

宏基因组学

Metagenomics

李余动

lyd@zjsu.edu.cn



Our world is full of microbes (微生物无处不在)

human



soil



ocean



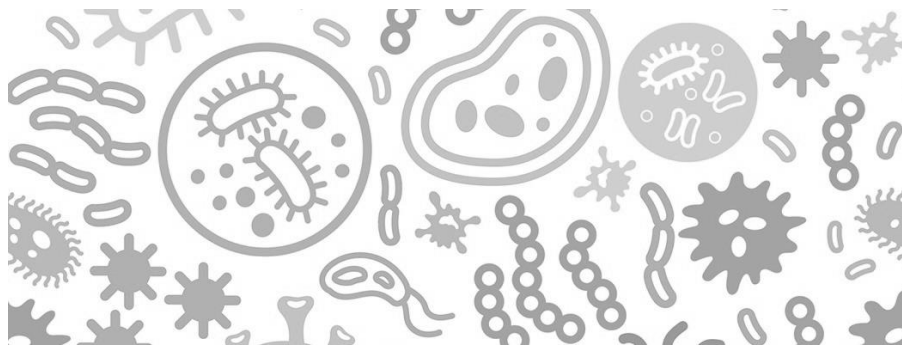
hot springs



(~10⁹⁻¹⁰个微生物/1克土)



Upon a closer look...



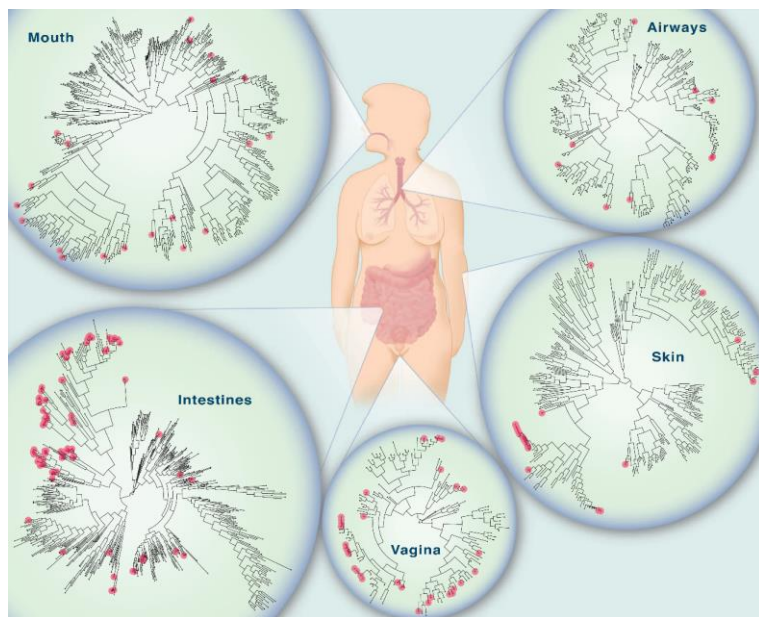
Microbes

- bacteria
- archaea
- fungi
- protists (原生生物)
- viruses
-

微生物(microbes)是指“一切肉眼看不见或看不清的微小生物的总称”。

Human Microbiome (人体微生物组)

- 在人体内及体表生活着大量的微生物，这些微生物群及其遗传信息的总和被称为人体微生物组，主要分布于肠道、皮肤、口腔、呼吸道、泌尿生殖道。
- 现代人生活方式的改变，导致肠道微生物消失，这影响身体的健康状态，是人类在21世纪面临的巨大挑战。



添加干酪乳杆菌活性菌
源自养乐多专利菌种



微生物组研究的核心：测序 + 数据分析

贵

- 设备与试剂贵，不易获得

繁

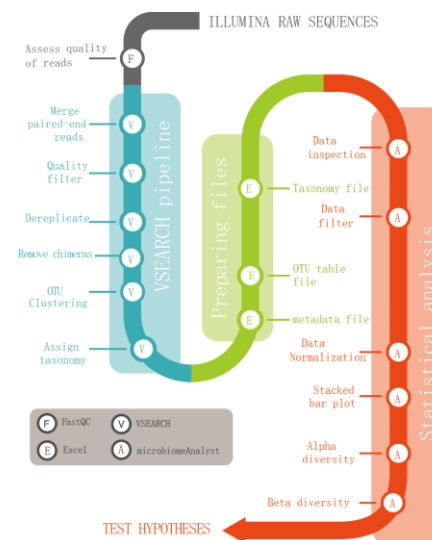
- 操作步骤多，易出错

难

- 数据量大，分析难度大



微生物组仿真实验提供学生模拟实验操作机会，使学生直观感受微生物组研究过程。

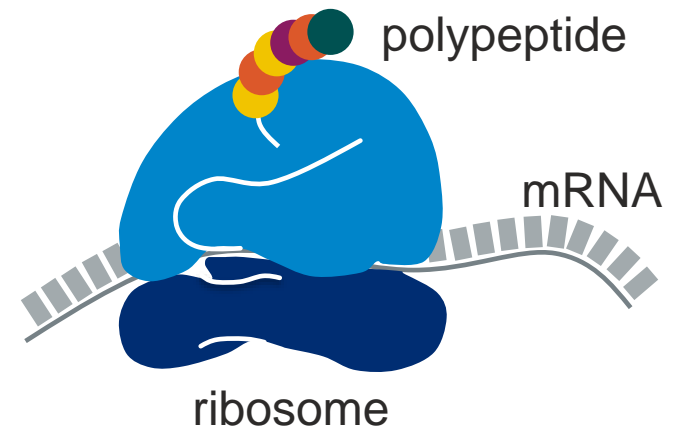


How do we identify members of microbiomes (如何鉴定微生物) ?

How can we study it by DNA sequencing (测序) ?

The 16S ribosomal RNA as a microbial fingerprint

- Fingerprints are *both* **universally present** on all people *and* **unique**
- The **ribosome** (核糖体) is essential for survival across all kingdoms of life and is thus **highly conserved**



核糖体是蛋白质翻译的场所

16S rRNA: 细菌的“分子化石”

- Specifically, the **16S rRNA** component of the ribosome is highly **conserved** among bacteria/archaea, yet contains **hypervariable** regions.

23S rRNA
5S rRNA
31 proteins

16S rRNA
21 proteins



16S rRNA在细菌/古菌的进化过程中高度保守，又有高度变异的区域。

→ 16S rRNA contains 9 variable regions

■ conserved region (non-specific)

■ variable region (species-specific)



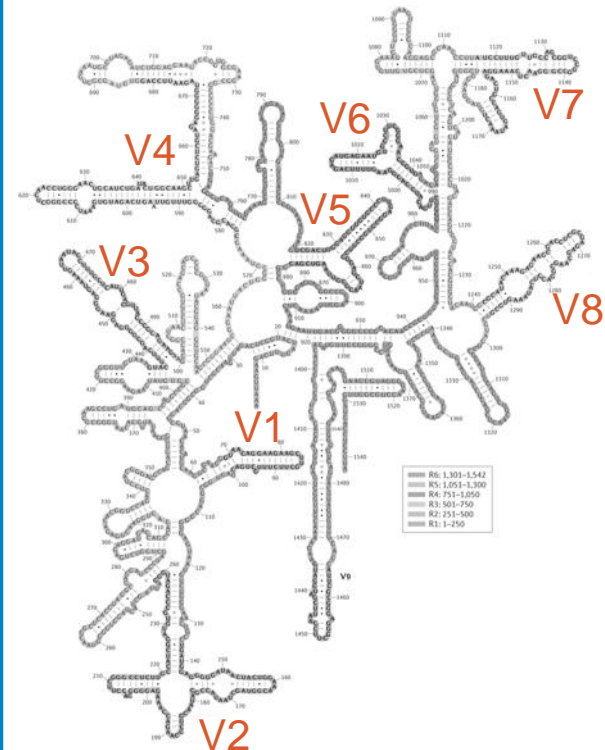
Hypervariable regions (可变区) can be used for species identification

→ More distantly related species exhibit more divergent

16S rRNA sequences

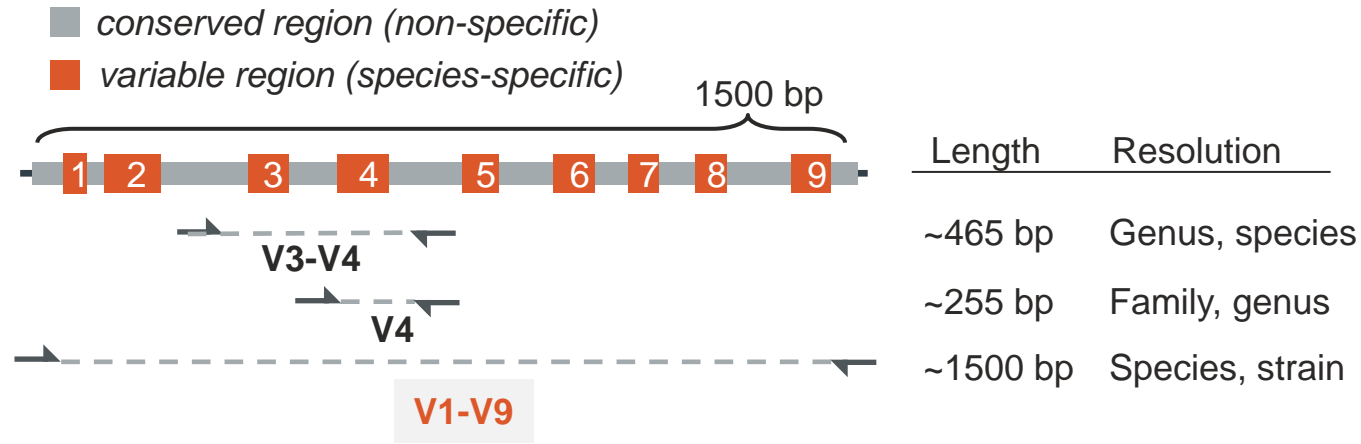
	70	80	90	100	110	120	130																																							
<i>E. coli</i>	C	G	G	T	A	A	C	A	G	A	G	C	T	T	G	C	T	G	A	C	G	G	G	T	A	T	G	T	C	T	G	G	A													
<i>M. iranicum</i> M05	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> HNTM87	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> GN10803	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> NJH	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> NLA001001296	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> FI05198	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> OPBG12013762	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> CCUG52297	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. iranicum</i> UM.TJL	C	G	G	---	A	---	C	C	T	T	T	---	G	G	G	T	---	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T							
<i>M. tuberculosis</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. aubagnense</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. fallax</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. mageritense</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. senegalense</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. farcinogenes</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. neworleansense</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. fortuitum</i> subsp. <i>acetamidolyticum</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. wolinskyi</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T
<i>M. phocaicum</i>	C	G	G	---	A	A	A	G	G	C	C	T	T	C	---	G	G	G	T	A	C	T	C	G	A	G	T	G	G	C	G	A	C	G	G	G	T	A	C	A	C	G	T	G	G	T

