2023 基因组学作业

作业名称：题目 1-3

姓名：徐子扬

学院：数学与统计学院

专业：统计学

学号：320200904951

2023 年 4 月 24 日

摘要

该作业报告完成了本学期基因组学课程布置的三道题目.

作业第 1 题是“介绍 Oxford Nanopore 的原理、优缺点、应用”. 对于原理，我主要介绍了该公司推出的首款纳米孔测序设备 MinION[3] 的技术原理，以及两种常见的 Basecaller 电流信号解码算法 Nanocall[5] 和 DeepNano[7]；对于优缺点，我讨论了长读长、高通量、实时靶向测序、直接检测碱基修饰等 6 种优点，以及准确性较低、高错误率等 3 种缺点，并与第一、二代测序进行比较；对于应用，我主要讨论了填补参考基因组中的空缺、建立新的参考基因组、识别大型结构变异等 6 种可能的应用场景.

作业第 2 题是“简述我知道的 RNA 的种类和功能”. 我简要概述了 9 种 RNA 的种类和功能. 包括：mRNA（信使 RNA）在基因表达过程中负责转导遗传信息；tRNA（转运 RNA）在蛋白质合成过程中传递氨基酸；rRNA（核糖体 RNA）作为核糖体的主要成分，参与蛋白质合成；snRNA（小核 RNA）和 snoRNA（小核仁 RNA）参与 RNA 剪接和修饰；miRNA（微小 RNA）调控基因表达，影响生物进程；lncRNA（长非编码 RNA）在转录调控、基因沉默等方面发挥作用；piRNA（Piwi-interacting RNA）参与生殖细胞中的转座子沉默；circRNA（环状 RNA）作为 miRNA 的海绵分子，调控基因表达.

作业第 3 题是“泛基因组学 (Pan-genome)：内容、应用场景、研究实例”. 对于内容，我主要介绍了泛基因组学的定义以及常见概念；对于应用场景，我给出作物基因组学、育种和进化研究、研究不同品种结构变异影响基因差异表达和结合 GWAS 数据捕获更完整的遗传变异信息共 3 种应用场景；对于研究实例，我选取了 2023 年 3 月在《Nature Communications》上发表的“亚洲水稻泛基因组倒位指数”文章《Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice》[1]，从研究背景、倒位指数、系统发育树、鉴定与评估等多个方面展开讨论，最后认为泛基因组学在大规模组学数据和科学计算时代可以发挥非常重要的作用.

该作业使用 LaTeX 排版，编译所需的全部文件已经上传到`https://github.com/StatXzy7/lzu_genomics_hw`.

关键词：基因组学; Oxford Nanopore; RNA; 泛基因组学

目录

1 介绍 Oxford Nanopore 的原理、优缺点、应用

牛津纳米孔技术 Oxford Nanopore 属于第三代 DNA 测序技术，其特点是单分子测序、实时测量和较长读长，从而大幅提升测序速度和准确性. 英国的牛津纳米孔技术有限公司（Oxford Nanopore Technologies, ONT）专注于开发和销售纳米孔测序产品（如便携式 DNA 测序仪 MinION，2014）[2][3]，用于单分子的直接、电子分析.

自 2014 年提供首台纳米孔测序仪 MinION 以来，纳米孔测序技术及其在基础和应用研究中的应用得到了显著增长. 纳米孔技术的快速发展带来了单个长 DNA 和 RNA 分子测序准确性、读长和吞吐量的大幅提升. 为充分利用纳米孔长读长研究基因组、转录组、表观基因组和表观转录组，需要广泛发展实验和生物信息学方法. 纳米孔测序应用于基因组组装、全长转录本检测、碱基修饰检测等领域，以及快速临床诊断和疫情监测等专门领域. 通过开发新纳米孔、碱基调用方法和为特定应用定制实验方案，仍有数据质量和分析方法的问题值得改进.

1.1 Oxford Nanopore 牛津纳米孔测序技术的原理

1.1.1 Minion 技术原理和测序步骤

具体来说，我们这里主要讨论 Oxford Nanopore 推出的首款纳米孔测序设备 MinION 的技术原理. 该技术依赖于纳米级蛋白质孔，即“纳米孔”，作为生物传感器，嵌入耐电性聚合物膜中. 在电解液中，施加恒定电压产生流过纳米孔的离子电流，从而使带负电荷的单链 DNA 或 RNA 分子从带负电荷的顺式侧向带正电荷的反式侧穿过纳米孔. 转移速率由一种马达蛋白控制，该马达蛋白以有序的方式推动核酸分子穿过纳米孔. 在易位期间，离子电流的变化与感应区中的核苷酸序列相关，并利用计算算法进行解码，从而实现单分子的实时测序. 除了控制转移速度外，马达蛋白还具有解旋酶活性，使双链 DNA 或 RNA-DNA 双链能够解离成通过纳米孔的单链分子.

图1展示了 MinION 纳米孔测序原理. 一个 MinION Flow 细胞包含 512 个通道，每个通道中有 4 个纳米孔，总计 2,048 个纳米孔用于 DNA 或 RNA 的测序. 纳米孔位于由连接到传感器芯片的微支架阵列支撑的耐电聚合物膜内. 每个通道与传感器芯片中的独立电极相连，并由专用集成电路（ASIC）单独控制和测量. 由于在膜上施加了恒定电压，离子电流通过纳米孔，膜的反面带正电荷. 在马达蛋白质的作用下，首先解开双链 DNA（dsDNA）分子（或 RNA-DNA 杂交双链），然后在电压驱动下，单链 DNA 或带负电荷的 RNA 穿过纳米孔. 当核苷酸通过纳米孔时，测量特征电流变化，并根据这些变化确定相应的核苷酸类型，每秒约可

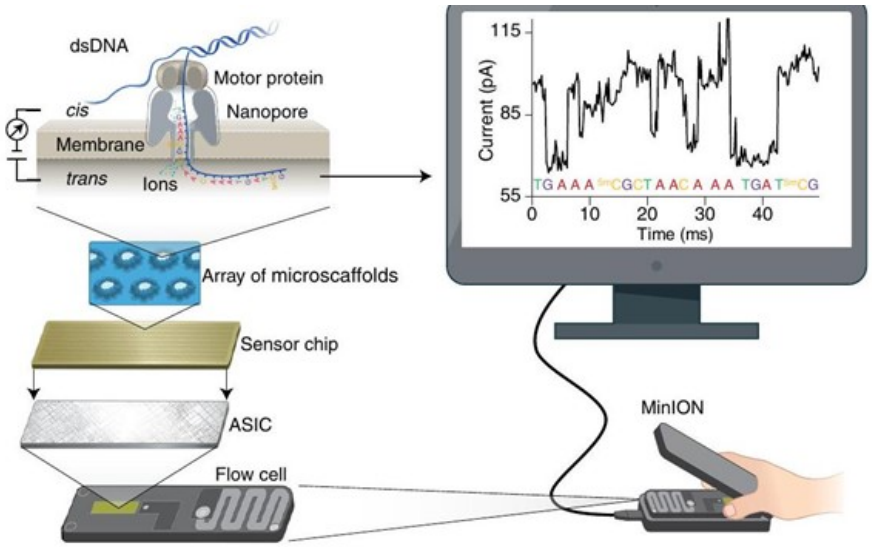

图 1: MinION nanopore sequencing 原理 [4]

识别 450 个碱基（R9.4 纳米孔）.

原理中涉及到的主要概念和作用：

1. 纳米孔：MinION 设备包含众多纳米孔，嵌入一层类细胞膜物质（通常为生物膜蛋白）. 每个纳米孔可检测通过的 DNA 分子.
2. 核酸分子穿孔：测序过程中，DNA 分子被引导穿过纳米孔. 需要对 DNA 分子进行处理，如加入适配器或线性化.
3. 电压梯度与电流信号：纳米孔两侧施加电压梯度，使带负电荷的 DNA 分子受电场驱动穿过纳米孔.DNA 分子穿过时，阻碍纳米孔内离子流，改变电流信号. 电流信号随 DNA 中不同碱基（A、T、C 和 G）的通过而变化.
4. 信号解码：MinION 设备记录纳米孔实时电流信号，用专门算法将信号转换为对应碱基序列.

图2展示了 DNA 通过纳米孔转位的步骤：（i）开放通道；（ii）双链 DNA（带有铅接头（蓝色）、结合分子马达（橙色）和发夹接头（红色））被纳米孔捕获；捕获后依次转位的是（iii）铅接头、（iv）模板链（金色）、（v）发夹接头、（vi）互补链（深蓝色）和（vii）拖尾接头（棕色）；最后（viii）状态恢复至开放通道. 在这个过程中，DNA 以特定顺序穿过纳米孔，产生的离子电流信号可以被检测并用于解码 DNA 序列.

