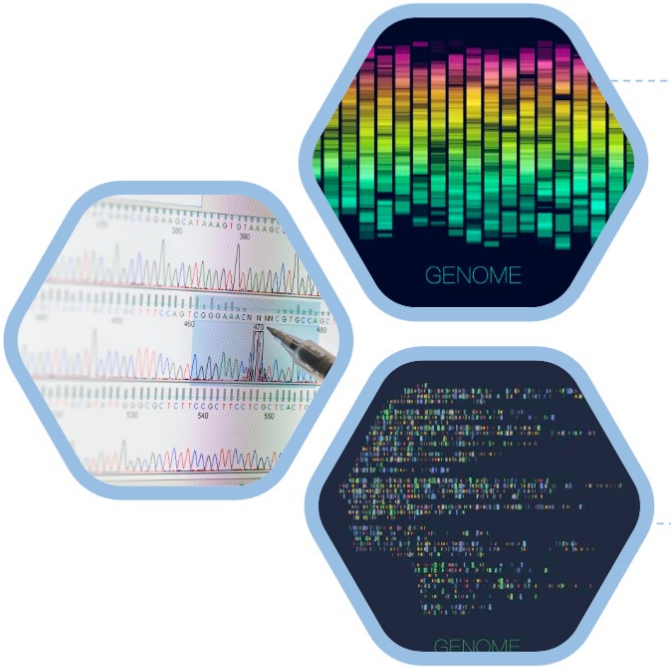
生物信息学 课程
Bioinformatics

基因组的组装、预测和注释

生物科学 101计划

1. 基因组的组装

基因组组装是将测序得到的读长拼接成叠连群，再通过末端配对测序的方法进行组装，形成支架。



2. 基因的预测

基因预测软件使用fasta格式的基因组序列文件作为输入，输出包括基因组索引文件、预测得到的基因序列文件和蛋白质序列文件。

3. 基因组的注释和分析

基因组分析的最后一步是对基因组进行注释和分析，包括GO、KEGG、NR等注释方法。

胡敏杰
生科院319室
Minjie-hu@zju.edu.cn

基因组的组装、预测和注释

为什么要组装基因组？

- **了解遗传信息的全貌**

基因组测序可以获得生物体全部的遗传信息，是理解生物遗传特征的基础。

- **揭示基因结构与功能**

通过组装后的完整基因组，可以定位基因、预测其功能，研究基因表达调控机制。

