间或者使用不同峰值调用工具得出的峰值数量可以直接比较。这是因为不同的算法在处理信号噪声、峰值合并以及确定统计显著性的策略上可能存在差异，从而导致结果难以直接互比。因此，在评估和对比不同实验或不同分析结果时，应充分考虑这些因素，并谨慎解读峰值调用结果。

ChIP-seq分析

Motif分析是ChIP-seq数据分析中的一项重要技术，它主要用于识别和解析与特定蛋白质或组蛋白修饰相关的DNA序列模式（motif）。这些模式通常代表了转录因子或其他调控蛋白的结合位点。

1. 实验验证：
    - 当已知ChIP实验所研究蛋白质的DNA结合基序时，motif分析能够通过在ChIP-seq数据中检测到该已知motif来验证实验的成功性。如果实验正确执行并捕获到了目标蛋白质-DNA复合物，那么在其结合区域应能发现预期的DNA序列特征。

2. 协同蛋白质识别：

    - 除了验证目标蛋白质的motif外，motif分析还能揭示与ChIPed蛋白质共同作用或形成复杂结构的其他未知蛋白质的DNA结合motif。这有助于理解复杂的转录调控网络，比如多个转录因子如何共同作用于相同的基因启动子区域以协调基因表达。

3. 表观遗传学应用：

    - 在对组蛋白修饰（如甲基化、乙酰化等）进行ChIP-seq分析时，motif分析也十分有用。即使开始并不知道具体哪种蛋白质或调控机制与特定的组蛋白修饰相关联，通过motif分析可以发现与这些表观遗传标记意外关联的DNA序列信号。这些新发现的motif可能指向尚未被充分认识的转录调控因素，从而推动我们对基因表达调控机制的深入理解。

目标预测（**Target prediction**） 是基于ChIP-seq数据推测转录因子（TF）可能调控的基因或基因启动子的过程。以下是对目标预测的不同方法及其特点的解释：

1. 简单方法 ：

- 在这类方法中，目标基因通常是根据在特定距离范围内（如TSS附近）是否存在转录因子结合峰来确定的。这意味着不考虑结合峰与基因之间的确切相对位置，而是将所有检测到的结合峰等同对待。
- 这种方法对识别的目标基因数量非常敏感，调用的峰值越多，潜在的目标基因数量也相应增加。然而，这种方法可能导致假阳性率较高，因为并非所有位于TSS附近的结合都代表真实的调控作用。

2. 改进型方法 ：

- 为了提高预测性能，可以考虑结合峰的数量以及它们在TSS周围的分布情况。例如，分析峰的高度、宽度以及多个峰是否聚集成簇等因素，可以帮助区分真阳性调控位点和偶然出现的结合事件。
- 此外，还可以通过分析结合峰相对于不同基因特征（如启动子区、增强子区或其他功能元件）的具体位置关系，以更精细的方式推断转录因子的作用模式和目标基因。

3. 整合**ChIP-seq**与**RNA-seq**数据 ：

- 结合ChIP-seq数据和RNA-seq数据能够更好地推断转录因子的靶标基因，并预测该转录因子具有激活还是抑制功能。通过比较转录因子结合区域与基因表达变化之间

的关联性，可以鉴定出那些结合了转录因子后其下游基因表达上调或下调的实例，从而判断转录因子是激活还是抑制相关基因的表达。

ChromHMM

Chromatin states method利用隐马尔可夫模型（HMM，Hidden Markov Model）是一种统计学方法，用于描述和推断染色质的复杂结构和功能状态。在表观遗传学中，染色质的不同修饰模式往往与特定的功能状态相关联，如活跃转录、抑制转录、启动子区、增强子区等。

在ChromHMM中，通常将不同的组蛋白修饰或DNA甲基化等表观遗传标记的数据整合在一起，并假设这些标记在一个基因组区域上的联合分布遵循一个隐藏的多状态过程。每个状态对应一种特定类型的染色质构象和功能活性。

HMM包含两个主要部分：可见状态（observed states）和隐藏状态（hidden states）。在表观遗传分析中，可见状态是实验测得的各种表观遗传标记数据，而隐藏状态则是我们想要预测的染色质状态。

通过训练HMM模型，可以学习到不同染色质状态之间的转换概率以及每种状态对应的表观遗传标记特征，从而将整个基因组划分为一系列具有生物学意义的染色质状态区域。这种方法有助于揭示染色质结构与基因表达调控的关系，进而理解复杂的细胞命运决定机制和疾病的发生发展。

ChromHMM利用隐马尔可夫模型（HMM）进行分析后，其主要输出结果是整个基因组的染色质状态注释图谱。具体来说，它将基因组按照一定的长度窗口划分，并对每个窗口赋予一个或多个最可能的染色质状态标签。这些标签通常代表了不同的功能和表观遗传特征，例如：

1. 活跃转录启动子（Active Promoter）
2. 强化激活区域（Strong Enhancer）
3. 预活跃增强子（Poised Enhancer）
4. 转录延伸区（Transcribed Region）
5. 活跃调控区（Regulatory Element）
6. 紧凑型/抑制性异染色质（Repressed or Heterochromatic）

十一、3D基因组

基因组三维空间结构 染色体由DNA与组蛋白共同组成 染色体从一级结构（绳珠 模型）到四级超螺旋折叠结构 DNA分子一共被压缩了8400倍左右，形成了复杂的三维空间结构 正是这些折叠和压缩，使得基因在细胞中的分布复杂而又有序。

基因组三维空间结构与功能的研究简称三维基因组学 在考虑基因组序列、基因结构及其调控元件的同时, 对基因组序列在 细胞核内的三维空间结构, 及其对基因转录、复制、修复和调控等生物过程中功能的研究。

为什么要研究三维基因组？ 科学家们发现，调控元件在空间结构上并不是在染色体上呈线性地一字依次排开，这些离散的调控元件 并不能有效地解释很多基因的调控结果和机制。由此，猜测其与基因组的三维空间结构相关。

随着测序分辨率的增加，研究者发现基因组的三维空间分为：

- 染色体疆域
    - 染色体疆域（Chromosome Territories）：指每条染色体都占据着一个独特的区域，同一染色体上的交互频率高于不同染色体之间的交互频率。交互频率随着基因座之间的线性距离增加而呈指数级下降。
    - 基因丰富的染色体偏向位于中间
- 区室
    - 区室（Compartments A/B）：根据互作图谱，能够将基因组近似分为A（常染色质-转录活跃区域），B（异染色质-转录非活性区）。A/B区室在染色体上相间分布，并且在不同细胞状态下可以互相转变。
- 拓扑关联结构域
    - 拓扑关联结构域 （Topologically Associating Domains, TAD ）：是区室下的亚结构，长度为 300Kb-1Mb，具有 TAD 内部互作频率高， TAD 间互作频率低的特点。其边界富集 CTCF、持家基因、 tRNAs、SINE 反转录转座子等DNA元件。
- 染色质环
    - 染色质环（Chromatin loop）：也可称为交互峰（interaction peaks），由染色体上相距较远的两个片段构成，其在线性空间中虽相距较远，但在三维空间结构中却具有显著的近距交互作用（成环）。调控元件（如增强子）便可以通过这种结构远距离调控基因的表达。

研究的问题是 基因组 3D 结构的检测？ 基因组 3D 结构的形成和维持机制？ 基因组 3D 结构如何调控基因表达？ 基因组 3D 结构的动态变化和作用？（发育，疾病，病毒感染等？）

三维基因组基本实验技术

FISH

FISH的基本步骤

1. **制备探针**：首先制备带有荧光标记的核酸探针，这些探针与目标DNA或RNA序列互补。
2. **样本制备**：样本细胞固定在载玻片上，通常需经过去脂、脱水等预处理。
3. **杂交**：将荧光标记的探针与样本细胞共孵育，使其与目标序列特异性结合。
4. **洗涤**：去除未结合的探针。
5. **检测**：使用荧光显微镜观察并分析荧光信号。

2D-FISH与3D-FISH的区别

- **2D-FISH**：这是传统的FISH方法，通常在二维平面上观察和分析染色体。它适用于检测染色体上特定区域的异常，如染色体重排、缺失或扩增。
- **3D-FISH**：与2D-FISH相比，3D-FISH在空间结构上提供了更多信息。它允许在三维空间中观察染色体的结构和定位，能够更精确地分析基因的空间组织和表达。这对于理解染色体在细胞核中的排列和功能有重要意义。