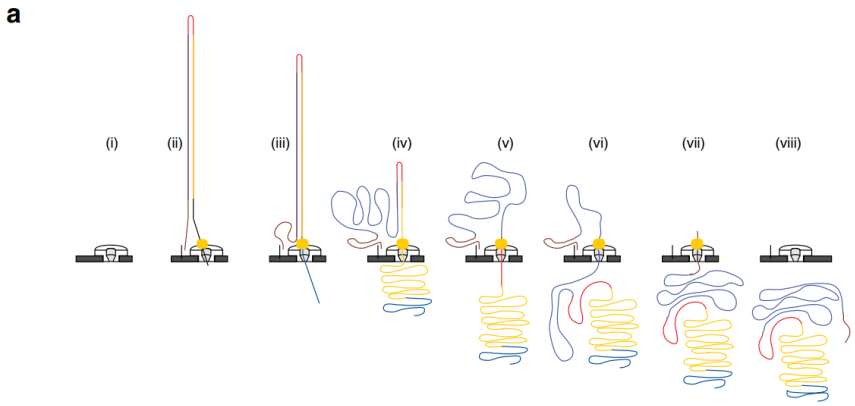

图 2: MinION nanopore sequencing 步骤 [3]

1.1.2 Basecaller 电流信号解码算法

当 DNA 分子穿过纳米孔时，不同的碱基（A、T、C 和 G）会引起电流信号的特异性变化. 这些变化取决于每种碱基与纳米孔之间的相互作用，在实际情况中，单个碱基的信号解析很困难，因为测序过程中 DNA 是以 k-mer（一般是 3-6 个连续碱基）的方式通过纳米孔的. 因此，在实际应用中，通常会观察到由连续几个碱基组成的 k-mer 引起的复合信号. 这些信号随着不同 k-mer 组合的通过而发生变化，可以通过特定的算法进行解码以获取原始的碱基序列. 如何精确地将 ONT 生成的原始电信号翻译为序列信息（即 basecalling）也就是科学家们关注的重点，当然也是我这个专业的研究者更关心的，所以还是得专门来讨论一下.

牛津纳米孔公司似乎提供了 MinION 的云计算服务 Metrichor，该方法主要基于隐马尔可夫模型（HMM），但属于闭源软件. 为了开发离线的替代方案，一些研究者着手研究开源算法，下面我们将主要讨论在文献 [3] 中提到的两种知名的算法 Nanocall 和 DeepNano.

Nanocall[5] 是一种基于隐马尔可夫模型 HMM 的基本呼叫器 Basecaller，它在本地执行高效的 1D 基本呼叫，无需互联网连接，其精度可与牛津纳米孔公司提供的基于 Metrichor 的 1D 基本呼叫相媲美.

Nanocall 的输入是一组存储在 ONT 特定的 FAST5 文件中的分段事件序列.Nanocall 分别处理每个输入文件，具体步骤如下. 首先，在发现发夹结构时，将模板链和互补链分为单独的事件序列. 接下来，它估算孔模型的缩放参数. 可选地，Nanocall 可以使用期望最大化算法进行多轮训练，以更新缩放参数，并使用标准的 Baum-Welch 算法 [6] 更新状态转移参数. 最后，Nanocall 执行标准的 Viterbi 解码，以找到隐藏状态的路径，其中状态是纳米孔中的 6-mer. 对于上述模型中的一些经典算法，应当可以在任何一本随机过程或者生物信息学教材中找到，因此不再赘述.

具体地说，状态转移是从一个状态转移到另一个状态的先验概率. 默认的状态转移是基于两个参数计算的：'stay' 概率 p_{stay} 和'skip' 概率 $p_{\text{skip.}}$ 前者，p_{stay}，表示两个连续事件来自同一上下文/状态的概率. 这对应于分割错误，即引入了错误的事件分割. 后者，p_{skip}，表示两个连续事件来自状态差异超过一个 kmer 位移的概率. 这对应于分割错误或测序错误（即 DNA 通过孔太快，无法捕捉可检测的事件 events），其中一个或多个事件丢失. 一个稍微复杂的问题是，通过增加跳过次数，总是有多种从一个状态到另一个状态的方式. 例如，ACGTGT 可以通过一次或三次跳过后在后面接上 GTGTAC. 下面对状态转移的计算也考虑到了这一点：

$$\begin{aligned}\tau\left(k_1, k_2\right) =& \delta_{k_1=k_2} \cdot p_{\text{stay}} + \delta_{\text{suffix }(k_1,5)=\text{prefix}(k_2,5)} \cdot p_{\text{step}} \cdot \frac{1}{4} \\ &+ \sum_{i=2}^{5} \delta_{\text{suffix }(k_1,6-i)=\text{prefix}(k_2,6-i)} \cdot {p_{\text{skip }1}}^{i-1} \cdot \frac{1}{4^i} + \sum_{i>5} {p_{\text{skip }1}}^{i-1} \cdot \frac{1}{4^6}.\end{aligned}$$

这里，δ 是标准示性函数；$\text{prefix}(k,i)/\ \text{suffix}(k,i)$ 是长度为 i 的 k 的前缀/后缀；$p_{\text{step}} := (1 - p_{\text{stay}} - p_{\text{skip}})$；而 $p_{\text{skip }1} = p_{\text{skip}} / (1 + p_{\text{skip}})$ 对应恰好一个跳过的概率.

在包含隐状态的状态转移矩阵训练之后，Nanocall 运行 Viterbi 解码算法，计算生成观察到的事件序列的最可能的状态序列. 之后通过迭代添加在连续状态之间转换所需的最少碱基数来构建最终的碱基序列. 例如，连续状态 ACTCTC 和 CTCTCA 生成碱基序列 ACTCTCA，而不是 ACTCTCTCA. 特别地，由于这种启发式方法和纳米孔中不变状态的特性，所调用的碱基序列不会包含比 kmer 大小（6 bp）更长的同源重复序列. 然而，仍能检测到长度大于 1 的重复序列. 因此，Nanocall 读取在大小为 1 的重复序列附近可能产生系统性（非随机）错误.

DeepNano[7] 是一个基于循环神经网络 (recurrent neural network)[8] 框架的算法，用于执行基调用，其精度优于基于 HMM 的方法. 在互联网连接有限的情况下进行现场测序时，能够执行本地、离线基地呼叫非常有用.

在图3的 DeepNano 中，给定一组输入向量 $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_t\}$. 它的预测是一组输出向量 $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_t\}$. 在这里，我们需要预测 DNA 序列，因此每个输入向量 $\vec{x}_i$ 包含每个事件（即长度为 k 的电流信号）的均值、标准差和长度，输出向量 $\vec{y}_i$ 给出了呼叫碱基的概率分布. 在处理每个输入向量 $\vec{x}_i$ 时，循环神经网络计算两个向量：其隐藏状态 $\vec{h}_i$ 和输出向量 $\vec{y}_i$. 这两者都取决于当前输入向量和之前的隐藏状态：$\vec{h}i = f\left(\vec{h}_{i-1}, \vec{x}_i\right), \vec{y}_i = g\left(\vec{h}_i\right)$. 通常，通过使用具有多个隐藏层的神经网络可以提高预测准确性，其中每个层使用来自前一层的隐藏状态. 我们使用具有三

层或四层的网络. 对于三层的计算如下：

$$\vec{h}_i^{(1)} = f_1\left(\vec{h}_{i-1}^{(1)}, \vec{x}_i\right)$$
$$\vec{h}_i^{(2)} = f_2\left(\vec{h}_{i-1}^{(2)}, \vec{h}_i^{(1)}\right)$$
$$\vec{h}_i^{(3)} = f_3\left(\vec{h}_{i-1}^{(3)}, \vec{h}_i^{(2)}\right)$$
$$\vec{y}_i = g\left(\vec{h}_i^{(3)}\right)$$

