第六章 基因组学及宏基因组学

编者：宁康、窦岩梅、王明钰、王涛、周宇

审读：李兴旺

本章内容

- 首先，定义了基因组学，并讨论了获得和解读基因组序列的方法，并介绍了高通量测序技术在动植物和微生物DNA测序中的应用。
- 其次，详细介绍了基因组组装的算法、基因预测技术和基因组注释的方法。
- 第三，深入探讨了序列变异检测的基本原理和技术。
- 第四，宏基因组学部分概述了微生物组学的概念、数据分析方法，以及在健康和环境领域的应用。
- 最后，对基因组和宏基因组的未来发展方向进行了展望。

目录大纲

01 **基因组学概述**

研究基因组序列、结构和功能的科学。

02 **基因组组装与注释**

基因组组装算法、基因预测技术和基因组注释方法。

03 **序列变异检测**

单碱基替换、短插入缺失和结构变异的检测技术。

04 **宏基因组学**

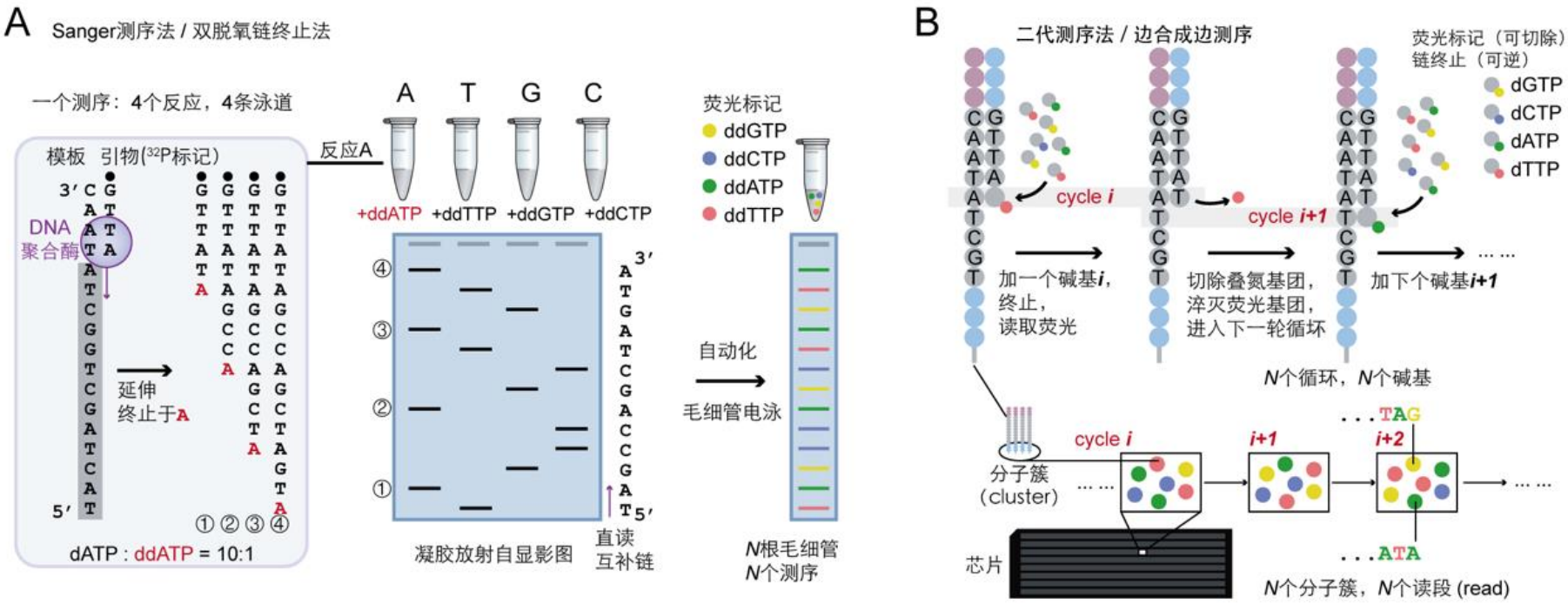
微生物组学概念、数据分析方法，以及在健康和环境领域的应用。

基因组学概述

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

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

测序原理及方法



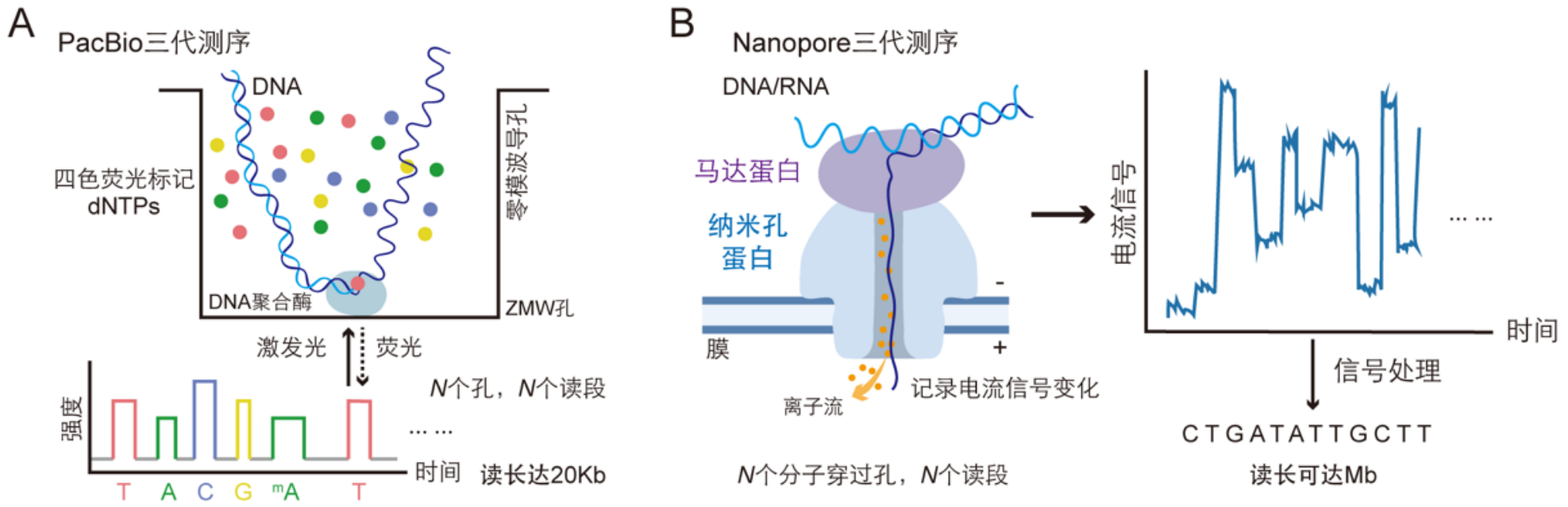
(A)Sanger测序法基本原理及其毛细管电泳自动化。(B)二代高通量并行测序原理。

测序原理及方法

	仪器成本（$）	测序成本($/Mbase)	读长（bp）	通量	测序质量（Q30）	特点
HiSeq (NovaSeq XPlus)	~100万	~0.02	2x150	每日～8Tb	≥85%	通量高、单Read成本低
BGISEQ -T20	~100万	~0.01	2x150	每小时900Gb	≥80%	成本效益高

两种常用二代测序技术的比较

测序原理及方法



三代测序技术原理示意图。(A)PacBio长读长测序法基本原理。(B)Nanopore长读长测序基本原理。

基因组学概述

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

基因组学概述

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

常用工具及其作用

基因组学概述

基因组学数据可视化

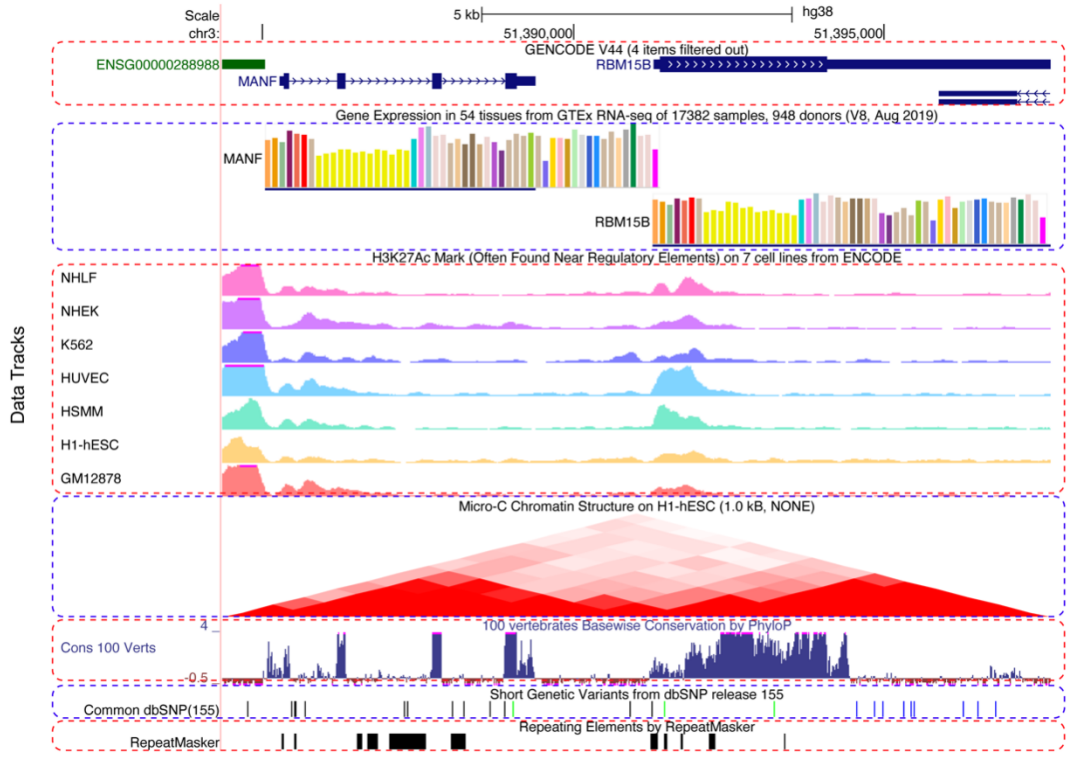
- 常见的基于网络的集成基因组浏览器包括：UCSC和Ensembl Genome Browser、NCBI Genome Data Viewer等，提供了涵盖多个物种的基因组数据和工具。

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

基因组学概述

基因组学数据可视化

https://genome.ucsc.edu/



UCSC Genome Browser数据界面示例

目录大纲

01 **基因组学概述**

研究基因组序列、结构和功能的科学。

02 **基因组组装与注释**

基因组组装算法、基因预测技术和基因组注释方法。

03 **序列变异检测**

单碱基替换、短插入缺失和结构变异的检测技术。

04 **宏基因组学**

微生物组学概念、数据分析方法，以及在健康和环境领域的应用。

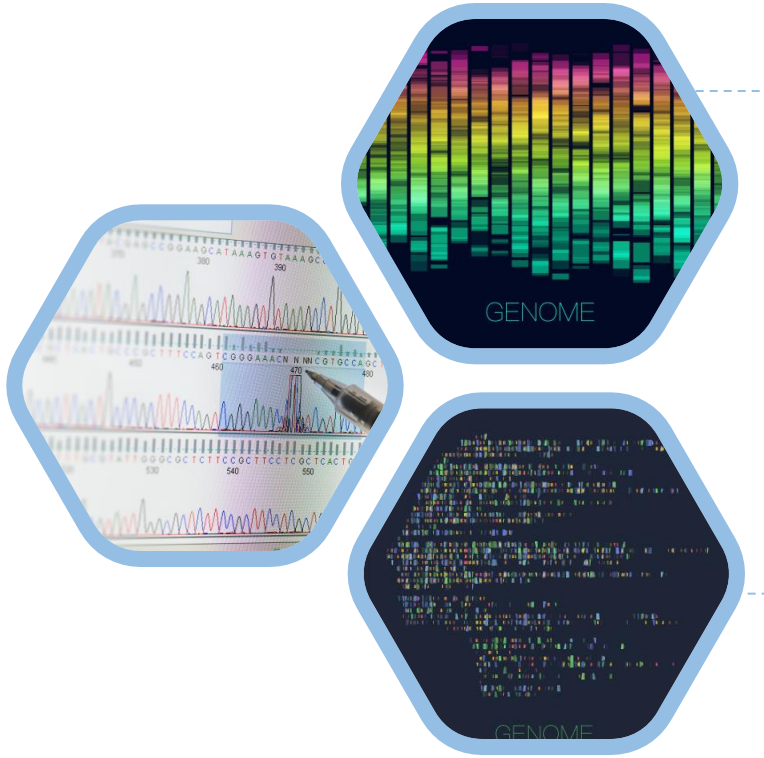
基因组的组装、预测和注释

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

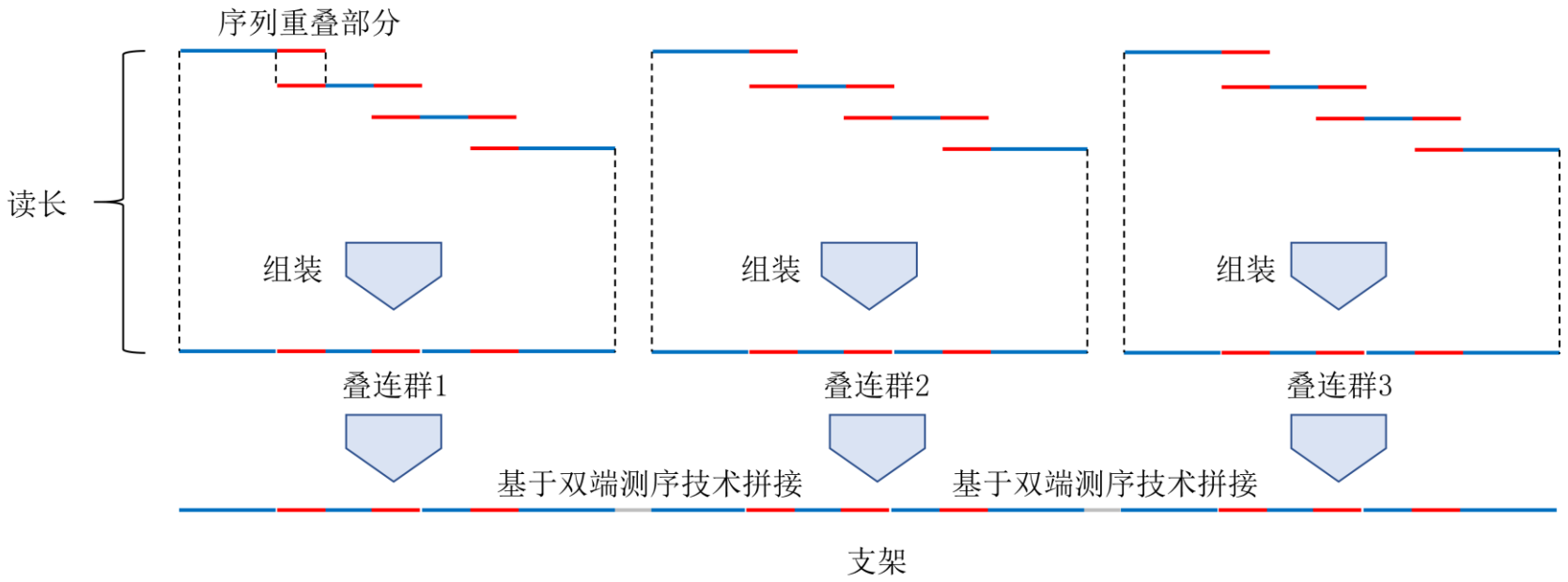
基因预测软件使用fasta格式的基因组序列文件作为输入，输出包括基因组索引文件、预测得到的基因序列文件和蛋白质序列文件。