染色体构象捕获(3C)

染色体构象捕获（Chromosome Conformation Capture，简称3C）技术是一种用于研究染色体在三维空间中的组织和相互作用的分子生物学方法。

3C技术的基本原理

1. **交联**：首先，使用甲醛将细胞中的DNA和相互接触的蛋白质交联固定，从而“冻结”染色体在核内的空间结构。
2. **消化**：然后使用限制性核酸内切酶切割交联的DNA，产生大量的DNA片段。
3. **连接**：这些DNA片段经过适当的处理后，可以通过DNA连接酶相互连接。只有在原始细胞核中空间接近的DNA片段才能有效地连接。
4. **反交联和扩增**：将交联反转，释放DNA，然后使用PCR等方法放大特定连接的DNA片段。
5. **分析**：通过分析这些连接的DNA片段，可以确定在细胞核中物理上接近的染色体区域。

4C技术（Circular Chromosome Conformation Capture）

4C技术是3C技术的一个扩展，它允许从单一基因座（锚点）出发，识别与其相互作用的所有染色体区域。这使得科学家能够制作特定基因座的“相互作用图”，揭示它与基因组其他部分的空间关系。

主要步骤：

1. **交联**：与3C一样，首先使用甲醛固定细胞，使DNA和蛋白质交联。
2. **消化与连接**：使用限制酶切割DNA，然后连接成环状结构。
3. **反交联与扩增**：反交联后，使用特定于锚点的引物进行PCR扩增。
4. **测序与分析**：通过测序确定与锚点相互作用的染色体区域。

5C技术（Chromosome Conformation Capture Carbon Copy）

5C技术进一步扩展了3C技术，允许同时检测数以万计的DNA-DNA相互作用。5C通过使用大量的特异性引物混合物来检测特定染色体区域（如一个染色体臂）的所有潜在相互作用。

主要步骤：

1. **交联、消化与连接**：与3C和4C相同的前处理步骤。
2. **特异性引物混合物**：使用成对的引物，对连接的DNA片段进行选择性扩增。
3. **扩增与测序**：大规模扩增并通过测序分析。
4. **数据处理**：利用生物信息学工具处理和解释相互作用数据。

在解释5C技术结果时，我们通常关注的是染色体区域间的相互作用，尤其是在特定的基因座附近。以α-珠蛋白（a-globin）基因为例，当我们在GM12878（一种人类淋巴细胞系）和K562（一种白血病细胞系）细胞中研究这个基因座时，我们关注的是在这些细胞类型中α-珠蛋白基因座的活跃状态和其周围大约500kb区域的染色质相互作用。

α-珠蛋白基因座的活跃与静止状态

- **GM12878细胞中的α-珠蛋白基因座**：在这种细胞类型中，α-珠蛋白基因座可能是非活跃的或静止的，这意味着该基因在这些细胞中不是主要的表达基因。
- **K562细胞中的α-珠蛋白基因座**：相比之下，在K562细胞中，α-珠蛋白基因座是活跃的，这意味着它在这些细胞中是被积极表达的。

5C结果的解释

当使用5C技术分析这两种细胞中α-珠蛋白基因座及其周围500kb区域时，我们可以观察到以下几点：

1. **相互作用模式**：比较两种细胞类型中α-珠蛋白基因座的相互作用模式。在K562细胞中，由于基因的活跃状态，可能会观察到更多与基因表达相关的染色体区域的相互作用。
2. **染色质结构差异**：在GM12878和K562细胞中，α-珠蛋白基因座及其周围区域的染色质结构可能存在显著差异，这反映了不同细胞类型中基因表达调控的复杂性。

Hi-C技术的基本原理

1. **交联（Cross-linking）**：首先，使用甲醛或其他交联剂处理细胞，将空间上接近的DNA片段和蛋白质固定在一起。这一步骤“冻结”了染色质在细胞核内的三维结构。
2. **消化（Digestion）**：然后使用限制性内切酶切割交联的DNA，产生大量的碎片。
3. **连接（Ligation）**：这些DNA片段经过适当处理后，被连接酶连接成环状。只有在原始细胞核中空间接近的DNA片段才能有效地连接。
4. **反交联（De-crosslinking）和扩增（Amplification）**：接下来，将交联反转，释放DNA，并通过PCR等方法放大。
5. **测序（Sequencing）和数据分析（Data Analysis）**：最后，通过高通量测序技术对连接的DNA片段进行测序，并使用生物信息学方法分析数据，构建整个基因组的三维结构图。

Hi-C技术是3C的一个扩展，可以全面地分析整个基因组内所有区域之间的相互作用，而不仅限于特定的基因区域。

- **全基因组覆盖**：Hi-C技术能够捕获整个基因组范围内的染色体-染色体之间的相互作用，提供全面的3D基因组地图。
- **操作步骤**：与3C类似，Hi-C也包括交联、切割、连接等步骤，但在连接之后，通常使用高通量测序而不是PCR来分析整个基因组的DNA-DNA相互作用。
- **应用**：用于研究基因组的整体三维结构，包括染色质领域（如TADs）、染色体领域（如A/B区）等，以及这些结构如何影响基因表达和细胞功能。

Hi-C技术输出图像是一种常见的基因组交互作用热图。这种热图通常用于表示染色体内部或不同染色体之间的相互作用频率。

图像特征解释

1. **横纵坐标表示染色体**：在您描述的Hi-C热图中，横纵坐标均表示同一个染色体（在这个例子中是染色体14）。这意味着图像显示的是染色体14上不同区域之间的相互作用。
2. **左上到右下的对角线**：这条对角线代表染色体上与自身相互作用的区域。对角线上深红色的区域表示DNA序列间距离较近，相互作用频率高。这通常是染色体内部

紧密相互作用区域的特征，反映了染色体的局部结构和折叠。
3. **方形区域**：沿着对角线的方形区域表示距离相近的染色体区域之间存在高频率的相互作用。这些方形区域通常表示染色质领域或拓扑关联域（TADs），这些是基因组内高度自相互作用的区域，对基因表达调控具有重要意义。
4. **左下到右上的对角线颜色较浅**：这表明在染色体的远端区域之间的相互作用频率较低。在Hi-C热图中，距离较远的染色体区域通常显示为颜色较浅的区域，因为它们在空间中的接触几率较低。

图像的不同展示方式

1. **整个正方形展示**：在一些Hi-C热图中，整个正方形被展示出来，这样可以同时观察染色体的所有内部和外部相互作用。这种展示方式可以同时显示染色体的长臂和短臂之间的相互作用。
2. **转换为三角形展示**：另一种常见的展示方式是将正方形热图转换为一个三角形。这是通过只展示对角线上方或下方的一半来实现的。由于染色体内部相互作用的对称性（即图像的左上半部分和右下半部分基本相同），这样做可以简化图像而不丢失信息，使得图像更加清晰易读。

ChIA-PET 技术

ChIA-PET（Chromatin Interaction Analysis by Paired-End Tag Sequencing）是一种高级分子生物学技术，用于研究染色质之间的相互作用及其与特定蛋白质的关联。这种技术结合了染色质免疫沉淀（ChIP）和3C（Chromosome Conformation Capture）技术，使研究者能够同步检测染色质相互作用和与之相关的特定DNA结合蛋白。

ChIA-PET的基本步骤包括：

1. **交联**：使用甲醛或其他交联剂将蛋白质与DNA相互连接。
2. **染色质免疫沉淀**：使用特定抗体沉淀与目标蛋白（例如CTCF）结合的DNA区域。
3. **3C步骤**：切割并连接交联的DNA，形成DNA连接点，这些连接点代表物理上接近的染色质区域。
4. **测序**：通过高通量测序技术对连接点进行测序。
5. **数据分析**：分析测序数据，识别特定蛋白（如CTCF）介导的染色质-染色质相互作用。

空间分辨率

- **空间分辨率**：指的是在染色质相互作用分析中，能够区分两个独立相互作用事件的最小物理距离。在ChIA-PET实验中，这通常受限于所使用的限制性内切酶识别位