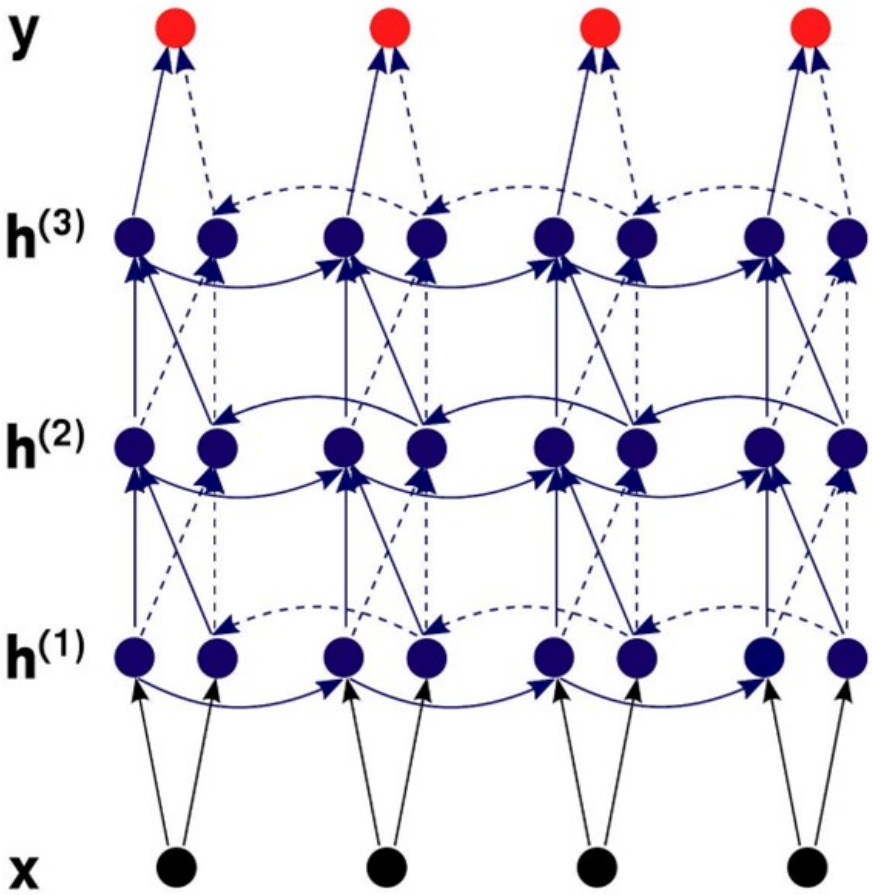

图 3: deepnano 循环神经网络模型 [7]

当然，对于上述两种算法循环神经网络和隐马尔可夫模型，实际上我个人认为两者的构造还是非常接近的，因为都存在相应的时序特征需要训练而且都共享训练权重，因此处理 DNA 序列肯定是有效的. 不过如果现在让我设计的话，肯定是要加入预训练和 Transformer 模型了，时代在不断变化.

1.1.3 原理总结

Oxford Nanopore 的纳米孔测序技术基于单个 DNA 分子通过纳米孔的电信号变化来实现.A、C、G、T 四种不同的碱基在电场作用下通过纳米孔时，会产

生不同的电信号，这些信号可以被检测和记录下来. 这些信号可以被转换成 DNA 序列，并且可以在实时中进行读取和分析.

1.2 Oxford Nanopore 牛津纳米孔测序技术的优缺点

1.2.1 优点

1. 长读长：纳米孔测序技术的长读长为 DNA 或 RNA 序列比对和匹配带来了优势，有助于获得高质量、更完整的基因组组装. 长读长在植物基因组等具有复杂结构和高度重复区域的研究中显著，简化了从头组装过程，改进参考基因组，实现 phasing，加快宏基因组物种鉴定. 对于 RNA，长读长有助于全长转录本表征，及异构体、剪接变体和融合转录本的量化和分析. 由电检测提供的读长具有很高上限，依赖核酸转位物理过程，2018 年实现了高达 2.273 兆碱基（Mb）的读长 [9].

2. 高通量：ONT 测序技术具有较高吞吐量，满足不同项目规模需求. 吞吐量提高得益于活跃纳米孔数量增加、DNA/RNA 转移速度提升及运行时间延长. 早期 MinION 用户报告每个流动池产量为数百兆碱基，而现吞吐量已增至约 10-15 千兆碱基（Gb），得益于更快的化学反应（R6 纳米孔速度约每秒 30 个碱基，提高至 R9.4 纳米孔每秒约 450 个碱基）及引入 Rev D SIC 芯片后更长运行时间. 后续设备，如 PromethION，运行更多流动池，每个流动池纳米孔数量更多 [10].

3. 实时靶向测序：这是在短时间内获取和分析 DNA 或 RNA 序列的有效方法，尤其适用于临床应用.MinION 平台因其小型、低成本、简易的文库准备和便携性，使实时分析成为可能 [11]. 通过在测序过程中实时捕获和分析 DNA 链，可以迅速积累目标片段的读取. 实时靶向测序技术有助于显著缩短从生物样本收集到数据分析所需的时间，对于现场和临床护理应用具有重要意义.

4. 直接检测碱基修饰：第二代 NGS 技术无法直接检测原生 DNA 中的碱基修饰. 然而，纳米孔技术可以对原生 DNA 和 RNA 的单个核苷酸进行单分子测序，从而检测其中的修饰. 比如有研究发现，纳米孔系统能够准确识别五种不同类型的胞嘧啶 [12]（包括 C、5-甲基胞嘧啶、5-羟甲基胞嘧啶、5-甲酰胞嘧啶和 5-羧胞嘧啶），准确率达 92% 至 98%.

5. 可直接对 RNA 表达分析：ONT 技术可以直接测序 RNA 分子，提供更真实的转录本信息，避免了将 RNA 逆转录成 cDNA 的步骤. 而 NGS 测序 cDNA

拷贝的片段较短，导致全长转录本的组装和 RNA 剪接异构体的准确表征困难.MinION 平台可进行全长 cDNA 读取，例如研究利用 MinION 成功检测了果蝇中四个基因的 RNA 剪接变体和异构体，其中一个复杂基因的 7000 多种异构体的比对同一性高达 90%. 这在 450 个碱基长度的 NGS 读取中是无法实现的 [13].

6. 便携性、简化性与低成本：ONT 的便携设备如 MinION 和 Flongle 使纳米孔测序技术能在实验室之外的环境中使用，如现场实验和偏远地区. 同时，ONT 的测序流程无需 PCR 扩增，避免了 PCR 偏差和可能的错误. 此外，ONT 的设备和测序试剂价格相对较低，特别适合小规模实验和个体研究者.

1.2.2 缺点

1. 准确性较低：ONT 纳米孔测序的一个主要缺点是其较低的测序准确性. 与其他测序平台相比，ONT 的单次读取准确性通常在 85%-90%之间. 这可能会影响某些应用，例如单核苷酸多态性（SNP）检测和低丰度变异体的识别. 比如图4中显示的结果.

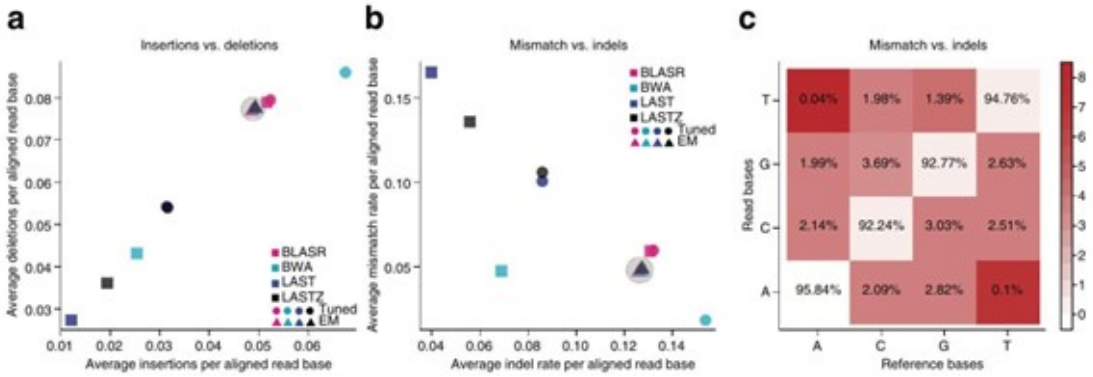

图 4: MinION 测序与参考序列比对结果 [3]

2. 高错误率：ONT 测序数据中的错误主要是插入和缺失错误，这可能会对某些应用产生负面影响，例如测序短的重复序列或串联重复序列. 比如我们之前提到的 Nanocall[5] 在大小为 1 的重复序列附近可能产生的系统性（非随机）错误.

3. 数据分析难度大：由于测序产生的大量长读长数据而且还是电信号，数据分析需要较高的计算资源和专业知识，需要使用到比较先进的算法，比如我们之前提到的 HMM,RNN，当然之后肯定还需要更先进的深度学习算法.

1.2.3 与第一、二代测序比较

1. 读长（Read Length）：
   ONT 测序技术可以生成非常长的读长（超过 1 Mb），而第一代（例如 Sanger 测序）和第二代测序技术（例如 Illumina）通常具有较短的读长（数百到数千个碱基）.

2. 准确性（Accuracy）：
   ONT 测序的准确性相对较低（约 85％-90％），而 Sanger 测序的准确性高达 99.99％，Illumina 测序的准确性也在 99％以上.

3. 通量（Throughput）：
   ONT 测序通量适中，可以根据不同设备满足不同项目规模的需求.Illumina 测序具有非常高的通量，适合大规模项目. 而 Sanger 测序的通量相对较低.

4. 实时性（Real-time）：
   ONT 测序技术具有实时分析的能力，可以在测序进行时获取数据. 而 Sanger 测序和 Illumina 测序在完成整个测序过程后才能获取数据.