- **解析进化关系**

对不同物种的基因组进行比较，可揭示物种起源、进化分化与系统发育关系。

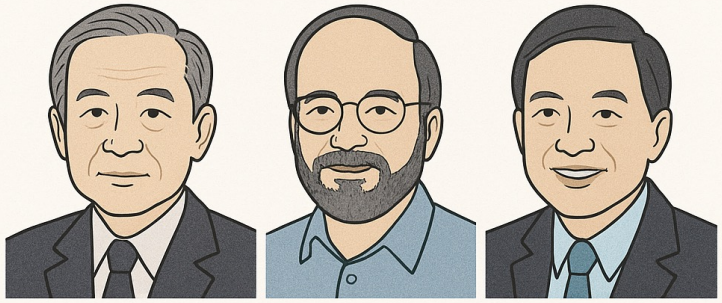
- **发现新基因与非编码序列**

高质量组装能发现未知基因、转座子、重复序列等，完善对基因组结构的理解。

。。。

The Nobel Prize in Chemistry 2008

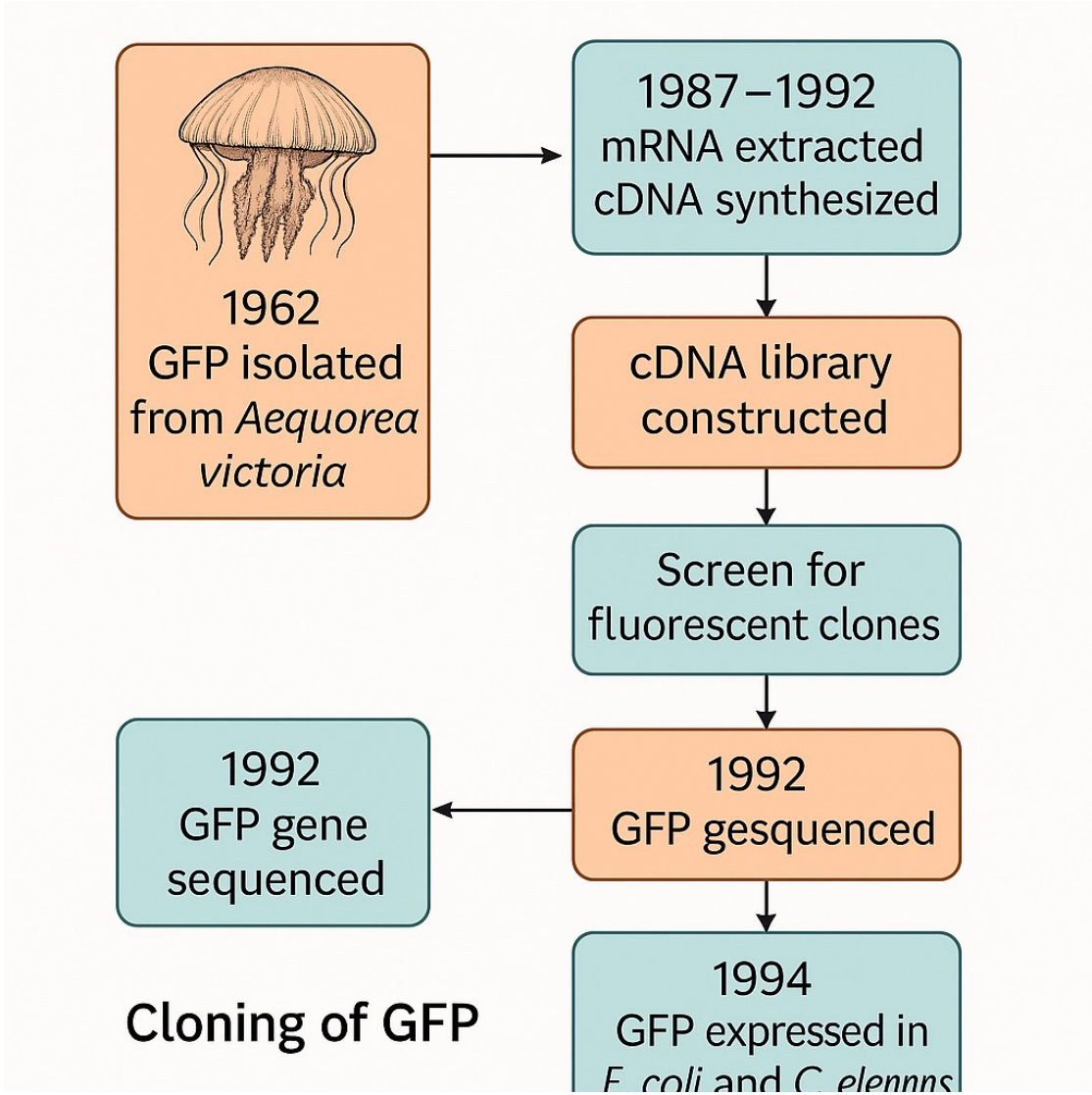
for the green fluorescent protein



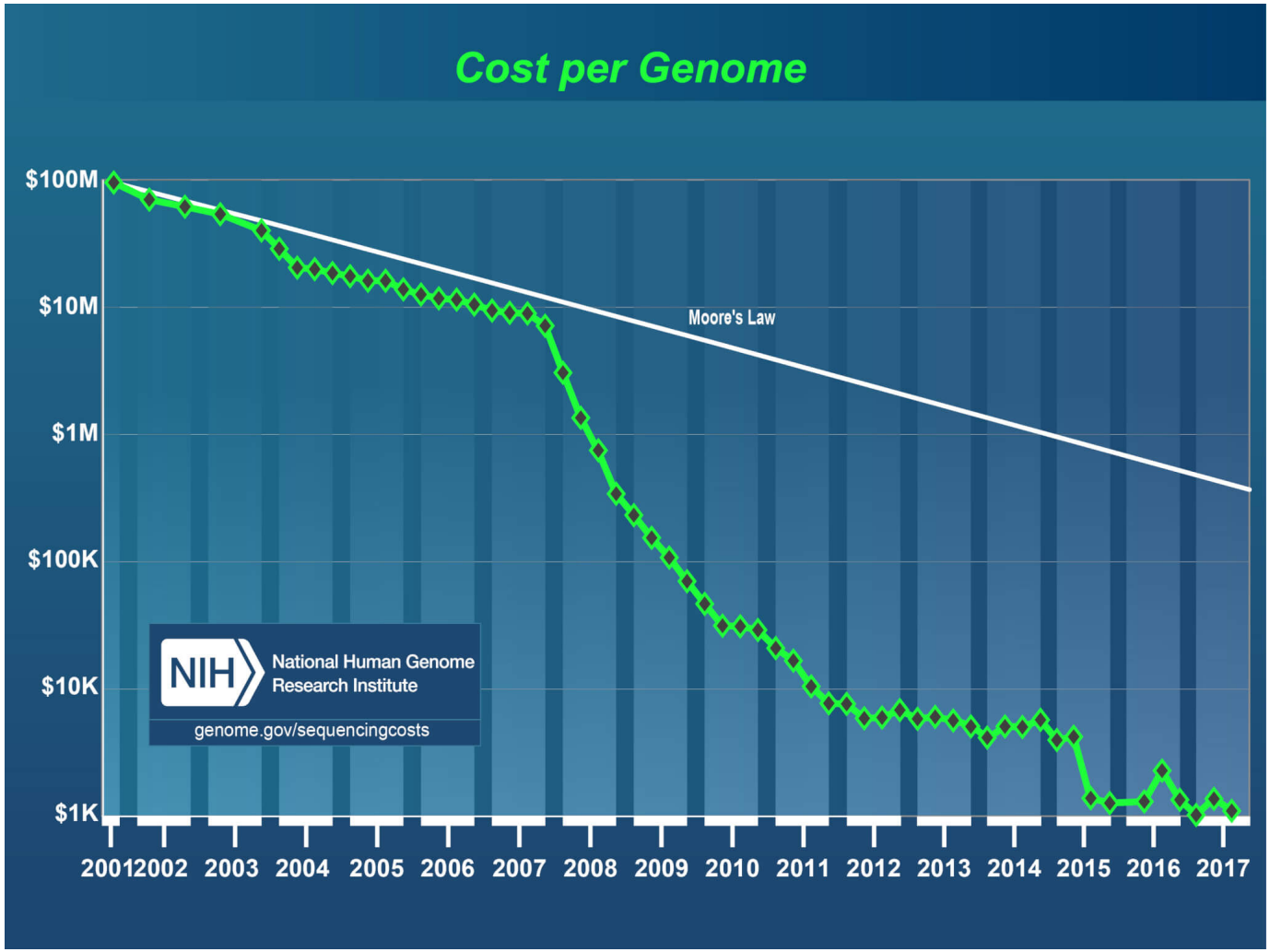
Osamu Shimomura (b. 1928)

Martin Chalfie (b. 1947)

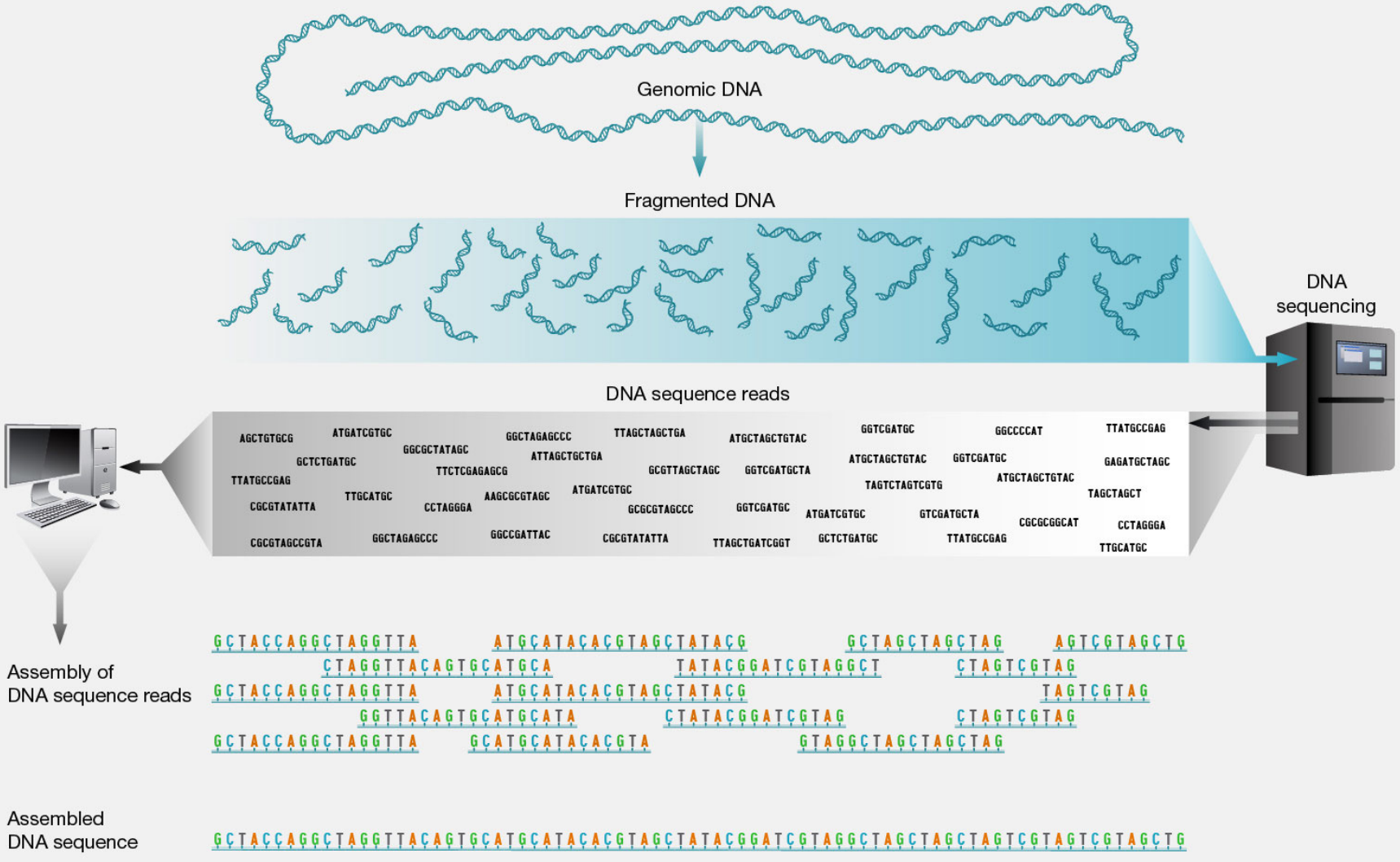
Roger Y. Tsien (b. 1952)