点的频率。

- **100bp左右的分辨率**：意味着ChIA-PET技术能够区分约100个碱基对距离的相互作用。这是一个相对高的分辨率，使得实验能够精细地分析染色质结构和相互作用。

RNA与染色质作用

GRID-seq（Global RNA Interactions with DNA by deep Sequencing）是一种先进的生物技术，用于在全基因组范围内检测RNA和染色质之间的相互作用。通过利用这项技术，科学家们能够在细胞内揭示RNA分子（包括但不限于mRNA、非编码RNA等）与DNA的物理相互作用，从而绘制出全局的RNA-染色质交互组（RNA–chromatin interactome）图谱。

GRID-seq技术的主要特点和步骤：

1. **交联**：首先使用甲醛将RNA分子与DNA及相关蛋白质在细胞内交联固定，以捕获它们在生理条件下的相互作用。
2. **RNA和DNA的分离与连接**：然后将RNA与DNA分离，并通过特定的方法连接这些RNA分子的末端到相应的DNA位点。
3. **逆转交联与纯化**：接下来逆转交联处理，释放RNA-DNA杂交分子，并进行纯化。
4. **测序**：最后，使用高通量测序技术对这些RNA-DNA杂交分子进行测序。
5. **数据分析**：通过生物信息学分析，绘制出全局的RNA-染色质交互图谱，揭示RNA分子与基因组特定区域的相互作用模式。

4D基因组学

基因表达和调控随着时间的变化而变化，研究时间动态变化下的基因组三维结构和功能，称之为4D基因组学

十二、基因组突变

医学基因组学

全外显子测序，全基因组测序（WES，WGS） 基于序列技术的医学检测技术 无创产前检测 NIPT 植入前检测 PGT 症状前检测 PST 精准医学和 21 世纪的医学 法医检测（亲子鉴定等）

单基因病的定位

gene mapping: 确定一个确切的表型相关的基因的位置。

1. 遗传分析定位 linkage analysis
2. 细胞遗传学定位
3. 原位杂交 DNA FISH
4. 外显子组测序

遗传分析原理是利用家系追踪。

基因组突变分析

1. 点突变与 InDel
2. 结构变异 SV
3. 基因扩增 CNV

结构变异指的是拷贝数改变，朝向改变（发生倒换），基因转位。

GWAS 分析

GWAS（全基因组关联研究）是一种研究方法，用来识别基因组中与特定疾病、性状或生物学特征相关联的遗传变异。这种方法通过比较不同个体（例如疾病患者和健康对照组）的基因组序列，寻找频率在两组之间显著不同的遗传标记，通常使用的遗传标记是SNP。GWAS不依赖于关于疾病或性状的先验生物学假设，因此能够揭示之前未知的遗传风险因素。通过检测群体中成千上万个SNP的频率分布，GWAS能够识别出与特定性状或疾病相关联的遗传区域。这些关联并不直接表明因果，但可以为进一步研究遗传机制和病理过程提供线索。

GWAS 技术通过比较无关个体：（ 1 ）患者和正常人群；（ 2 ）随机人群队列的基因组多态位点，找出疾病特异性的遗传标记，从而识别出某种疾病或表型的发病机制或相关位点。理论上，检测到的多态位点越多，识别关键位点的可能性就越大。

1. 每个**SNP**是一个独立测试：

- SNP（单核苷酸多态性）是指在某个基因位点上，不同个体间DNA序列的一种最小的差异，通常是单个核苷酸的变化。在GWAS中，研究者会检测成千上万个SNP，以找出与特定疾病或特征相关的SNP。
- 例子：假设我们在研究心脏病，我们会检查参与者基因组的每一个SNP，看看哪些SNP在心脏病患者中出现得更频繁。

2. 通过比较病例和对照组中每个等位基因的频率来测试关联性：

- 等位基因是指一个基因位点上的不同版本。在GWAS中，研究者比较患有某种疾病的人（病例）和健康人（对照组）在每个SNP位点的等位基因频率。
- 例子：如果某个SNP的一个特定等位基因在心脏病患者中出现得比在健康人中更频繁，那么这可能表明这个等位基因与心脏病有关。

3. **Odds Ratio（比值比）- 关联强度**：

- 比值比是一种表示特定等位基因出现在病例组和对照组的比例之比的统计量。它用于衡量特定等位基因与疾病之间关联的强度。
- 例子：如果一个SNP的特定等位基因在心脏病患者中的出现比率是健康人的2倍，那么它的比值比是2，表明携带这个等位基因的人患心脏病的风险是普通人的两倍。

4. **P值：卡方检验**：

- P值是一种衡量结果发生的概率的统计指标。在GWAS中，卡方检验用于确定特定SNP的等位基因频率在病例组和对照组间的差异是否具有统计学意义。
- 例子：如果一个SNP的P值非常小（比如小于0.05），这意味着病例组和对照组在这个SNP位点的等位基因频率差异极有可能不是偶然发生的，而是真正与疾病有关。

多重检验校正是统计学中处理多次假设检验可能带来的错误发现率增加的一种方法。在生物学和其他科学研究中，我们经常一次性测试许多假设，例如在全基因组关联研究（GWAS）中检测成千上万个单核苷酸多态性（SNP）。这时，需要使用多重检验校正来控制错误发现的概率。

1. 多重假设检验问题：

- 在传统的假设检验中，我们拒绝零假设（即没有效应或差异）的标准是观察到的数据在零假设下出现的概率很低（例如P值小于0.05）。
- 当同时检验多个假设时（比如检测多个SNP与疾病的关联），即使零假设是真的（即实际上没有关联），观察到罕见事件（即低P值）的机会也会增加。这意味着错误拒绝零假设（即错误地认为有关联）的风险增加。

2. 假发现率（**False Discovery Rate, FDR**）：

- FDR是一种在进行多重比较时衡量类型I错误（错误拒绝真实的零假设）的比率的方法。
- 通过控制FDR，研究者可以减少由于多重比较导致的错误发现。

3. **Bonferroni** 校正：

- Bonferroni校正是一种常见的多重检验校正方法。它通过将个别假设的显著性水平从α调整为α/m来进行校正，其中α是整体希望达到的错误率（比如0.05），m是假设的数量。
- 例子：如果有20个假设需要测试，并且希望整体错误率保持在0.05，那么每个假设的显著性水平将被调整为0.05/20=0.0025。这意味着只有当P值小于0.0025时，我们才认为该假设显著，这样做可以减少由于多重比较导致的错误发现。

曼哈顿图（Manhattan Plot）是一种在全基因组关联研究（GWAS）中常用的散点图，用于展示与某种特质或疾病显著关联的单核苷酸多态性（SNP）。这种图的命名来源于其独特的外观，类似于曼哈顿的天际线。

1. **X**轴：

- X轴代表基因组坐标，通常是按照染色体的顺序排列。每个染色体上的SNP都按照其在染色体上的位置被展示出来。

2. **Y**轴：

- Y轴显示的是每个SNP关联性的负对数P值（-log10(P值)）。负对数转换的目的是为了更直观地展示小P值。因为在GWAS中，小P值（比如小于0.05或更严格的阈值）表示SNP与研究特质的关联具有统计学意义。
- 例如，如果一个SNP的P值为1×10−151×10−15，其负对数转换值为15。这意味着该点在Y轴上的位置会很高，表示这个SNP与特质的关联非常显著。

3. 图像解读：

- 曼哈顿图上的点代表SNP。点的高度（在Y轴上的位置）表示了其与特定特质的关联强度。越高的点表示更小的P值，即更强的统计学意义。
- 通常，这种图会有一个设定的阈值线。超过这个阈值线的点被认为是显著的，可能与特质相关。
- 每个染色体的SNP通常用不同的颜色表示，以便于区分。

GWAS 优势． 全基因组关联分析。 对多个个体在全基因组范围的遗传变异（标记）多态性进行检测，获得基因型。 将基因型与可观测的性状，即表型，进行群体水平的统计

学分析。 根据统计量或显著性 P 值筛选出最有可能影响该性状的遗传变异（标记），挖掘与性状变异相关的基因． GWAS 优势： 分辨率高（单碱基水平）。研究材料来源广泛，可捕获的变异丰富。节省时间