5. 样品准备（Sample Preparation）：
   ONT 测序通常不需要 PCR 扩增，可以避免引入 PCR 偏差和可能的错误. 而 Sanger 测序和 Illumina 测序通常需要进行 PCR 扩增.

6. 成本（Cost）：
   ONT 测序设备和试剂相对较便宜，尤其适合小规模实验和个体研究者.Sanger 测序成本适中，但通量较低.Illumina 测序在大规模项目中成本效益较高，但对于小规模实验成本可能较高.

7. 便携性（Portability）：
   ONT 的设备（如 MinION 和 Flongle）非常小巧便携，适合现场实验和远程地区. 而 Sanger 测序和 Illumina 测序设备较大，通常需要固定在实验室中使用.

1.3 Oxford Nanopore 牛津纳米孔测序技术的应用

牛津纳米孔测序技术在基础研究、临床应用和现场应用等领域有着广泛的应用. 这些应用根据 ONT 测序的优势，如长读长、原生单分子和便携性等特点，可以进一步细分为多个具体主题.

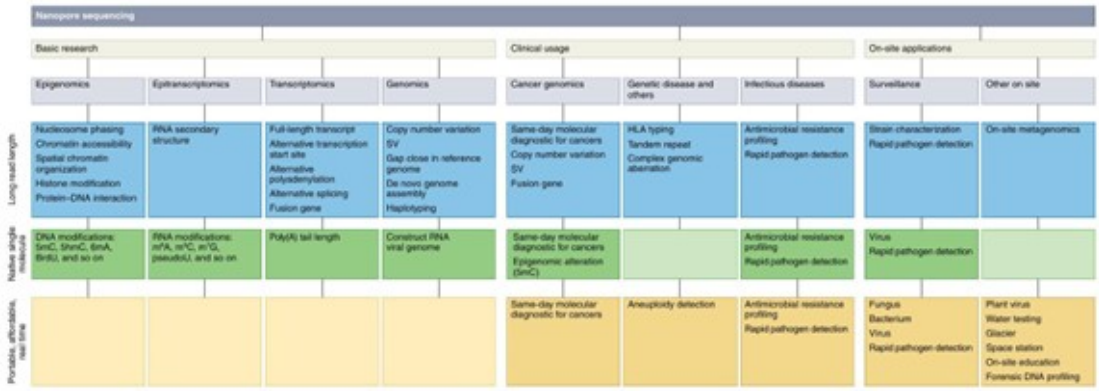

图 5: Oxford Nanopore 牛津纳米孔测序技术的应用表 [4]

1.3.1 填补参考基因组中的空缺

牛津纳米孔测序技术在基因组组装方面具有显著应用价值，特别是在填补参考基因组中的空缺和识别重复区域方面. 人类基因组、线虫参考基因组等已经通过 ONT 长读长成功地填补了空缺和扩展. 在其他模式生物、近缘物种和非模式生物方面也取得了类似的进展，有助于改善这些物种基因组的连续性和完整性. 例如，ONT 读取已用于关闭人类参考基因组中的 12 个缺口（每个缺口 > 50 kb），测量端粒重复的长度，以及组装人类 Y 染色体的中心粒区域 [14]. 此外，ONT 使得首次实现了人类 X 染色体的无缺口端粒至端粒组装，包括重建约 2.8 Mb 的中心粒卫星 DNA 阵列和关闭所有剩余的 29 个缺口（总计 1.1 Mb）[15].

1.3.2 建立新的参考基因组

牛津纳米孔测序技术在非模式生物初始参考基因组的组装中发挥了重要作用. 例如，通过仅使用 ONT 数据或结合其他技术，成功组装了根腐病菌、澳大利亚最大淡水鱼、普通小丑鱼等物种的基因组.ONT 直接 RNA 测序也被用于构建 RNA 病毒基因组，无需传统的反转录步骤. 在 SARS-CoV-2 大流行期间，ONT 测序通过 cDNA 和直接 RNA 测序重建了全长 SARS-CoV-2 基因组序列，为病毒的生物学、进化和致病性研究提供了重要信息.

1.3.3 识别大型结构变异 (SV)

牛津纳米孔测序技术在识别大型结构变异方面具有显著优势，特别是在生物医学背景中. 例如，它已成功应用于乳腺癌细胞系、急性髓系白血病患者以及具有先天异常的个体.ONT 技术还有助于在人类基因组中识别大量结构变异.

1.3.4 表征全长转录组和复杂转录事件

ONT cDNA 测序技术在表征全长转录组和复杂转录事件方面展现了显著的潜力. 与 PacBio 长读长相比，它在识别基因亚型方面具有相似的性能. 虽然在评估基因丰度和检测剪接位点方面存在一定的限制，但准确性和吞吐量的改进正在改善这些分析.ONT 技术还用于研究不同生物中的 RNA 分子的 poly(A) 尾长以及人类环状 RNA 的全长亚型.

1.3.5 检测 RNA 修饰

ONT 直接 RNA 测序技术为检测具有关键生物学功能的 RNA 修饰和 RNA 编辑提供了新的可能性. 通过 ONT 测序，研究者在不同生物中检测了各种 RNA 修饰 [16]，如 m6A、m7G 和假尿苷. 此外，结合人工化学修饰，ONT 直接 RNA 测序可以用于探测 RNA 的二级结构. 同时，通过标记新生 RNA 并进行 ONT 直接 RNA 测序，RNA 代谢动态也得到了分析.

1.3.6 癌症研究治疗

ONT 测序技术已广泛应用于各种癌症类型，以识别复杂的基因组变异. 这种技术可以快速检测白血病患者中的突变、染色体易位和断点. 同时，ONT 测序还能分析具有高重复序列和大基因大小的癌症易感基因. 通过直接检测 DNA 修饰 [17]，ONT 数据可同时捕获基因组和表观基因组变异.ONT 测序技术还可以在短时间内为癌症提供多模式的快速分子诊断. 例如，它已成功应用于检测临床样本中的融合基因 [18]. 总之，ONT 测序技术在癌症研究中具有显著的应用价值，为更深入地了解癌症相关基因突变和诊断提供了有力工具.

2 简述我知道的 RNA 的种类和功能

在生物学中，RNA（核糖核酸）是一类具有多种功能的生物大分子，涉及基因表达调控、蛋白质合成和基因组稳定等重要生物过程. 了解 RNA 的种类和功能对于深入理解生物学的基本原理和各种生命活动至关重要. 许多教材和专著都详细描述了不同类型的 RNA 及其作用，而本章节将简要介绍我所知道的所有 RNA 种类及其功能. 我的回答主要参考了教材 [19]，并结合了其他相关资料.

2.1 mRNA（信使 RNA）

mRNA（信使 RNA）在基因表达过程中发挥着关键作用，负责将 DNA 中的遗传信息传递到蛋白质.mRNA 的生成、加工和翻译调控是生物学中基因表达调控的核心环节.

在生物学定义上，mRNA 是 DNA 和蛋白质之间的信息传递桥梁.DNA 中的基因在转录过程中被转录成 mRNA，mRNA 携带着来自 DNA 的遗传信息. 这一过程由 RNA 聚合酶催化完成，生成初始的前 mRNA.

在功能上，mRNA 经过一系列修饰和加工，如 5’ 帽子添加、3’ 尾部加多聚腺苷酸（Poly-A）尾、选择性剪接等，形成成熟的 mRNA. 这些修饰有助于 mRNA 的稳定性和翻译效率，同时影响 mRNA 在细胞内的定位和运输.

在生物学过程中，mRNA 从细胞核转移到细胞质，在核糖体上进行翻译，指导蛋白质的合成. 核糖体依次识别 mRNA 上的密码子，tRNA 携带对应的氨基酸与 mRNA 的密码子配对，逐渐将氨基酸连接成蛋白质链. 不同物种和生物过程中的 mRNA 表达受到多层次的调控. 如转录后调控（mRNA 降解、局部翻译等）、非编码 RNA（如 miRNA）的作用等，这些调控机制共同影响蛋白质的合成和细胞功能.

2.2 tRNA（转运 RNA）

tRNA（转运 RNA）是蛋白质合成过程中的核心组成部分，它负责将特定的氨基酸带到核糖体上，并与 mRNA 上的密码子配对，实现遗传信息的解码和蛋白质合成的指导.

在生物学定义上，tRNA 是一类小分子 RNA，其具有独特且高度保守的二级结构，称为“三叶草”结构. 这一结构具有抗核酸酶降解的稳定性，并使得 tRNA 能够在不同物种中起到相似的功能.tRNA 的一个端部携带氨基酸，而另一端包含一个与 mRNA 密码子互补的反密码子.

在功能上，tRNA 的主要作用是在蛋白质合成过程中将氨基酸带到核糖体上. 在翻译过程中，核糖体会读取 mRNA 上的遗传密码，tRNA 的反密码子与 mRNA 上的密码子进行互补配对，将相应的氨基酸带到核糖体上，从而将氨基酸连接成蛋白质链.