16S rRNA secondary structure



Nature Reviews | Microbiology

Highly conserved region (高度保守区): easy to target across all bacteria

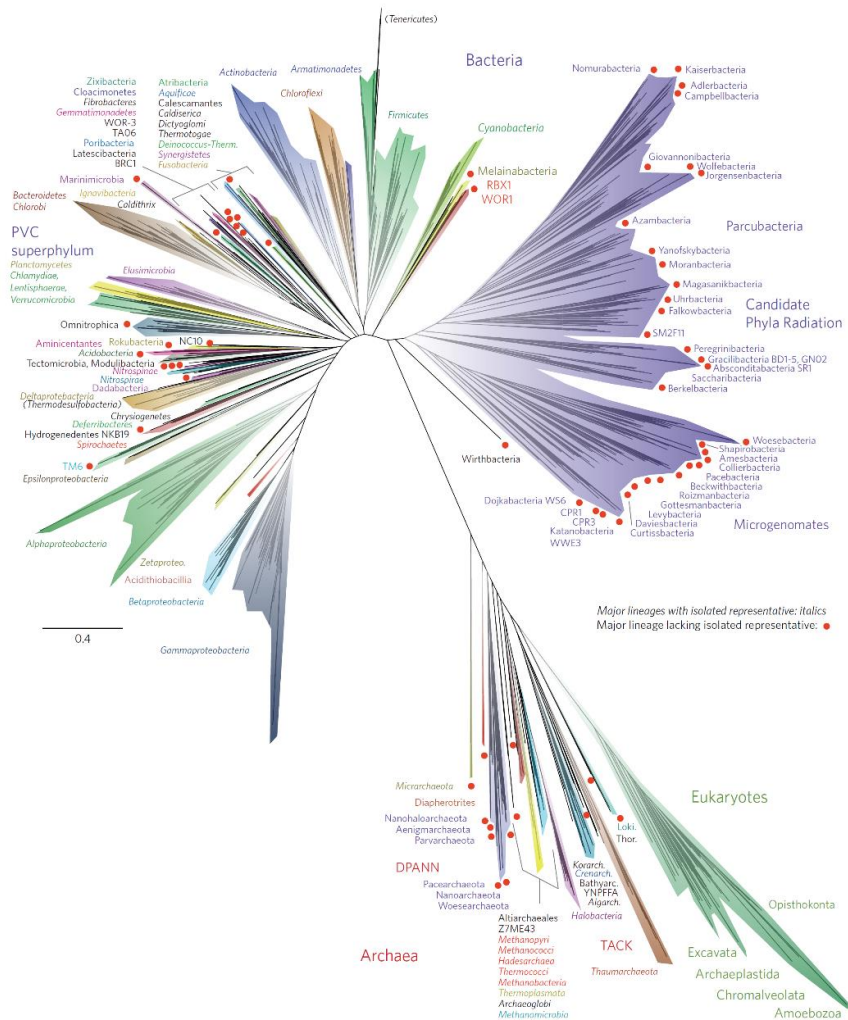


在16S保守区域设计**PCR引物**，可扩增不同的可变区，其中V3/V4区其特异性好，数据库信息全，适合短读段的二代测序。

分类单元
↓
界Kingdom
门Phylum
纲Class
目Order
科Family
属Genus
种 Species

Tree of Life: 16S rRNA gene

生命三域：细菌、古菌与真核生物（Woese and Fox, 1977）



Carl Woese (1928-2012)

(Hug *et al.*, 2016, Nature Microbiology)

16S rRNA基因为什么作为分子标记？

- 在生物体间普遍分布，序列有高度保守性
- 又有可变区，在不同生物中有一定变化，而且有稳定的突变速率。
- 分子大小适中(1500bp)，可进行测序分析
- 在细胞中含量高(rRNA基因拷贝数多)，易分离纯化

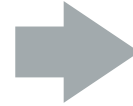
How to analyze microbial communities(微生物群落)?

How can we better understand our microbiomes?

Microbial genomics suffers from lack of cultivation approaches(纯培养方法)



Isolate

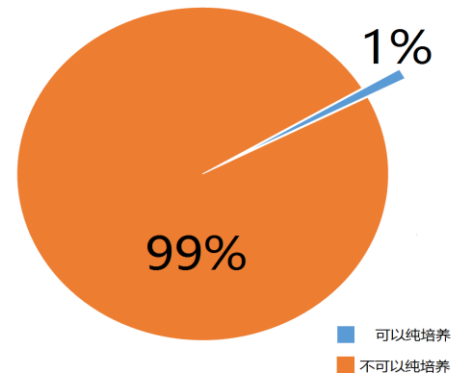


Genomics

多种因素导致大部分细菌目前无法培养

“The estimate that fewer than 1% of the prokaryotes in most environments can be cultivated in isolation has produced a quandary: what is the significance of the field of modern microbial genomics if it is limited to culturable organisms?”

Schloss et al, Genome Biology, 2005



宏基因组学(Metagenomics)

- 宏基因组学 (Metagenomics) 又称环境微生物基因组学，是指不经过微生物培养阶段，采用直接提取环境中总DNA的方法，对微生物基因总和进行研究的一门新学科。
 - Metagenomics is the study of genetic materials recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics.
- 宏基因组(Metagenome) 是由 Handelsman等1998年提出的新名词，其定义为“the genomes of the total microbiota found in nature”，即生境中全部微生物基因组的总和。
 - 包含可培养的和不可培养的微生物，目前主要指环境样品中的细菌和真菌。

Crosstalk R245

Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman¹, Michelle R Rondon¹, Sean F Brady², Jon Clardy² and Robert M Goodman¹



Cultured soil microorganisms have provided a rich source of natural-product chemistry. Because only a tiny fraction of soil microbes from soil are readily cultured, soil might be the greatest untapped resource for novel chemistry. The concept of cloning the metagenome to access the collective genomes and the biosynthetic machinery of soil microflora is explored here.

Addresses: ¹Department of Plant Pathology, University of Wisconsin—Madison, 1630 Linden Drive, Madison, WI 53706, USA. ²Department of Chemistry and Chemical Biology, Baker Laboratory, Cornell University, Ithaca, NY 14853, USA.

Chemistry & Biology October 1998, 5:R245–249
<http://biomednet.com/elecref/10745521005R0245>

Correspondence: Jo Handelsman
E-mail: joh@plantpath.wisc.edu

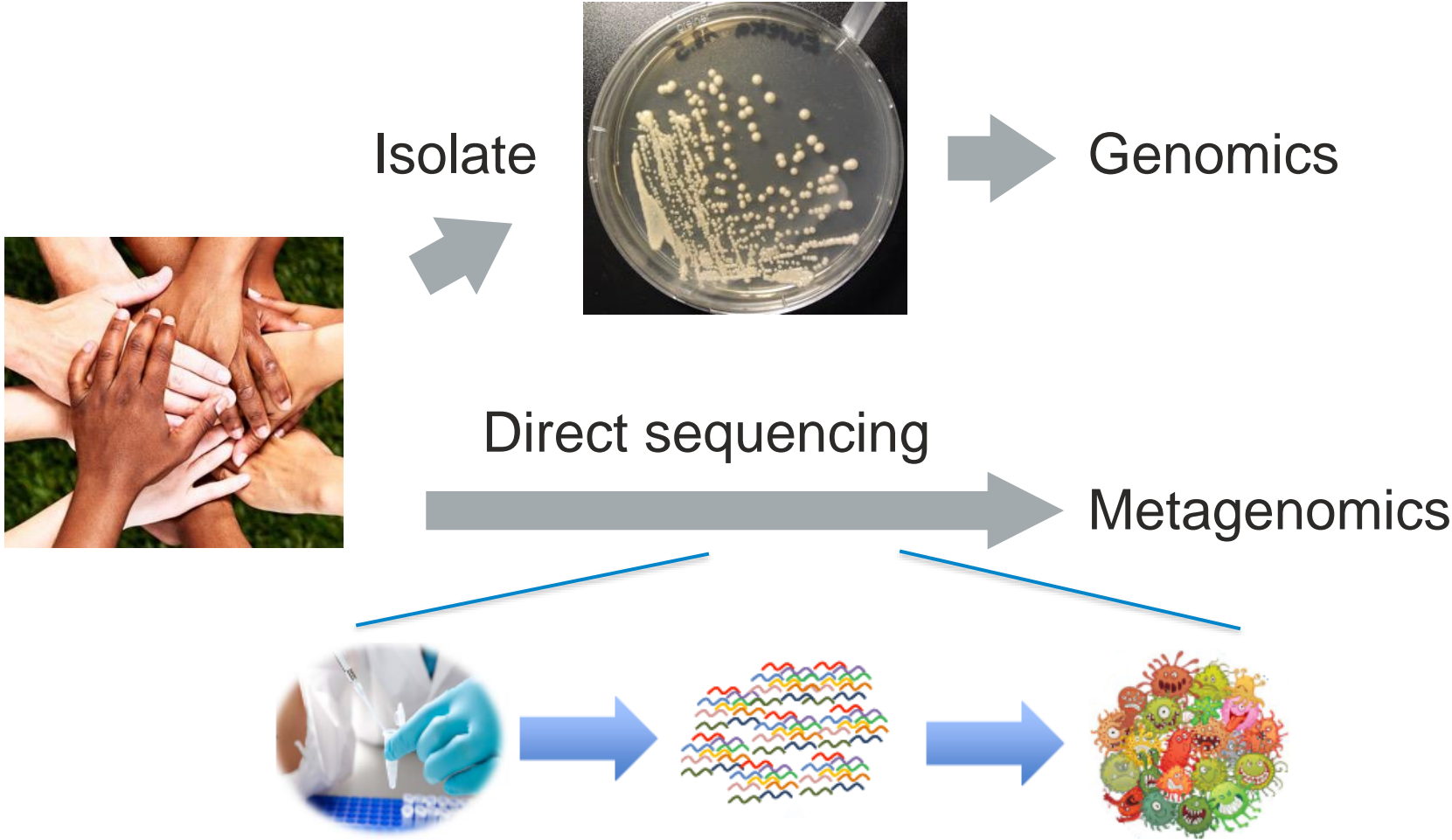
© Current Biology Ltd ISSN 1074-5521

Despite being familiar and useful, soil is also one of the least understood habitats on earth. The last 25 years of research have revealed that culturing is an excellent method to learn a lot about a tiny proportion of the microorganisms on earth [2–7]. Many lines of evidence show that fewer than 0.1% of the microorganisms in soil are readily cultured using current techniques [8–10]. And, most impressively, the other 99.9% of soil microflora is emerging as a world of stunning, novel genetic diversity. New groups of bacteria have been identified in soil that appear to diverge so deeply from the cultured bacteria that they could represent new phyla, or even new kingdoms of life [11–13]. Groups of *Archaea* related to those found thus far only in the open ocean are soil inhabitants around the world [14,15]. Estimates are that a gram of soil might contain 1,000–10,000 species of unknown prokaryotes [8]. There is likely to be