3. 基因组的注释和分析

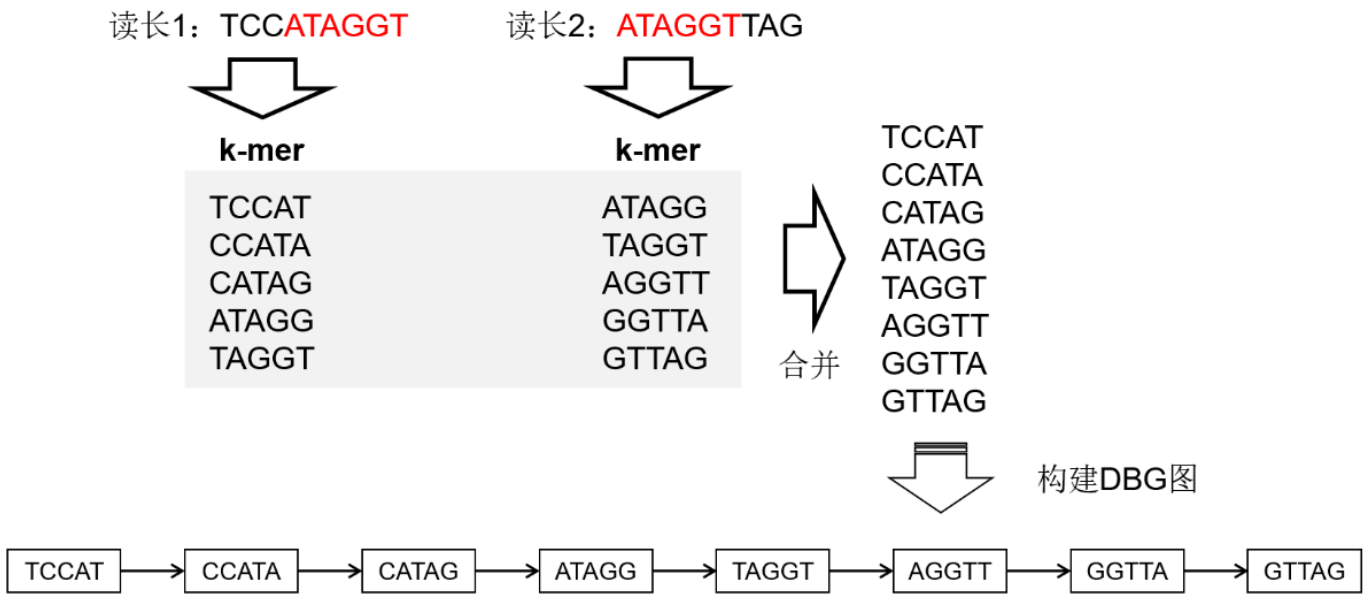
基因组分析的最后一步是对基因组进行注释和分析，包括GO、KEGG、NR等注释方法。

基因组的组装



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

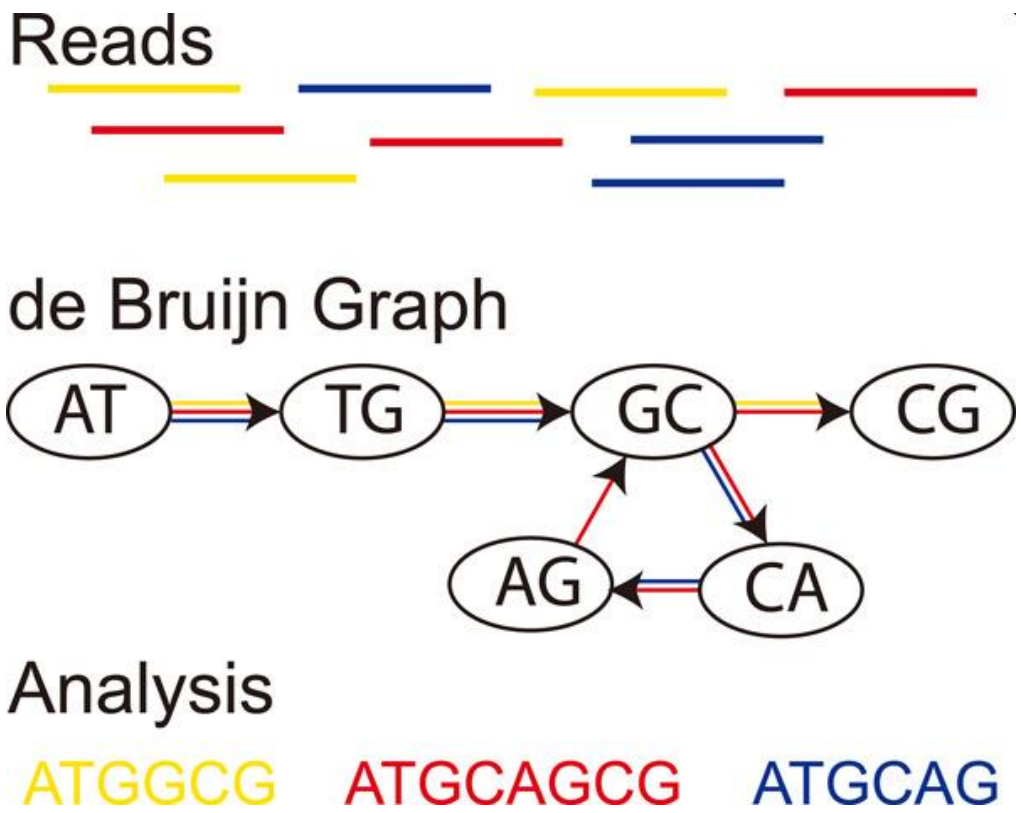
基因组的组装



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



De Bruijn graph（DBG）图构建示例

基因组的组装、预测和注释

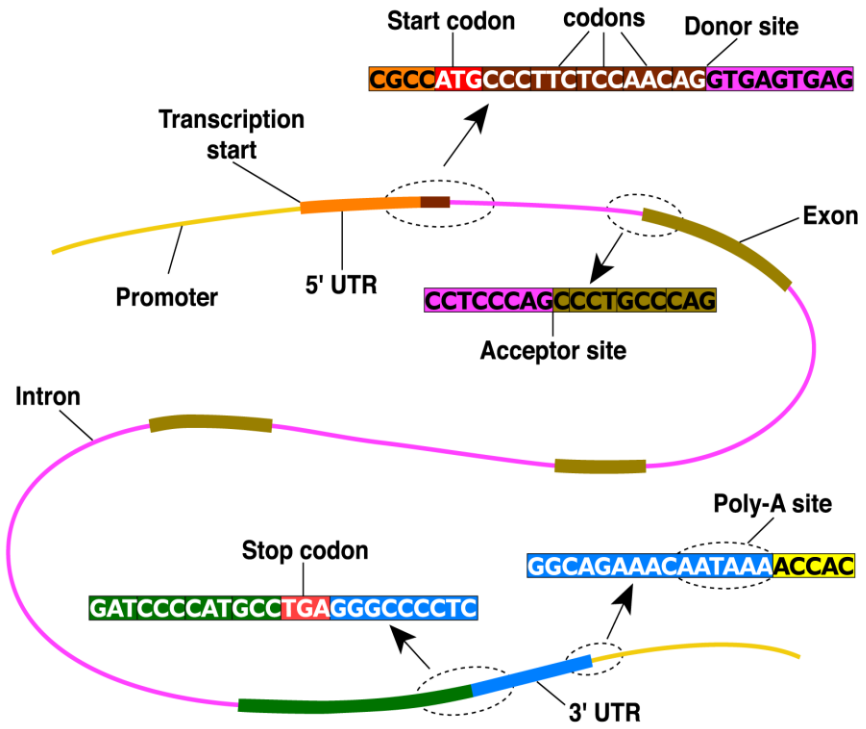
基因组的组装

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

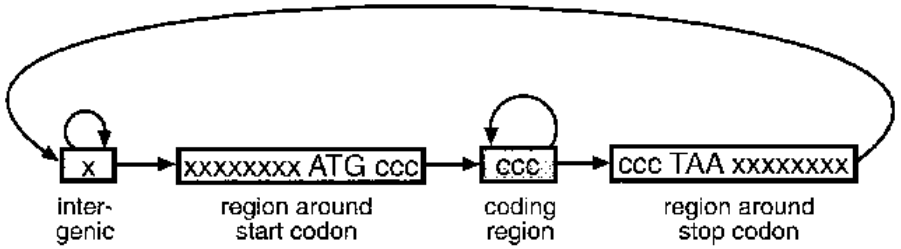


基因预测原理

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:



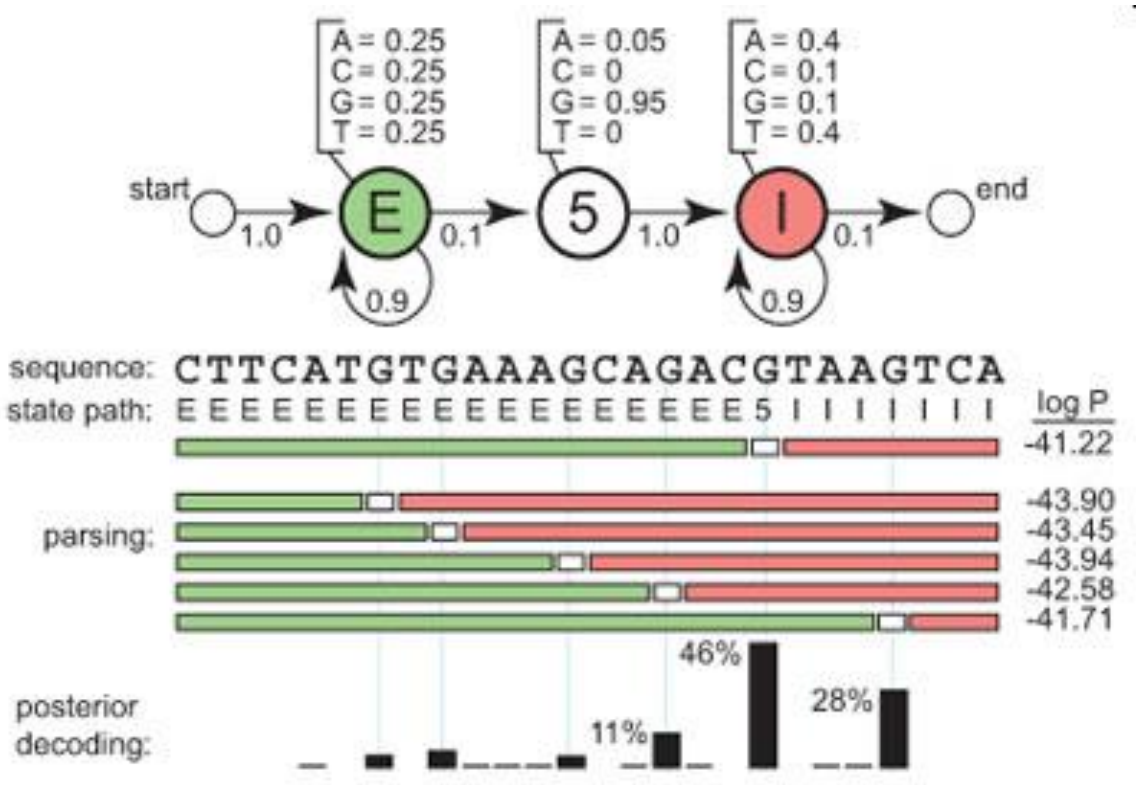
AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