基因组的组装、预测和注释

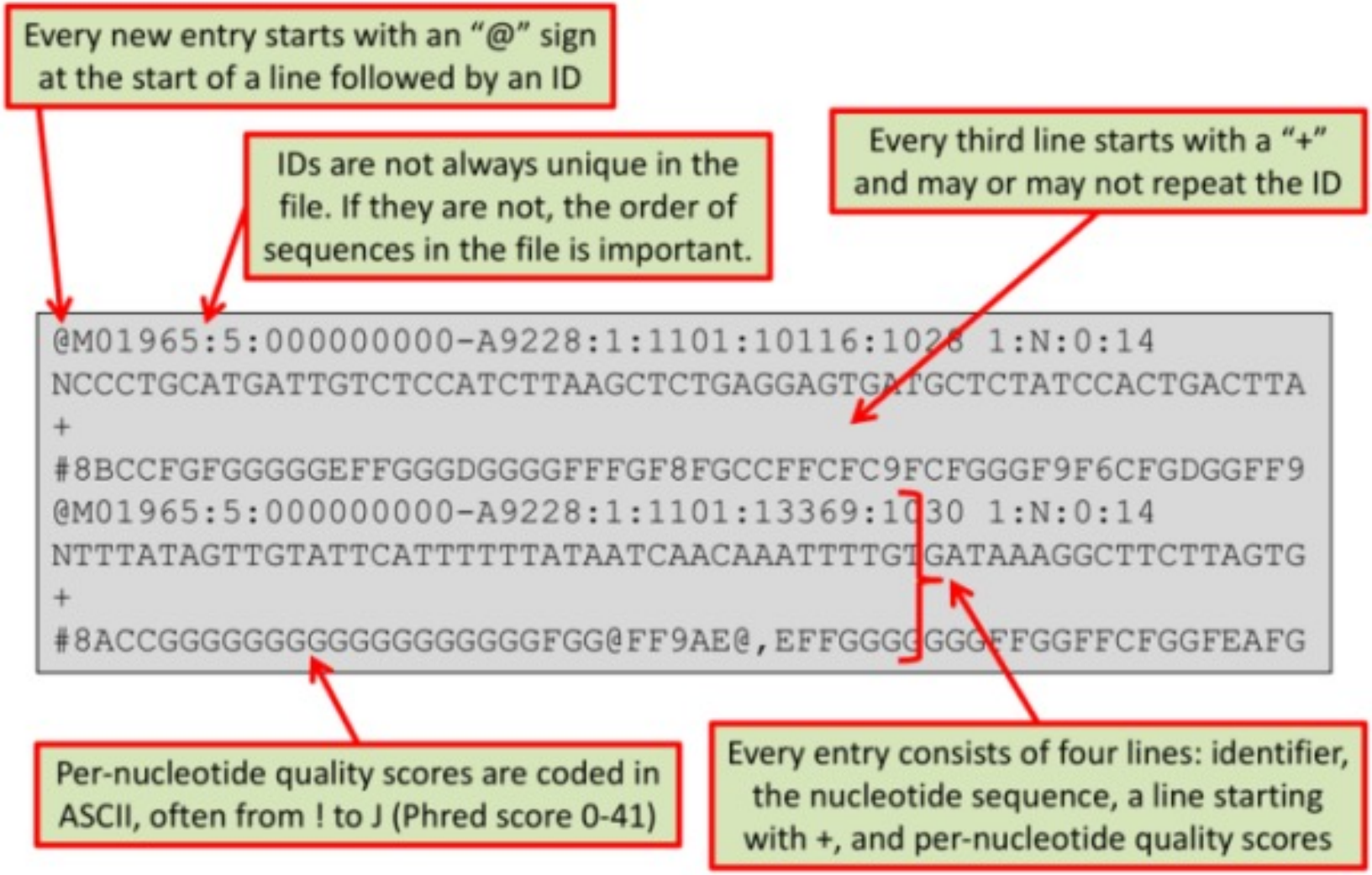


Overview


Genomic DNA
Fragmented DNA
DNA sequencing
DNA sequence reads
Assembly of
DNA sequence reads
Assembled
DNA sequence