在生物学过程中，tRNA 的作用对生物的生长、发育和生存至关重要. 通过将氨基酸正确地带到核糖体上，tRNA 使得蛋白质能够按照 mRNA 上的遗传密码正确地合成，从而保证了生物体的正常生理功能. 例如，在人类肌肉细胞中，tRNA 对肌纤维蛋白的合成起着关键作用，影响着肌肉的功能和力量.

2.3 rRNA（核糖体 RNA）

核糖体 RNA（rRNA）在生物体内具有重要的功能. 它是核糖体的主要组成成分，起到关键作用，确保生物体内蛋白质合成过程的顺利进行.

在生物学定义上，rRNA 是一类非编码 RNA，不参与蛋白质的编码过程，但在细胞功能中起着至关重要的作用.rRNA 分子具有一定的保守性，这使得它们在不同物种中能够发挥相似的功能. 核糖体是生物体内进行蛋白质合成的场所，负责将遗传信息从 mRNA 转化为蛋白质. 在核糖体中，rRNA 与蛋白质共同组成核糖体的亚基，参与蛋白质合成的关键步骤.

在功能层面，rRNA 在核糖体结构和功能中发挥核心作用. 它参与蛋白质合成过程中的多个关键步骤，如 mRNA 的辨认、tRNA 的结合以及肽键的形成等. 例如，在核糖体的 A 位、P 位和 E 位上，rRNA 与 tRNA 形成稳定的相互作用，确保氨基酸被正确地添加到生长中的肽链上. 此外，rRNA 在肽链延长和终止阶段也发挥作用，促使新合成的蛋白质从核糖体中释放.

rRNA 对于生物的生长、发育和生存具有重要意义. 核糖体数量和活性与生物体的生长速度密切相关，而 rRNA 的合成和稳定性是核糖体生物合成的关键因素. 在一些疾病中，如核糖体功能障碍症，rRNA 的异常表达或处理可能导致细胞功能障碍和发育异常.

2.4 snRNA（小核 RNA）

snRNA 是长度在 20 到 300 个核苷酸之间的小分子 RNA，在生物学中具有重要功能.snRNA 与蛋白质结合，形成剪接体，参与 mRNA 剪接过程.

在真核生物中，剪接是 mRNA 成熟过程中的一个关键环节. 在该过程中，mRNA 中的内含子（非编码区）被去除，而外显子（编码区）被连接在一起，形成成熟的 mRNA. 这个过程的正确进行是蛋白质合成和基因表达的关键.snRNA 分子与蛋白质结合形成剪接复合物，参与识别和移除内含子，从而使 mRNA 正确剪接.

2.5 snoRNA（小核仁 RNA）

snoRNA 是长度在 60 到 300 个核苷酸之间的小分子 RNA，在生物学中发挥关键作用. 主要参与 rRNA 和 tRNA 的化学修饰，如 2’-O-甲基化和假尿苷化等.

snoRNA 通过与蛋白质结合，形成核仁小颗粒（snoRNP），在 rRNA 和 tRNA 的修饰过程中起作用. 例如，在酵母中，U14 snoRNA 参与 18S rRNA 的加工过程，对核糖体的生物合成具有关键作用.snoRNA 与 rRNA 和 tRNA 中特定位点互补配对，引导相关酶进行正确的化学修饰.

2.6 miRNA（微小 RNA）

miRNA 是长度约为 22 个核苷酸的非编码 RNA，在生物学中具有重要功能. 通过与 mRNA 的 3’ 非翻译区域（3’ UTR）结合，调控基因表达，影响生长、分化和凋亡等细胞过程.

miRNA 的生物合成包括 pri-miRNA 的转录、剪切生成 pre-miRNA 和加工成成熟 miRNA. 成熟的 miRNA 与 RNA 诱导沉默复合体（RISC）结合，与目标 mRNA 完全或部分互补配对，导致 mRNA 降解或翻译抑制，从而调控基因表达.

miRNA 在生物学过程中发挥关键作用，调控基因表达具有重要意义. 例如，miR-34 家族在哺乳动物中发挥抗肿瘤作用，通过抑制细胞生长和促进细胞凋亡来抑制肿瘤发生. 这些调控机制对维持生物体内稳态及响应环境变化至关重要.

2.7 lncRNA（长非编码 RNA）

lncRNA 是长度超过 200 个核苷酸的非编码 RNA. 与 mRNA 不同，它们不直接参与蛋白质合成. 然而，在基因调控、染色质修饰和 X 染色体失活等方面，lncRNA 具有多种功能.

作为基因调控因子，lncRNA 可以在转录和翻译水平上调控基因表达. 例如，HOTAIR 通过与 PRC2 互作，在染色质上引导 H3K27me3 修饰，抑制靶基因表达. 此外，lncRNA 还可作为”miRNA 海绵”，调节 miRNA 对靶基因的调控.

在染色质修饰方面，lncRNA 参与组蛋白修饰，影响染色质状态. 例如，Xist 与 X 染色体失活过程密切相关. 在雌性哺乳动物中，Xist 通过与染色质修饰因子相互作用，导致一个 X 染色体的失活，平衡雌性与雄性之间 X 染色体上基因的剂量.

2.8 piRNA（Piwi-interacting RNA）

piRNA 是与 Piwi 家族蛋白结合的小分子 RNA，长度约为 26-31 个核苷酸.piRNA 主要在生殖细胞中发现，参与转座子沉默和基因调控.

piRNA 与 Piwi 蛋白结合，形成 piRNA-Piwi 复合物，起到转座子沉默作用. 转座子是可在基因组内移动的遗传元素，它们的不受控制的活动可能导致基因突变和基因组不稳定.piRNA 通过引导 piRNA-Piwi 复合物结合到转座子的转录物，阻止其转录和复制，维护生殖细胞基因组的稳定性.

除转座子沉默作用外，piRNA 还参与基因表达调控. 在果蝇中，piRNA 可以引导 piRNA-Piwi 复合物结合到 mRNA 分子，导致 mRNA 的降解，影响基因表达. 此外，piRNA 还参与染色质修饰.

2.9 circRNA（环状 RNA）

circRNA 是一类环形结构的非编码 RNA，长度可变，通常比线性 RNA 更加稳定. 它们在生物学中具有多种功能，如基因表达调控、作为 miRNA 的海绵和编码蛋白质等.

circRNA 可以调控基因表达. 一些 circRNA 通过与转录因子结合，影响转录因子的活性，进而调节基因的转录. 此外，circRNA 还可以与 RNA 结合蛋白相互作用，影响 mRNA 的加工和稳定性. 尽管 circRNA 主要被认为是非编码 RNA，但一些 circRNA 具有编码蛋白质的潜能. 这些 circRNA 包含开放阅读框（ORF），在特定条件下，可通过转录和翻译产生蛋白质，从而影响生物学过程.

3 泛基因组学 (Pan-genome)：内容、应用场景、研究实例

泛基因组代表了一个物种内的所有基因集合，由核心基因组（包含物种所有个体间共享的序列）和“可选”基因组组成.2005 年，泛基因组这一概念首次应用于细菌物种研究 [20]，当时在对数株链球菌（Streptococcus agalactiae）的测序中发现，核心基因组包含了 80% 的链球菌基因，而其余 20% 的基因在至少一个菌株中缺失. 从那时起，研究人员一直在努力揭示许多物种的泛基因组，如 2010 年进行的人类泛基因组研究 [21]. 泛基因组研究表明，依赖于单一的参考基因组可能会对我们对物种多样性的理解产生负面影响. 例如，在植物物种中 [22]，与农艺性状相关的许多重要基因通常位于可选基因组中. 随着基因组测序技术的不断改进，特别是长读取测序的发展，研究者们开始构建更复杂物种的泛基因组. 通过泛基因组研究，我们发现单一参考基因组无法充分代表物种内群体的遗传多样性，因此泛基因组研究在未来植物育种、疾病抗性和环境适应性等方面具有广泛的应用前景.

3.1 内容

泛基因组（pangenome 或 supragenome）是指分类单元（clade）内所有菌株 (strains, 实际上也可以推广到物种) 的所有基因集合. 它可分为核心泛基因组（core pangenome，所有个体中存在的基因）、壳层泛基因组（shell pangenome，两个或更多菌株中存在的基因）和云泛基因组（cloud pangenome，仅在单个菌株中发现的基因），如图6中所示. 云基因组也称为附属基因组，包含部分菌株中存在的可有可无的基因和菌株特异性基因. 研究泛基因组的领域称为泛基因组学.