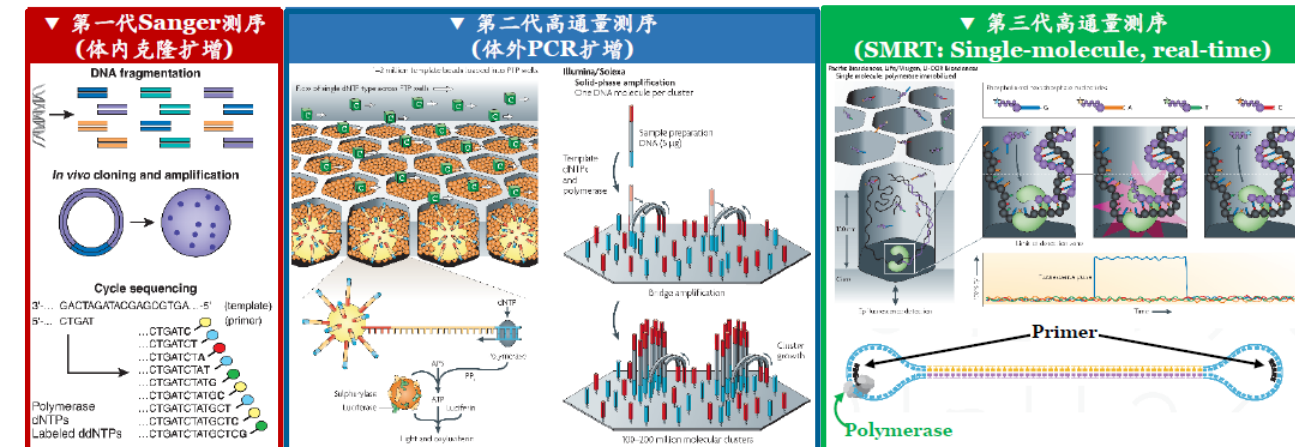
Jo Handelsman

Metagenomics has revolutionized microbiome studies



下一代测序技术催生了宏基因组学

Next Generation Sequencing (NGS) technology



Method	Generation	Read length (bp)	Single pass error rate	No. of reads per run	Time per run	Cost per million bases
Sanger ABI 3730xl	1st	600-1000	0.001%	96	0.5-3 h	\$500
454 (Roche) GS FLX+	2nd	700	1%	1×10^9	23 h	\$8.57
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1%	8×10^9 (paired)	7-60 h	\$0.03
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1%	1.2×10^9 (paired)	1-6 days	\$0.04
Ion Torrent	2nd	200	1%	8.2×10^7	2-4 h	\$0.1
SOLiD 5500xl	2nd	2×60	5%	8×10^8	6 days	\$0.11
PacBio RS II: P6-C4	3rd	Avg. 10-15 k	13%	$3.5-7.5 \times 10^4$	0.5-4 h	\$0.40-0.80
Oxford Nanopore MinION	3rd	Avg. 2-5 k	38%	$1.1-4.7 \times 10^4$	50 h	\$6.44-17.90

高通量测序

(Shendure & Ji, 2008, Nature Biotechnology; Metzker, 2010, Nature Reviews Genetics; Rhoads & Au, 2015, Genomics Proteomics Bioinformatics)

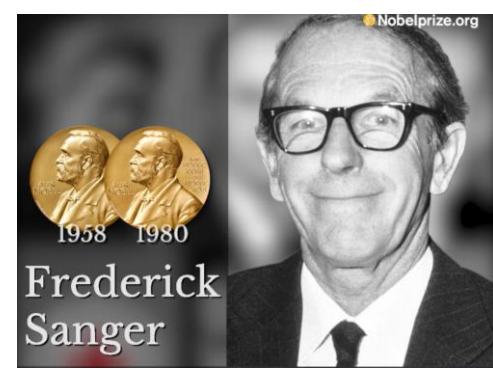


Next generation sequencer determines the bases of every DNA molecule.

Principles of Sanger Sequencing

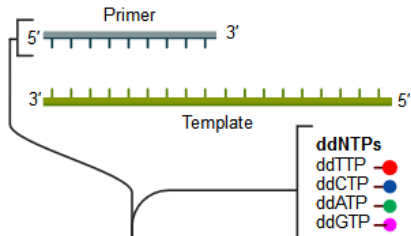
(第一代测序)

双脱氧核苷酸链终止法

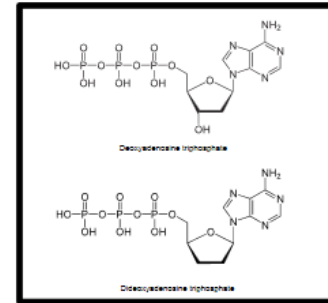
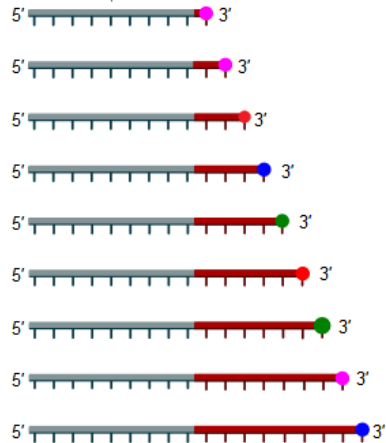


① Reaction mixture

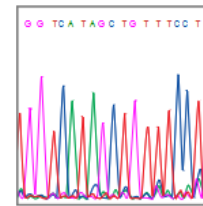
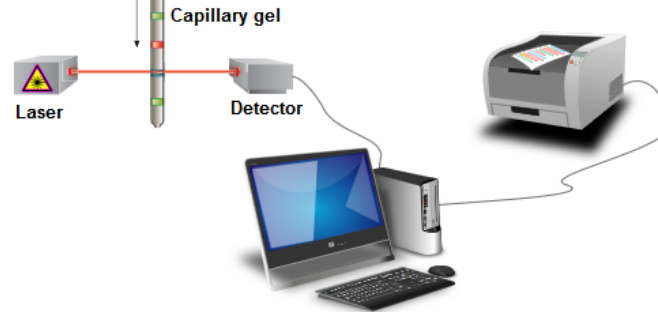
- Primer and DNA template
- DNA polymerase
- ddNTPs with flouochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)



② Primer elongation and chain termination



③ Capillary gel electrophoresis separation of DNA fragments

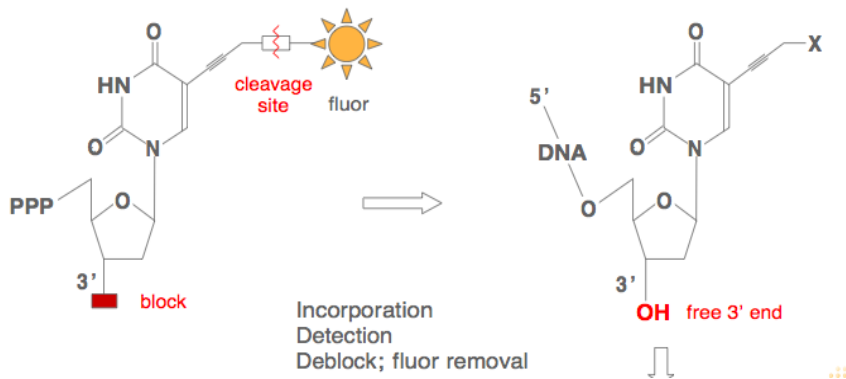


④ Laser detection of flouochromes and computational sequence analysis

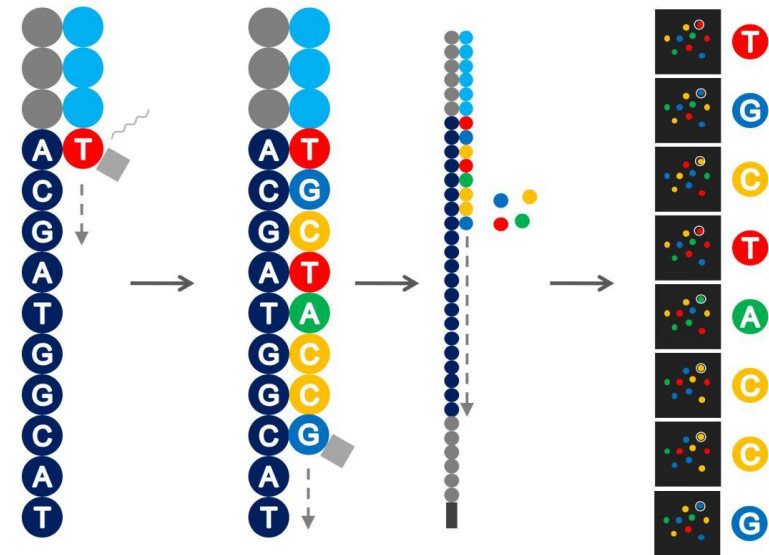
Principles of Illumina Sequencing (第二代测序)

- All 4 labelled nucleotides in 1 reaction
- Higher accuracy

Sequencing by Synthesis 边合成边测序



3'-blocked reversible terminator (可逆屏蔽终结子)



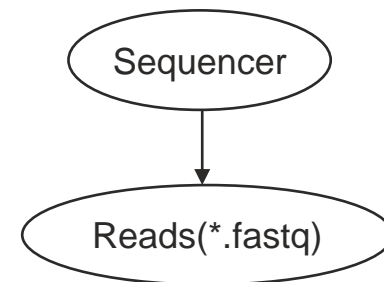
通过给不同的dNTP加上不同的荧光基团，再与固定到测序芯片上的DNA片段进行合成反应，通过观测发出的荧光信号颜色，判断这一步合成的核苷酸类型。

测序数据

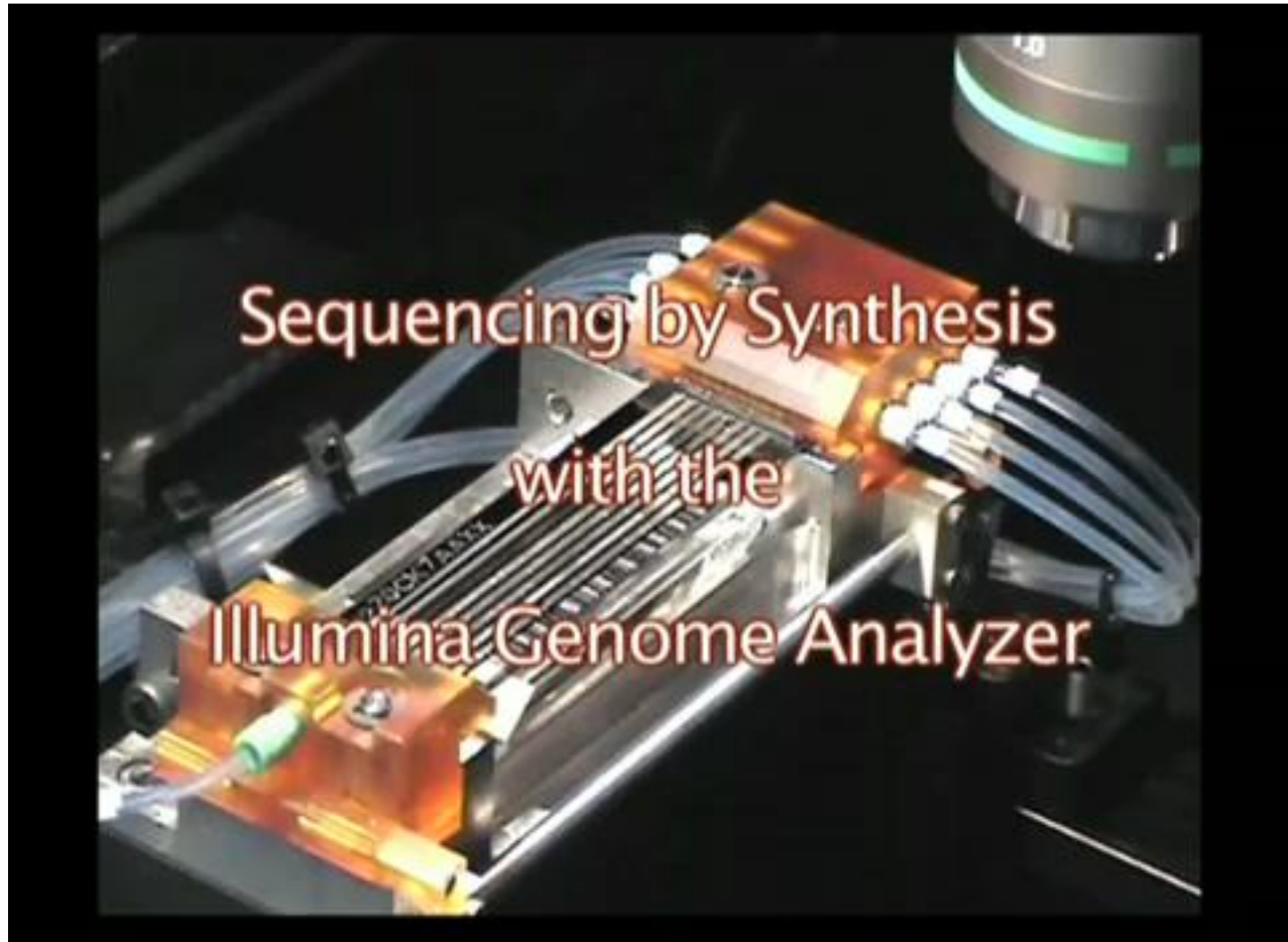
- Read (读段) : A short DNA fragment which is read out by sequencer.
 - DNA sequence
 - Quality information(质量值以ASCII码表示, 一般要Q>30)

```
@HISEQ2000:404:C73LWACXX:2:1101:1487:1876 1:N:0:CGATGT  
NCCCTCTTGA ACTCTCTCTTCAAAGTTCTTTCAACTTTCCTTACGGTACTTGTTGACTATCGGTCTCGTGCAGATCGGA  
+  
#4=DB?:DF?ADCFDGD>BHCEB9F3AAACEFHC>@BBFFFGD@??BF??D9B?FGDFFGDGGB@;AE>ED25;). . . ; ; ;
```