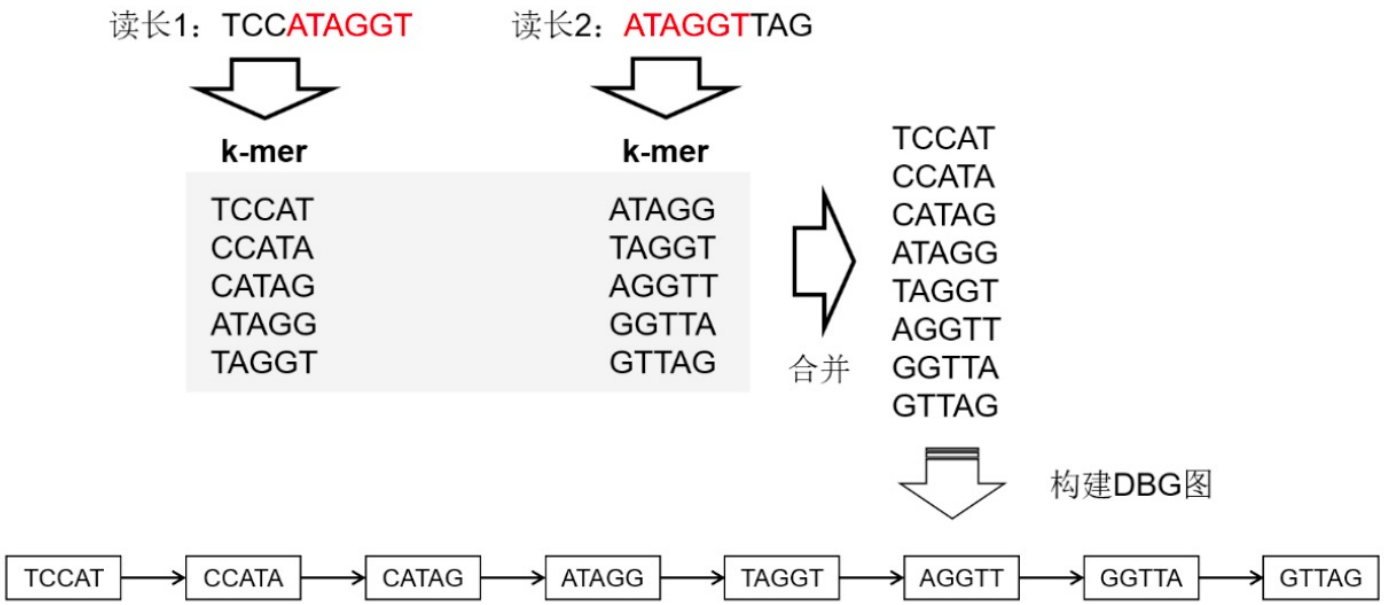
基因组的组装、预测和注释

FASTQ File



基因组的组装

K-mer=5

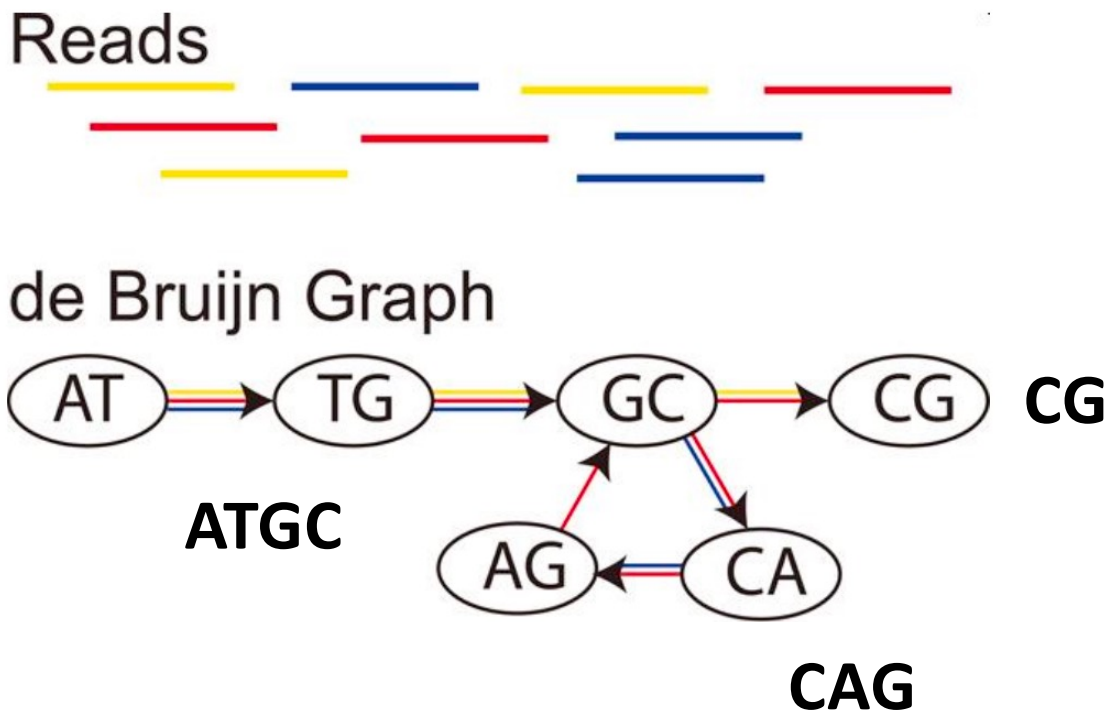


De Bruijn graph（DBG）图构建示例

基因组的组装、预测和注释

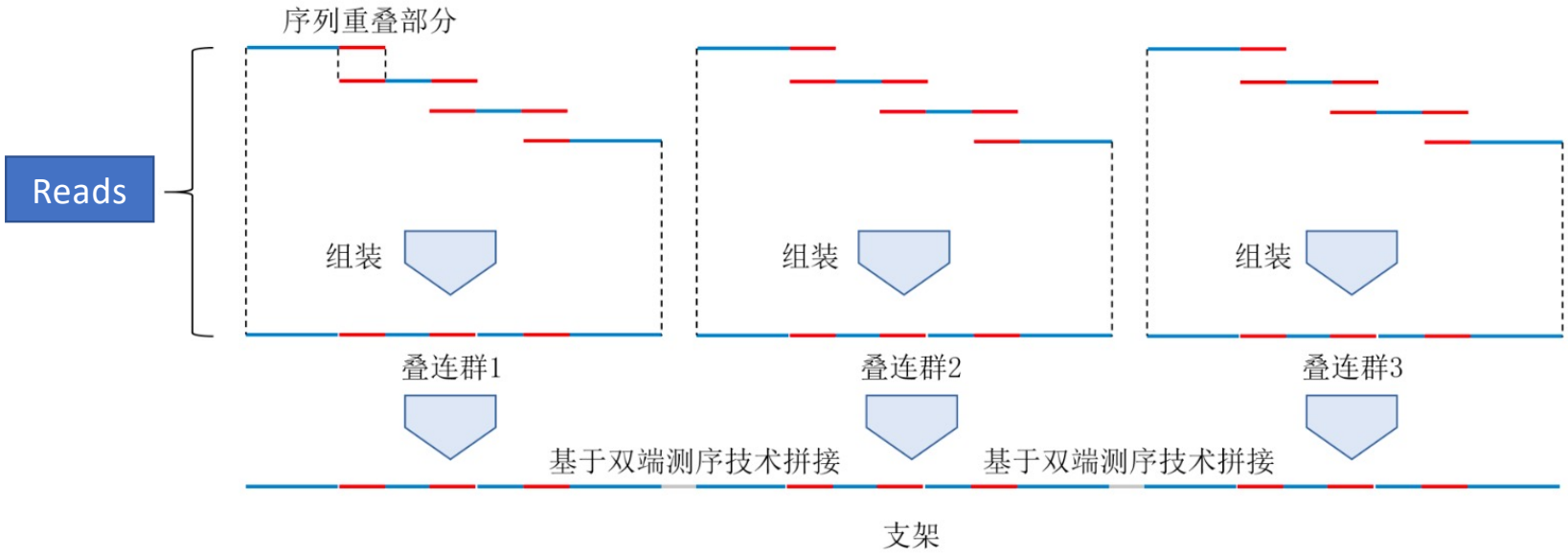
基因组的组装

K-mer=2



基因组的组装、预测和注释

基因组的组装



读长（read）、叠连群（contig）和支架（scaffold）的关系示意图

为什么会产生不连续的contig？　　重复序列

怎么获得更长的Scaffold？

1. 利用配对末端测序 (Paired-end / Mate-pair Reads)
这是最经典、最基础的 Scaffolding 方法，主要用于二代测序（如 Illumina）。
•**Mate-pair Reads** 原理类似，但它们的 Insert Size 更大（通常在 2 kb 到 40 kb），因此能跨越更长的距离，连接更远的 Contigs，对于构建高质量的 Scaffold 非常重要。

2. 利用三代长读长 (Long Reads)
PacBio 或 Oxford Nanopore 产生的长读长（**可达几十 kb 甚至上 Mb**）为 Scaffolding 提供了极其强大的信息。
•**原理**: 一条足够长的三代读长可能**自身就跨越了多个 Contigs。**
•**过程**: 将这些长读长比对回 Contigs 上。如果一条长读长的前半段比对上了 Contig A，后半段比对上了 Contig B，那么它就直接提供了 A 和 B 的顺序、方向和它们之间的序列信息。
•**优势**: 不仅能完成 Scaffolding（确定顺序和方向），还能在很大程度上 **填补 (Gap Filling)** Contigs 之间的缺口，因为长读长直接测通了原来的未知区域。用长读长完成 Scaffolding 和补洞的过程通常被称为“基因组Polish（润色）”。

3. 利用染色体构象捕获技术 (Hi-C)