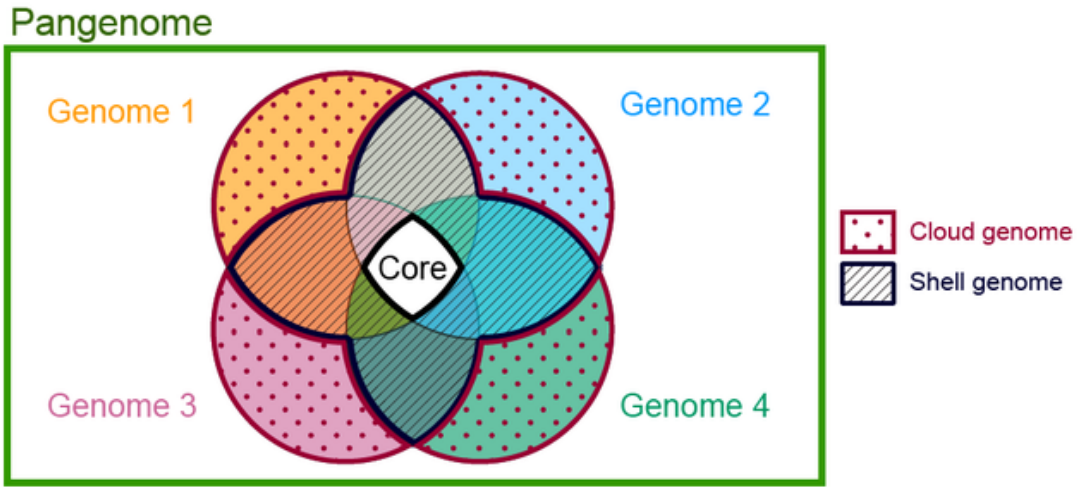

图 6: 泛基因组中的不同类型

细菌物种的基因组比单个菌株的基因含量大得多. 有些物种具有开放泛基因组，而另一些具有封闭泛基因组. 人口规模和生态位多样性被认为是决定泛基因组大小的主要因素. 泛基因组最初为细菌和古菌物种构建，近年来发展了真核生物泛基因组，特别是植物物种. 泛基因组动态与可移动元件有关，具有重要的进化背景意义，与宏基因组学相关，也在更广泛的基因组学背景中使用.

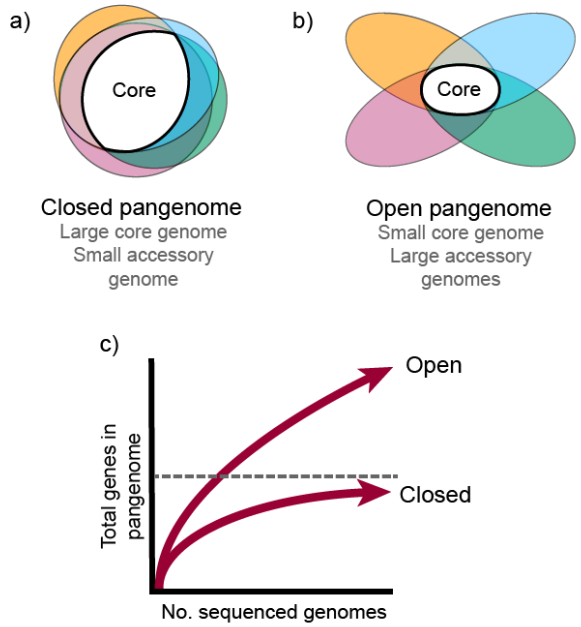

图 7: 封闭和开放泛基因组. 封闭泛基因组趋于渐近，可以预测完整大小.

3.2 应用场景

想要深入理解泛基因组学的应用场景首先得明白泛基因组学的优点. 泛基因组学的优点在于它能够更全面地揭示物种内的基因组多样性，揭示了整个基因组在不同物种中的重要性. 因此，泛基因组学为我们提供了一个更全面的视角来研究物种内的遗传多样性，并为发掘有益基因提供了丰富的资源.

3.2.1 作物基因组学、育种和进化研究

泛基因组可以支持植物育种和进化研究，探讨基因存在和缺失变异的起源. 第一个植物泛基因组研究是基于七个野生大豆个体的全基因组对比，发现与种子成分、开花和成熟时间、器官大小和生物量相关的可变基因以及在野生大豆 Glycine soja 中存在、而在家养大豆 Glycine max 中不存在的病害抗性基因 [23]. 另一个研究基于三个不同水稻品种，发现一个栽培品种中 S5 杂种不育座位的缺失以及 Sub1A 水生状耐受基因的 PAV[24].

3.2.2 研究不同品种结构变异影响基因差异表达

结构变异会影响基因的转录调控和表达，例如基因剂量变化、可变剪接和转录调控因子结合位点的改变. 大豆泛基因组研究中，研究者通过对 26 个具有代表性的野生和栽培大豆品种进行 de novo 基因组组装，构建了一个高质量的基于图的大豆泛基因组. 这一研究揭示了许多通过短序列直接映射到单一参考基因组上难以发现的遗传变异. 基于图谱基因组的 2,898 个加入物的结构变异和来自代表性的 26 个加入物的 RNA 测序（RNA-seq）数据有助于将遗传变异与负责重要性状的候选基因联系起来. 这一泛基因组资源将促进大豆的进化和功能基因组学研究.[25].

3.2.3 结合 GWAS 数据捕获更完整的遗传变异信息

通过将泛基因组应用于 GWAS 分析，研究人员可以更全面地了解植物基因组的多样性，从而有助于揭示与重要性状相关的遗传变异，推动植物遗传改良和功能基因组学研究. 传统 GWAS 分析在缺失参考基因组中的功能基因时，可能无法准确关联表型. 然而，采用泛基因组作为参考，将结构变异纳入 GWAS 分析，有助于解决这一问题.

以一项关于油菜（Brassica napus）的研究 [26] 为例，研究者通过测序、de novo 组装和注释 8 个油菜品种，揭示了大量小型变异和存在及缺失变异（PAV）.PAV 基因组广泛关联研究（PAV-GWAS）成功确定了与籽荚长度、种子重量和开

花时间相关的因果结构变异，这些变异在基于单核苷酸多态性的 GWAS（SNP-GWAS）中未被检测到. 此研究表明，PAV-GWAS 可补充 SNP-GWAS，识别与性状相关的关联，并为深入了解油菜基因组结构和遗传改良提供资源.

3.3 研究实例：亚洲稻米泛基因组倒位指数

下面，我将介绍一篇关于亚洲稻米泛基因组学研究的文章. 这篇文章于 2023 年 3 月发表在《Nature Communications》上，题目为“亚洲水稻亚种群结构中泛基因组倒位指数揭示的进化启示”(Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice)[1]. 本文的主要作者是 Yong Zhou 等人，来自 KAUST 等多个机构. 这项研究旨在探索亚洲稻米的基因多样性，并通过泛基因组学方法提供进化方面的见解.

3.3.1 研究背景和内容

为了应对 2060 年至 2070 年预期将达到 100 亿的全球人口增长，稻米研究社区正寻求创新方法，培育出营养丰富、可持续、能适应气候变化的新品种. 在亚洲稻米的基因组中，有一些基因片段的顺序与其他个体不同，这种现象被称为倒位（inversion）[27]. 倒位是一种重要的基因组结构变异，它可以影响基因的表达和功能.

本研究使用了 73 个高质量亚洲水稻（Oryza sativa）基因组，覆盖其亚种群结构，以及 2 个野生近缘种（O. rufipogon 和 O. punctata）. 研究人员基于这些基因组构建了一个包含 1769 个非冗余倒位的泛基因组倒位指数，涵盖了约 29％的 O. sativa cv. Nipponbare 参考基因组序列.

通过该倒位指数，研究者估计亚洲水稻倒位发生率约为每百万年 700 次，这一速率是以往植物研究估算的 16 至 50 倍. 对这些倒位的详细分析表明，它们对基因表达、重组率和连锁不平衡具有显著影响. 该研究揭示了亚洲水稻泛基因组中大型倒位（≥100 bp）的普遍性和规模，并暗示了其在功能生物学和作物性状方面尚未充分研究的潜在影响.

3.3.2 水稻倒位指数概述

图8中，**a**, **b** 代表重采样置换检验，用于确定基因组数量与所有倒位和非基因组特异性倒位之间的关系；**c** 倒位区域密度；**d** Bionano 对大于 1 Mb 的倒位进行验证，即 Clu-INV0100180，Clu-INV0100660 和 Clu-INV0600550. 在每个面板中，顶部线条显示作为参考的光学图谱，底部线条显示具有倒位的品种的基因组

组装. 灰色线条连接对齐的限制位点（蓝色区域），而黄色片段显示未对齐区域. 黑色框突出显示每个倒位的位置. 源数据作为源数据文件提供.