有了GWAS结果，可以针对性的制作药物。

EWAS

EWAS （Epigenome-wide association study，表观基因组关联分析）．与 GWAS 形成互补，将表观遗传学变异和复杂疾病或性状进行关联，在表观遗传学层面对复杂疾病的致病原因或性状关联进行解读，找到与致病原因相关或复杂性状相关的表观遗传学变异位点。．通过检测整个基因组成千上万特异表观修饰差异（比如 DNA 核苷酸上甲基分布），来鉴别出疾病中常见的表观突变或与复杂性状密切相关的表观修饰。

癌症基因组学

定义：基因组学中一门新兴的子学科，通过高通量测序技术将基因与癌症关联起来。 目标：通过鉴定新的原癌基因或者抑癌基因来为癌症诊断、癌症临床结果预测和癌症靶标治疗提供新的方法。癌症基因组学的应用导致了诸如伊马替尼、曲妥珠单抗和安维汀等的癌症靶标治疗的成功。

癌症由于原癌基因和抑癌基因突变导致。例如著名p53是抑癌基因，调节细胞周期。Ras癌基因家族。

癌症的精准医学

精准医学：通过分析癌症基因组学信息为肿瘤患者量身定制更好的诊断和治疗策略现已开发出多种方法来对抗癌症，例如：抑制癌细胞异常生长和生存的酶．阻断癌细胞异常基因表达．停止癌细胞中过度驱动的分子信号通路

作物基因组学

研究作物的基因组。 GWAS分析作物的病害基因。

寻找基因组突变

变异检测（Variant Calling）是基因组学中的一个核心过程，用于从高通量测序数据（如来自下一代测序技术的数据）中识别基因组序列变异，如单核苷酸多态性（SNP）和插入缺失（Indels）。以下是变异检测的一般策略，包括对测序读段（Reads）的处理和质量评估方面的考虑：

1. **Reads堆叠在每个感兴趣的碱基上**：

- 首先，测序得到的读段（Reads）需要与参考基因组进行比对（Mapping）。这一步是为了确定每个读段在基因组上的确切位置。
- 然后，这些读段在基因组上堆叠起来。对于基因组中的每个位置，都会有多个读段覆盖。这样可以观察到在特定位置上各个读段的碱基是否一致。

2. **每个碱基的质量和映射质量**：

- 每个读段中的每个碱基都有一个质量分数，这个分数表明了该碱基被正确识别的置信度。质量分数高意味着更大的置信度。
- 除了碱基质量，读段的映射质量也很重要。映射质量是指读段与参考基因组比对的可信度。高映射质量意味着读段很可能准确地映射到了正确的位置。

3. **变异检测**：

- 在完成读段的堆叠和质量评估后，接下来的步骤是识别变异。这通常涉及到比较堆叠的读段和参考基因组序列。
- 如果在某个位置，大多数读段显示的碱基与参考基因组不同，且这些读段的质量都很高，那么这可能表明这个位置存在一个变异。

4. **变异的确认和注释**：

- 一旦检测到潜在的变异，接下来就是确认这些变异是否真实。这可能涉及到查看特定位置的读段深度（即覆盖该位置的读段数量），以及考虑可能的测序错误或者比对误差。
- 确认变异后，通常会对这些变异进行注释。注释包括确定变异的类型（如SNP、Indel）、变异的位置（位于编码区、内含子、调控区等）、以及可能的功能影响（如是否导致氨基酸改变）。

5. **质控和过滤**：

- 变异检测的另一个重要方面是质量控制和过滤。这包括去除质量较低的读段，排除映射质量低的区域，以及可能的重复序列区域，因为这些区域更容易发生比对错误。

- 此外，还可能根据特定的研究需求或标准来过滤变异。例如，可能只关注那些在特定群体中罕见的变异，或者只关注已知与疾病相关的变异。

NGS data 由于测序时可能是从一条染色体上测，可能是另一条。

1. **P(reads|A/A, read mapped) = P(C observed | A/A, read mapped)** :

- 这个公式表示在基因型为A/A（即该位点上两个等位基因都是A）的情况下，观测到C碱基的概率。这可能反映了测序错误或者其他技术误差，因为在A/A的情况下不应该观测到C。

2. **P(reads|A/C, read mapped) = P(C observed | A/C, read mapped)** :

- 这个公式计算的是在杂合子基因型A/C（即该位点上一个等位基因是A，另一个是C）的情况下，观测到C碱基的概率。这种情况下观测到C是符合预期的，因此这个概率通常会比A/A情况下的概率高。

3. **P(reads|C/C, read mapped) = P(C observed | C/C, read mapped)** :

- 最后这个公式代表的是在基因型为C/C（即该位点上两个等位基因都是C）的情况下，观测到C碱基的概率。这种情况下，观测到C是完全符合预期的，因此这个概率通常是最高的。

而NGS测序测出多个读段后可以根据观察值推测基因型，即贝叶斯公式。即多次重复测序可能测到了3个C，两个A，参考基因组上是C，问你测序的这个生物这个碱基的基因型。根据测序结果可能认为这是A/C杂合子的结果。根据先验概率和观察值写出后验概率。

GATK（Genome Analysis Toolkit）是一套由Broad Institute开发的软件，专门用于对高通量测序（HTS）数据进行基因组分析。这套工具集中的工具主要用于对来自下一代测序技术（如Illumina和PacBio平台）的数据进行处理和分析。GATK广泛应用于生物信息学和基因组学研究，尤其在变异检测和基因型分析方面非常流行。以下是GATK的一些关键特点和应用：

1. 变异检测 :

- GATK包含多个工具和流程，用于从测序数据中检测单核苷酸多态性（SNPs）和插入缺失（Indels）等变异。它提供了一系列标准化流程，帮助研究者有效地识别遗传变异。

2. 数据预处理 :

- GATK包括用于数据预处理的工具，如读段映射后的数据质量控制、重复序列的标记、局部重排算法等。这些步骤有助于提高变异检测的准确性。

3. 精确和高效：

- GATK以其高精度和效率而闻名。它通过使用复杂的统计模型来提高变异检测的准确性，并优化算法以处理大规模数据集。

4. 灵活性和可扩展性：

- GATK可以自定义和扩展，支持各种分析流程。它提供了丰富的选项和参数，允许研究者根据自己的需求调整分析。

5. 社区和文档支持：

- GATK拥有一个活跃的用户社区和详细的文档，这对于初学者和专家都非常有用。Broad Institute定期更新其文档，并提供培训和支持。

6. 应用范围广泛：

- GATK不仅用于基础的基因组学研究，还广泛应用于临床研究和精准医疗领域，帮助科学家和医生理解遗传变异与疾病之间的关系。

十三、单细胞组学

单细胞组学是一种研究单个细胞功能和分子状态的技术，它使我们能够深入了解细胞的内部工作机制。这个领域关注于识别和解析单个细胞内的分子组成，如DNA、RNA和蛋白质，从而揭示它们如何影响细胞的功能和状态。（关注转录组）

在单细胞组学中，“不同功能”指的是单个细胞在生物体中扮演的不同角色。例如，一些细胞可能专门用于传递信息（如神经细胞），而其他细胞则可能参与免疫反应或组织修复。通过单细胞组学，科学家可以详细地了解这些不同细胞是如何执行其特定功能的。

“分子状态”则涉及细胞内部各种分子（如RNA和蛋白质）的活动和水平。这些分子状态可以反映细胞的健康状况、活动阶段（如细胞周期的不同阶段）或对外界刺激的响应。例如，一个细胞在感受到炎症信号时可能会表达不同的基因和蛋白质，与它在静态或非应激状态下的表达模式截然不同。

单细胞组学的技术，如单细胞测序，使科学家能够在单个细胞水平上分析这些复杂的过程，从而提供比传统的组织水平或多细胞水平研究更精细的生物学洞察。

单细胞测序（Single Cell Sequencing）

1. **分辨率**：最大的优势在于其高分辨率。它能够揭示单个细胞内的基因表达模式，这对于理解细胞间的异质性至关重要。
2. **应用**：特别有助于研究那些由细胞异质性驱动的生物过程，如癌症发展、组织发育、免疫响应等。
3. **数据复杂性**：产生的数据比较复杂，需要特殊的生物信息学工具来分析和解释。
4. **成本和技术要求**：相对于群体测序，单细胞测序成本更高，技术要求也更严格。