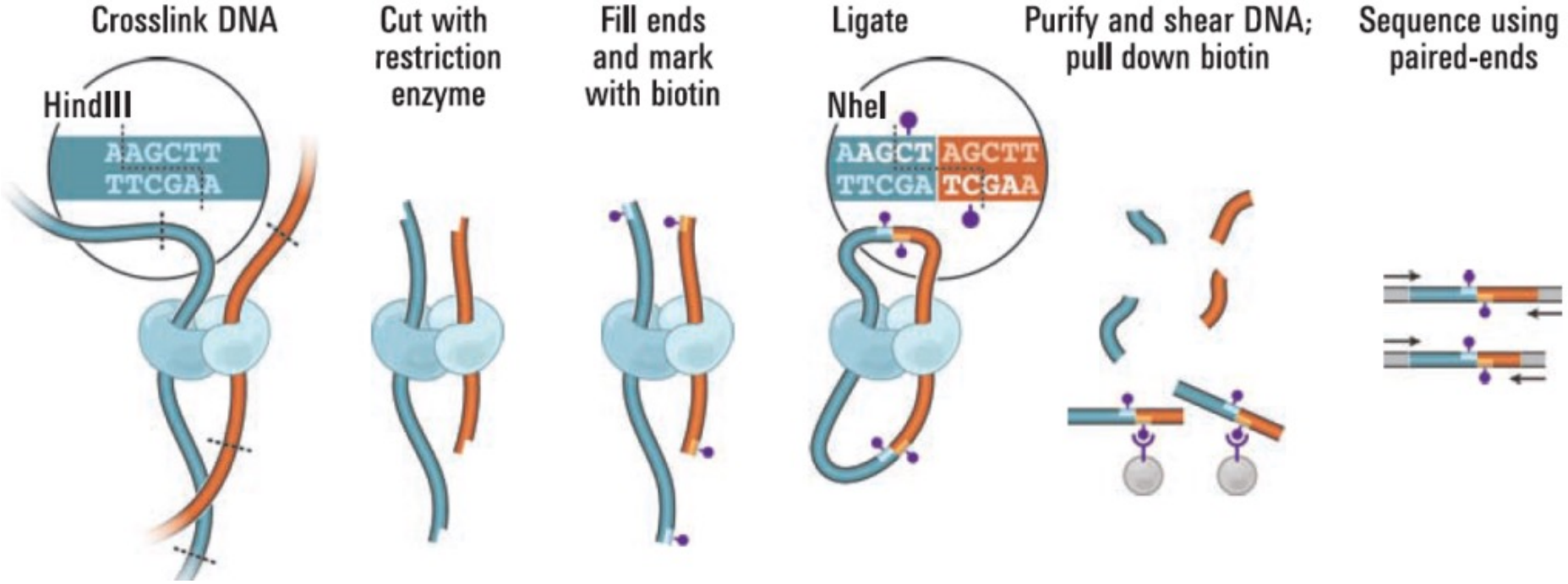
这是一种用于构建染色体级别组装的“终极武器”。

•**过程**:

- 通过 Hi-C 实验，获得海量的“染色质交互”读长对。
- 将这些读长对比对回已有的 Contigs 或 Scaffolds 上。
- 统计任意两个 Contig 之间存在交互读长对的数量。
- 算法会根据交互频率矩阵，将 Contigs 进行聚类、排序和定向，最终将它们锚定到代表染色体的不同组中。

•**优势**: Hi-C 提供了超长距离的连接信息（可达整条染色体的尺度），是目前将基因组组装从 Scaffold 提升到 **染色体 (Chromosome) 级别** 的最主要技术。

Hi-C技术



在同一条染色体上线性距离越近的DNA片段，在三维空间中相互接触的频率就越高。

3. 利用染色体构象捕获技术 (Hi-C)

这是一种用于构建染色体级别组装的“终极武器”。

•**过程**:

- 通过 Hi-C 实验，获得海量的“染色质交互”读长对。
- 将这些读长对比对回已有的 Contigs 或 Scaffolds 上。
- 统计任意两个 Contig 之间存在交互读长对的数量。
- 算法会根据交互频率矩阵，将 Contigs 进行聚类、排序和定向，最终将它们锚定到代表染色体的不同组中。

•**优势**: Hi-C 提供了超长距离的连接信息（可达整条染色体的尺度），是目前将基因组组装从 Scaffold 提升到 **染色体 (Chromosome) 级别** 的最主要技术。

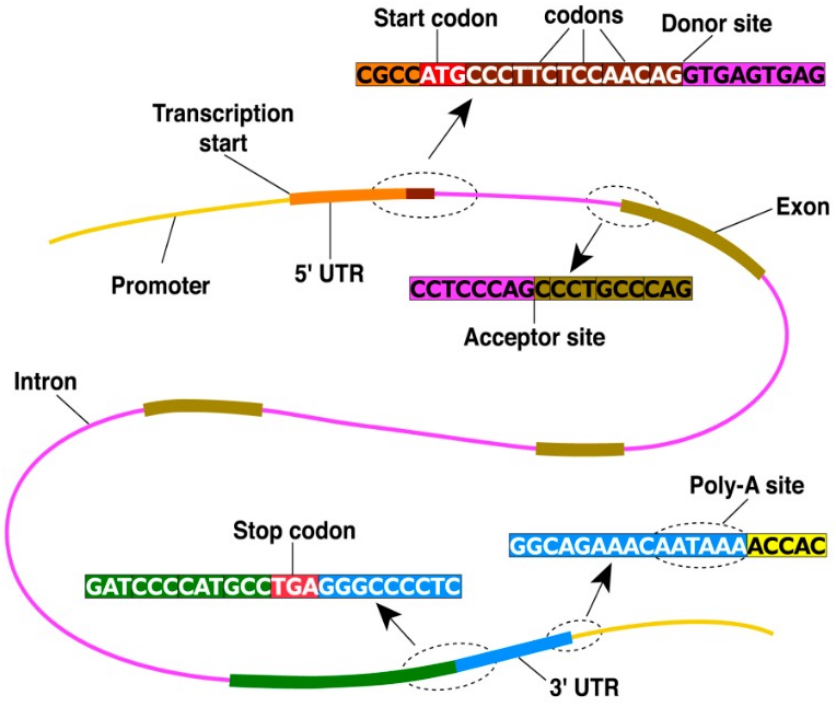
基因组的组装

软件	应用	备注
SOAPdenovo	二代测序数据基因组组装	可应用于所有生物
SPAdes	二代测序数据基因组组装	主要应用于原核生物
Flye	三代测序数据基因组组装	
canu	三代测序数据基因组组装	

部分代表性基因组组装软件

基因组的组装、预测和注释

基因的预测

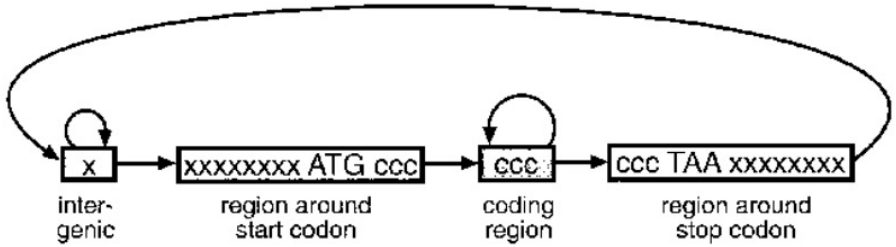


基因预测原理

基因的预测

- Nucleotides *{A,C,G,T}* are the observables
- Different states generates generate nucleotides at different frequencies

A simple HMM for unspliced genes:



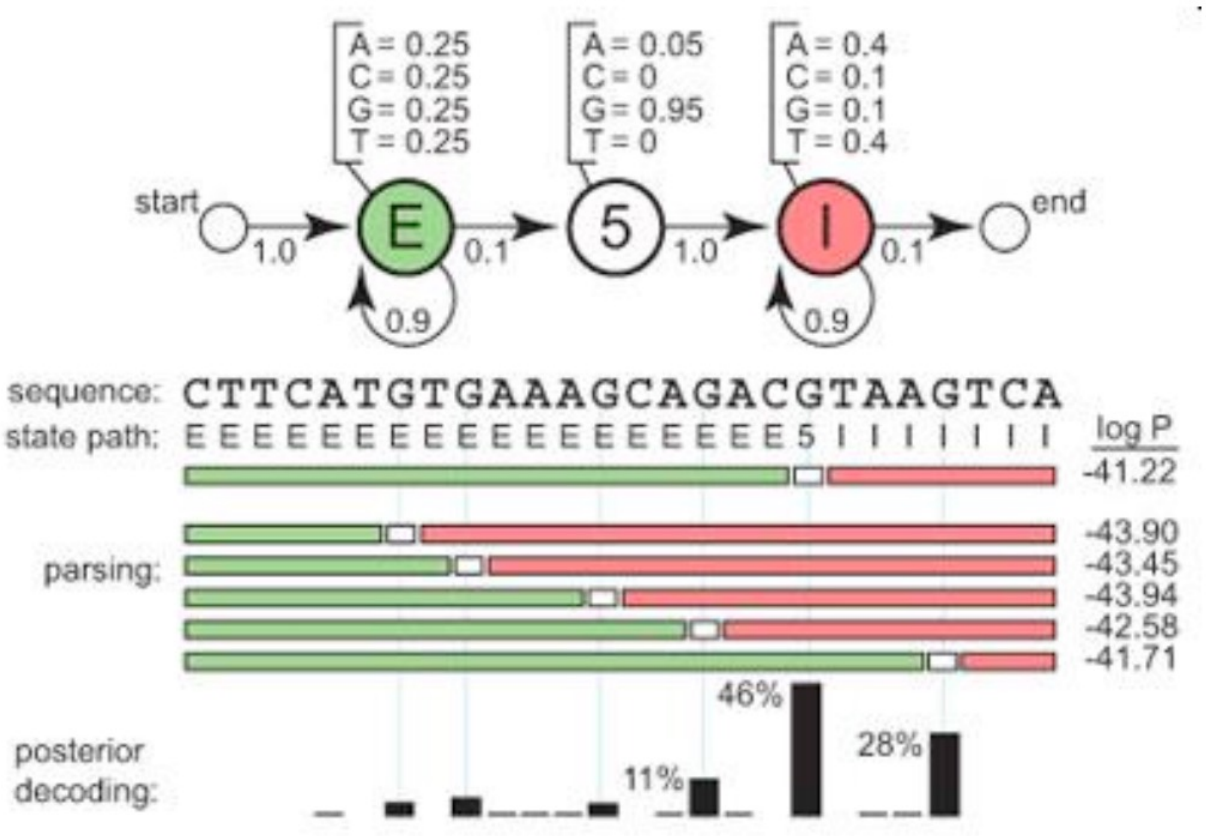
AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

- The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

基因预测原理（HMM）

基因组的组装、预测和注释

基因的预测



基因预测原理（HMM）

优点:

- 不需要任何先验知识，可以发现全新的、独有的基因。
- 对于没有近缘物种或转录组数据的基因组是唯一的方法。

缺点:

- 准确率相对较低，容易产生假阳性（将非基因预测为基因）和假阴性（错过真实基因）。
- 对于非典型的基因结构（如不遵循GT-AG法则的剪接）识别能力差。

基因的预测

软件	应用
GeneMarkS	原核生物基因预测
GeneMarkES	真核生物基因预测
Prokka	原核生物基因预测
Augustus	真核生物基因预测
prodigal	原核生物基因预测

部分代表性基因预测软件

2. Homology-based / Evidence-based (同源或证据依赖预测法)

这种方法的原理是：**利用已知的生物学证据来定位基因**。

A. 同源序列比对 (Homology)

将目标基因组序列与公共数据库中已有的、来自其他物种（通常是近缘物种）的已知基因或蛋白质序列进行比对。

B. 转录组数据比对 (Transcriptomic Evidence)

通过对特定组织或条件下的细胞进行**RNA测序 (RNA-Seq)**，可以获得该物种在当下所有被激活和转录的基因序列。

将RNA-Seq产生的读长（reads）比对回基因组，可以直接“看”到哪些区域被转录了。

跨越内含子的reads（spliced reads）可以极其精确地确定外显子-内含子的边界。

代表软件: Exonerate, GenomeThreader (用于比对), StringTie, Cufflinks (用于基于RNA-Seq的转录本组装)。

优点:
准确率非常高，因为有直接的生物学证据支持。
能够精确确定基因的结构，特别是可变剪接产生的不同转录本。

缺点:
严重依赖外部数据库的完整性和相关性。如果一个基因是该物种特有的，或者在测序的组织/条件下不表达，这种方法就无法找到它。
无法发现新基因。

3. 整合/混合预测法 (Integrated/Hybrid Approach)

证据收集: 首先，将所有可用的证据（同源蛋白、ESTs、RNA-Seq数据）比对到基因组上，得到一组高可信度的基因模型。

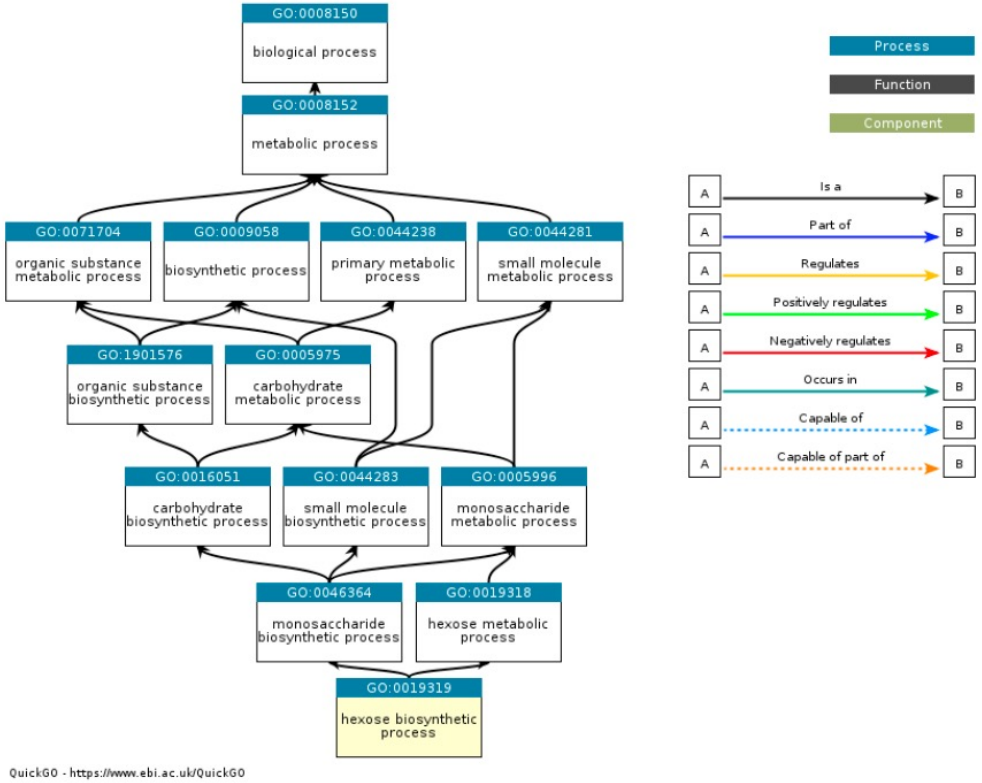
模型训练: 使用这组高可信度的基因作为“训练集”，来优化和训练 ab initio 模型的参数（例如，特定物种的密码子偏好性、剪接位点信号等）。