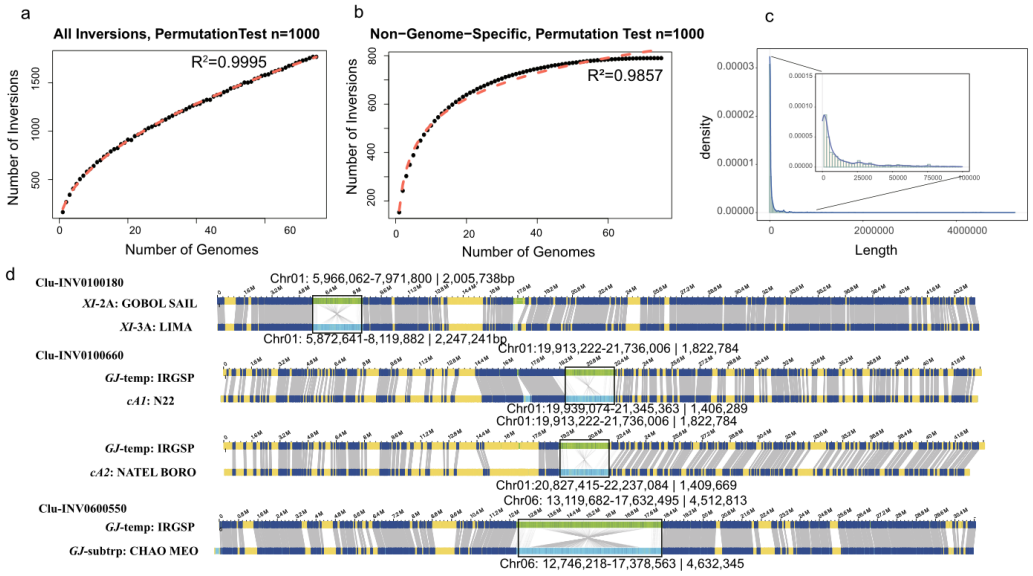

图 8: 水稻倒位指数概述图

这部分中，研究人员使用了一些典型的数据处理方法：

1. 重采样置换检验（Resampling permutation test）：这是一种统计方法，通过对数据进行随机重新排列，来评估观察到的结果是否具有显著性. 在本研究中，研究者使用重采样置换检验来确定基因组数量与所有倒位和非基因组特异性倒位之间的关系.

   在 n 次置换后，对于重采样置换检验的 p 值计算可使用以下公式:

   $$p = \frac{1 + \sum_{i=1}^{n} I\left(T_i \geq T_{obs}\right)}{n + 1}$$

   其中，I 是示性函数，T_i 表示第 i 次置换的统计量，T_{obs} 表示观察到的统计量，n 表示置换次数.

2. 倒位区域密度（Density of inversion regions）：这是一种用于描述倒位在基因组上分布密度的指标. 通过计算倒位区域的密度，研究者可以更好地了解倒位在水稻基因组中的分布特征.

   为了估计亚洲稻基因组倒位率（IR）,本文考虑了具有最近共同祖先（MRCA）分歧时间估计的种群或基因组对，并将倒位总数除以两倍的最近共同祖先时间（TMRCA，对应于两个节点上的谱系总分支长度），即使用以下方程计算估计值：

$$IR = \frac{Number\ \text{of}\ INVs}{2 \times TMRCA}$$

3. Bionano 验证：Bionano 是一种基于光学映射技术的基因组结构变异检测方法. 本研究中，研究者使用 Bionano 技术验证了大于 1 Mb 的倒位，以确保倒位检测结果的准确性. 在图中，顶部线条显示作为参考的光学图谱，底部线条显示具有倒位的品种的基因组组装. 通过比较这些数据，可以直观地了解倒位的位置和特征.

3.3.3 系统发育树分析

使用全基因组倒位指数对 75 个高质量水稻基因组进行系统发育树分析，如图9. 使用 UPGMA 方法（算术平均无加权配对群组法）推断用于创建全基因组倒位指数的 75 个高质量基因组的系统发育关系. 系统发育树上的分支长度代表了进化距离，即基因组之间的相似性或差异程度. 研究者通过对比不同基因组之间的进化距离，可以了解它们之间的亲缘关系. 此外，研究者还对两个野生近缘种（O. rufipogon 和 O. punctata）进行了特殊标注，以突显它们在系统发育树中的位置.

文章中提到的 GJ、cB、XI 和 cA 组是水稻基因组中不同的亚群. 研究者通过对这些亚群进行不同颜色的标注，可以更直观地展示它们在系统发育树上的分布. 此外，研究者还使用 Mantel 检验来评估 SNP 和 INV 多态性距离矩阵之间的相关性. 结果表明，这两者之间存在显著的相关性（r = 0.79，p = 1e-6），这意味着基因组中的 SNP 变异和倒位变异之间存在一定的关联.

这部分中，研究人员使用了一些典型的数据处理方法：

1. UPGMA（算术平均无加权配对群组法）是一种常用的聚类方法. 它基于距离矩阵进行操作，通过计算不同样本间的成对距离，从而推断它们之间的系统发育关系.UPGMA 是一种层次聚类方法，其基本思想是逐步将距离最近的样本或样本群组合并，直到所有样本都被归入一个大的群组中.

2. Mantel 检验是一种数据处理方法，其计算公式如下：

$$r = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(X_{ij}-\bar{X}\right)\left(Y_{ij}-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(X_{ij}-\bar{X}\right)^2\left(Y_{ij}-\bar{Y}\right)^2}}$$

其中，X_{ij} 和 Y_{ij} 分别表示两个距离矩阵中的元素，$\bar{X}$ 和 $\bar{Y}$ 分别表示距离矩阵的平均值，n 是观测值的数量.

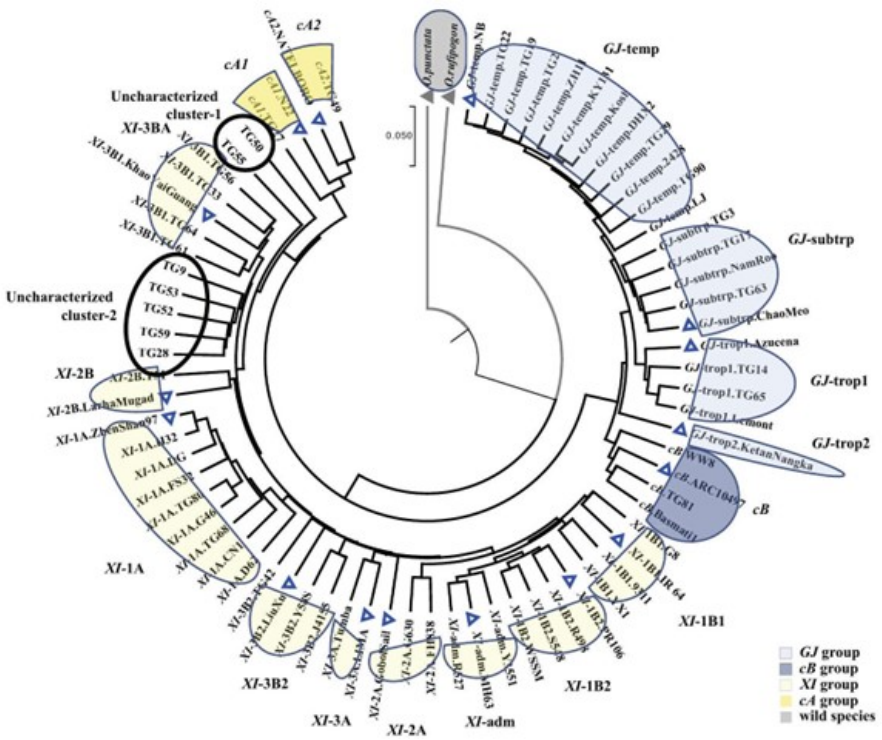

图 9: 使用全基因组倒位指数对 75 个高质量水稻基因组进行系统发育树分析

3.3.4 其他分析和结果概要

由于本报告篇幅有限，无法详细分析所有结果，以下简要概述其他部分：

1. 分析了亚洲水稻及其两个野生近缘种（O. punctata 和 O. rufipogon）中的物种特异性、群体特异性和共享倒位. 提出了特异性倒位和倒位速率的模型.

2. 研究了水稻全基因组倒位指数中的转座子元素分布. 发现三个转座子元素家族（LTR-RT Ty1-copia、Ty3-gypsy 和 DNA-TE MULE）在倒位断点处的频率较高.

3. 分析了位于倒位断点处基因的转录丰度. 展示了 MH63（XI-adm）基因组中位于倒位末端的两个 OsNAS 基因拷贝，以及倒位对其 5’UTR 区域的影响. 同时，研究了 OsNAS 基因在根组织中的转录丰度以及 Fbox 基因编码序列在 MH63（XI-adm）中如何被倒位破坏.

4. 对大倒位的群体连锁不平衡进行了分析. 研究了倒位导致的连锁不平衡（LD）区块破坏，展示了倒位如何破坏两个 LD 区块. 同时，给出了一个实例，说明高 LD 的 SNP 区块被倒位破坏，以及 IRGSP RefSeq（GJ-temp）与 Azucena（GJ-trop1）单倍型区块不连续时，INV030410 区域的 Azucena（GJ-trop1）单倍型区块被破坏.