- 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）

基因组的组装、预测和注释

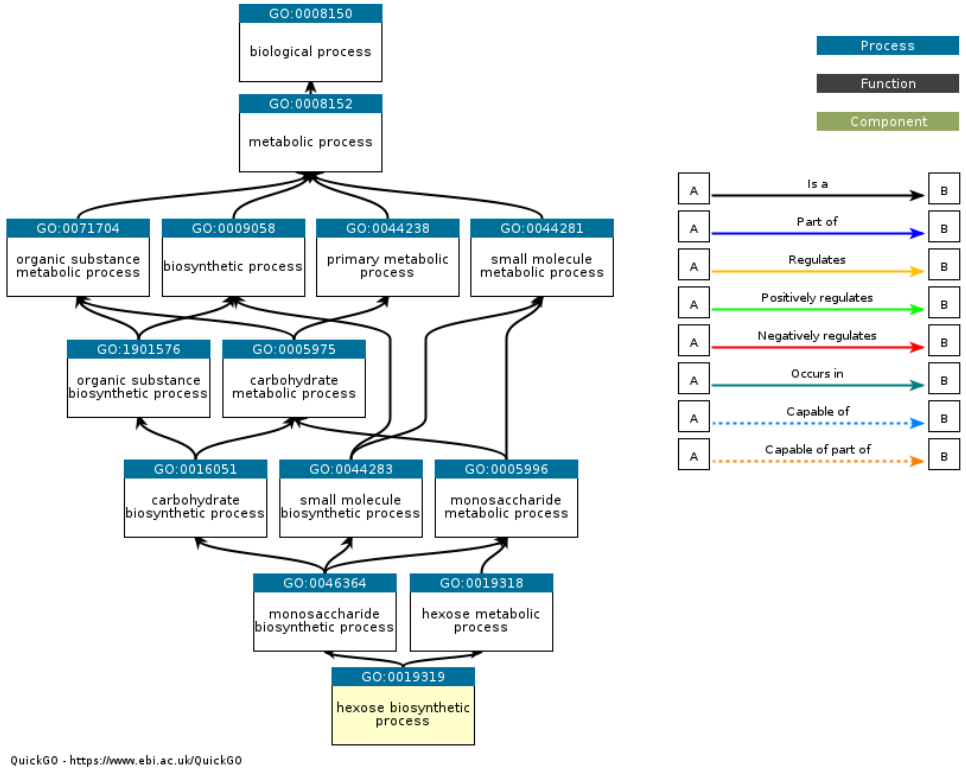
基因的预测

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

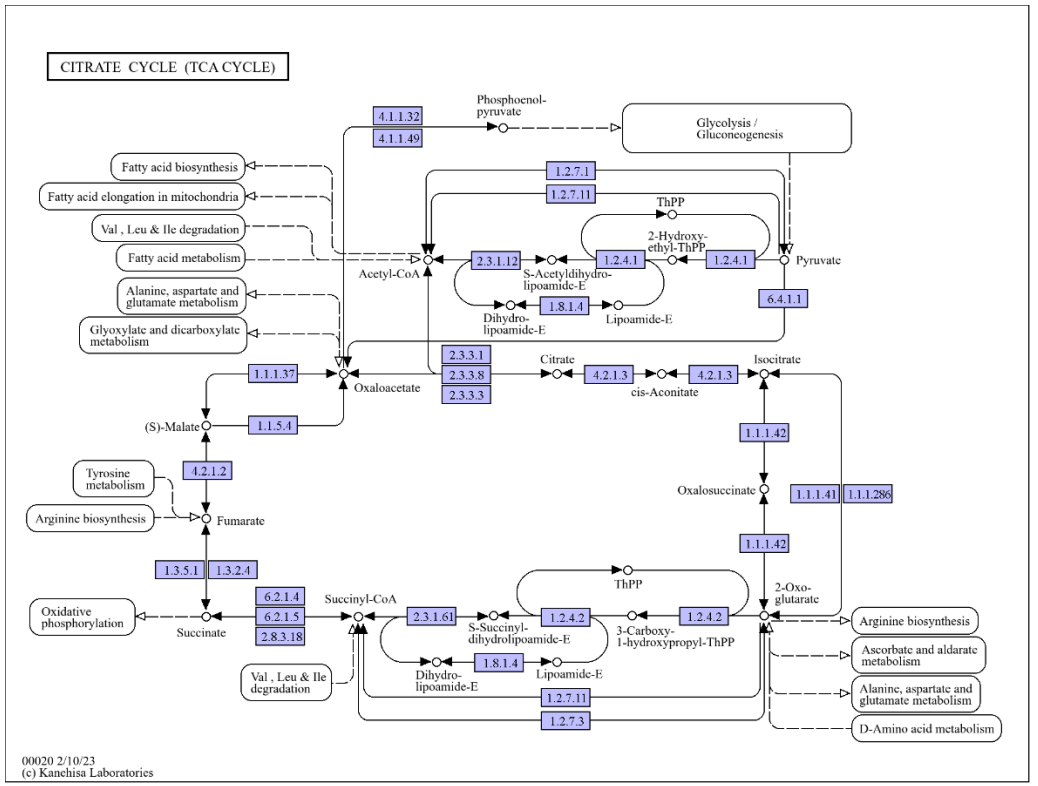
部分代表性基因预测软件

基因组的组装、预测和注释

基因组的注释和分析



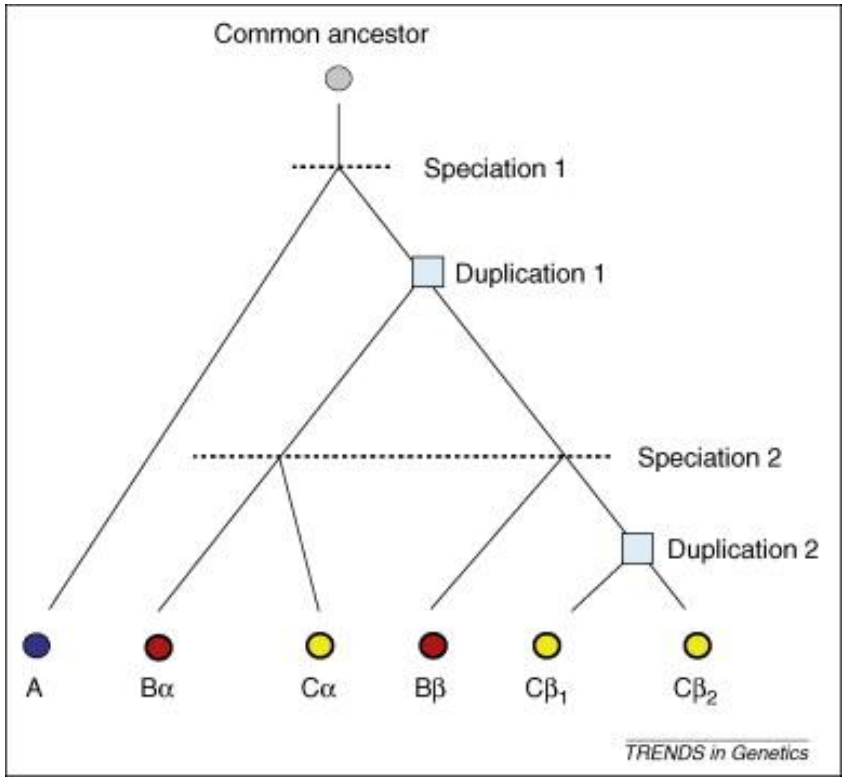
GO基因注释的示例（GO:0019319）

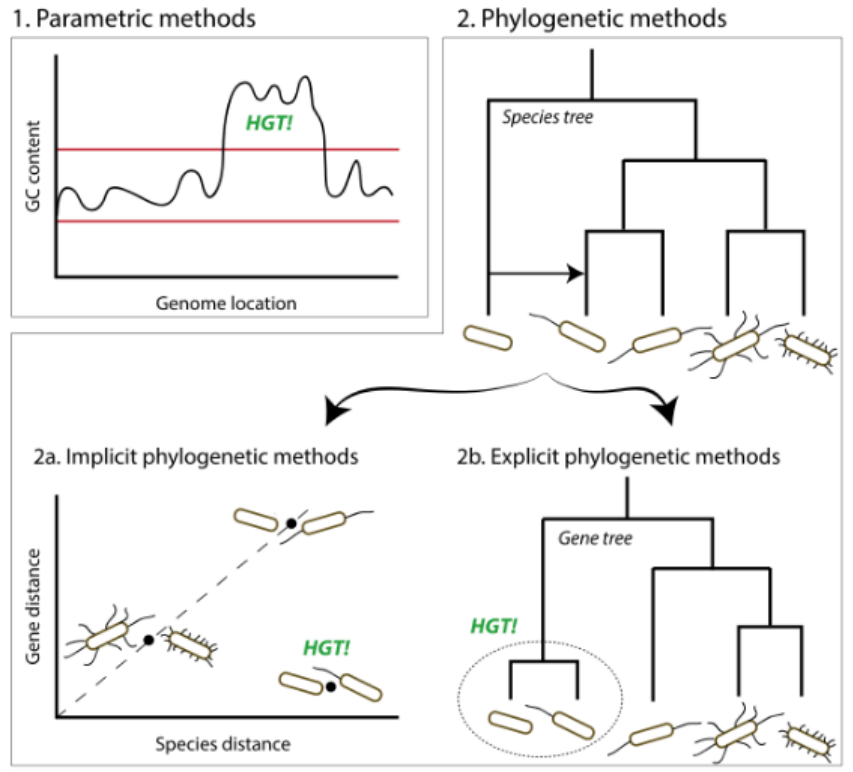


KEGG基因注释的示例（TCA循环）

基因组的组装、预测和注释

基因组的注释和分析

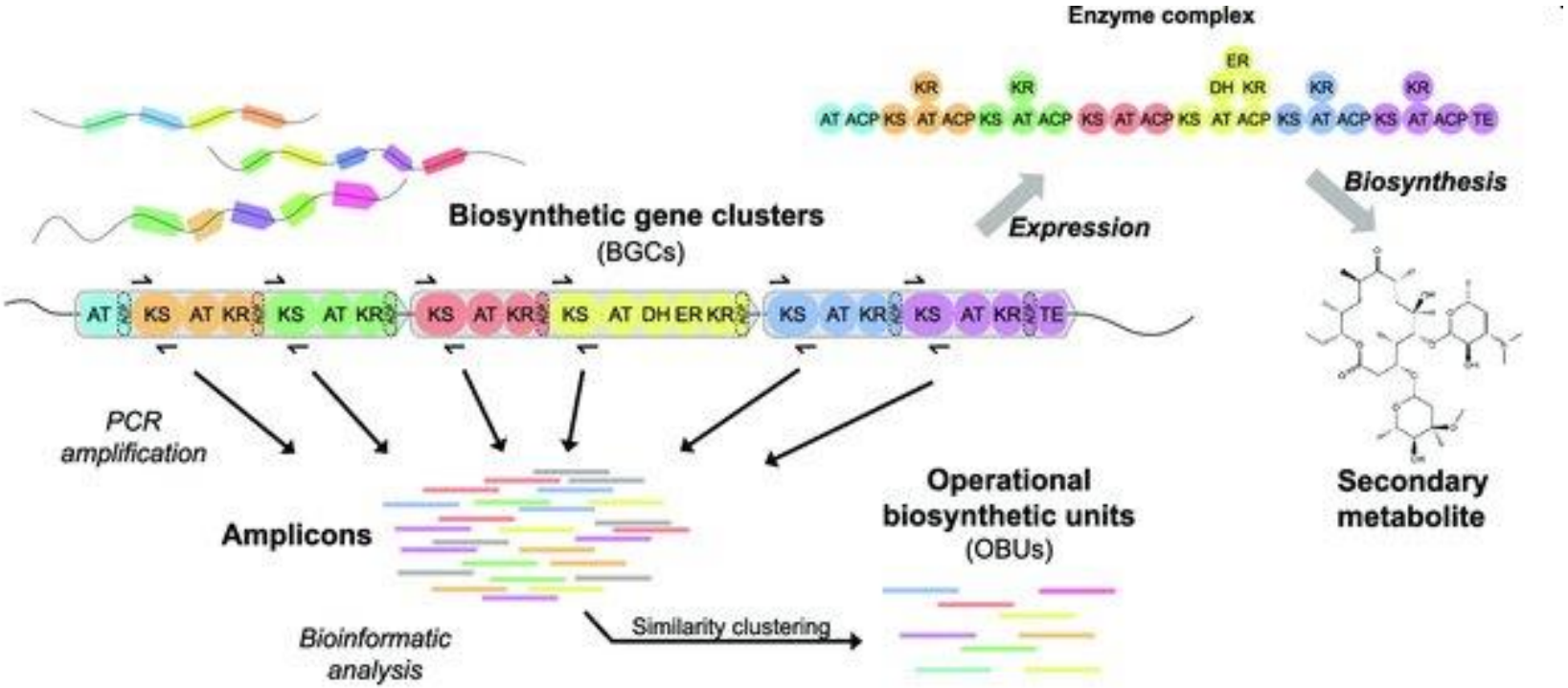




同源基因&HGT

基因组的组装、预测和注释

基因组的注释和分析

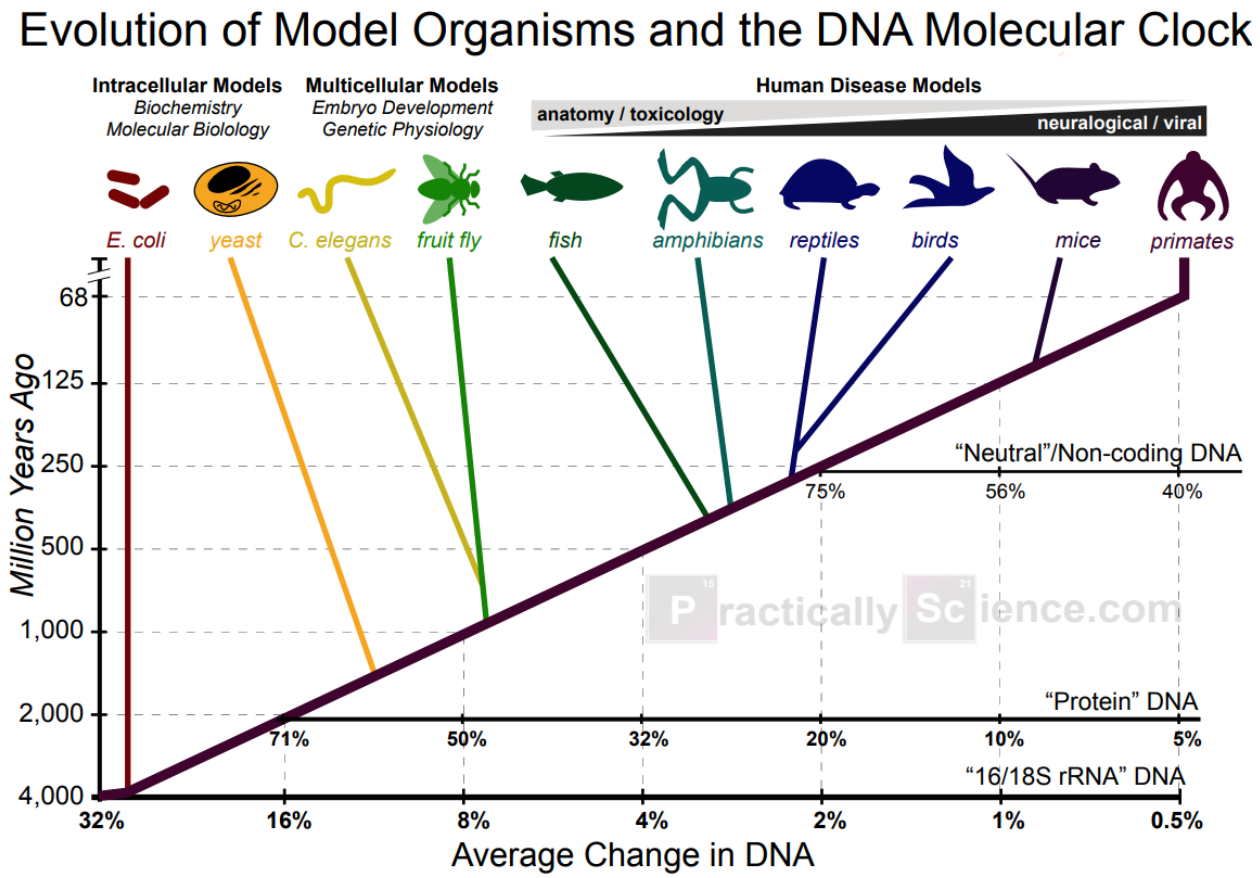


Biosynthetic gene clusters (BGCs) 生合成基因簇

目录大纲

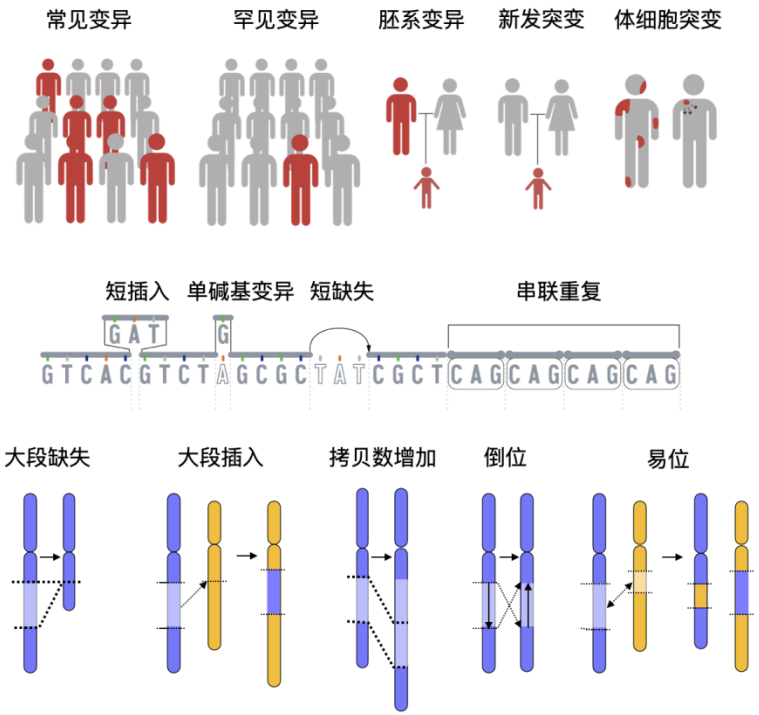
序列变异检测

序列变异检测原理与技术



序列变异检测

序列变异检测原理与技术



各种类型的基因组变异。其中主要包括单碱基变异、短插入缺失以及各种结构变异

序列变异检测

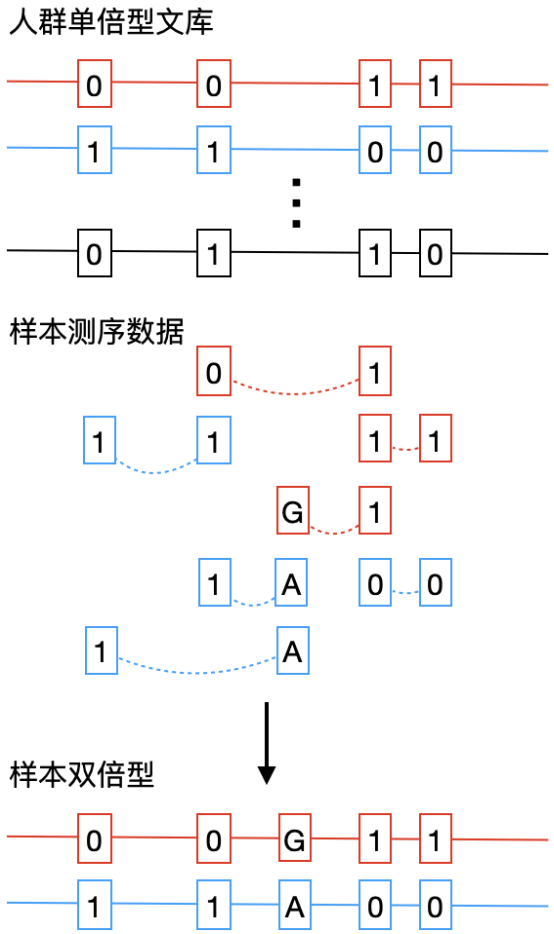
序列变异检测原理与技术

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

检测单碱基序列变异的常用工具

序列变异检测

序列变异检测原理与技术



用人群单倍型信息和测序数据中多变异位点等位基因连锁关系进行单倍型分析和基因型推断简图。其中样本测序数据中的弧形虚线连接的两端reads是一对read pair。

序列变异检测原理与技术

参考基因组序列

CGTATGATCTAGCGCGCTAGCTAGCTAGC

五种罚分等同的缺失发生位置

CGTATGATCTA - - GCGCTAGCTAGCTAGC
CGTATGATCTAG - - CGCTAGCTAGCTAGC
CGTATGATCTAGC - - GCTAGCTAGCTAGC
CGTATGATCTAGCG - - CTAGCTAGCTAGC
CGTATGATCTAGCGC - - TAGCTAGCTAGC