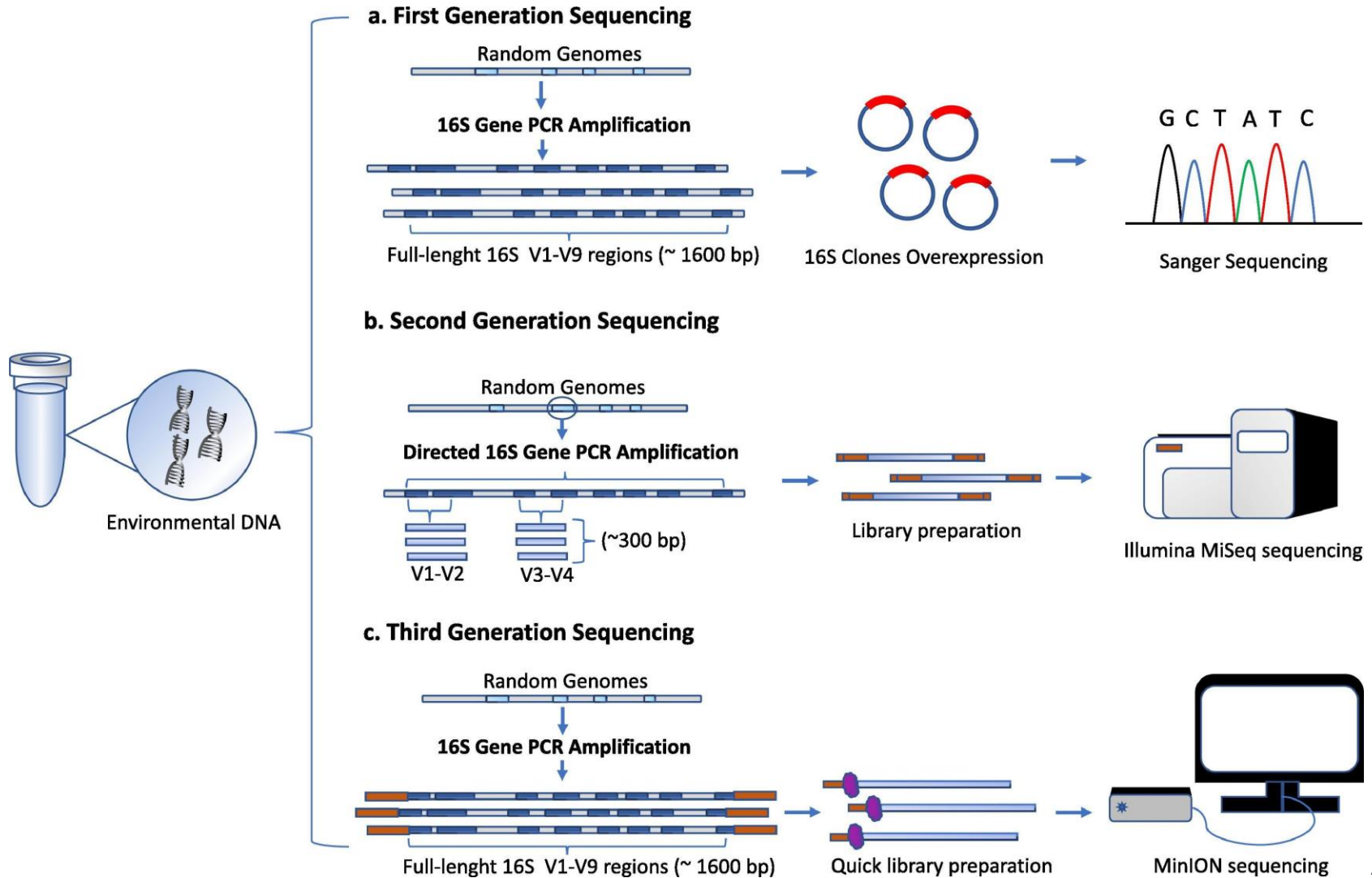
- 高通量测序的序列数据一般存储在FASTQ格式文件, 文件后缀一般为“.fastq”, “.fq”等



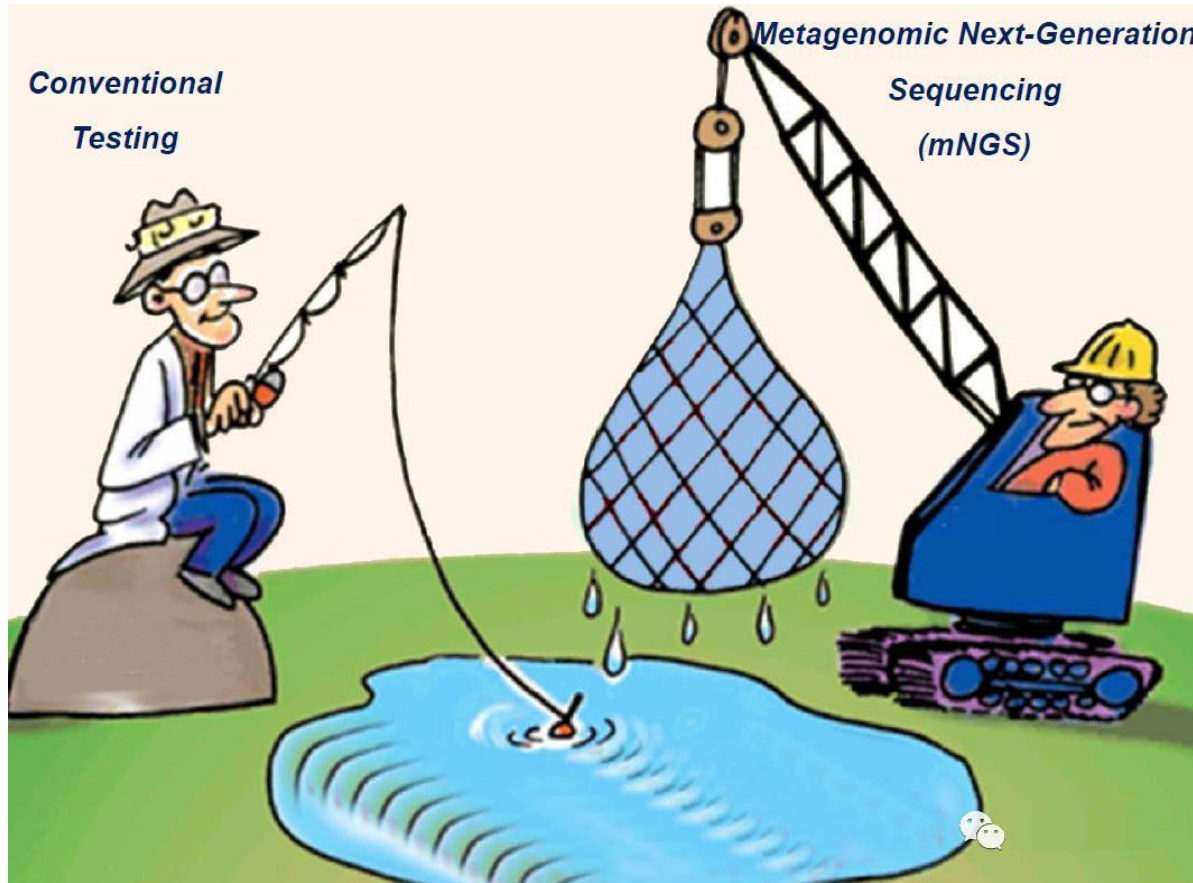
第二代测序视频



Three metagenomic strategies for each sequencing technology generation

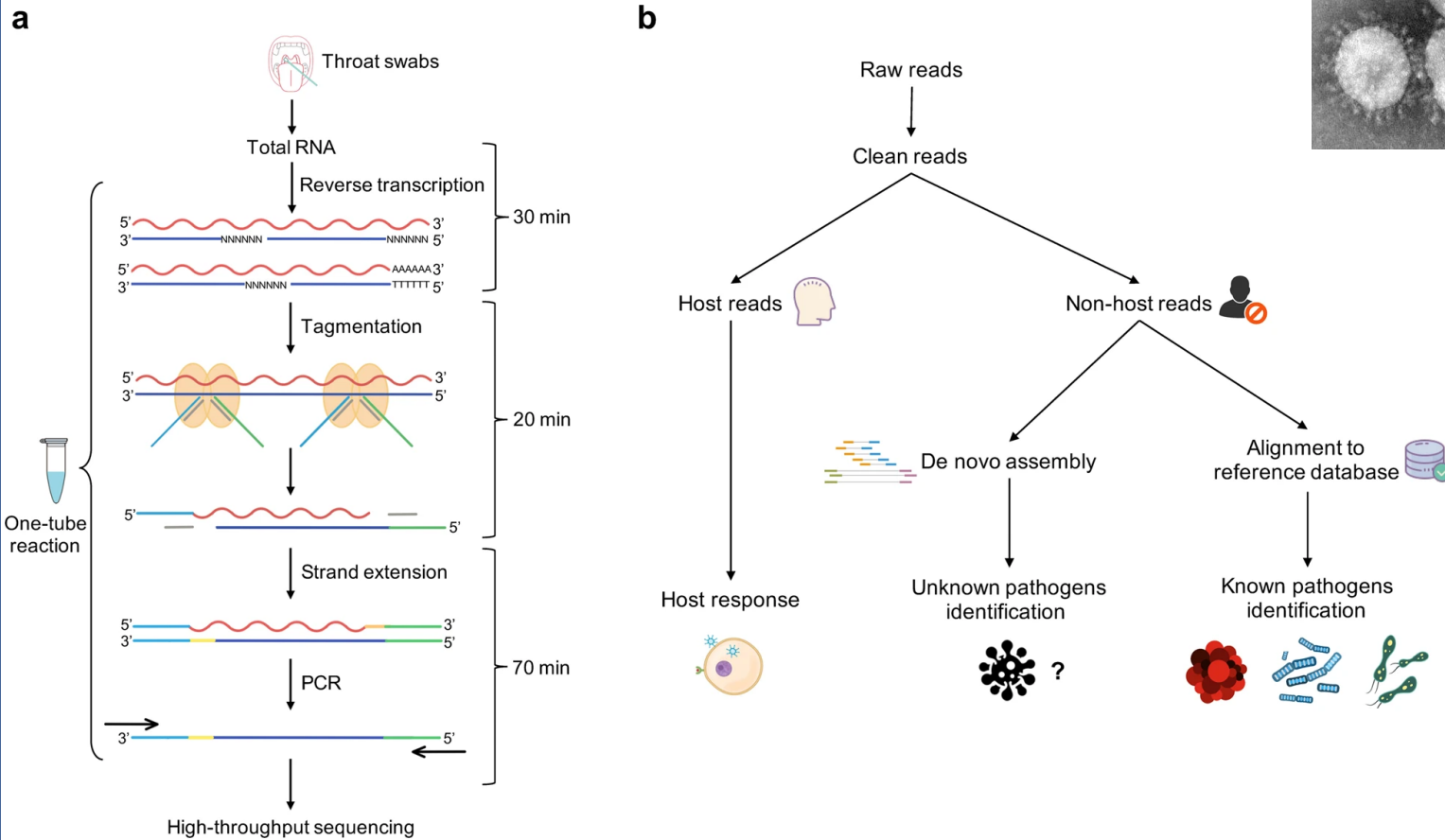
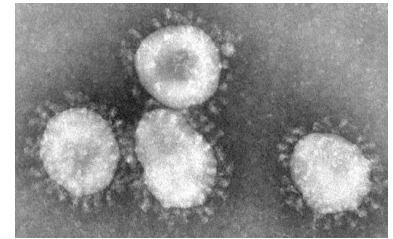