3.3.5 基因组倒位的鉴定与评估方法

本文的研究者旨在识别大型倒位（>100 bp），在两个基因组（GJ-temp: IRGSP-1.0 和 XIadm: MH63）上测试了四种分析流程.

工作流程 1：MH63 基因组切成约 10 倍覆盖率的 50 Kb 重叠片段，使用 NGMLR 映射到 IRGSP-1.0，并用 SVIM 调用倒位. 保留深度大于 6 且通过筛选标准的倒位.

工作流程 2：与工作流程 1 类似，但使用 Sniffles 调用倒位.

工作流程 3：MH63 基因组与 IRGSP-1.0 使用 Minimap2 对齐. 筛选出大于 90% 同源性和长度大于 100 bp 的结果. 使用 SyRI 调用倒位.

工作流程 4：与工作流程 3 类似，但使用 MUMmer 的 Nucmer 进行对齐.

本文的研究者验证后最后选择了工作流程 4，使用 MUMmer 将 74 个基因组序列与 IRGSP RefSeq 对齐，筛选最小 90% 同源性和最小 100 bp 长度. 用 MUMmer 的“showcoords”功能获取坐标. 最后，使用 SyRI 工具调用倒位，生成包含 ID、参考和查询基因组坐标起始、结束位置的 VCFs，用于下游全基因组比较.

关于这部分内容，研究者开源了代码，我也尝试了在 Linux 系统下安装 numcer,syri 等软件进行序列的比对和评估测试，不过由于原文亚洲水稻基因的数据量确实比较大 (一个数据集大约 100 多 M，其实不算大，但是也不小了)，暂时还没有来得及复现文章的结果，因此也不再赘述了.

3.3.6 研究实例总结和展望

本研究通过全面分析亚洲稻米种群结构层面的倒位变异，揭示了倒位在水稻中的重要性. 倒位作为一种重要的结构变异类型，对于基因重组抑制、适应性特征选择、生殖隔离和物种形成具有显著作用. 通过发现 1769 个非冗余倒位并估计了不同层次的倒位速率，本研究为倒位研究提供了新的见解.

展望未来，亚洲稻米泛基因组倒位指数的建立仅是揭示稻米中所有自然变异的第一步. 下一步研究者将建立亚洲稻米的水稻数字基因库，将超过 100,000 份测序数据映射到水稻种群参考组. 稻米基因组结构变异的研究和应用，将为未来稻米育种和农业发展提供支持. 通过这一研究实例的学习和分析，我们也认识到泛基因组学在大规模组学数据和科学计算时代可以发挥非常重要的作用.

参考文献

[1] Y. Zhou, Z. Yu, D. Chebotarov, K. Chougule, Z. Lu, L. F. Rivera, N. Kathiresan, N. Al-Bader, N. Mohammed, A. Alsantely *et al.*, “Pan-genome inversion index

reveals evolutionary insights into the subpopulation structure of asian rice," *Nature Communications*, vol. 14, no. 1, p. 1567, 2023.

[2] D. Deamer, M. Akeson, and D. Branton, "Three decades of nanopore sequencing," *Nature biotechnology*, vol. 34, no. 5, pp. 518–524, 2016.

[3] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The oxford nanopore minion: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, pp. 1–11, 2016.

[4] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, "Nanopore sequencing technology, bioinformatics and applications," *Nature biotechnology*, vol. 39, no. 11, pp. 1348–1365, 2021.

[5] M. David, L. J. Dursi, D. Yao, P. C. Boutros, and J. T. Simpson, "Nanocall: an open source basecaller for oxford nanopore sequencing data," *Bioinformatics*, vol. 33, no. 1, pp. 49–55, 2017.

[6] L. E. Baum *et al.*, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[7] V. Boža, B. Brejová, and T. Vinař, "Deepnano: deep recurrent neural networks for base calling in minion nanopore reads," *PloS one*, vol. 12, no. 6, p. e0178751, 2017.

[8] C. L. Giles, G. M. Kuhn, and R. J. Williams, "Dynamic recurrent neural networks: Theory and applications," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 153–156, 1994.

[9] L. Gong, C.-H. Wong, J. Idol, C. Y. Ngan, and C.-L. Wei, "Ultra-long read sequencing for whole genomic dna analysis," *JoVE (Journal of Visualized Experiments)*, no. 145, p. e58954, 2019.

[10] S. M. Nicholls, J. C. Quick, S. Tang, and N. J. Loman, "Ultra-deep, long-read nanopore sequencing of mock microbial community standards," *Gigascience*, vol. 8, no. 5, p. giz043, 2019.

[11] M. Loose, S. Malla, and M. Stout, "Real-time selective sequencing using nanopore technology," *Nature methods*, vol. 13, no. 9, pp. 751–754, 2016.

[12] Z. L. Wescoe, J. Schreiber, and M. Akeson, “Nanopores discriminate among five c5-cytosine variants in dna,” *Journal of the American Chemical Society*, vol. 136, no. 47, pp. 16 582–16 587, 2014.

[13] M. T. Bolisetty, G. Rajadinakaran, and B. R. Graveley, “Determining exon connectivity in complex mrnas by nanopore sequencing,” *Genome biology*, vol. 16, pp. 1–12, 2015.

[14] M. Jain, H. E. Olsen, D. J. Turner, D. Stoddart, K. V. Bulazel, B. Paten, D. Haussler, H. F. Willard, M. Akeson, and K. H. Miga, “Linear assembly of a human centromere on the y chromosome,” *Nature biotechnology*, vol. 36, no. 4, pp. 321–323, 2018.

[15] K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon *et al.*, “Telomere-to-telomere assembly of a complete human x chromosome,” *Nature*, vol. 585, no. 7823, pp. 79–84, 2020.

[16] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland *et al.*, “Highly parallel direct rna sequencing on an array of nanopores,” *Nature methods*, vol. 15, no. 3, pp. 201–206, 2018.

[17] P. Euskirchen, F. Bielle, K. Labreche, W. P. Kloosterman, S. Rosenberg, M. Daniau, C. Schmitt, J. Masliah-Planchon, F. Bourdeaut, C. Dehais *et al.*, “Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing,” *Acta neuropathologica*, vol. 134, pp. 691–703, 2017.

[18] W. R. Jeck, J. Lee, H. Robinson, L. P. Le, A. J. Iafrate, and V. Nardi, “A nanopore sequencing–based assay for rapid detection of gene fusions,” *The Journal of Molecular Diagnostics*, vol. 21, no. 1, pp. 58–69, 2019.

[19] 基因组学. 高等教育出版社, 2019. [Online]. Available: https://books.google.co.jp/books?id=1kWzzQEACAAJ

[20] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin *et al.*, “Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial “pan-genome”,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 39, pp. 13 950–13 955, 2005.

[21] R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, G. Zhou, X. Zhu, H. wu, J. Qin, X. Jin, D. Li, H. Cao, X. Hu, H. Blanché, and J. Wang, "Building the sequence map of the human pan-genome," *Nature biotechnology*, vol. 28, pp. 57–63, 12 2009.

[22] P. E. Bayer, A. A. Golicz, A. Scheben, J. Batley, and D. Edwards, "Plant pan-genomes are the new reference," *Nature Plants*, pp. 1–7, 2020.

[23] Y.-h. Li, G. Zhou, J. Ma, W. Jiang, L.-g. Jin, Z. Zhang, Y. Guo, J. Zhang, Y. Sui, L. Zheng *et al.*, "De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits," *Nature biotechnology*, vol. 32, no. 10, pp. 1045–1052, 2014.

[24] M. C. Schatz, L. G. Maron, J. C. Stein, A. H. Wences, J. Gurtowski, E. Biggers, H. Lee, M. Kramer, E. Antoniou, E. Ghiban *et al.*, "Whole genome de novo assemblies of three divergent strains of rice, oryza sativa, document novel gene space of aus and indica," *Genome biology*, vol. 15, pp. 1–16, 2014.

[25] Y. Liu, H. Du, P. Li, Y. Shen, H. Peng, S. Liu, G.-A. Zhou, H. Zhang, Z. Liu, M. Shi *et al.*, "Pan-genome of wild and cultivated soybeans," *Cell*, vol. 182, no. 1, pp. 162–176, 2020.

[26] J.-M. Song, Z. Guan, J. Hu, C. Guo, Z. Yang, S. Wang, D. Liu, B. Wang, S. Lu, R. Zhou *et al.*, "Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of brassica napus," *Nature Plants*, vol. 6, no. 1, pp. 34–45, 2020.

[27] "The rice genome revolution: From an ancient grain to green super rice," *Nature Reviews Genetics*, vol. 19, no. 8, pp. 505–517, Aug. 2018.