重复序列区间的插入缺失，可见“GC”或“CG”缺失发生在多个位置罚分等同

序列变异检测

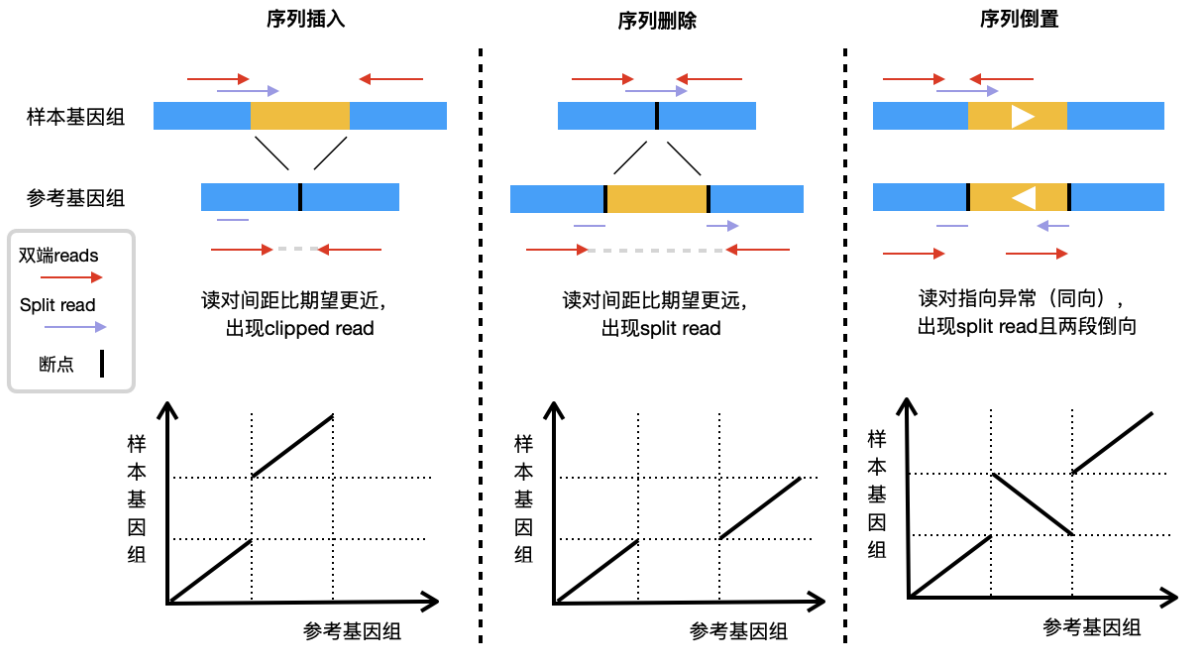
结构变异的检测

突变机制	非等位基因同源重组	非同源末端连接	复制叉停滞和模板交换	转座子逆转录转座
结构变异类型	重复、缺失、倒位	重复、缺失	重复、缺失、倒位、复杂结构变异	序列插入
断点处侧翼是否有同源序列	是	否	否	否
断点序列特征	内部同源性	额外的插入或缺失，或微同源序列	微同源序列	无特别特征
结构变异内部序列特征	任何	任何	任何	转录本序列

同变异机制以及其引发的结构变异类型和特征

序列变异检测

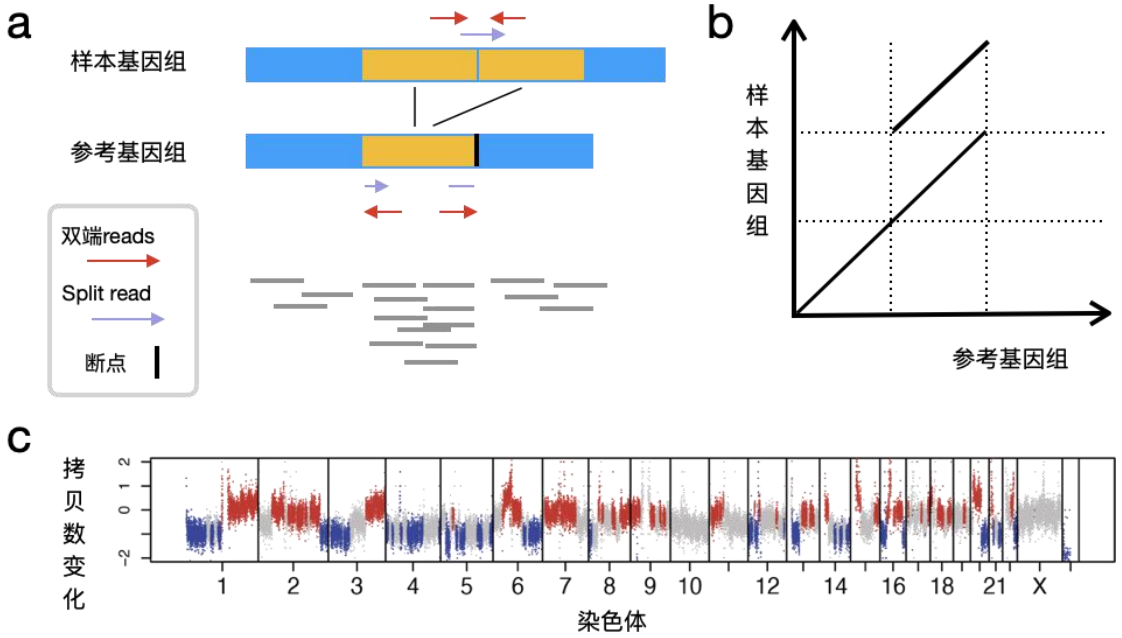
结构变异的检测



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

序列变异检测

结构变异的检测



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

目录大纲

宏基因组学

01 微生物组介绍

02 微生物组高通量测序

03 微生物组测序数据和基本分析流程

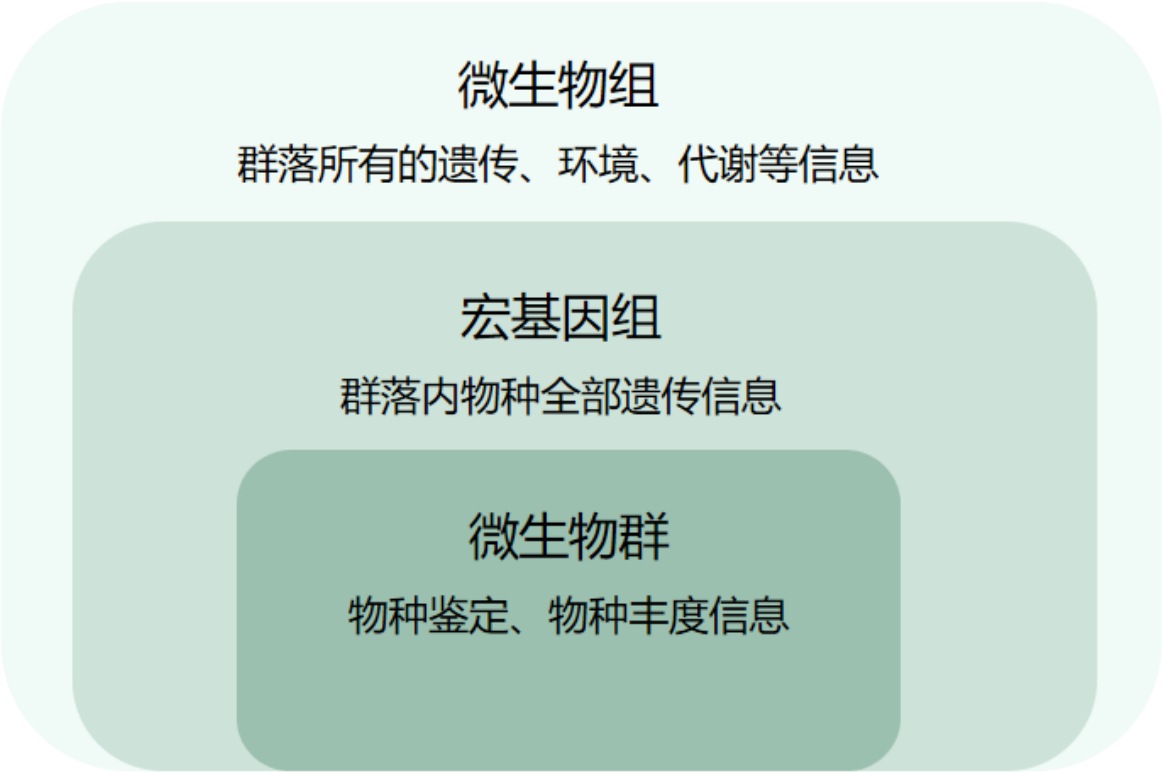
04 微生物组大数据与人工智能

4.1

微生物组介绍

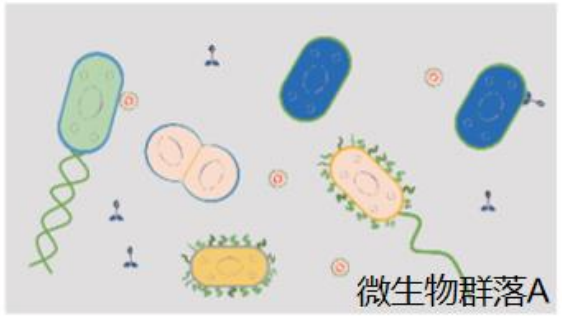
微生物群、宏基因组和微生物组

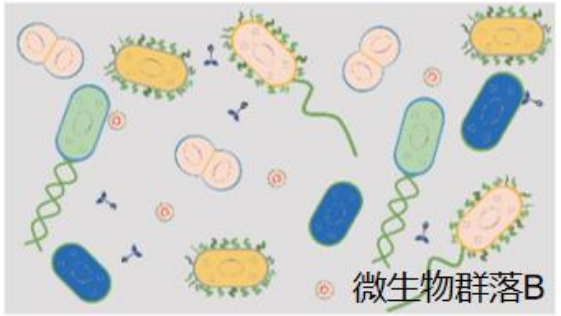
微生物组的研究范围最广，包含了微生物研究中的各种信息



微生物群

微生物群（microbiota）既包括植物体上共生或病理的微生物生态群体，也包括在土壤、水体和空气等环境中自由生存的细菌、古菌、原生动物、真菌和病毒，在宿主的免疫、代谢和激素等方面非常重要



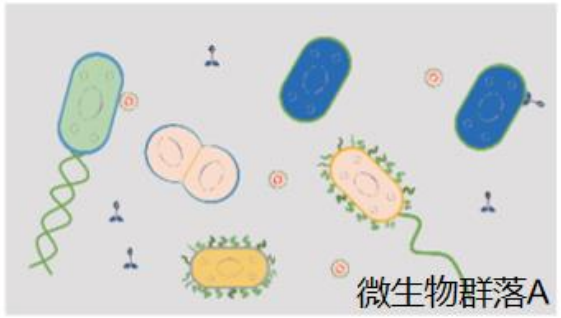


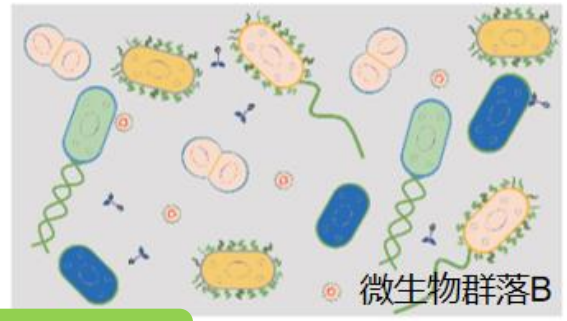
对于不同微生物群之间的差异度量主要是基于微生物群内微生物的组成和分布进行的，这是形成一个特定群落的基础，也是群落之间差异的来源。为了度量这种差异，多种度量指标被提出，其中最常用的有α-多样性、β-多样性和γ-多样性

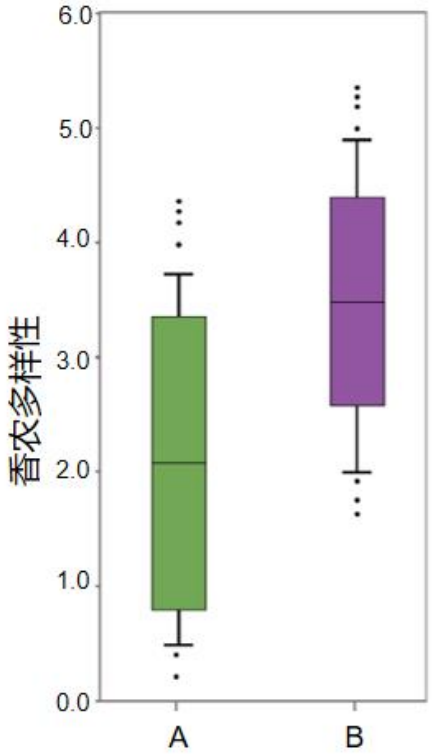
微生物群

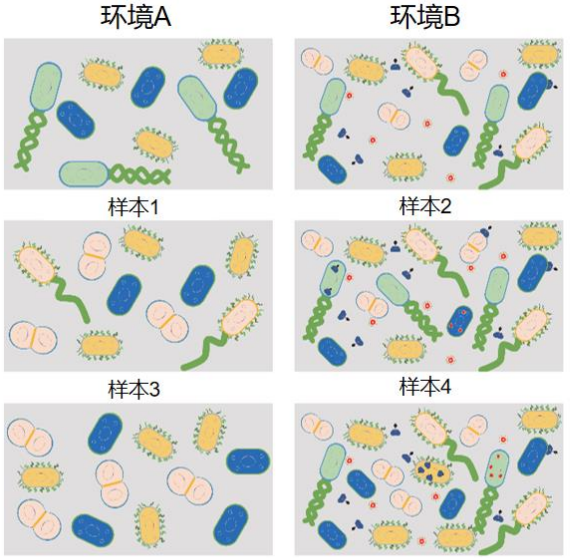
α-多样性表征一个群落内物种的个数（species richness，丰富度）和每个物种的数量及分布（evenness，均匀度）。