宏基因组下一代测序技术(mNGS)



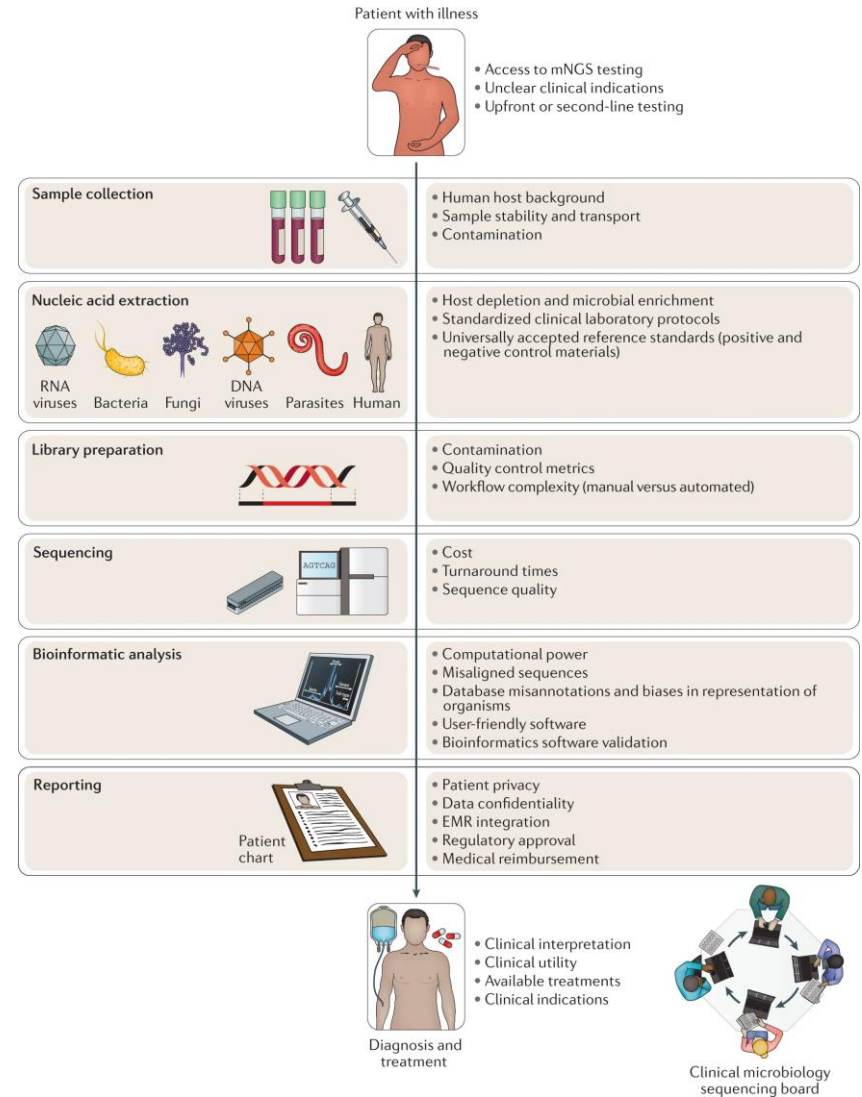
Metagenomic strategy for detecting SARS-CoV-2

传染病病原体鉴定时间：
2003年非典病毒(半年) → 2019年新冠病毒(一周)



Clinical metagenomics

- Clinical metagenomic next-generation sequencing (mNGS), the comprehensive analysis of microbial and host genetic material (DNA and RNA) in samples from patients, is rapidly moving from research to clinical laboratories.
- The capacity to detect all potential pathogens – bacteria, viruses, fungi and parasites – in a sample and simultaneously interrogate host responses has great potential utility in the diagnosis of infectious disease.



How does metagenomic sequencing work (宏基因组测序流程)?

What is a typical experimental & computational workflow?

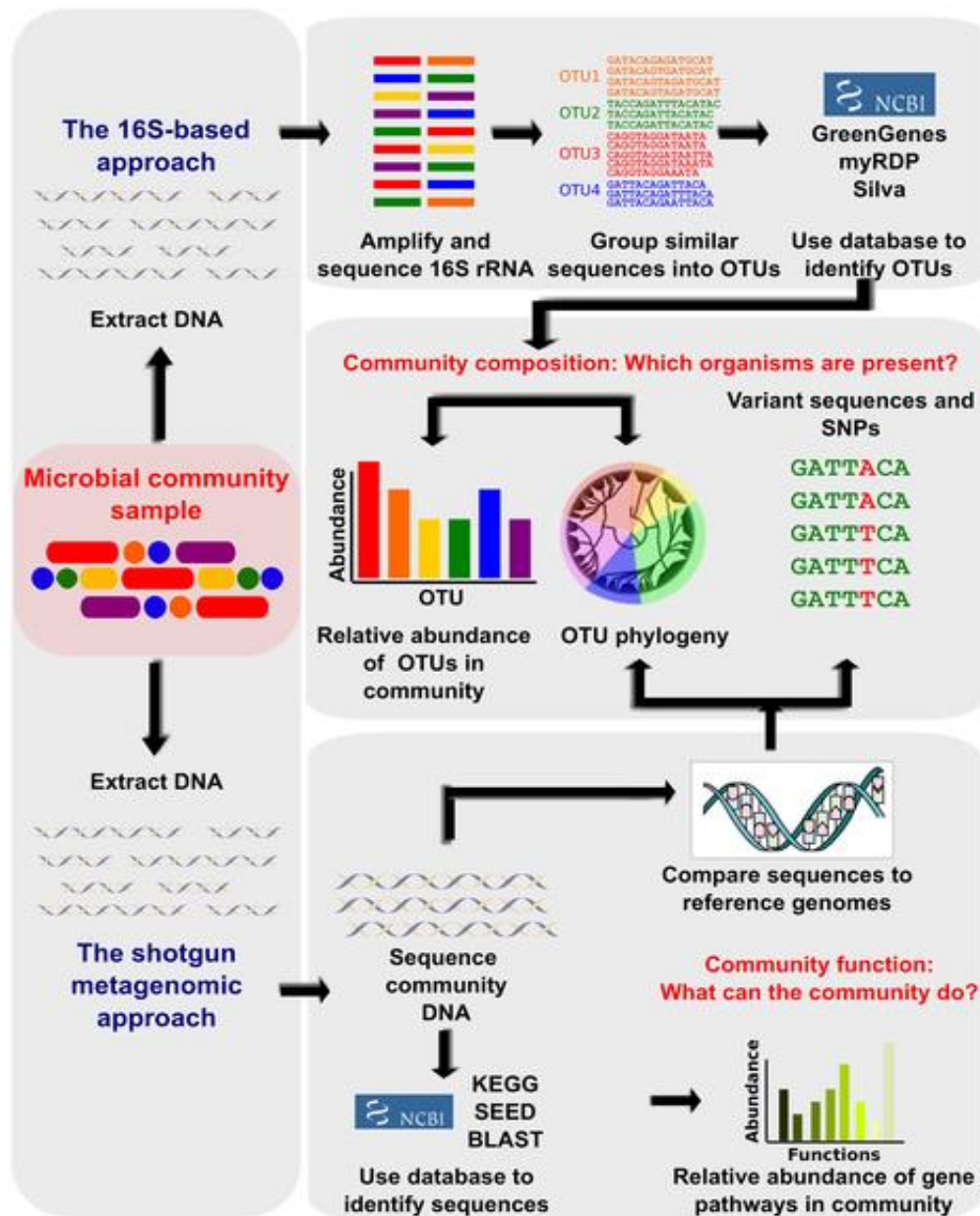
宏基因组学两种测序策略

扩增子测序

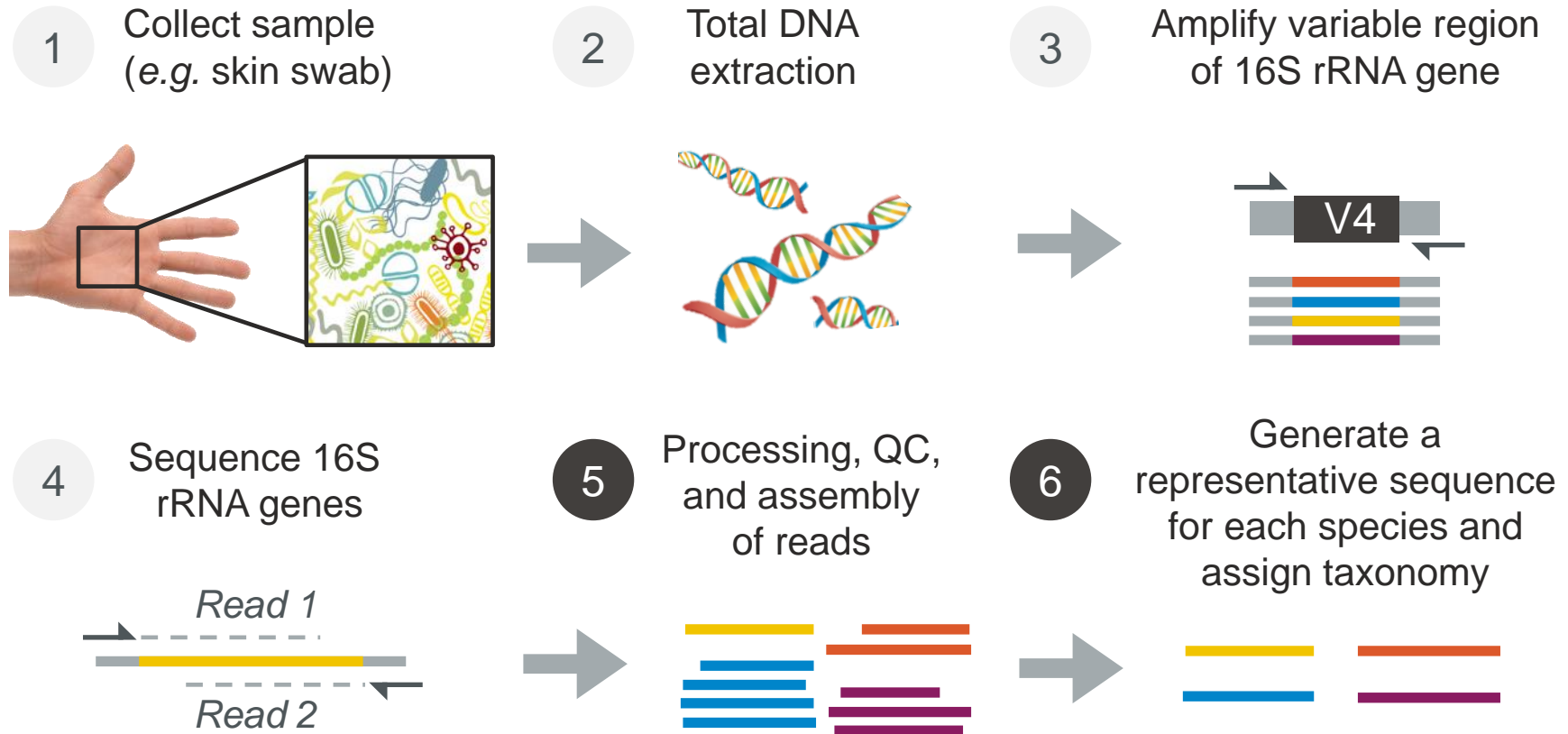
- 对不同生物中保守的标记基因（16S/18S/ITS），进行PCR扩增，再测序分析