群体测序（Bulk Sequencing）

1. **分辨率**：群体测序提供的是多个细胞混合样本的平均信号，无法揭示单个细胞的特性。
2. **应用**：更适合于研究整体水平上的基因表达模式，例如在特定处理或条件下细胞群体的响应。
3. **数据处理**：数据分析相对简单，因为它处理的是群体平均值，而非单个细胞的复杂数据。
4. **成本和技术要求**：成本较低，技术要求也不如单细胞测序严格。

单细胞组学的应用

1. **鉴定异质细胞群体**：在许多生物过程中，即使是在看似均一的细胞群体中，也存在着显著的细胞间异质性。使用单细胞测序，可以鉴定并描述这些异质细胞群体的特征，这对于理解组织功能和疾病发展至关重要。
2. **发现新的细胞标志物和调控途径**：单细胞数据可以揭示特定细胞类型或状态特有的分子标志，如特定蛋白质或RNA分子。这些新发现的标志物和调控途径对于疾病诊断和治疗策略的开发极为重要。
3. **揭示新的细胞类型、细胞状态和稀有细胞类型**：通过单细胞分析，可以发现之前未被识别的细胞类型或状态，包括在生物体中数量极少但可能具有重要生物学功能的稀有细胞。
4. **重建发育层级和揭示谱系关系**：单细胞技术能够追踪细胞在发育过程中的转化，从而重建发育过程中的细胞谱系关系。这对于理解组织发育和再生、以及癌症等疾病的发生机制具有重要意义。
5. **对健康和疾病组织及器官进行分析**：单细胞数据可以用于比较健康和疾病状态下的组织和器官，揭示疾病发展的分子机制。这对于疾病的早期诊断、预后判断和个性化治疗策略的制定具有极大的价值。通过分析病变组织中的细胞组成和功能状态，可以更好地理解疾病的本质和进展。

单细胞测序技术

单细胞测序技术（scRNA-seq）一般方法

1. 分离单个细胞

- **从细胞群体中分离单个细胞**：这通常通过流式细胞仪、显微操作或基于微流控的技术实现。目的是从一个多细胞的样本中获得单独的、孤立的细胞。

2. 提取单细胞转录组

- **RNA的提取和反转录**：从每个单独的细胞中提取RNA，并将其转录成cDNA。这一步是必要的，因为DNA比RNA更稳定，更适合后续的测序步骤。

3. 扩增和准备测序

- **cDNA的扩增**：由于单个细胞中的RNA量很少，因此需要扩增cDNA以获得足够的材料进行测序。

4. 测序和数据分析

- **高通量测序**：使用高通量测序技术（如Illumina测序平台）对扩增的cDNA进行测序。
- **生物信息学分析**：对生成的大量数据进行分析，以识别基因表达模式和细胞类型。

Drop-seq方法

Drop-seq是一种大规模单细胞转录组测序方法。它通过将单个细胞和一个独特的分子条形码封装到微液滴中，实现高通量的单细胞基因表达分析。

原理和步骤

1. 单细胞和微珠的制备

- 细胞样本被制备成悬浮液。
- 微珠（微小的聚合物珠子）涂覆有独特的分子条形码和寡核苷酸序列，这些条形码在后续分析中用于标记每个细胞的RNA。

2. 生成微液滴

- 在Drop-seq中，微流控技术被用于生成包含单个细胞和单个微珠的微液滴。每个微液滴都是一个独立的反应单位，确保每个细胞的RNA只与一个微珠接触并被其上的独特分子条形码标记。

3. 逆转录

- 在微液滴中，细胞被裂解，释放出RNA。这些RNA分子与微珠上的寡核苷酸序列结合，然后进行逆转录成cDNA。此过程中，每个cDNA分子会获得与其原始细胞相对应的分子条形码。

4. **打破微液滴并收集cDNA**

- 完成逆转录后，微液滴被打破，释放出带有条形码的cDNA。这些cDNA随后被收集并准备用于测序。

5. **扩增和测序**

- 为了获得足够的材料进行测序，cDNA经过扩增。之后，通过高通量测序技术（如Illumina平台）进行测序。

6. **数据分析**

- 测序产生的大量数据需要经过复杂的生物信息学处理。首先，读取（reads）会基于其分子条形码被追踪回原始的细胞。然后，对这些读取进行比对、定量和其他分析处理，以识别和比较不同细胞的基因表达模式。

Chromium Next GEM Single Cell 3’ Reagent Kits v3.1 是由10x Genomics公司开发的一套用于单细胞转录组测序的试剂盒。这种试剂盒利用了10x Genomics独特的微流控技术和GEM（Gel Bead-in-Emulsion）技术，使研究人员能够高效地对单个细胞的RNA进行捕获和测序。以下是关于这个试剂盒的一些关键信息：

关键特征和工作原理

1. **GEM技术**

- GEM技术是通过将单个细胞和一颗带有寡核苷酸的凝胶微珠（Gel Bead）封装在一个微液滴（Emulsion）中来实现的。每个凝胶微珠带有独特的分子条形码，允许后续将测序数据追踪回其来源的单个细胞。

2. **单细胞捕获和RNA反转录**

- 单细胞被分离并与凝胶微珠一起封装在微液滴中。在这些微液滴内，细胞被裂解，释放出RNA，随后RNA被微珠上的寡核苷酸捕获并进行逆转录成cDNA。

3. **分子条形码和UMI（Unique Molecular Identifiers）**

- 每个凝胶微珠上的寡核苷酸不仅包含用于识别源细胞的分子条形码，还包括UMI，用于识别和量化单个RNA分子，从而减少PCR扩增过程中的偏差。

4. 高通量和高分辨率

- 试剂盒设计用于处理数千到数万个

细胞，提供高通量的数据输出。它能够捕获并分析多种细胞类型和状态，从而提供高分辨率的单细胞基因表达数据。

5. 数据分析和生物信息学工具

- 10x Genomics提供了专门的生物信息学软件（如Cell Ranger），用于处理、分析和可视化由Chromium Next GEM Single Cell 3’ Reagent Kits v3.1生成的数据。

Smart-seq2是一种用于单细胞RNA测序的技术，它是Smart-seq（Switching Mechanism At 5' end of RNA Template）方法的改进版本。这种技术特别强调捕获全长RNA分子的能力，从而提供更加全面和详细的基因表达信息。Smart-seq2在生物医学研究中尤其重要，因为它能够提供关于单个细胞水平上基因表达的深入洞察。

Smart-seq2的主要特点和步骤

1. 高效的逆转录

- Smart-seq2使用一种特殊的逆转录酶和一种优化的引物混合物，这使得从RNA到cDNA（互补DNA）的转换更加高效和完整。这一步骤是捕获全长RNA的关键。

2. 全长**RNA**的捕获

- 与其他单细胞测序技术相比，Smart-seq2能够更有效地捕获RNA分子的全长。这意味着它能够提供更全面的基因表达信息，包括基因的各个亚型和剪接变体。

3. 提高数据质量

- 通过提高逆转录的效率和准确性，Smart-seq2产生的cDNA具有更高的质量和更低的偏差，这对于后续的数据分析至关重要。

4. 扩增和测序准备

- 完成逆转录后，生成的cDNA经过扩增，然后进行高通量测序。Smart-seq2的这一步骤被优化以最小化扩增偏差和提高扩增效率。

整合单细胞多组学

整合单细胞多组学（Integrated Single-Cell Multiomics）是一种新兴的生物信息学技术，它结合了单细胞水平上的多种组学数据，如基因组（genomics）、转录组（transcriptomics）、蛋白质组（proteomics）、表观遗传组（epigenomics）等。这种方法允许科学家在单个细胞水平上同时分析不同的生物分子层面，提供了一个更全面的视角来理解细胞的生物学特性和功能。

"Barcoding Chromatin in Individual Cells with Oil Droplets" 是一种用于研究单细胞层面上染色质结构的先进技术。这个方法利用微液滴技术和分子条形码来标记和分析单个细胞中的染色质信息。它在单细胞表观遗传学研究中具有重要意义，尤其是在探索细胞内染色质动态和细胞状态之间的关系方面。