全基因组预测: 用训练好的模型对整个基因组进行扫描，预测所有的基因，包括那些没有直接证据支持的新基因。

代表软件: MAKER, BRAKER。这类软件本质上是自动化的流程，它内部集成了多种预测和比对工具，并负责训练、预测和整合的全过程。

基因组的组装、预测和注释

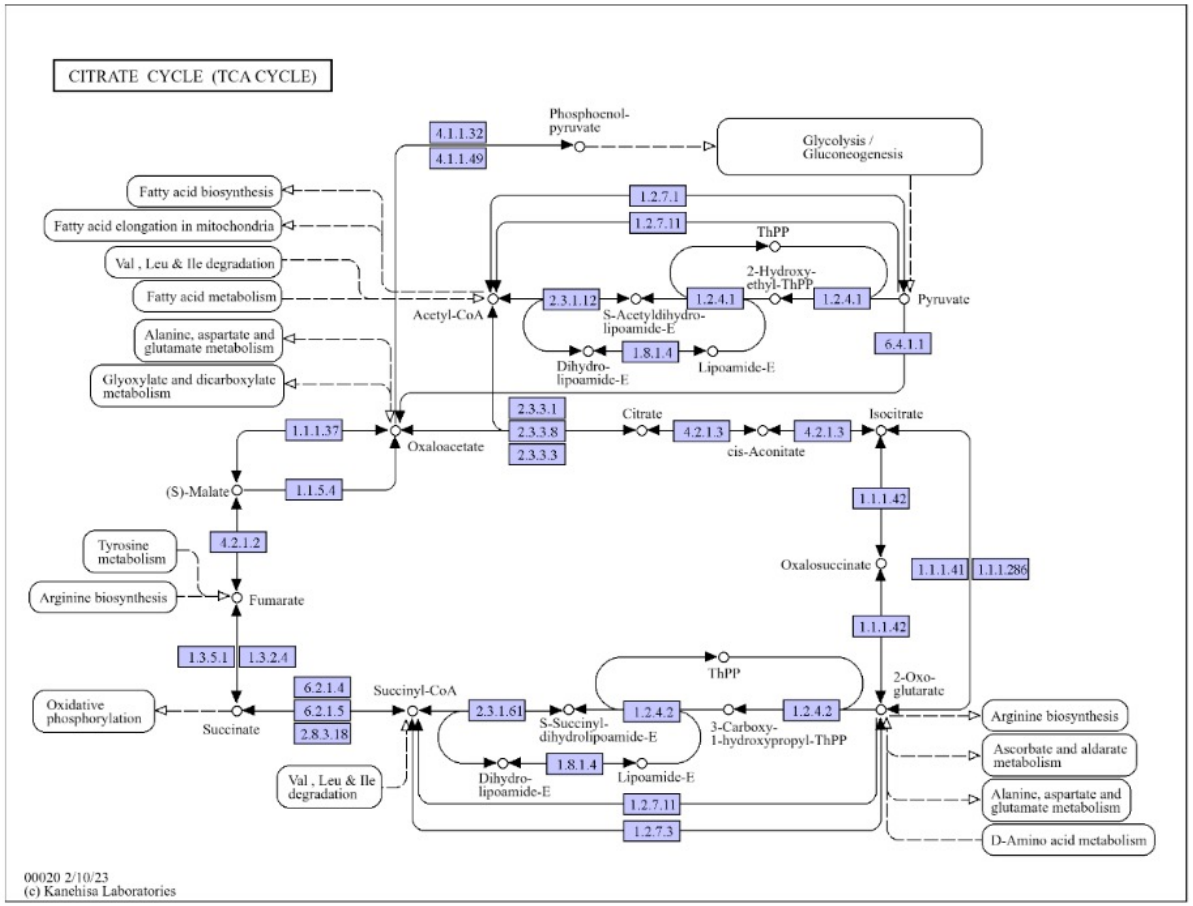
基因组的注释和分析



GO基因注释的示例（GO:0019319）

Tool: Eggnog

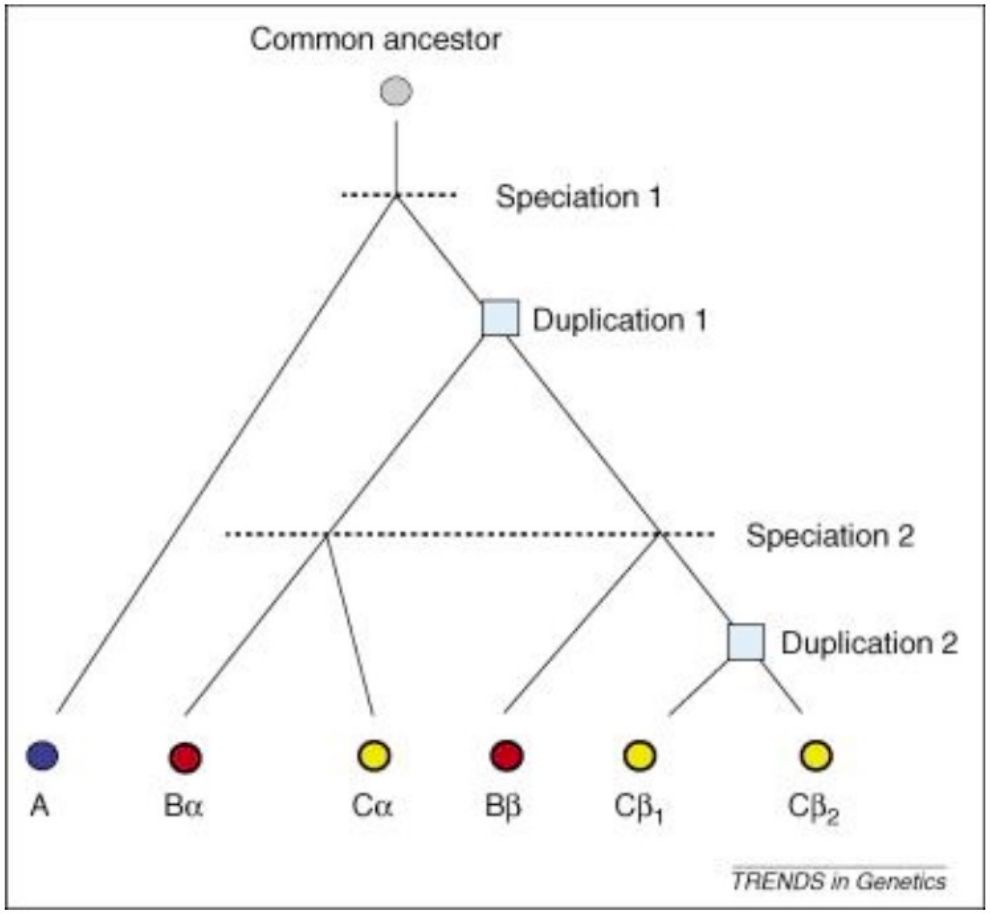
基因组的组装、预测和注释



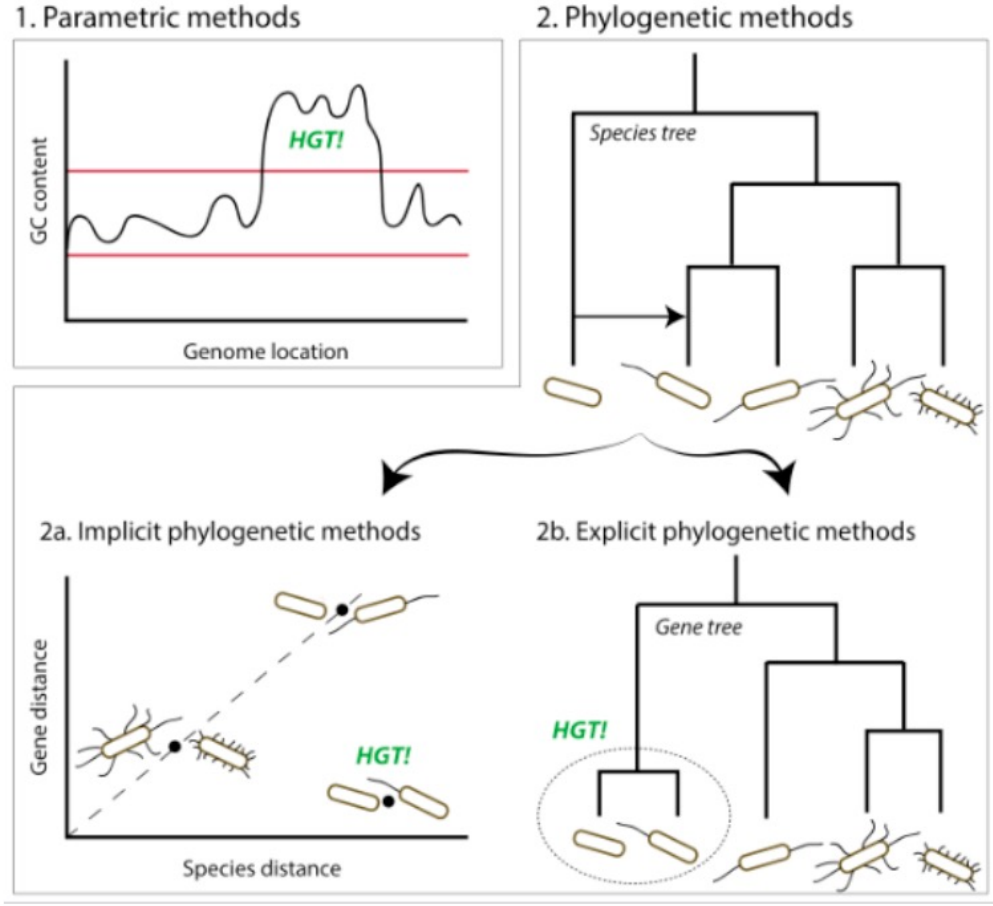
KEGG基因注释的示例（TCA循环）

Tool: Eggnog