宏基因组测序

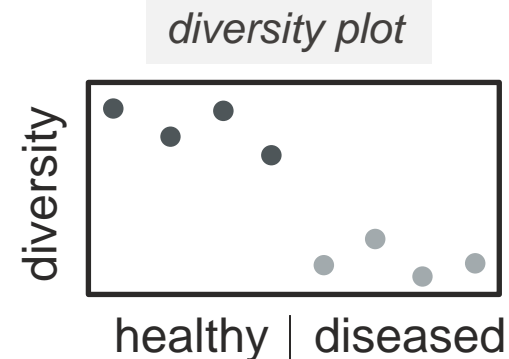
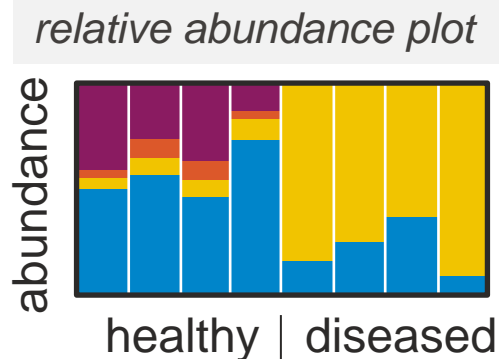
- 直接提取样品的全部基因组DNA，进行测序分析



A workflow for 16S amplicon sequencing

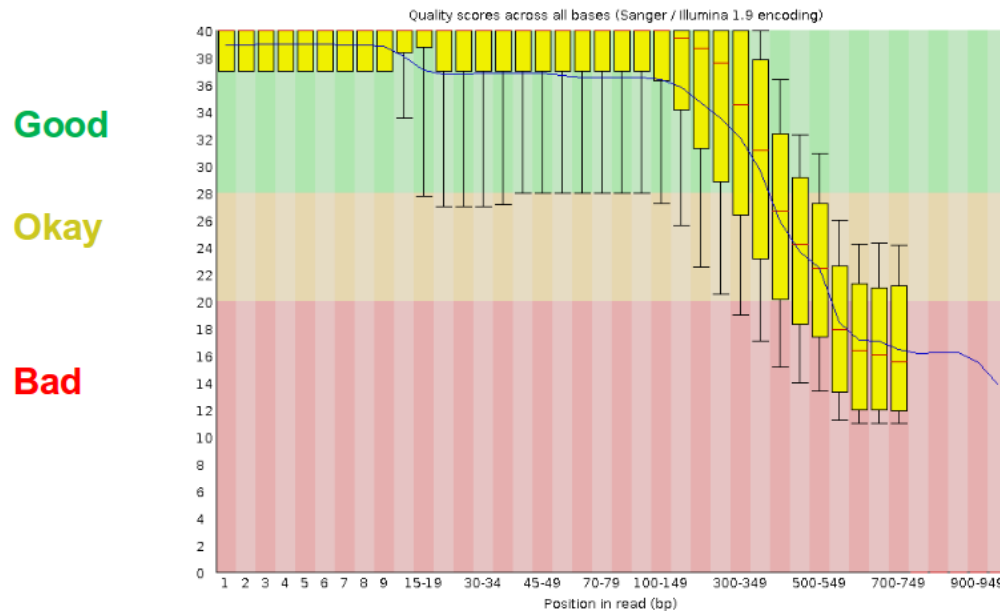


7
Who? How much?



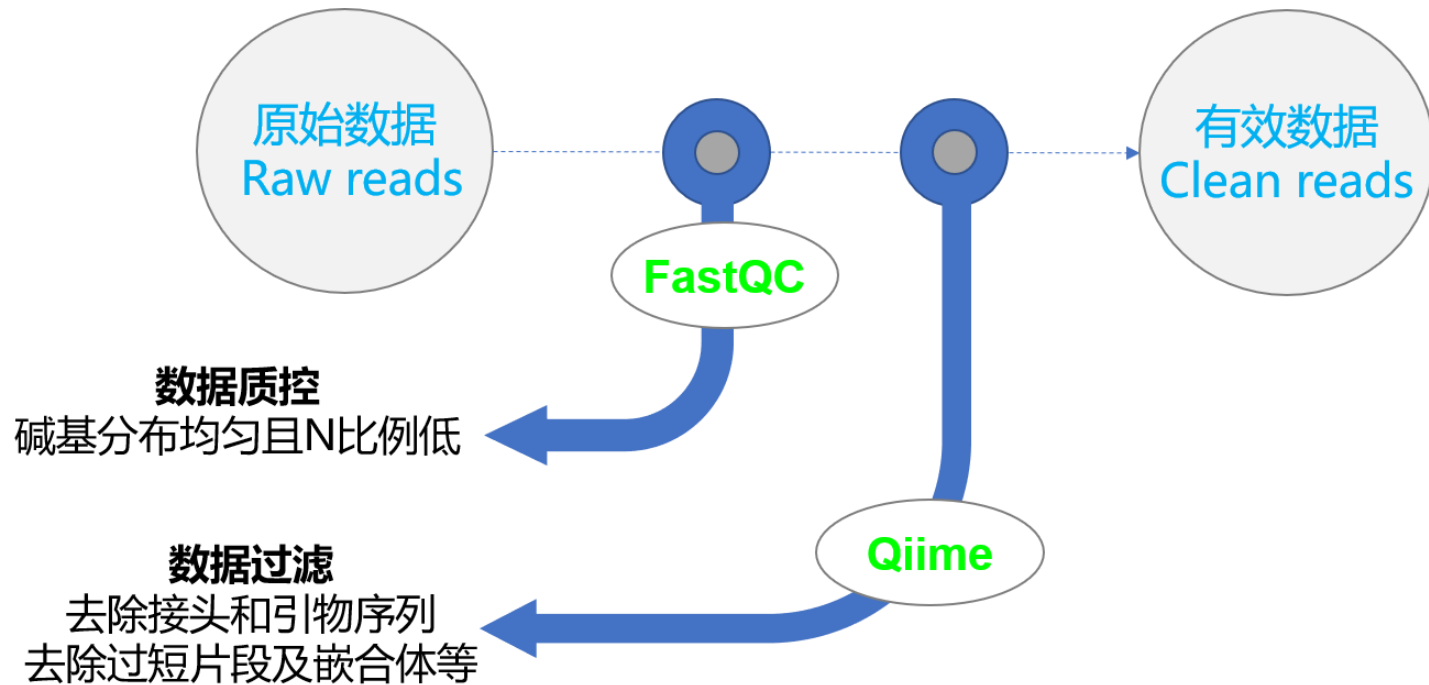
Sequence processing (测序数据处理)

5 Processing and QC of reads (测序数据质控)



Reads quality analyzed by FastQC

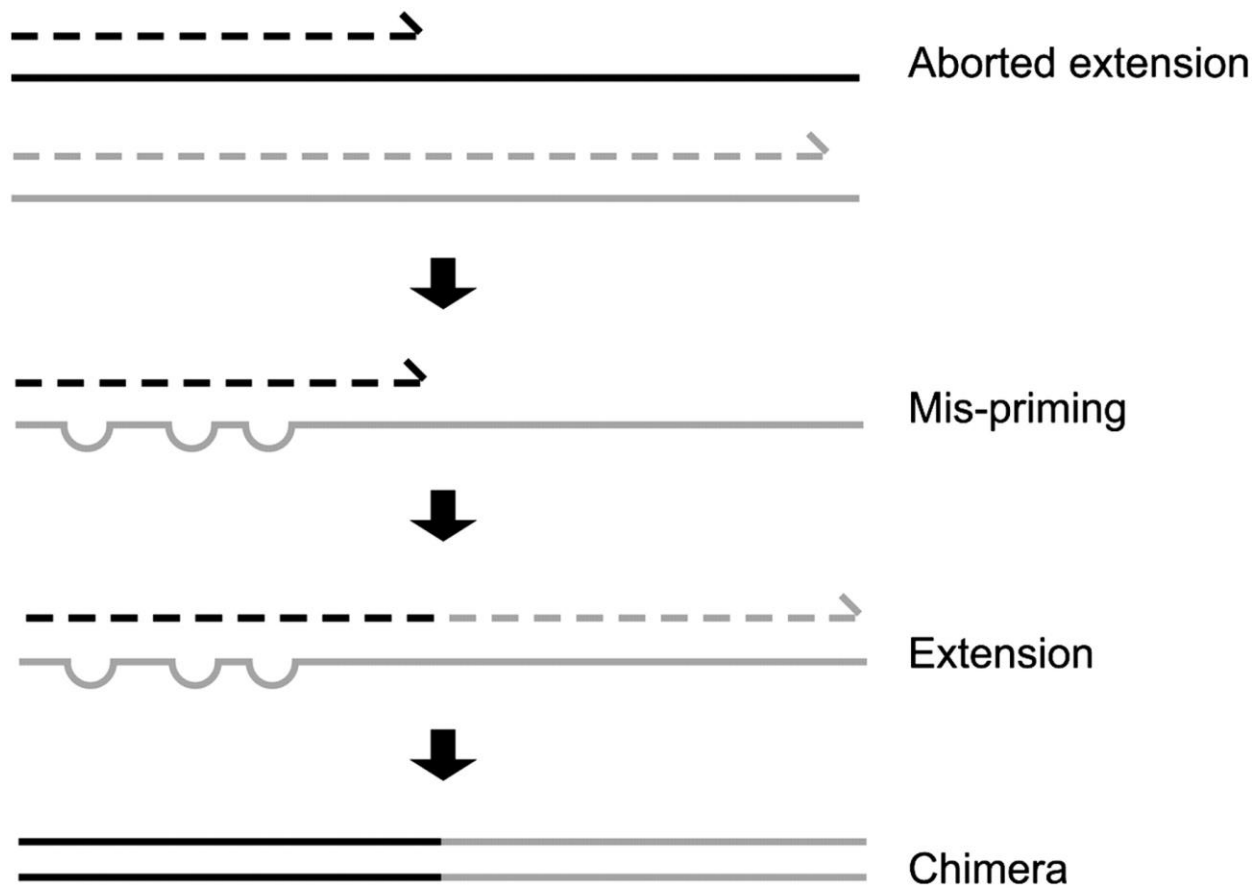
Sequence processing(测序数据处理)



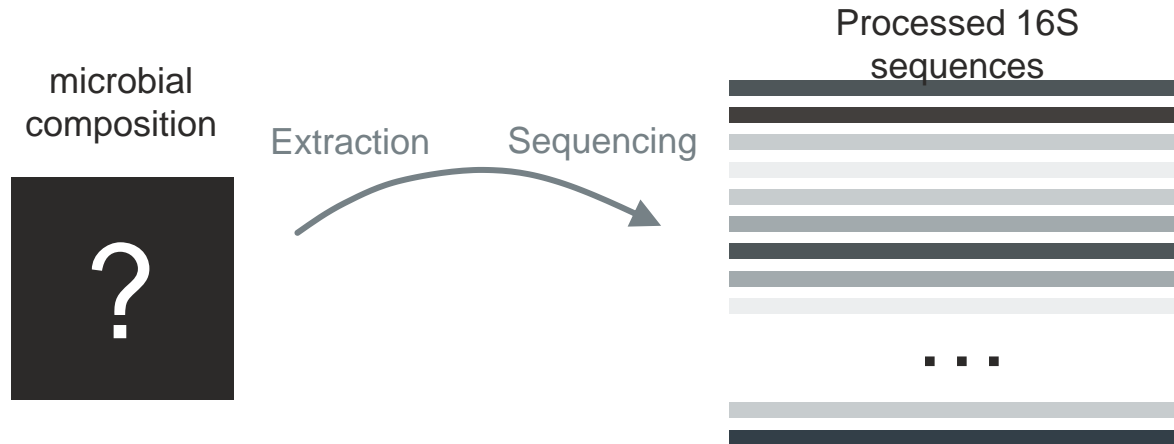
There are a lot of ways to filter and trim your data:
(i) low quality bases ($Q < 20$)
(ii) Remove chimeras (嵌合体)

Chimera Removal(去除嵌合体):

During PCR multiple sequences can combine to form a hybrid.