基本原理和步骤

1. 微液滴技术

- 使用油性微液滴作为反应空间，每个微液滴内包含单个细胞或细胞的一部分。这种技术允许将来自单个细胞的染色质信号隔离和捕获。

2. 染色质分离和处理

- 细胞被裂解，释放染色质。然后染色质被特定的酶（如微球核酸酶）处理，生成较短的染色质片段。

3. 分子条形码引入

- 将带有独特分子条形码的寡核苷酸引入到每个微液滴中。这些条形码用于标记染色质片段，确保后续分析可以追踪到它们来自哪个细胞。

4. 测序和数据分析

- 带有条形码的染色质片段被提取、扩增并进行测序。利用生物信息学工具，可以通过分子条形码将染色质数据追溯到特定的细胞。

“一个理想的实验工作流程将观察细胞的所有方面，包括其完整的分子状态历史、空间位置和环境相互作用。”

多模态和整合方法用于单细胞分析是一系列技术的集合，它们允许在单个细胞水平上同时获取多种生物学数据。这些方法结合了不同的生物信息学技术，如单细胞转录组学（scRNA-seq）、单细胞表观遗传学（如单细胞ATAC-seq或单细胞DNA甲基化测序）、单细胞蛋白质组学等，以提供关于单个细胞的全面生物学信息。

单细胞多模态测量

a. 基于荧光的蛋白质和RNA测量

- 在这种方法中，使用荧光标记技术来测量单个细胞中特定蛋白质的水平和特定RNA的表达。例如，可以使用荧光原位杂交（FISH）来检测特定RNA分子，同时使用荧光标记的抗体来检测目标蛋白。

b. 溶解并分离不同的细胞组分

- 这种策略包括将单个细胞裂解并将其不同的生物分子组分（如蛋白质、RNA等）分离，以便进行独立分析。这使得可以从同一细胞获取多种类型的生物学信息。

单细胞多模态测量 2

细胞表面蛋白

- 通过将多聚腺苷酸（polyadenylated）标签连接到抗体上，可以对细胞表面蛋白进行特异性检测。这些带有标签的抗体结合到细胞表面特定蛋白，通过测定这些条形码，可以推断出蛋白的存在和丰度。

基因敲除信息

- 通过计数sgRNA，可以获得基因编辑或基因敲除的信息。sgRNA的数量和类型反映了细胞内特定基因的编辑状态。

细胞谱系编码

- 通过分析细胞内的条形码序列中累积的编辑事件，可以追踪细胞谱系和发育历史。

单细胞多模态测量 3

体细胞突变

- 通过分析单个细胞的DNA，可以检测特定的体细胞突变。这种分析对于理解肿瘤进化、遗传异质性以及细胞个体之间的遗传差异非常重要。

保留内含子

- 在RNA测序数据中，可以检测到保留的内含子。这些保留的内含子可以用来估算转录丰度的变化速率，即所谓的RNA速度（RNA velocity）。RNA速度是指在特定时间点RNA分子的合成和降解的动态平衡，可以用来推断细胞状态的变化趋势。

单细胞测序应用于病毒追踪。

计算分析

在单细胞RNA测序（scRNA-Seq）研究中，计算分析是一个关键步骤，它涉及使用各种生物信息学工具和算法来处理和解释大量的测序数据。这些数据通常是高维的，并且包含了来自成千上万个单独细胞的复杂信息。

scRNA-tools数据库

- **目的**：scRNA-tools数据库是一个收录用于分析单细胞RNA测序数据的工具和软件的目录。
- **功能**：它提供了一个广泛的资源列表，包括数据预处理、标准化、维度降低、聚类、差异表达分析等工具。

关键预处理步骤

1. 对齐和分子计数（**Alignment and Molecular Counting**）
    - 这一步涉及将测序得到的读段（reads）对齐到参考基因组，以识别它们来源的基因。然后，对每个细胞中检测到的每个基因的RNA分子进行计数。
2. 细胞过滤和质量控制（**Cell Filtering and Quality Control**）
    - 移除低质量的细胞数据，如那些具有异常高或低基因表达数量的细胞，或者那些基因表达过于均匀或过于离散的细胞。
3. 双体评分（**Doublet Scoring**）
    - 在单细胞实验中，有时两个细胞可能一起被捕获，形成“双体”（doublet）。双体评分是识别这些双体的过程，以便从后续分析中排除它们。
4. 细胞大小估计（**Cell Size Estimation**）
    - 估计每个细胞的大小，这可能影响RNA分子的计数。在某些情况下，需要根据细胞大小调整RNA计数。
5. 基因方差分析（**Gene Variance Analysis**）
    - 识别哪些基因在不同细胞间表现出显著的表达变异。这些基因可能在后续的聚类和差异表达分析中特别重要。

关键分析步骤

1. 降至中等维度空间（**Reduction to a Medium-Dimensional Space**）
    - 通过主成分分析（PCA）等方法将数据从高维空间降至中等维度，以简化数据结构并减少计算复杂性。
2. 流形表示（**Manifold Representation**）
    - 使用t-SNE、UMAP等技术进一步降维并在二维或三维空间中可视化数据，揭示细胞之间的复杂关系。

3. 聚类和差异表达（**Clustering and Differential Expression**）
    - 基于表达模式将细胞分组（聚类），然后识别不同细胞群体（如不同细胞类型）间的表达差异。
4. 轨迹推断（**Trajectories**）
    - 识别细胞的发育或分化路径。通过分析细胞在低维空间中的排列，可以推断它们的发展或分化轨迹。
5. 速度估计（**Velocity Estimation**）
    - 估计每个细胞的基因表达变化速度，这有助于理解细胞状态的动态变化。

单细胞RNA测序（scRNA-seq）的数据分析流程包括多个关键步骤，每个步骤都对最终结果至关重要。以下是这个流程的概述：

Step 1: Read Processing and Quantification

1. **Read Processing**
    - 这一步涉及将测序仪生成的原始数据（通常是FASTQ格式的文件）通过基因组比对软件（如STAR或Hisat2）与参考基因组进行比对。目的是确定每个读段（read）的来源（即它来自哪个基因）。
2. **Quantification**
    - 接着，对每个单细胞和每个基因的表达水平进行量化。这通常是通过计算每个基因在每个细胞中的读段数（或UMI计数）来实现的。

Step 2: Quality Control (QC) and Filtering

1. **Quality Control**
    - 对单细胞数据进行质量控制，以检测和去除低质量的细胞和读段。这可能包括检查读段数、基因的检测数、线粒体基因表达的比例等。
2. **Filtering**
    - 根据质量控制的结果，过滤掉质量不佳的细胞和基因，例如去除那些检测到的基因数目过少或过多的细胞。

Step 3: Data Integration

- 这一步骤特别重要于多个样本或实验条件下的数据整合。它旨在消除技术变异和批次效应，确保不同样本之间的数据可比性。

Step 4: Clustering and Visualization

1. **Clustering**
    - 使用统计或机器学习方法（如K-means、层次聚类、DBSCAN等）对细胞进行分组。目的是基于基因表达模式将具有相似特征的细胞聚集在一起。

2. **Visualization**
    - 通过降维技术（如PCA、t-SNE、UMAP）在二维或三维空间中可视化细胞，以便更容易地识别和解释聚类结果。

Step 5: Differential Gene Expression Analysis

- 在不同细胞群体（如不同的聚类）之间进行差异表达基因分析。这有助于识别哪些基因在特定细胞类型或状态中上调或下调。

Step 6: Cell Type Annotation

- 根据差异表达基因的结果，结合已知的细胞标志物和生物学知识，对聚类的细胞进行类型标注。这一步骤是将生物学意义赋予数据分析的结果。

空间转录组学

空间转录组学（Spatial Transcriptomics）是一种新兴的生物技术，它结合了传统的转录组学分析和组织形态学，以在组织、器官或细胞水平上研究基因表达的空间分布。这项技术允许科学家不仅仅了解哪些基因在特定条件下被激活或抑制，还能了解这些基因表达事件在组织或细胞中的具体位置。