同源基因

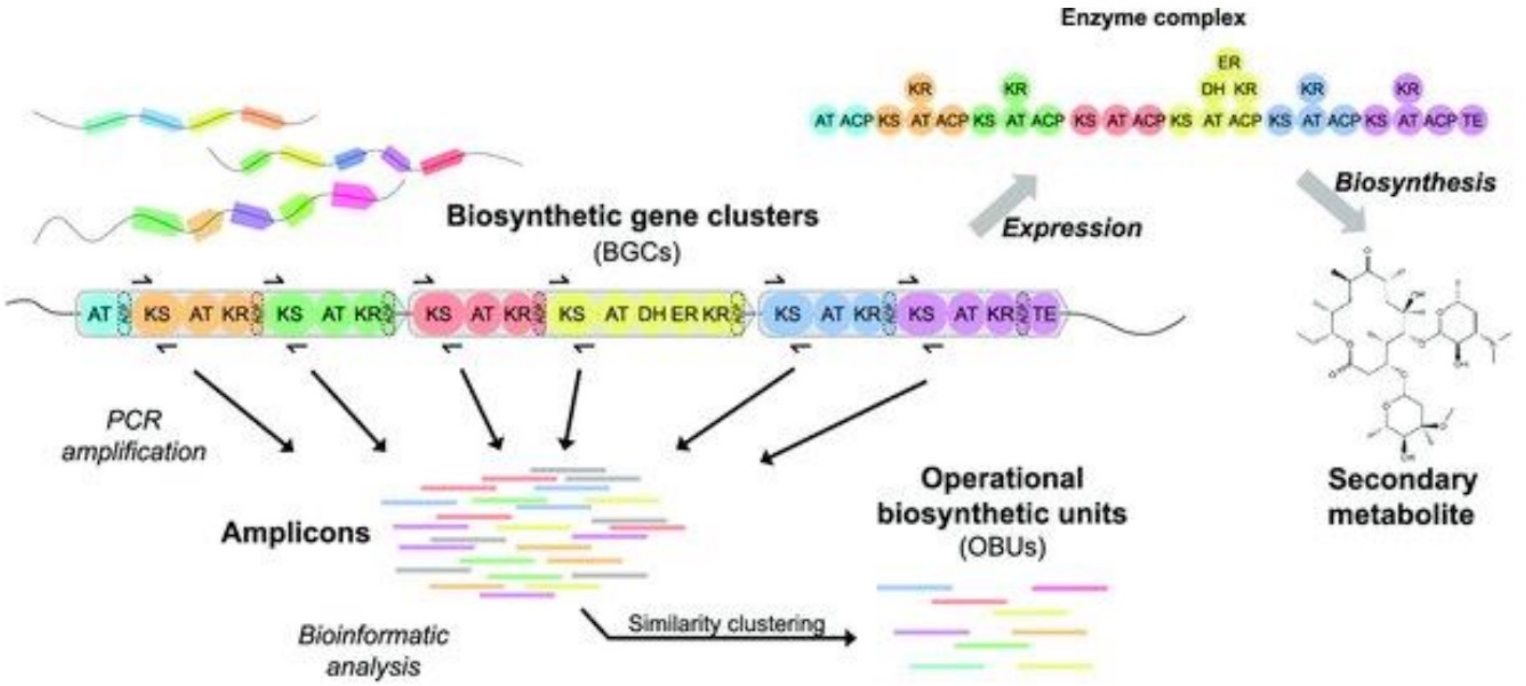

Common ancestor
Speciation 1
Duplication 1
Speciation 2
Duplication 2
A
Bα
Cα
Bβ
Cβ1
Cβ2
TRENDS in Genetics

基因的水平转移（horizontal gene transfer， HGT）



基因组的组装、预测和注释

基因组的注释和分析



Biosynthetic gene clusters (BGCs) 生合成基因簇

Tool: Antismash

基因组的组装、预测和注释

实验课教程网址：

https://genomics.sschmeier.com/ngs-assembly/