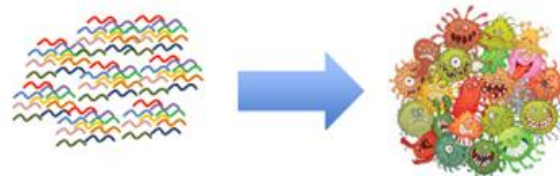


Where do we want to go next?



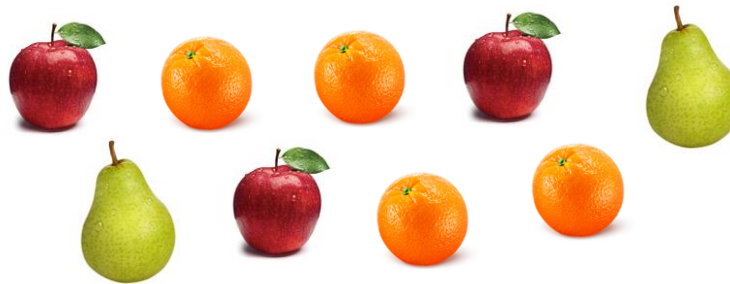
The challenge of metagenomics is that the sample is mixed!

→ Which 16S sequence came from which bacterium?



It is trivial to catalog identical objects

mixed sample



cataloged



Cataloging variable objects is hard

mixed
sample



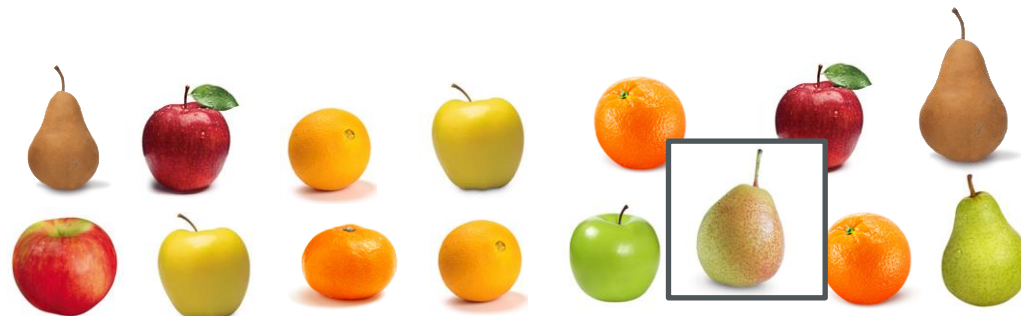
Cataloging variable objects is hard



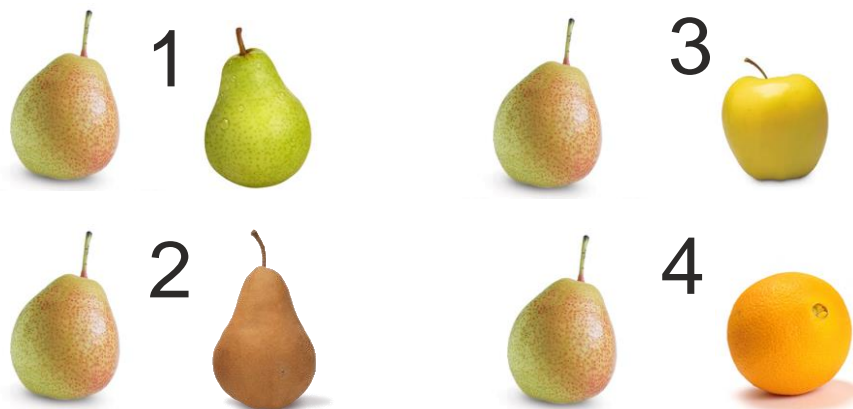
假设我们都不认识这些水果，如何把标星的梨挑选出来？

We catalog variable objects by iterative pairwise comparison(两两比较)