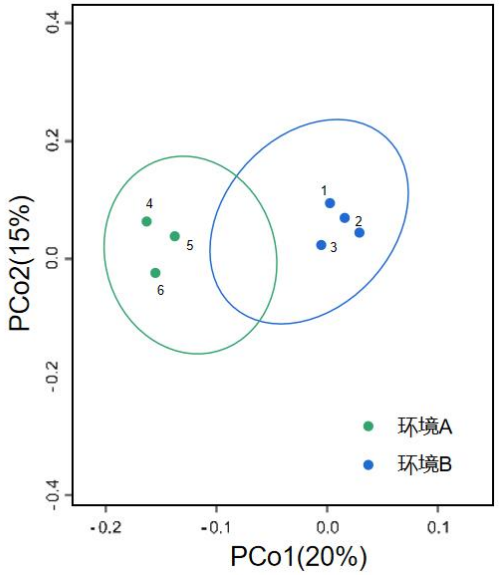
β-多样性是一种评估不同微生物组样本之间的特征差异的方法，其衡量的是不同样本或群落之间的物种多样性差异。











MOOC

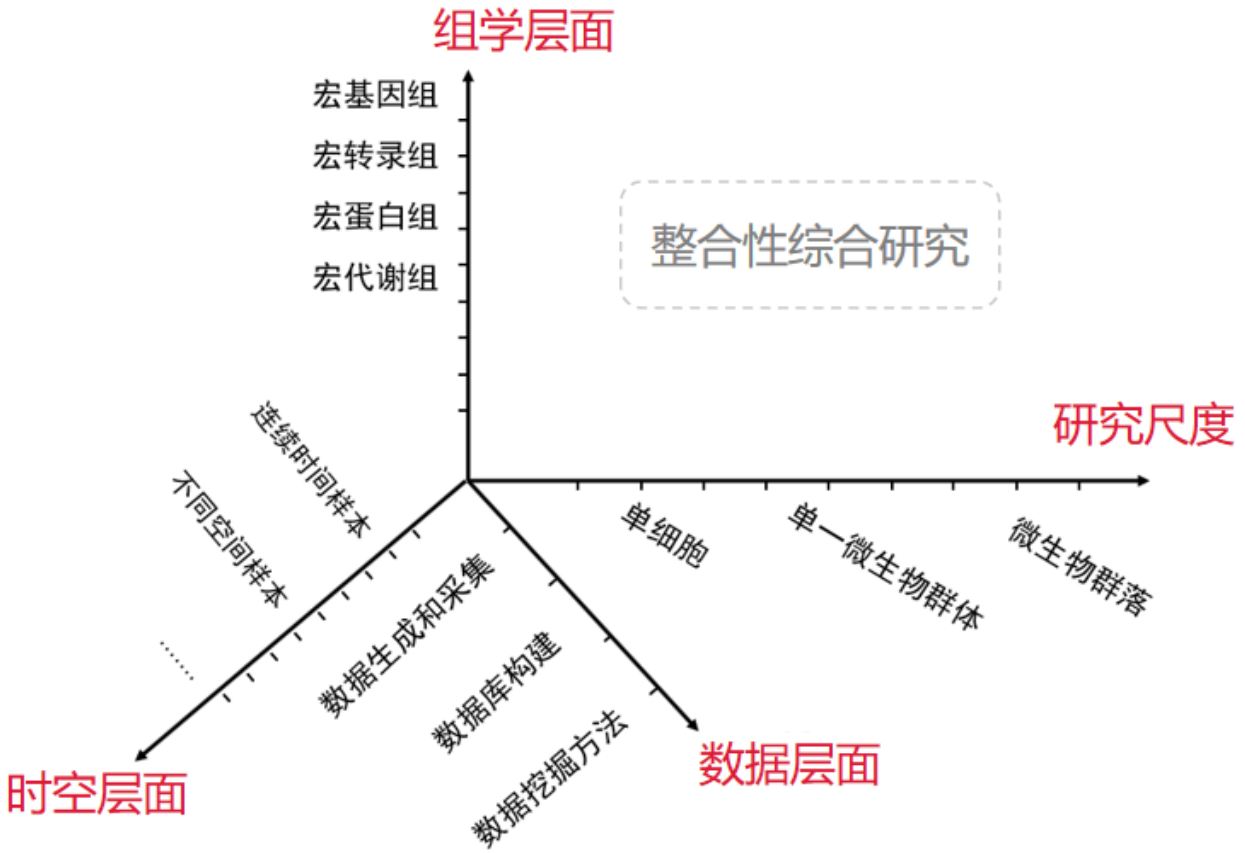
宏基因组

- 宏基因组又称元基因组或微生物环境基因组（metagenome），其定义是“生境中全部微生物遗传物质的总和”。它直接将包含了可培养的和未可培养的微生物的微生物群落的所有遗传物质作为研究对象，广义来说其包括环境基因组、生态基因组学和群体基因组学。

- 传统的微生物学和微生物基因测序依赖于单克隆的培养，早期环境基因测序通过克隆16s rRNA基因等特定基因来确定自然样品中的生物多样性，但是此方法将会漏掉大量未被培养的微生物。因此，近期研究常采用鸟枪法或PCR直接测序方法来获得样品群体中所有成员无偏好的基因，这类方法可以展现从前无法发现的微生物多样性。

- 宏基因组学或元基因组学（metagenomics） 的研究对象是环境样品中的微生物群体基因组，也就是上文所说的宏基因组。其主要的研究手段是功能基因筛选或测序分析，通过一系列分析可以获得微生物多样性、种群结构、进化关系、功能活性、相互协作关系及与环境之间的关系等有用信息。

宏基因组


组学层面
宏基因组
宏转录组
宏蛋白组
宏代谢组
整合性综合研究
研究尺度
单细胞
单一微生物群体
微生物群落
数据层面
数据生成和采集
数据库构建
数据挖掘方法
时空层面
连续时间样本
不同空间样本
……

微生物组

微生物组（microbiome）包括微生物（细菌、古细菌、低等或高等真核生物和病毒）的基因组，以及其周围环境，也就是说微生物组既包括微生物物种，又包括各个物种的基因组以及相关环境因素和代谢产物。微生物组是结合了宏基因组学、代谢组学、宏转录组学、以及宏蛋白组学等和临床/环境数据的集合。

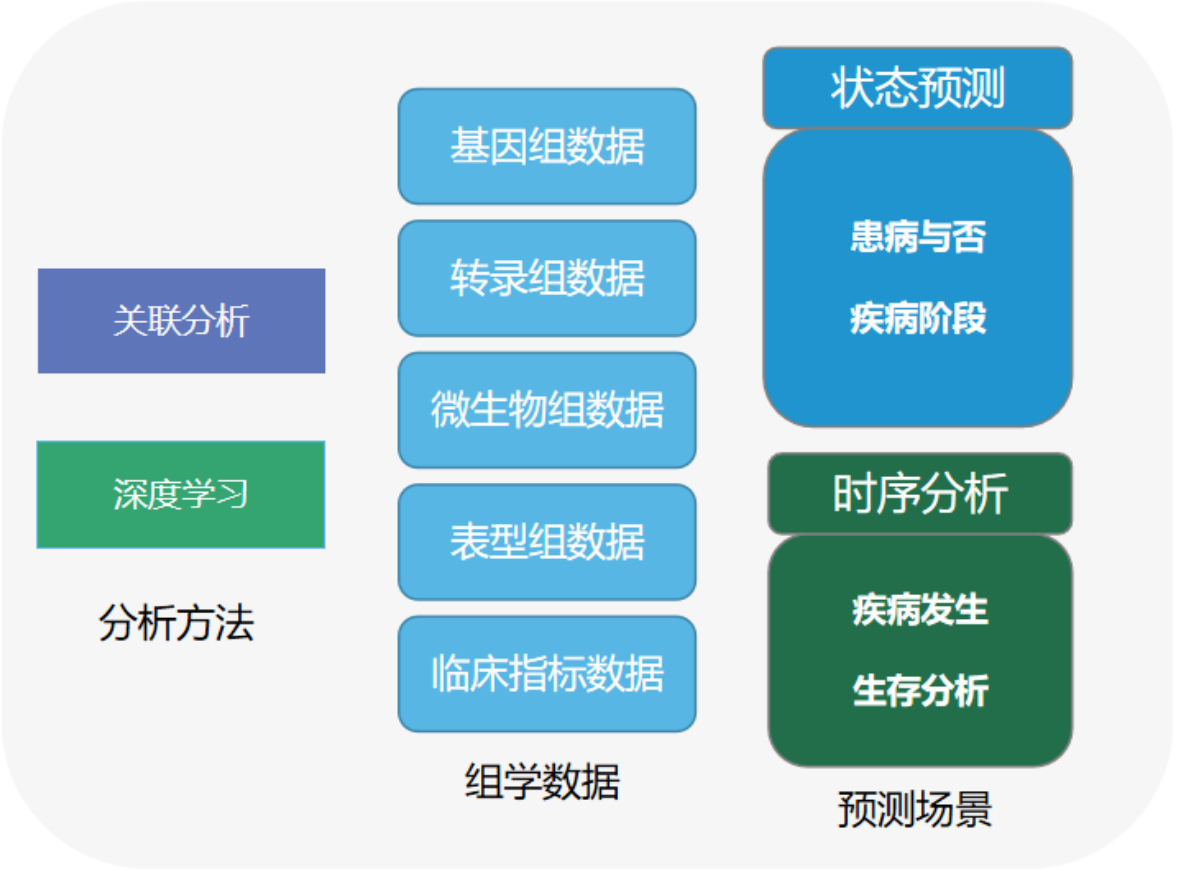
微生物组学研究内容：

①微生物培养，这是了解微生物形态结构和生理功能最直接的方法，但是微生物培养一般费时费力，且许多微生物是不可培养的，基于高通量测序可以解决这些问题。

②微生物测序，高通量测序技术的进步极大地促进了有关微生物的研究，基于高通量技术的微生物研究平台主要包括增子测序技术和宏基因组测序技术等。

③多组学研究，基因测序方法难以鉴定微生物中的关键功能分子，单独使用时无法回答何种成员微生物通过何种方式影响宿主等关键问题。而微生物组学与代谢组学等多组学联用的优势逐渐突出，其关联分析在宿主生理、疾病病理、药物药理等领域已取得众多进展，具有良好应用前景。

微生物组


关联分析
深度学习
分析方法
基因组数据
转录组数据
微生物组数据
表型组数据
临床指标数据
组学数据
状态预测
患病与否
疾病阶段
时序分析
疾病发生
生存分析
预测场景

4.2

微生物组高通量测序

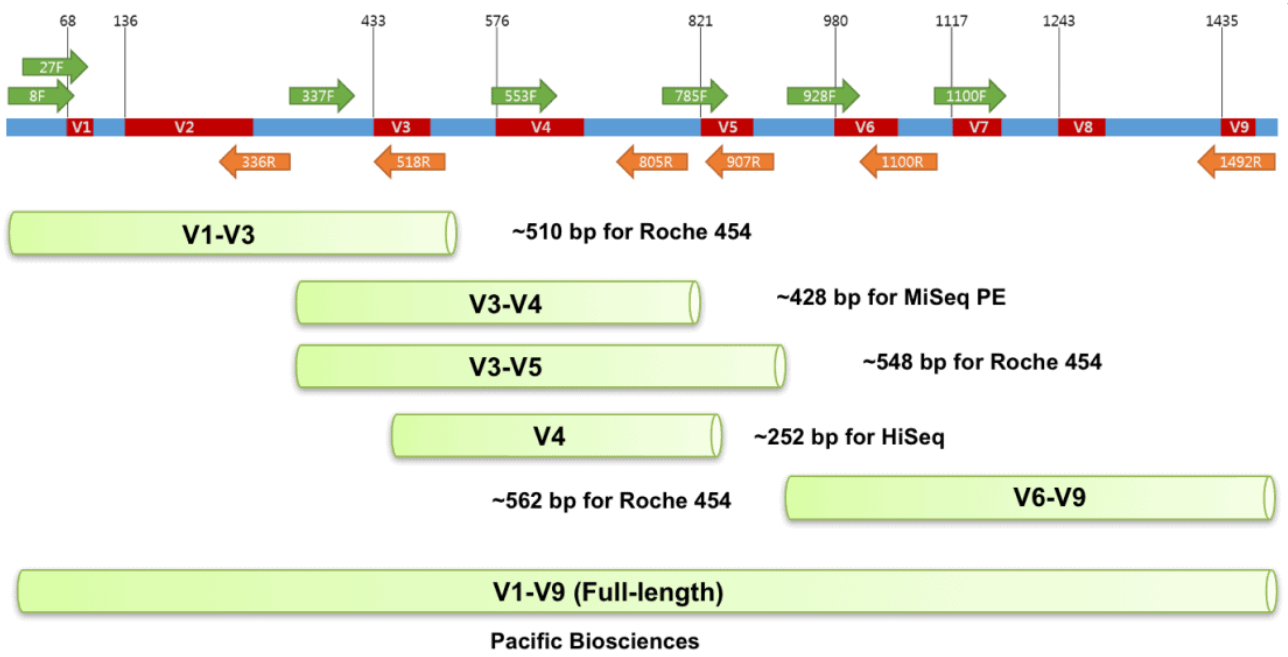
微生物组高通量测序

对微生物群体进行高通量测序又称微生物群落测序，是指通过序列来分析特定环境中微生物群体的构成情况或基因的组成以及功能。对微生物群落进行测序包括两类：一类是通过16s rDNA、18s rDNA、ITS区域进行扩增测序，进而分析微生物的群体构成和多样性；另一类是宏基因组测序，是不经过分离培养微生物，而对所有微生物DNA进行测序，从而分析微生物群落构成和基因构成，挖掘有应用价值的基因资源。

微生物组分析方法主要分为两类：一类为表型功能筛选，即利用模式微生物表型的变化筛选某些目的基因；另一类为序列基因型分析，即对文库中所有或部分的DNA 进行测序分析，以应用于生态学研究。

扩增子测序

以16s rDNA扩增进行测序分析主要用于微生物群落多样性和构成的分析，而目前的生物信息学分析也可以基于16s rDNA的测序，对微生物群落的基因构成和代谢途径进行预测分析，大大拓展了我们对于环境微生物的微生态认知。



扩增子测序

1. 基础概念

（1）16S rDNA（或16S rRNA）：编码原核生物核糖体小亚基的基因，长度约为1500bp，其分子大小适中，突变率小，是细菌系统分类学研究中最常用和最有用的标志。

（2）操作分类单元（operational taxonomic units，OTU）：提取样品的总基因组DNA，利用16S rRNA或ITS的通用引物进行PCR扩增。不同的 16S rRNA序列的相似性高于97%就可以把它定义为一个OTU，每个OTU对应于一个不同的16S rRNA序列，也就是每个OTU对应于一个不同的细菌（微生物）种。通过OTU分析，可以知道样品中的微生物多样性和不同微生物的丰度。

（3）测序区段：由于16s rDNA较长，只能对其中经常变化的区域，也就是可变区进行测序。16s rDNA包含9个可变区，分别是V1~V9。研究中，一般对V3-V4双可变区域进行扩增和测序，也偶尔会对V1-V3区进行扩增和测序。