技术原理

1. **空间定位**：在传统的转录组学分析中，细胞是在被分离出组织后进行分析的，这意味着失去了它们在原始组织中的空间信息。空间转录组学通过在细胞原位（即在它们自然存在的位置）进行基因表达分析，保留了这些重要的空间信息。
2. **高通量测序**：使用高通量测序技术，可以同时分析成千上万个基因的表达。
3. **图像分析**：结合显微镜图像分析，科学家能够将基因表达数据与细胞和组织的具体结构对应起来。

10x Genomics的空间捕获技术是一种创新的空间转录组学方法，它允许研究者在组织样本中精确地映射基因表达。这种技术结合了传统的转录组学、显微成像和先进的分子生物学技术，使得可以在组织的特定位置进行基因表达分析。以下是这一技术工作原理的简要概述：

技术流程

1. 条形码点阵（**Barcoded Spots**）
    - 使用的载玻片上布满了成千上万个微小的条形码点（spots）。每个点都包含大量的捕获探针，这些探针上带有独特的分子条形码。

2. 组织样本制备
    - 取新鲜冻存（Fresh-Frozen）或福尔马林固定石蜡包埋（FFPE）的组织样本，放置在这些带有条形码的点上。
3. 组织染色和成像
    - 对新鲜冻存的组织进行染色和显微成像，以便于在后续的分子分析中参照组织的形态结构。
4. 组织固定和渗透
    - 将组织固定并使其渗透，以便释放出RNA分子。
5. **RNA捕获**
    - 组织中释放的RNA与载玻片上的捕获探针结合。由于每个探针都带有独特的条形码，因此捕获的RNA也随之带有位置信息。
6. **cDNA合成**
    - 从捕获的RNA合成带有相应条形码的cDNA。
7. 测序准备
    - 将合成的cDNA制备成测序文库，然后使用标准的DNA测序方法进行测序。

空间转录组学的限制

1. 二维空间分析：

- 组织被切成薄片（通常是微米级厚度）放置在载玻片上，因此，所获得的空间信息仅限于切片平面内。

2. 高度信息丧失：

- 在细胞裂解和RNA释放的过程中，确实会丧失细胞在垂直于切片平面的方向上的空间信息。这意味着分析得到的空间图像主要是二维的。

Visium是一种由10x Genomics开发的空间转录组学平台，它提供了两种主要的信息类型：组织图像和基因表达数据。这两种信息的结合为研究者提供了组织中基因表达的空间分布图。

组织图像

1. 成像过程：在进行RNA测序之前，首先对组织样本进行显微成像。这一步骤的目的是捕获和记录组织的形态结构，包括细胞的位置、形态和组织的整体架构。
2. 空间参照：成像结果提供了一个空间参照框架，用于后续将基因表达数据映射回其在组织中的具体位置。

基因表达数据

1. **RNA捕获和测序**：组织样本被放置在包含成千上万个微小的、带有分子条形码的捕获探针的芯片上。样本中的RNA被释放并与这些探针结合，随后进行反转录和测序。
2. **空间分辨基因表达**：由于每个探针的位置是已知的，并且每个探针带有独特的条形码，因此可以确定每个RNA分子在组织中的原始位置。这使得研究者能够构建出组织中每个基因表达的空间分布图。

华大的技术

华大自主研发的时空组学技术Stereo-seq基于DNA纳米球（DNA Nano Ball，DNB）技术开发，是一种具有高通量、超高分辨率、大视场的原位全景式技术。这项技术能够在组织、细胞、亚细胞、分子“四尺度”上同时进行空间转录组分析。它通过时空芯片捕获组织中的mRNA，并利用时空条形码（Coordinate ID，CID）还原回空间位置，实现组织中基因空间表达的检测。Stereo-seq的这种方法为深入理解细胞的基因表达及其与形态和局部环境之间的关系提供了一个强大的研究基础

工作原理

1. **DNA纳米球阵列芯片**：

- 采用DNA纳米球阵列技术，每个DNA纳米球上都编码有独特的时空条形码。

2. **原位捕获**：

- 时空芯片被设计用来捕获组织中的mRNA分子。组织样本被放置在这些芯片上，细胞中的mRNA与芯片上的探针结合。

3. **空间标识测序**：

- 通过时空条形码，每个捕获的mRNA分子都能被精确地定位到其在组织中的原始位置。

4. **建库和测序**：

- 从捕获的mRNA合成cDNA并建立测序库。

5. **时空图谱构建**：

- 使用测序数据和条形码信息，构建组织中基因表达的空间图谱。

最新科技

NanoString的CosMx Spatial Molecular Imager是一种先进的空间分子成像技术，它可以在细胞和亚细胞水平上同时分析RNA和蛋白质

总结

单细胞转录组学的关键原理

1. 细胞条形码序列（**Cell Barcode Sequence**）
    - 在单细胞测序中，每个细胞被赋予一个独特的分子条形码，用于识别和区分不同的细胞。这些条形码在样本处理过程中引入，使得在后续的数据分析中可以追踪每个RNA分子回到其原始的细胞。
2. **mRNA**序列（**mRNA Sequence**）
    - mRNA序列是从细胞中捕获并测序的，它代表了基因的表达。通过测序这些mRNA分子，可以确定哪些基因在特定细胞中被激活或表达。
3. 唯一分子标识符（**UMI**）序列
    - UMI是一种独特的序列，用于标记单个RNA分子。它帮助区分由PCR扩增过程中产生的重复序列和真实的多个RNA分子，从而准确量化基因表达水平。

空间转录组学的关键原理

1. 细胞位置条形码序列（**Cell Location Barcode Sequence**）
    - 在空间转录组学中，除了识别细胞身份外，还需要确定细胞在组织中的具体位置。细胞位置条形码是一种特殊的序列，用于将RNA分子映射回它们在组织中的原始空间位置。这使得研究者可以了解基因表达在空间上的分布。

十四、深度学习

深度学习在基因组学中有很多应用。上面讲过的几乎所有都可以。

例如，预测基因的调控元件，预测表观遗传修饰出现在基因组的哪里，寻找疾病对应的变异，基因注释，TSS（转录起始位点预测） CASP 中的alpha-fold模型。

机器学习是致力于研究如何通过计算的手段，利用经验来改善系统自身 的性能的一门学科。 定义：假设用 P 来评估计算机程序在某任务类 T 上的性能，若一个程序通过利用经验 E 在 T 中任务上获得了性能改善，则我们就说关于 T 和 P ，该程序对 E 进行了学习。

传统方法是给定数据，给定规则，让程序输出结果。 机器学习是给定数据，给定结果，让程序预测规则。（监督）

集群分析（Cluster Analysis），或简称聚类，是一种将数据对象集（或观测值）划分为多个子集的过程。每个子集被称为一个“簇”，簇内的对象彼此相似，但与其他簇中的对象不相似。通过集群分析得到的簇集合可以称为一个聚类。在这个背景下，不同的聚类方法可能在同一数据集上生成不同的聚类结果。重要的是，这种划分不是由人类执行的，而是由聚类算法完成的。因此，聚类在数据中发现以前未知的群体方面非常有用。

1. 划分方法（Partitioning Methods）

- **特点：**
- 查找互斥的、通常是球形的簇。
- 基于距离。
- 可以使用平均值或中心点（medoid）等来代表簇中心。
- 对小到中等规模的数据集有效。

2. 层次方法（Hierarchical Methods）

- **特点：**
- 聚类是一个分层的分解过程（即，多个层级）。
- 一旦发生错误的合并或分裂，无法纠正。
- 可以结合其他技术，如微聚类，或考虑对象“链接”。

3. 基于密度的方法（Density-based Methods）

- **特点：**
- 能够找到任意形状的簇。
- 簇是空间中由低密度区域分隔的高密度对象区域。
- 簇密度：每个点在其邻域内必须有最小数量的点。
- 可能过滤掉离群点（outliers）。

4. 基于网格的方法（Grid-based Methods）

- **特点：**
- 使用多分辨率网格数据结构。
- 处理时间快（通常与数据对象的数量无关，但取决于网格大小）。

机器学习的一般方法

1. 数据收集（Collecting）

- 收集原始数据，这些数据可以来自多种来源，如在线数据库、实验或现实世界的观测。

2. 数据预处理（Preprocessing）

- 包括多个子步骤，如：
  - **规范化（Normalization）**：调整数据尺度以便于处理。
  - **离散化（Discretization）**：将连续特征转换为离散值。
  - **分割（Segmentation）**：将数据分割为训练和测试数据集。
  - **标注（Labeling）**：在监督学习中为数据添加标签。