扩增子测序

测序过程

（1）提取样品DNA：DNA可以来自土壤、粪便、空气或水体等。

（2）质检和纯化：一般16s rDNA扩增子测序对DNA的总量要求并不高，总量>100ng，浓度>10ng/uL大多可以满足要求。如果是来自和寄主共生的环境，如昆虫的肠道微生物，提取其DNA时可能混合了寄主本身的大量DNA，因此对DNA的总量要求会有所提高。

（3）测序：对完成PCR后的产物进行测序。目前，可以采用多种不同的测序仪，如罗氏的454、Illumina的MiSeq、Life的PGM或Pacbio的RSII三代测序仪进行16s rDNA测序。

（4）数据分析：

①聚类统计，同源聚类获得OTUs；

②样本构成丰度分析，稀释曲线、Rank-Abundance曲线

③多样性分析，PCoA、NMDS（非度量多维尺度分析）、PCA、LDA

④差异性菌群分析，功能上的差异。

⑤环境因子分析，RDA（Redundancy analysis）、CCA（canonical correspondence analysis）

微生物组高通量测序

宏基因组测序

- 不同于传统的先培养微生物再提取DNA的做法，宏基因组直接收集能够代表特定生物环境生物多样性的样品；然后利用各种理化方法破碎微生物，使其释放DNA，再利用密度梯度离心等方法进行分离纯化。
- 接着，对DNA 进行酶切或者超声打断处理，并将其与合适的载体DNA 进行连接，构建重组体。
- 将带有宏基因组DNA的载体通过转化方式转入模式微生物，建立各自的无性繁殖系。
- 最后，对宏基因组文库的DNA 进行分析。

宏基因组测序

测序过程

(1) 样品总DNA的提取及基因或基因组DNA的富集

(2) 宏基因组文库的建立

(3) 宏基因组文库的筛选

宏基因组测序技术的发展

1975年，Sange提出的双脱氧链终止法，开启了一代测序的时代；1990年，正式启动了规模宏大的人类基因组计划；1995年，测序得到了第一个完整的细菌基因组即嗜血流感菌；2001年，完成了人类基因组计划，生命科学研究开始进入基因组学时代。2005年，第一台二代测序仪454 GS20问世，Illumina也在2007年发布了二代测序仪。总的来说，在基因组测序领域已经进入了高通量测序时代，各项研究也逐渐从单一、局部的基因或基因片段的研究转变成了对整个基因组的研究。2011年，三代测序技术跨越了一、二代较短读长而直接对DNA单个分子进行测序实现又一突破，其应用日益广泛。

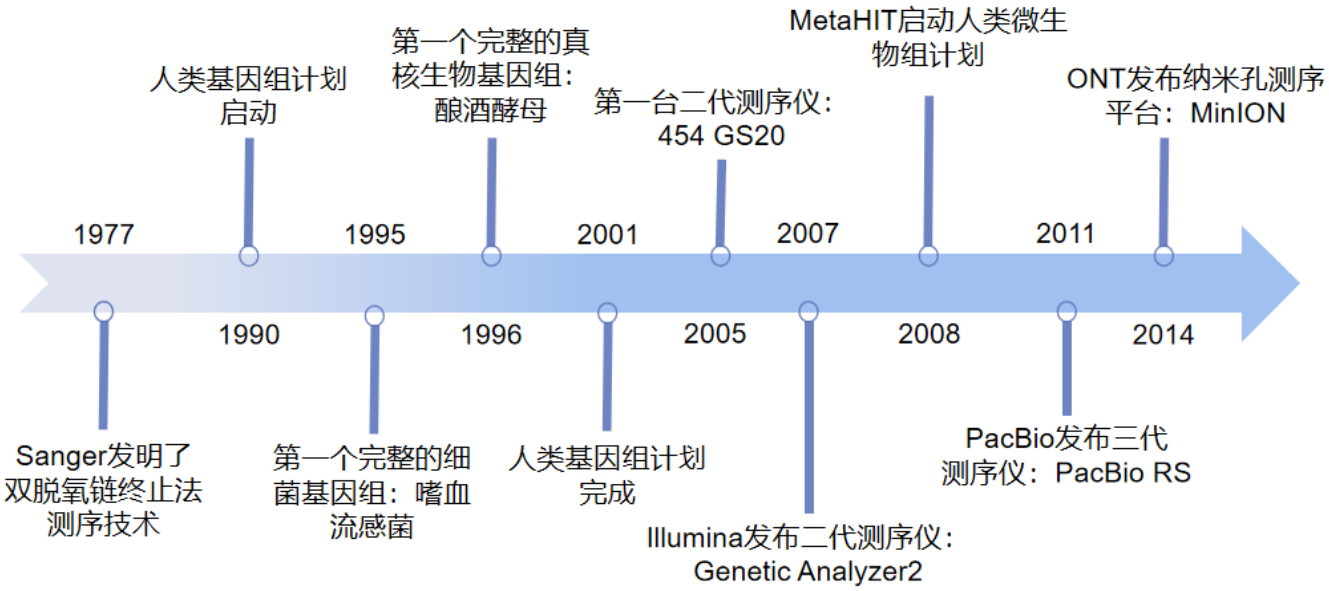


图1.4 宏基因组测序技术发展简史

MOOC

鸟枪法宏基因组测序的拓展研究

- 传统的鸟枪法宏基因组学的技术挑战仍然存在于计算密集的短读装配、菌株在群落内部的异质性以及低丰度微生物所需的覆盖深度等方面。

- 从这个角度来看，偶联稳定的同位素探测和基因组分辨的宏基因组学，将荧光激活的细胞分选方法应用于更大的微生物群落，将移动元件链接到宿主微生物细胞具有较好的发展前景。

- 毫无疑问，这些发展将推动基因组解析的宏基因组学方法的发展，将能更好地了解自然环境中未培养的微生物。

鸟枪法宏基因组测序的拓展研究

1. 耦合稳定性同位素示踪与基因组分辨的宏基因组学

DNA稳定同位素探测（DNA-stable isotope probing，DNA-SIP），可以根据同位素标记的底物（如$13_{C,}$ 15_{N}等）的摄取和掺入来有针对性地富集活性微生物。在该方法中，将从与同位素标记的化合物一起孵育的群落中提取的DNA沿氯化铯密度梯度分离成不同的馏分，含“重”同位素标记的DNA会在较高溜分。标记化合物的同化作用可由微生物DNA密度的变化推断，DNA-SIP是将微生物与特定代谢过程联系起来的强大工具。

基于16S rRNA标记基因的高通量测序，一些DNA-SIP研究可以探索系统发育与原位功能之间的联系。此外，与DNA-SIP相关的分级分离步骤还有助于从复杂的群落中整体回收基因组。通过在测序之前将群落DNA非随机地分为几十个部分，增加某些部分中稀有微生物的相对丰度，从而比散装DNA的鸟枪法测序覆盖范围更大。

鸟枪法宏基因组测序的拓展研究

2. 靶向探索“微型宏基因组”

在提取和测序DNA前，可以将复杂的微生物群落分为较小的亚组。荧光激活细胞分选（FACS）是一种复杂但更灵活、更精确随机或非随机地生成微型元数据的方法。例如，使用FACS从森林土壤中回收了一些未经培养的巨型病毒基因组，这些病毒基因组不能通过土壤样本在同一深度鸟枪测序方法中进行组装，并支持将复杂群落细分为低多样性的微观宏观基因组以恢复稀有成员的观点，并且使用传统的大规模宏基因组学方法可能会忽略这些稀有成员。

鸟枪法宏基因组测序的拓展研究

3. 链接可移动元件与微生物宿主

单分子读取技术可以将移动元件（特别是质粒）连接到宿主微生物细胞，改进了直接从环境中重建基因组的方法。质粒介导的水平基因转移影响微生物群落结构和进化，将独特的功能传递给微生物并在系统发生群之间交换基因。但由于对质粒的大小、结构、传播机制的多样性了解较少，从环境样品中直接分离质粒以及使用标准鸟枪法测序的准确计算预测存在局限性。

4.3

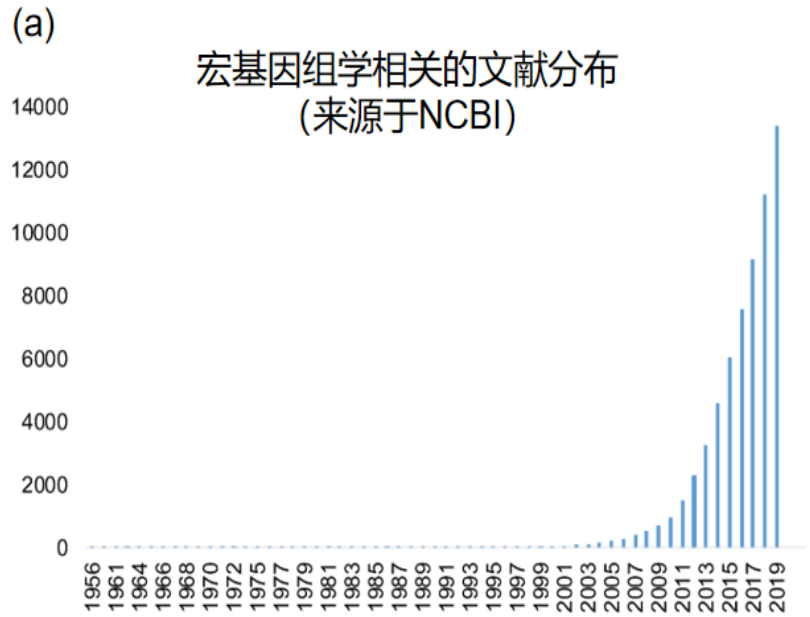
微生物组测序数据和基本分析流程

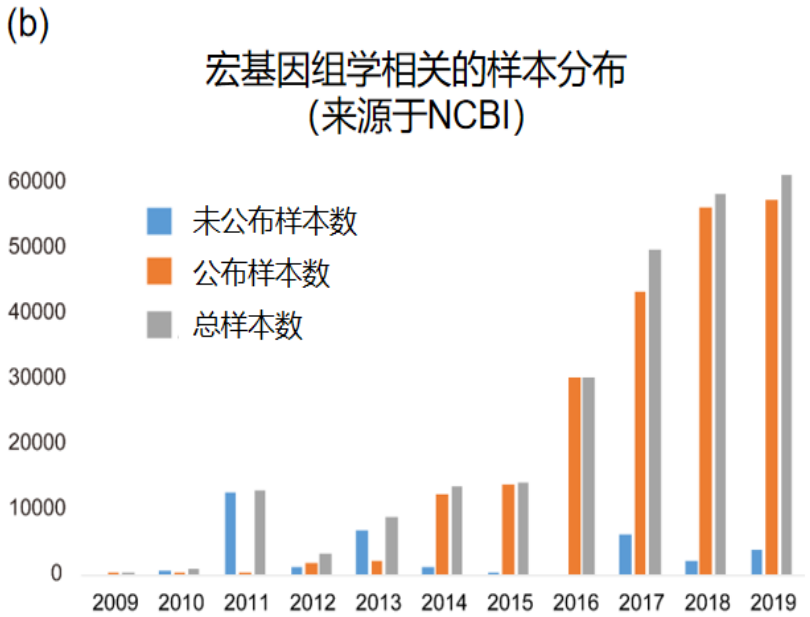
微生物组测序数据和基本分析流程

微生物组数据积累和整合

微生物组学数据的积累大大促进了微生物群落的研究，在过去十年间微生物相关的论文数量呈现指数增长，微生物组数据量每年以>100TB的速度增长。国际上已经建立了许多宏基因组相关的数据库，比如MG-RAST（http://metagenomics.anl.gov/）和CAMERA（http://camera.calit2.net/）等。

其中，NCBI的Sequence Read Archive （SRA）（http://www.ncbi.nlm.nih.gov/sra）、MG-RAST以及CAMERA2中公开的宏基因组项目超过10, 000个，包含超过1PB的数据。





MOOC

微生物组测序数据和基本分析流程

微生物组数据积累和整合

- 按照菌群来源的生存环境（biome）而组织起来的微生物群落样本和相关测序数据，是依据生存环境本体的组织架构，通过层级结构组织起来的。例如：截止2019年底，EBI MGnify的生存环境本体组织架构包括491个本体[26]，而人体大肠排泄物菌群的本体定位是“root > Host-associated > Human > Digestive system > Large intestine > Fecal”。这种本体结构非常有利于样本的分类。然而，目前这种本体的层级组织结构并非完全是树状的，而是具有一个本体属于多个本体的直接子本体的特征，例如“Fecal”就有多达5个以上的上一级本体信息。因此，每个微生物组数据的相关生存环境本体都有可能具有多标签（multi-label）。

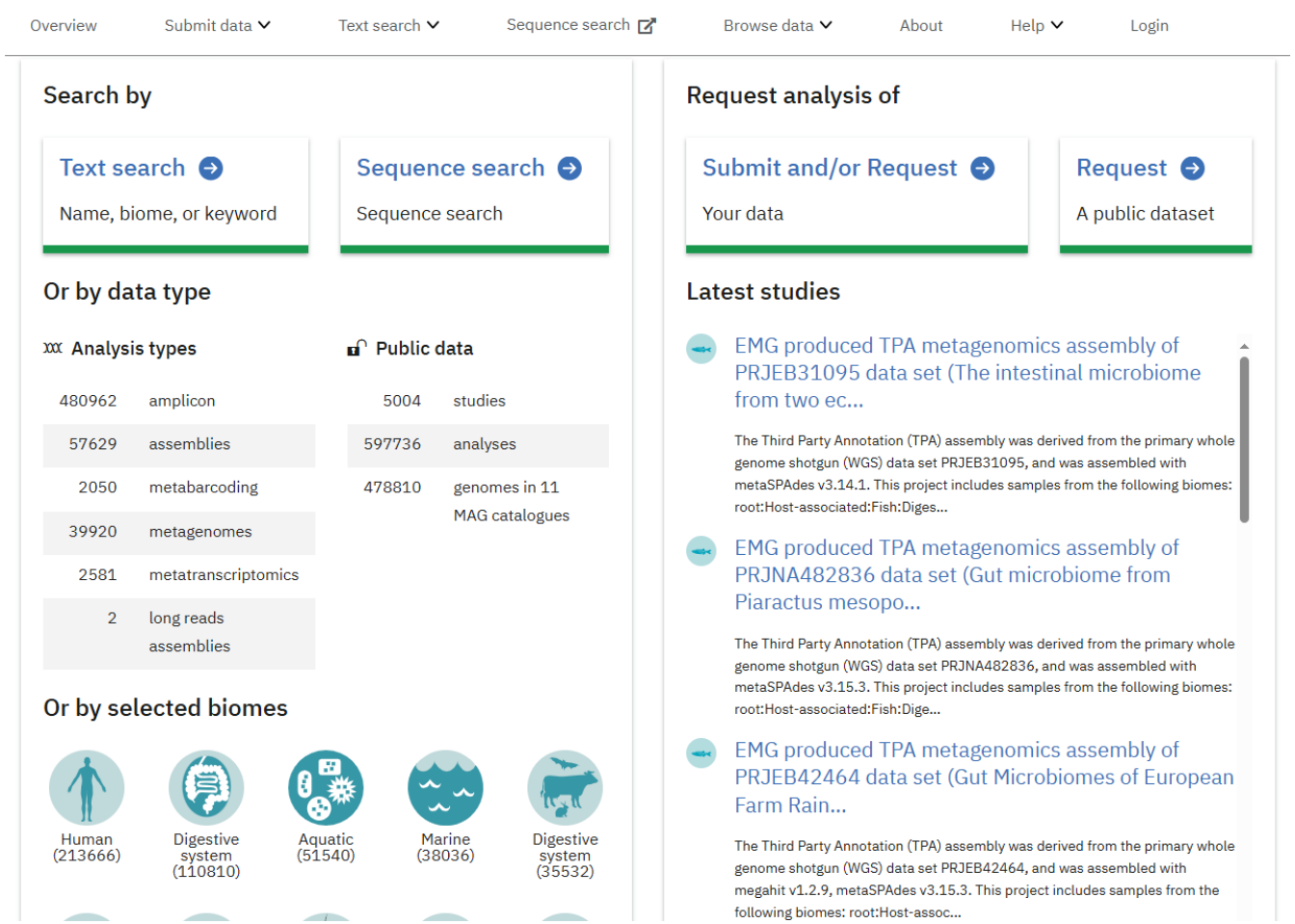
- 从一方面来说，微生物组数据的多标签属性，不利于样本的简单分类，造成了样本分类和比较方面的瓶颈。