3. 模型训练（Training）

- 选择合适的模型和算法。
- 进行超参数调整来优化模型。
- 使用训练数据集训练模型。

4. 模型评估（Evaluation）

- 使用测试数据集来评估模型性能。
- 应用各种评估指标，如准确度、ROC曲线等。
- 调整模型参数以改善性能，必要时进行迭代优化。

5. 模型应用（Usage）

- 将训练和优化的模型应用于新数据。
- 进行预测或决策。
- 在实际场景中使用模型以解决问题或提供洞察。

基本的机器学习模型有：决策树，SVM，回归，朴素贝叶斯，隐马模型，随机森林，循环神经网络，LSTM，CNN。

十五、合成生物学

从解读基因组走向书写基因组，即合成生物学

合成生物学发展史

19世纪合成了非基因生物材料，尿素。20世纪中合成了氨基酸和结晶牛胰岛素。 之后开始合成核苷酸和基因。 21世纪开始合成遗传装置，如基因开关，电路，基因传感器等。以及人工合成代谢通路，使得生物制品工业化。 2002年合成了第一个病毒基因组。脊髓灰质炎病毒。

科学，是认识世界获得知识的过程 工程，是应用知识设计和改造世界的过程。

应用举例： 合成生物振荡器调控基因表达。 合成生物感光机，可以写个hello world什么的 半合成青蒿素（Semi-synthetic Artemisinin）是一种重要的抗疟疾药物，它的开发代表了合成生物学和化学工程的一大突破。 半合成青蒿素的生产涉及了合成生物学和化学合成的结合。首先，利用工程化的微生物（如酵母）生产出青蒿素的前体物质——青蒿酸。然后，通过化学方法将青蒿酸转化为青蒿素。这种方法的优势在于生产过程可控、产量高，且相对环境友好。

基因组的设计与合成

合成基因组学的研究可分为两个方面：设计和合成。 “设计”是基于基因组等学科对生命奥秘的探索，以计算机辅助的技术手段“书写”一个组件、装置乃至系统，是合成基因组的核心技术。 “合成”则是综合运用化学、基因组学、分子生物学和“遗传工程”的基因操作技术，将设计的组件、装置或系统进行实体构建，并检验其生物学功能。

Synthia 的设计和合成

1. 将蕈状支原体进行全基因组测序，对基因组序列进行重新设计（如精简了 4 Kb 的序列，加入水印（ watermark ）序列、抗生素抗性基因序列、 lacZ基因序列等）
2. 用化学方法合成所需的寡核苷酸 (Oligo)
3. 用 Gibson 策略进行组装，形成约 100 Kb 的 DNA 大片段；
4. 将 DNA 片段连接成完整的人工基因组；
5. 将人工基因组导入山羊支原体的受体细胞进行培养；
6. 在培养液中加入 X-gal ， 使得表达beta-半乳糖苷酶的人工菌呈蓝色
7. 用抗生素选择含人工基因组的细胞。

技术细节

DNA的合成需要磷酸二酯键的合成，化学方法。 DNA的组装是两段DNA留重叠区域，使用限制酶产生粘性末端。再使用连接酶。或者DNA聚合酶。

基因组组分的设计与合成

合成生物学通过改造工程细胞的代谢通路，引进新的酶或其他物质基因，形成生产“工业化产物”的完整代谢通路。将改造的代谢通路转入“公用底盘细胞”（即含“最小基因组”的人造细胞）中构建“细胞工厂”，最终获得目标天然有机物。 改造细胞原有代谢通路的一般途径包括：通过测定与某个原产某一天然产物相关的基因组区段，分析相应代谢通路，并详细解析代谢途径中已有酶的结构、特性、调控机制及底物、产物的“反馈”调控（正向调节、负向反馈”）等关系，通过计算机辅助技术构建代谢模型：解析“底盘工程细胞”现有代谢通路相关酶，特别是注意对于合成目标产物缺少哪些酶的基因、是否有于扰合成目标产物代谢通路的酶或其他因子等。通过细胞精简技术将不利于合成目标产物的成分排除掉；人工合成或从别的生物（天然产生目标物质的生物）中克隆这叫弋谢酶基因，并“转”进“底盘工程细胞”

合成完整的基因组，再进一步合成“人工细胞”是合成生物学的终极目的。最小基因组(minimal genome) 是指能够维持细胞生命在最适环境条件（最丰富的营养，最适宜的温度、湿度、酸碱度等，无外界环境压力）下生存必需的最小数目基因的集合。科学层面上，最小基因组对于研究生命起源也有着重要的意义。从科学的角度来说，生物在长期进化过程中，通过基因水平转移〈 horizontal gene tran sfer 〉等途径，不断地摄入外源基因，并将其保留在生物自身基因组内，致使基因组出现了容量渐增化、结构复杂化、功能多效化等现象。如果能够去除基因组上大量的非必需基因从而简化基因组，也许在降低了基因组复杂度的同时，也可以减小非必需代谢途径的干扰等遗传噪音(genomic noises) ，提高细胞对底物和能量的利用效率。从技术的角度来说，生产出一种或几种非蛋白质的天然有机物或自然界中不存在的有机物，最理想的是有一个类似“工程细胞”的“底盘细胞”。这个“底盘细胞”的最重要的特占是具有最大程度的“兼容性”与“通用性”“兼容性”与“通用性”首先是物理学概念，即可以接纳多个大的“代谢通路”的基因后仍能行使正常的生物学功能：其次是生物学概念：这一近乎“万能”的“底盘细胞”含有最小的生存所必需的遗传物质，其与引入的代谢途径的相互干扰达到最小化。

装置与系统

生物装置，是以一定逻辑结构组合形成的基因单元，响应输入并输出特定信号。如报告基因GFP. 生物系统，将多个装置组合在一起得到更为复杂的调控结构或生物功能，形成具有完整生物学功能的代谢或调控网络便构成了“系统”。

基本元件

主要元件有：启动子(promoter) 、 RBS （ RibosomeBinding Site ， 核糖体结合位点）、 CDS （CoDing Sequence ， 编码序列）及报告基因(reporter) 、转录终止子(terminator) 、 RNA 稳定性元件、增强子(Enhancer) 、沉默子、绝缘子等。真核生物在转录本的非翻译区中存在着 RNA 酶靶位点、小RNA 靶位点等元件。通过这些元件可以实现基因转录与翻译的精细调控。同时，基因序列中非翻译区的存在也为人工元件、装置提供了插入位点。

标准化元件 生物砖像用标准元件组装一辆汽车或一架飞机一样，用标准的生物元件组装成一个完整的生命体。成立了生物砖基金会 （BioBricks Foundation ， BBF ）以建立和维护生物元件的标准并支持标准化的研究。标准化的生物砖大多来自于 iGEM 竞赛并储存于 PartsRegistry 数据库，储存内容包括生物砖编号、类型、来源、序列、性能等在内的信息，方便使用者查询和使用。

生物砖上有不同的四个剪切位点，用于把不同的元件结合起来。 BamHI: 识别序列：BamHI识别并切割具有以下序列的DNA：5'-GGATCC-3' / 3'-CCTAGG-5'。 切割方式：它在两个G的中间进行切割，产生“粘性末端”（sticky ends）。这种末端包含了单链的DNA，可以用于连接其他具有互补粘性末端的DNA片段。 BglII: 识别序列：BglII识别并切割具有以下序列的DNA：5'-AGATCT-3' / 3'-TCTAGA-5'。 切割方式：它在A和G之间进行切割，同样产生粘性末端。

这两个的粘性末端一模一样。所以切完了可以一块用。

人可以修改密码子，合成非天然氨基酸，非天然核苷酸。修改终止密码子。让终止密码子另作它用。

合成生物学的应用： 生物传感器：芳香烃污染物；砷离子化学污染 疾病监测系统：检测和杀死绿脓假单胞菌 生物能源：丁醇；藻类生物燃料 生物计算机 DNA 信息存储系统