- 另一方面，微生物组数据的多标签属性符合大数据研究的特征，利用机器学习或者深度学习等方法来处理将有望获得较好的结果。

微生物组测序数据和基本分析流程

微生物组数据积累和整合

https://www.ebi.ac.uk/metagenomics



代表性的微生物组数据库（MGnify）

微生物组测序数据和基本分析流程

微生物组数据分析挖掘

生物学问题

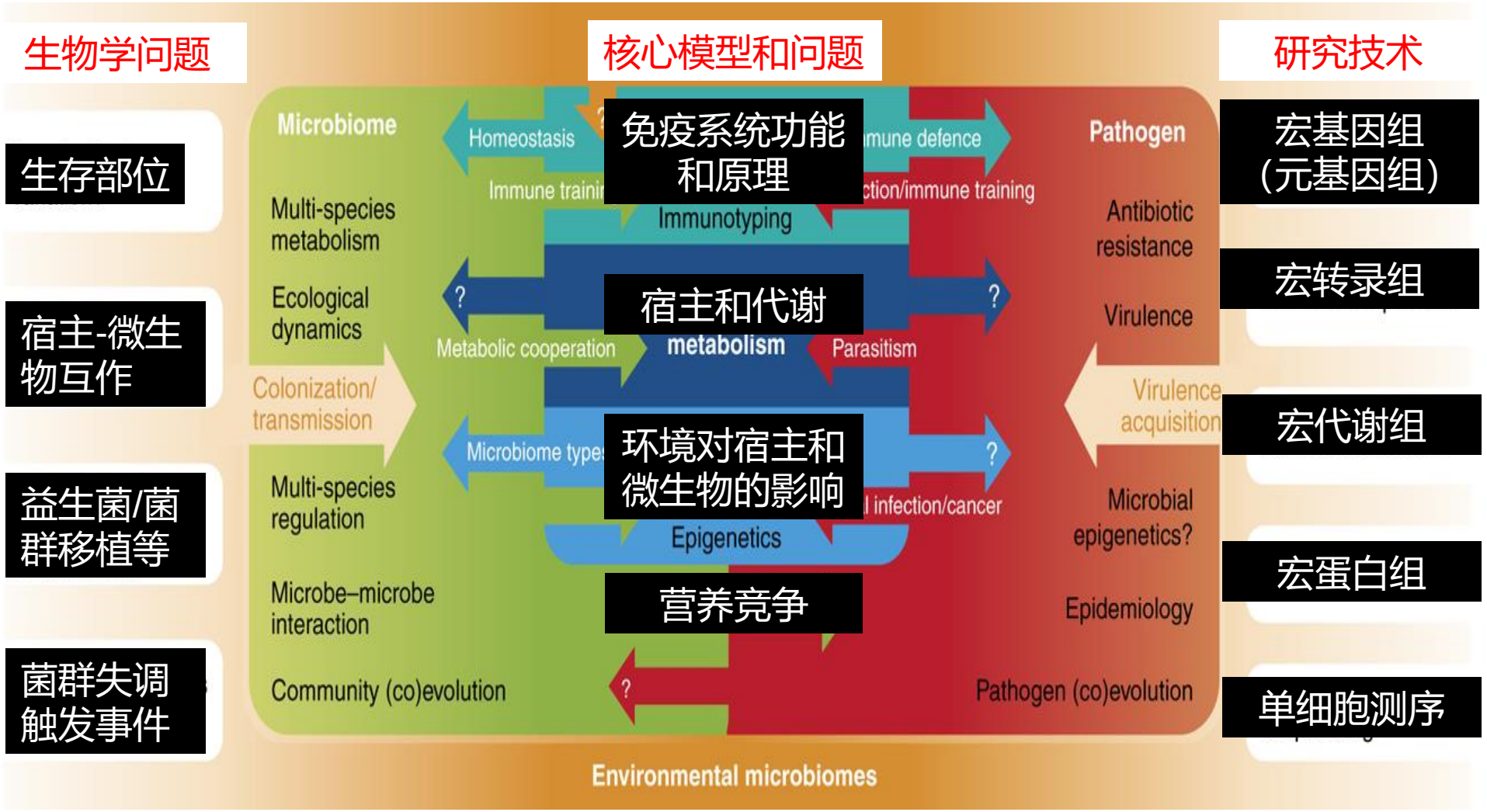
- 生存部位
- 宿主-微生物互作
- 益生菌/菌群移植等
- 菌群失调触发事件

核心模型和问题

- 免疫系统功能和原理
- 宿主和代谢
- 环境对宿主和微生物的影响
- 营养竞争

研究技术

- 宏基因组（元基因组）
- 宏转录组
- 宏代谢组
- 宏蛋白组
- 单细胞测序

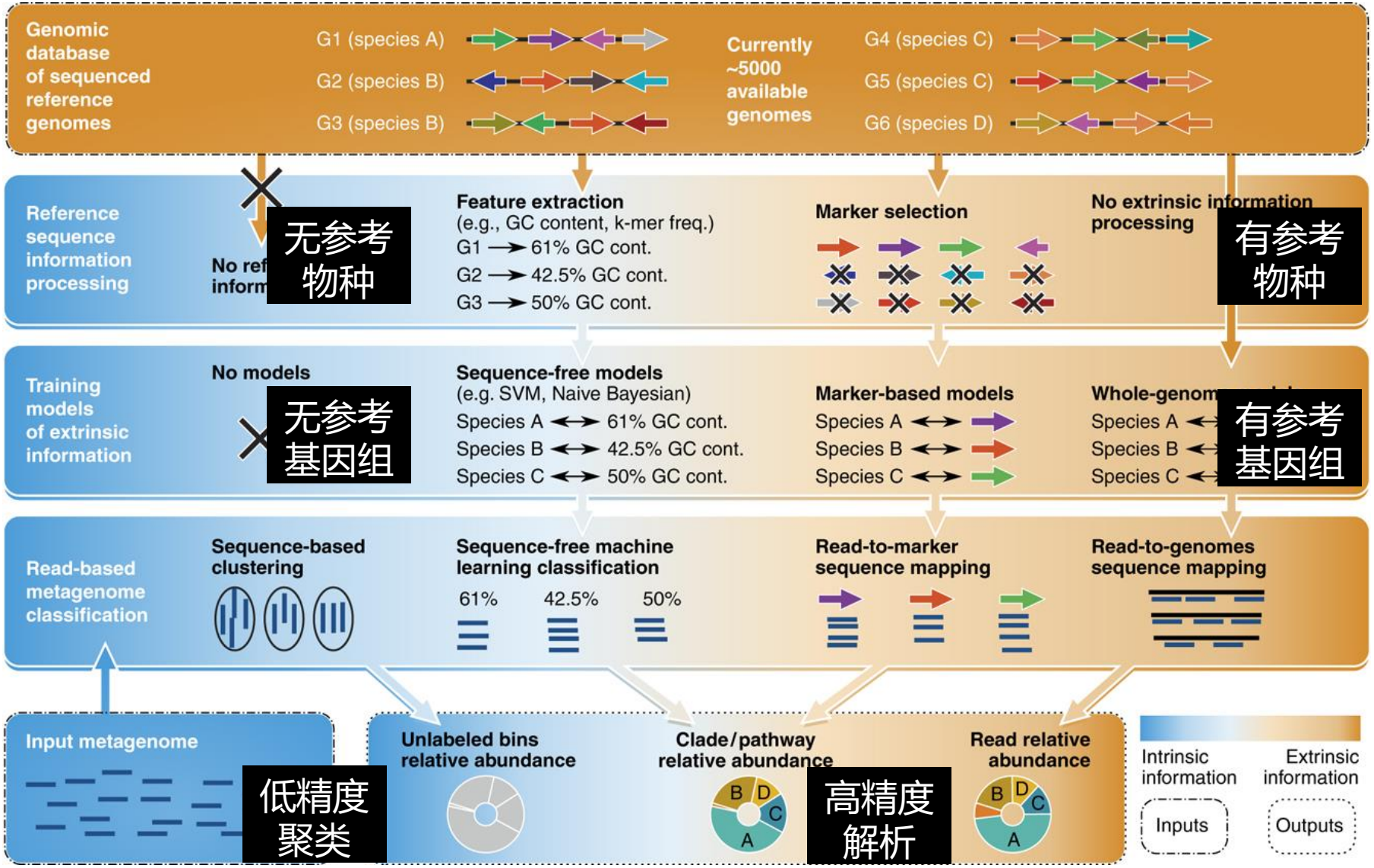


Huttenhower, et al., *Molecular Systems Biology*, 2013

MOOC

微生物组测序数据和基本分析流程

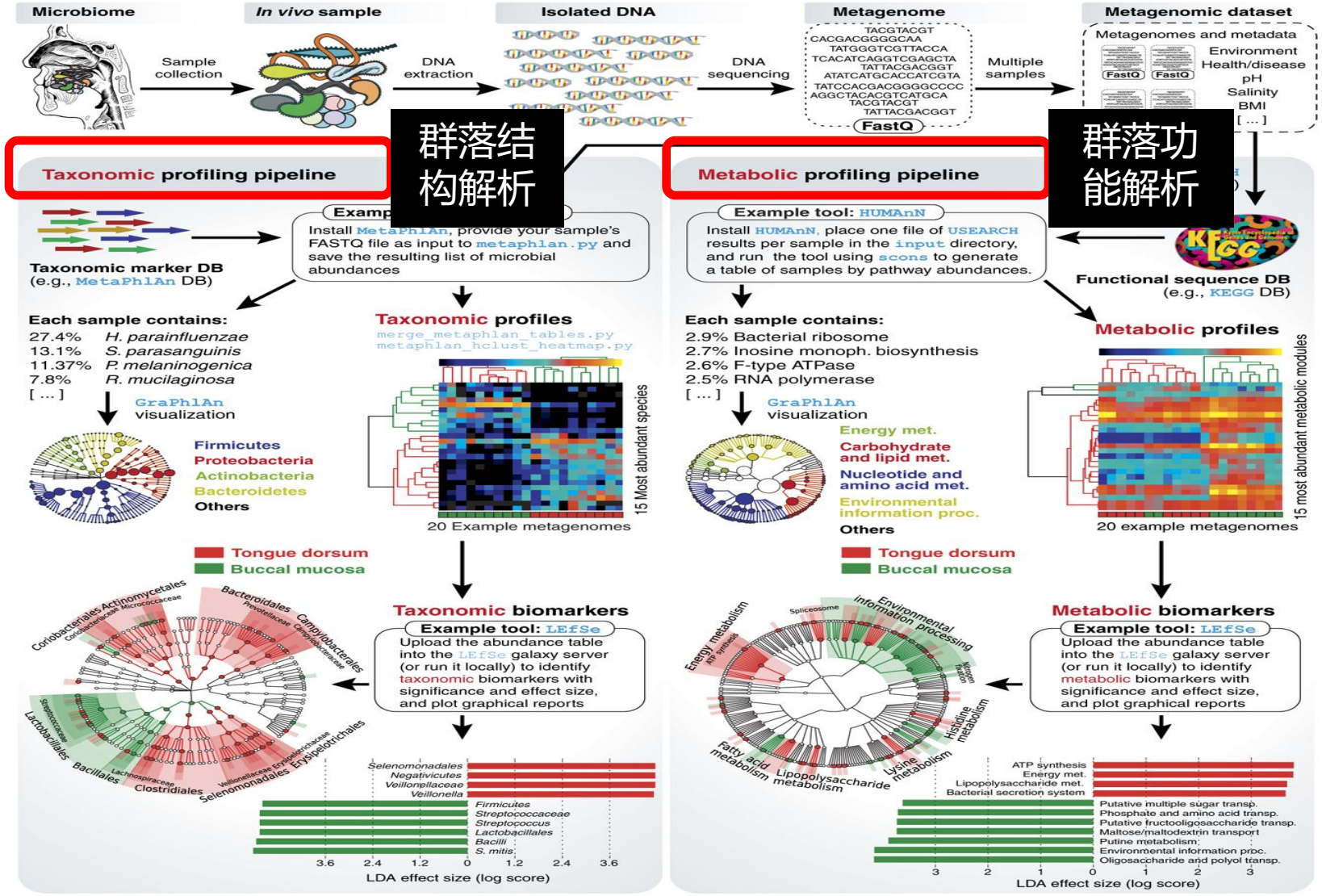
微生物组数据分析挖掘



MOOC

微生物组测序数据和基本分析流程

微生物组数据分析挖掘



MOOC

微生物组测序数据和基本分析流程

微生物组数据分析挖掘

- 随着海量微生物组数据的积累，涌现了大量的微生物组数据库，以及大量的微生物组数据分析方法和软件。

- 其中主流的微生物组数据库包括EBI MGnify，QIITA等通用微生物组数据库，以及针对抗性基因挖掘的CARD数据库、针对生合成代谢基因簇挖掘的antiSMASH等。

- 微生物组数据常用的分析方法和软件包括：针对测序数据质量控制的FastQC，针对微生物组测序数据分析（从测序数据到物种结构）的QIIME 2.0和MetaPhlAn，针对微生物组的功能谱分析的 HUMAnN 2.0，针对微生物组溯源分析的SourceTracker，针对微生物组功能基因挖掘的DeepARG、antiSMASH等方法。

微生物组测序数据和基本分析流程

微生物组数据分析挖掘

名称	简介	网址	参考文献
Trimmomatic	一种用于Illumina NGS数据低质量、引物和接头序列去除工具。		[23]
	MetaPhlAn用于从宏基因组中分析微生物群落的组成。	https://huttenhower.sph.harvard.edu/metaphlan2	[24]
HUMAnN2	HUMAnN 2.0可以高效、准确地分析一个群落中微生物路径存在/缺失和丰度。	https://huttenhower.sph.harvard.edu/humann2	[25]
MEGAHIT	超快、省内存的宏基因组组装软件。	https://github.com/voutcn/megahit	[26]
CD-HIT	构建非冗余基因集。	http://weizhongli-lab.org/cd-hit	[27]

代表性的宏基因组学数据分析方法和软件

微生物组测序数据和基本分析流程

微生物组数据分析挖掘

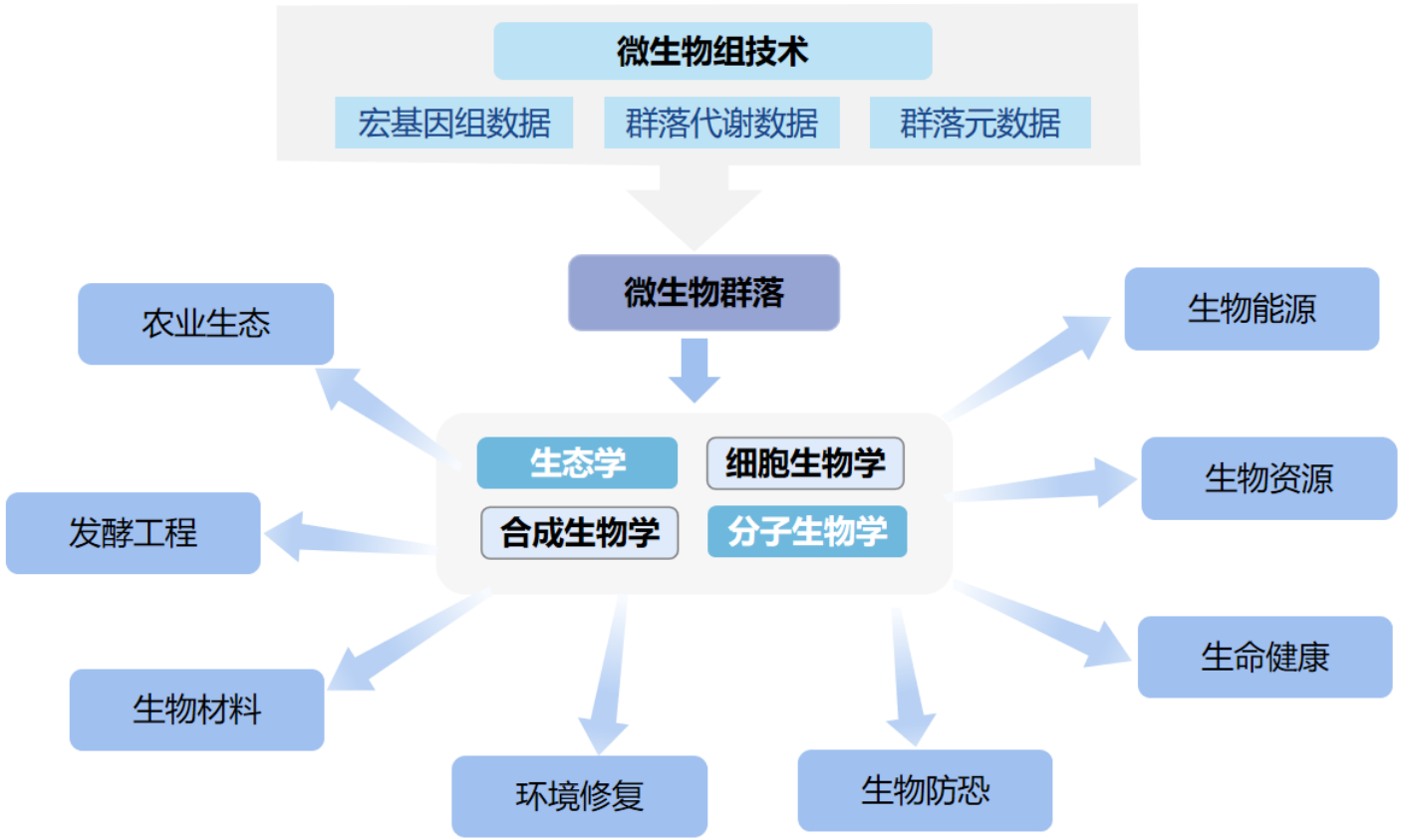
表 1.2 常用分析工具

软件（平台）	分析数据对象	分析结果	参考文献
MOCAT	宏基因组	物种结构、丰度和功能分类，以及物种之间的比较	[37]
MEGAN	16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较	[38]
MetaPhlAn	宏基因组	物种结构、丰度	[33]
PICRUSt	宏基因组，16S rRNA	物种结构和功能分类	[39]
antiSMASH	宏基因组	BGC 分析	[30]
CARMA	16S rRNA	物种结构和功能分类	[40]
Sort-ITEMS	16S rRNA	物种结构和功能分类	[41]
QIIME	16S rRNA	物种结构、丰度和功能分类	[42]
MG-RAST	宏基因组，16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较	[43]
CAMERA	宏基因组，16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较	[44]
IBDsite	宏基因组，16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较	[45]

4.4

微生物组大数据与人工智能

宏基因组学在健康和环境领域的应用



宏基因组学研究所涉及的学科以及在各个领域的广泛应用。基于微生物组技术获得宏基因组、代谢组等数据，基于这些数据研究发展成为各个学科，并应用到生物能源、生命健康、生物材料等各个领域

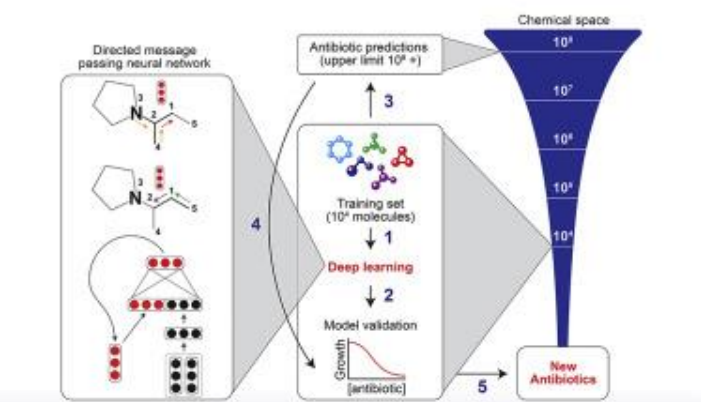
微生物组大数据与人工智能

微生物组大数据的复杂性

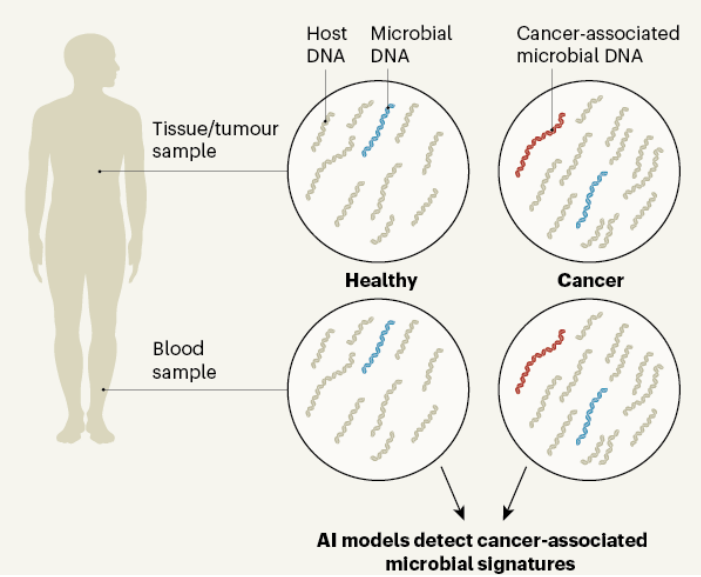
空间复杂性
时间复杂性
交互复杂性
多组学复杂性
……

MOOC

微生物组大数据与人工智能



A Deep Learning Approach to Antibiotic Discovery, Cell, 2020



AI finds microbial signatures in tumours and blood across cancer types, Nature, 2020

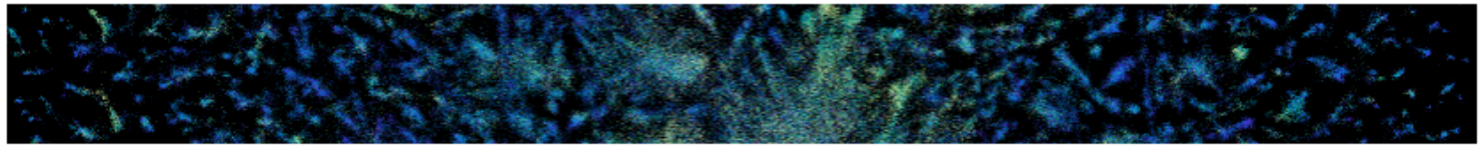
AI赋能的微生物组大数据挖掘

人工智能发掘微生物组数据特征
人工智能挖掘重要功能基因
人工智能解构时空动态变化模式
人工智能预测表型
人工智能预测疾病发生发展
……

微生物组大数据与人工智能

https://github.com/facebookresearch/esm

Evolutionary Scale Modeling



Update April 2023: Code for the two simultaneous preprints on protein design is now released! Code for "Language models generalize beyond natural proteins" is under examples/lm-design/. Code for "A high-level programming language for generative protein design" is under examples/protein-programming-language/.

This repository contains code and pre-trained weights for **Transformer protein language models** from the Meta Fundamental AI Research Protein Team (FAIR), including our state-of-the-art **ESM-2** and **ESMFold**, as well as **MSA Transformer**, **ESM-1v** for predicting variant effects and **ESM-IF1** for inverse folding. Transformer protein language models were introduced in the 2019 preprint of the paper "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". ESM-2 outperforms all tested single-sequence protein language models across a range of structure prediction tasks. ESMFold harnesses the ESM-2 language model to generate accurate structure predictions end to end directly from the sequence of a protein.

In November 2022, we released `v0` of the ESM Metagenomic Atlas, an open atlas of 617 million predicted metagenomic protein structures. The Atlas was updated in March 2023 in collaboration with EBI. The new `v2023_02` adds another 150 million predicted structures to the Atlas, as well as pre-computed ESM2 embeddings. Bulk download, blog post and the resources provided on the Atlas website are documented on this README.

代表性的微生物组大模型（基于宏基因组学数据）

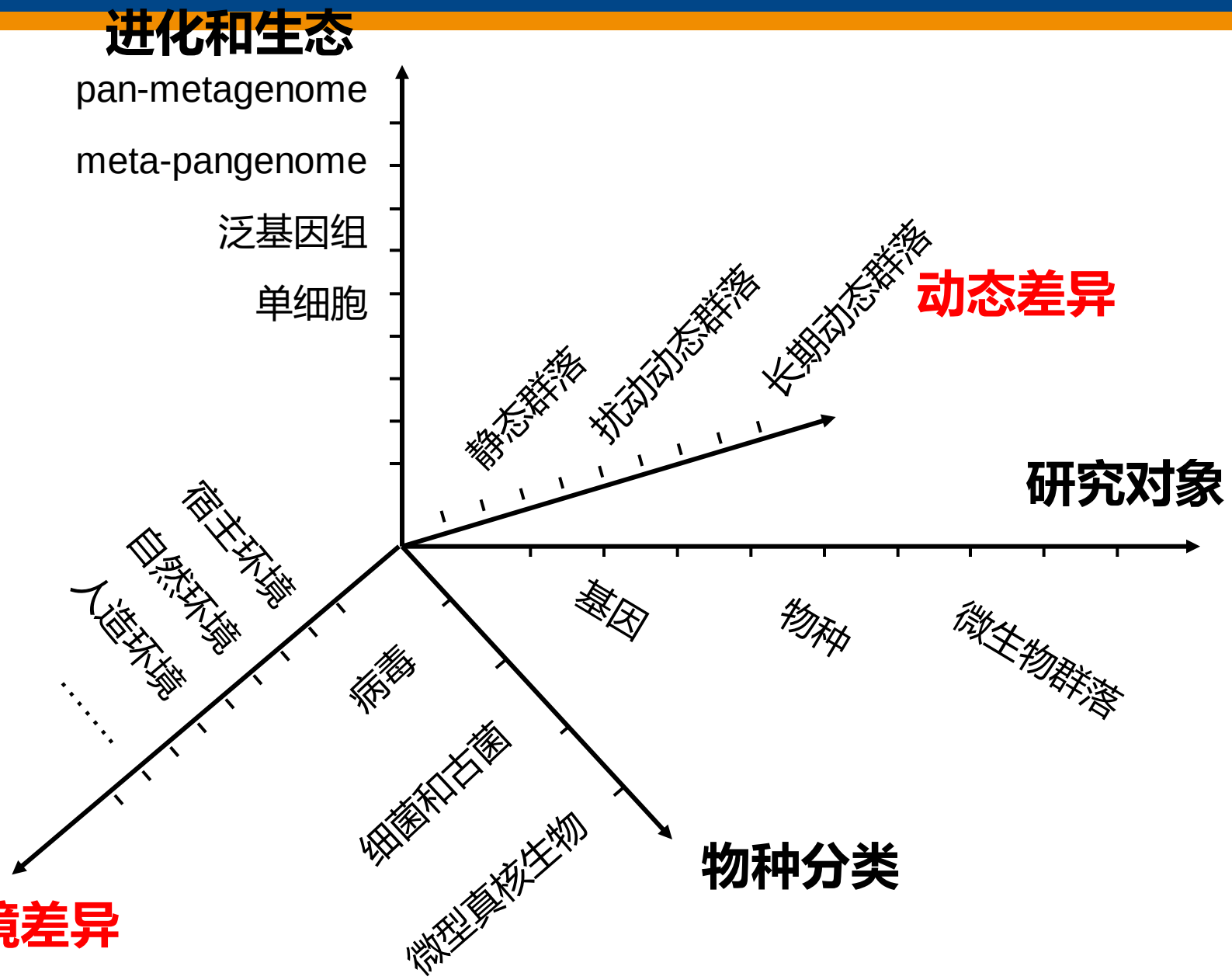
Available Models and Datasets

Pre-trained Models

Shorthand	`esm.pretrained.`	#layers	#params	Dataset	Embedding Dim	Mod
ESM-2	`esm2_t48_15B_UR50D`	48	15B	UR50/D 2021_04	5120	https:// esm/m
	`esm2_t36_3B_UR50D`	36	3B	UR50/D 2021_04	2560	https:// esm/m
	`esm2_t33_650M_UR50D`	33	650M	UR50/D 2021_04	1280	https:// esm/m
	`esm2_t30_150M_UR50D`	30	150M	UR50/D 2021_04	640	https:// esm/m
	`esm2_t12_35M_UR50D`	12	35M	UR50/D 2021_04	480	https:// esm/m
	`esm2_t6_8M_UR50D`	6	8M	UR50/D 2021_04	320	https:// esm/m
ESMFold	`esmfold_v1`	48 (+36)	690M (+3B)	UR50/D 2021_04	-	https:// esm/m
	`esmfold_v0`	48 (+36)	690M (+3B)	UR50/D 2021_04	-	https:// esm/m
	`esmfold_structure_module_only_*`	0 (+various)	various	UR50/D 2021_04	-	https:// esm/m

代表性的微生物组大模型（基于宏基因组学数据）

微生物组大数据与人工智能



4.5

总结与展望

总结与展望

- 基因组学部分详细阐述了基因组序列的获取、解读和应用，强调了高通量测序技术在推动基因组研究中的重要性。介绍了基因组组装、基因预测和基因组注释的主流方法，同时讨论了序列变异检测技术，包括单碱基替换、短插入缺失和结构变异的检测原理与技术。

- 宏基因组学部分则涵盖了微生物组的概述、数据及其分析方法，特别强调了宏基因组学在健康和环境领域的应用潜力，以及如何通过微生物组数据挖掘来揭示微生物群落的功能和作用。

- 随着测序技术的进步和生物信息学工具的不断发展，基因组学和宏基因组学将继续在疾病诊断、精准医疗、环境监测和生态系统研究中发挥关键作用。

- 人工智能技术，特别是深度学习，在这一过程中扮演着至关重要的角色。

谢谢！

Q&A

问题与解答

问题与解答

问题与解答

问题与解答