mixed
sample

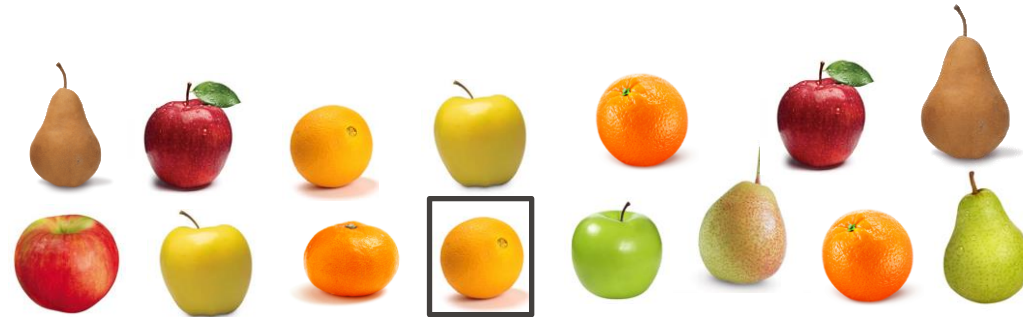


ranked
pairwise
comparisons

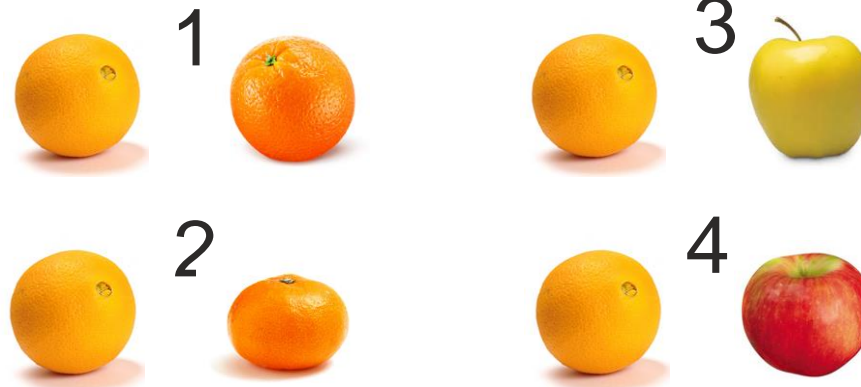


We catalog variable objects by iterative pairwise comparison

mixed
sample

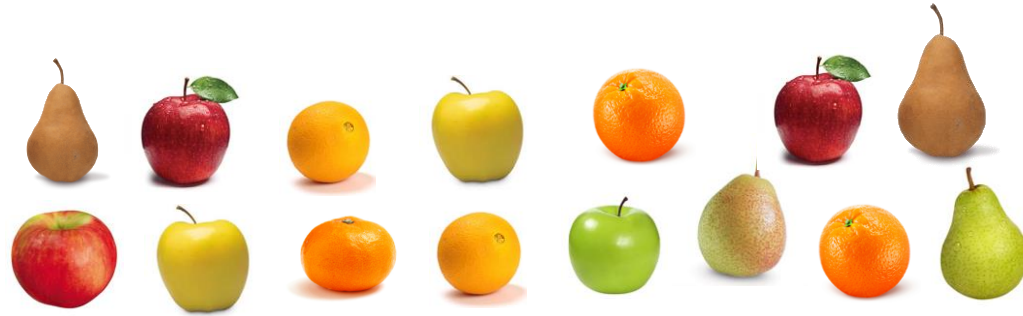


ranked
pairwise
comparisons

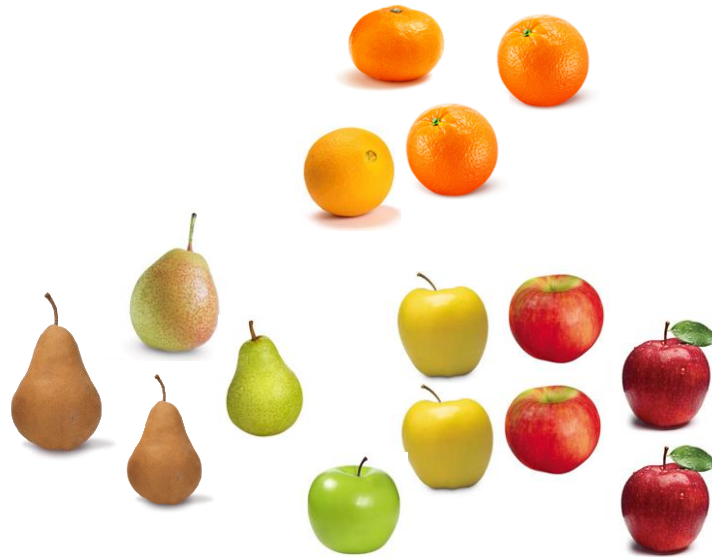


Clusters arise of similar items

mixed
sample

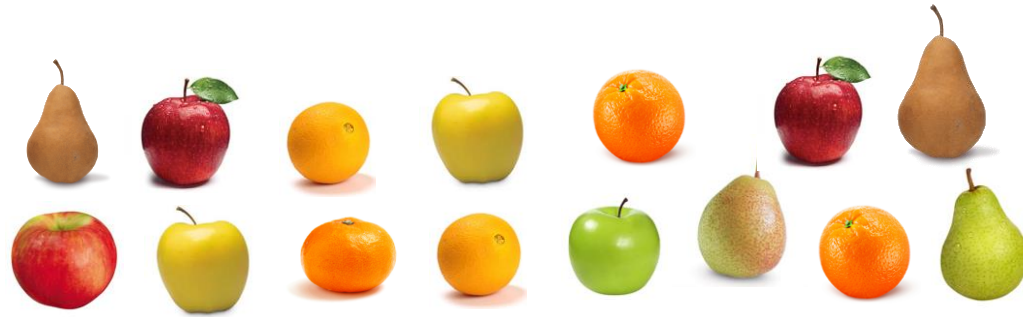


ranked
pairwise
comparisons

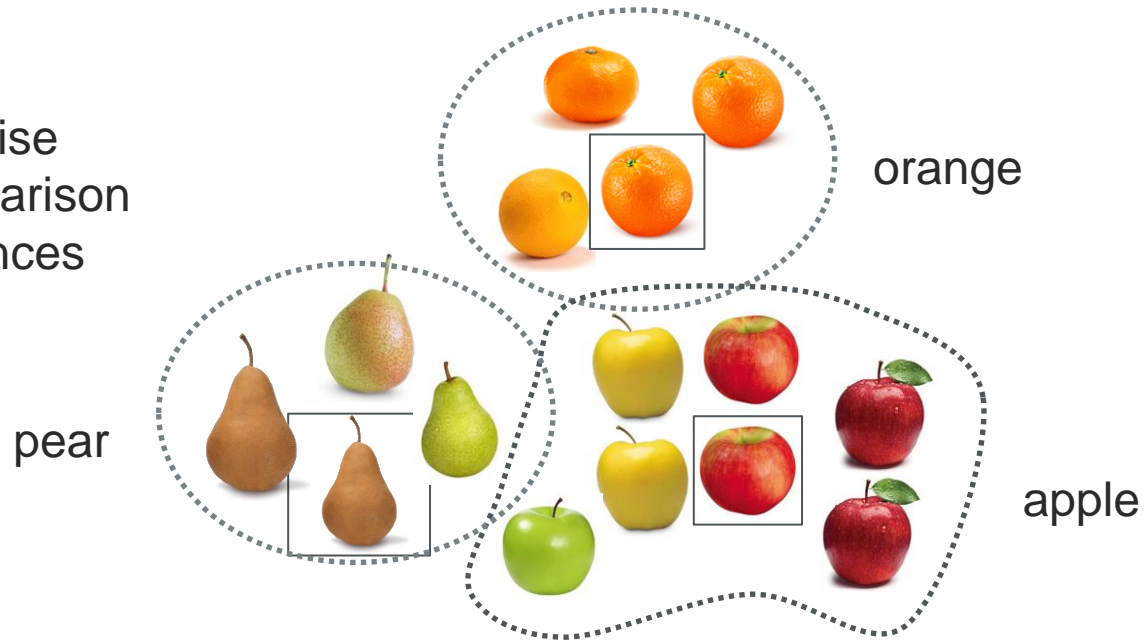


Similar fruits cluster together

mixed
sample

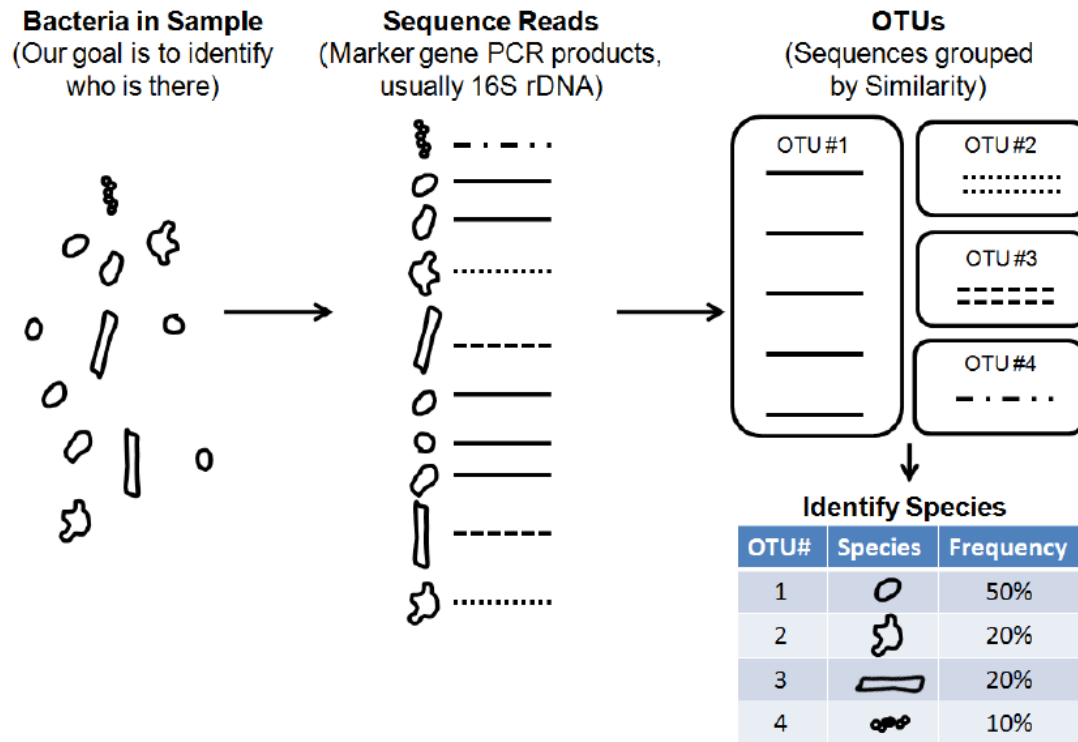


pairwise
comparison
distances



16S rRNA 基因序列的分类鉴定

- 1 Identify unique sequences
- 2 Use pairwise comparison to cluster into operational taxonomic units (OTUs)
- 3 Count how many sequences match each OTU



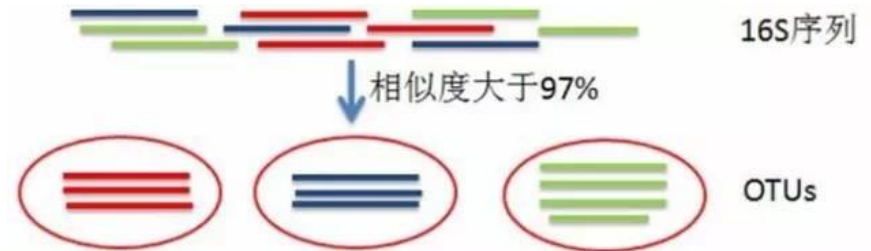
- OTU:可操作分类单元
 - 为便于进行分类分析，人为给某一个分类单元（品系、种、属等）设置的同一标志（Marker）
- OTU also refers to clusters of organisms, grouped by DNA sequence similarity of a **specific taxonomic marker** gene.
 - 原核生物16S rDNA
 - 真核生物18S rDNA/ITS

6

Sequences to OTUs to OTU abundance(丰度)

16S序列相似性达到97%以上的菌株为同一个种(Species)

OTU sequences are representative sequences chosen for each OTUs and are <97% similar compared to any other OTU sequences
 ~ **97% similarity = species**



To generate a **relative abundance table**, count the number of 16S sequences matching each OTU sequence

		Samples		
#OTU ID		F3D0	F3D141	F3D142
OTUs	OTU_6	749	535	313
	OTU_25	29	57	14
	OTU_1	613	497	312
	OTU_8	426	378	255
	OTU_31	149	38	10
	OTU_2	366	392	327

Counts

OTU丰度表：菌群组成

OTU taxonomy: 微生物分类注释


- OTU聚类后，挑选出每个OTU中的代表序列，与RDP、SILVA或GreenGenes等数据库进行比对，进行物种注释。

- Berkeley lab
- August 2013
- 202,421 entries

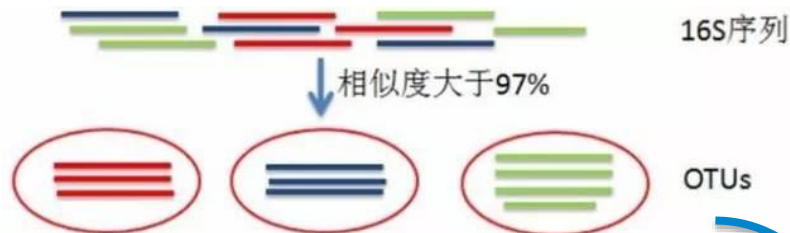


16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov





- Max Planck Institute
- July 2015
- 172,418 entries



high quality ribosomal RNA databases



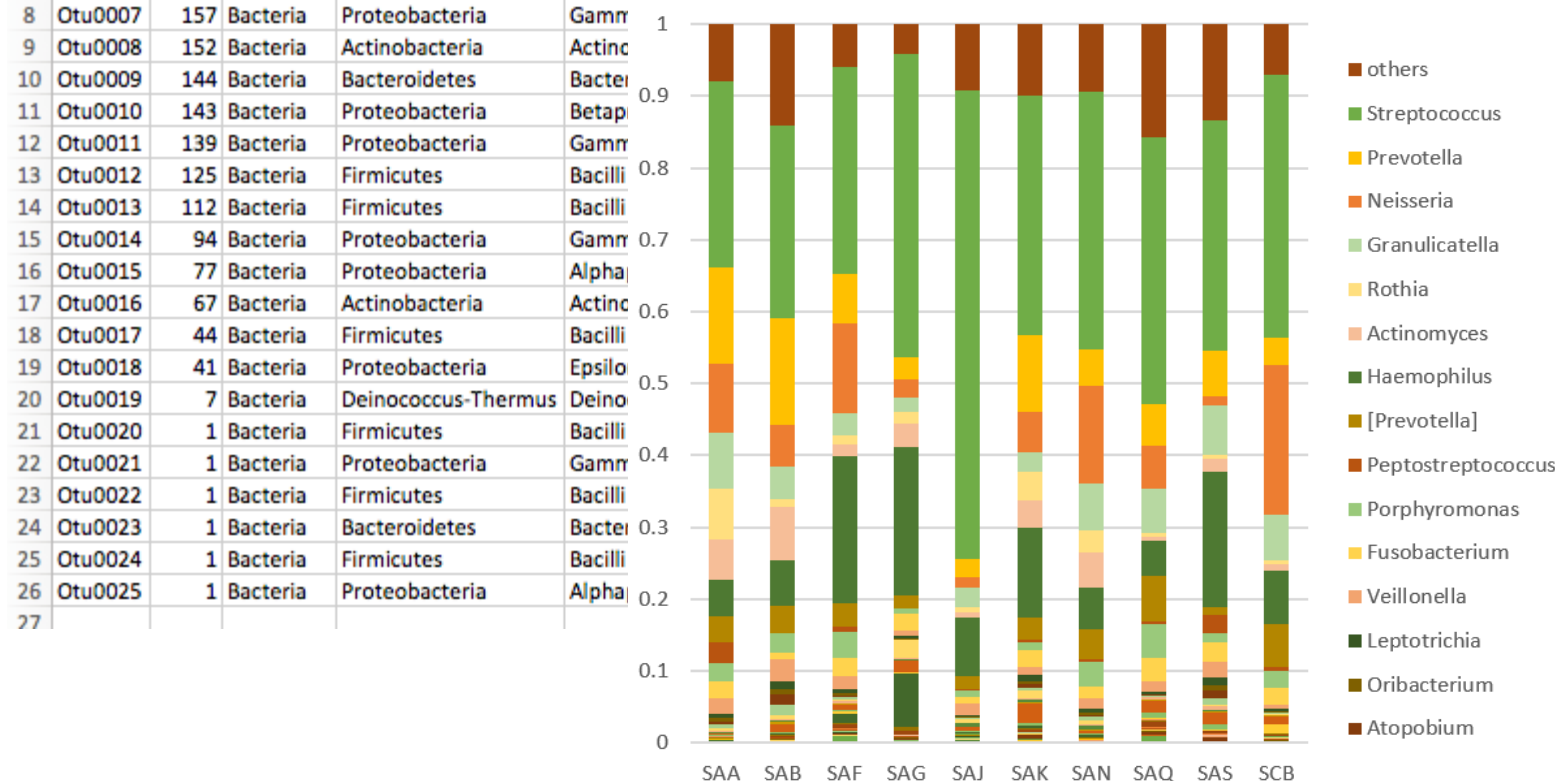
rDNA数据库(Ribosomal Database Project):
<http://rdp.cme.msu.edu/>

OTU	Count	Species
	3	Accumulibacter
	11	Unkown
	3	Competibacter
	1	Bacillus anthracis

OTU table

物种OUT聚类表: OTU table

	A	B	C	D	E	F	G	H
1	OTU	Reads	Taxonomy					
2	Otu0001	342	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
3	Otu0002	265	Bacteria	Firmicutes	Bacilli	Bacillales	Listeriaceae	Listeria
4	Otu0003	222	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
5	Otu0004	191	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
6	Otu0005	184	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
7	Otu0006	170	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium



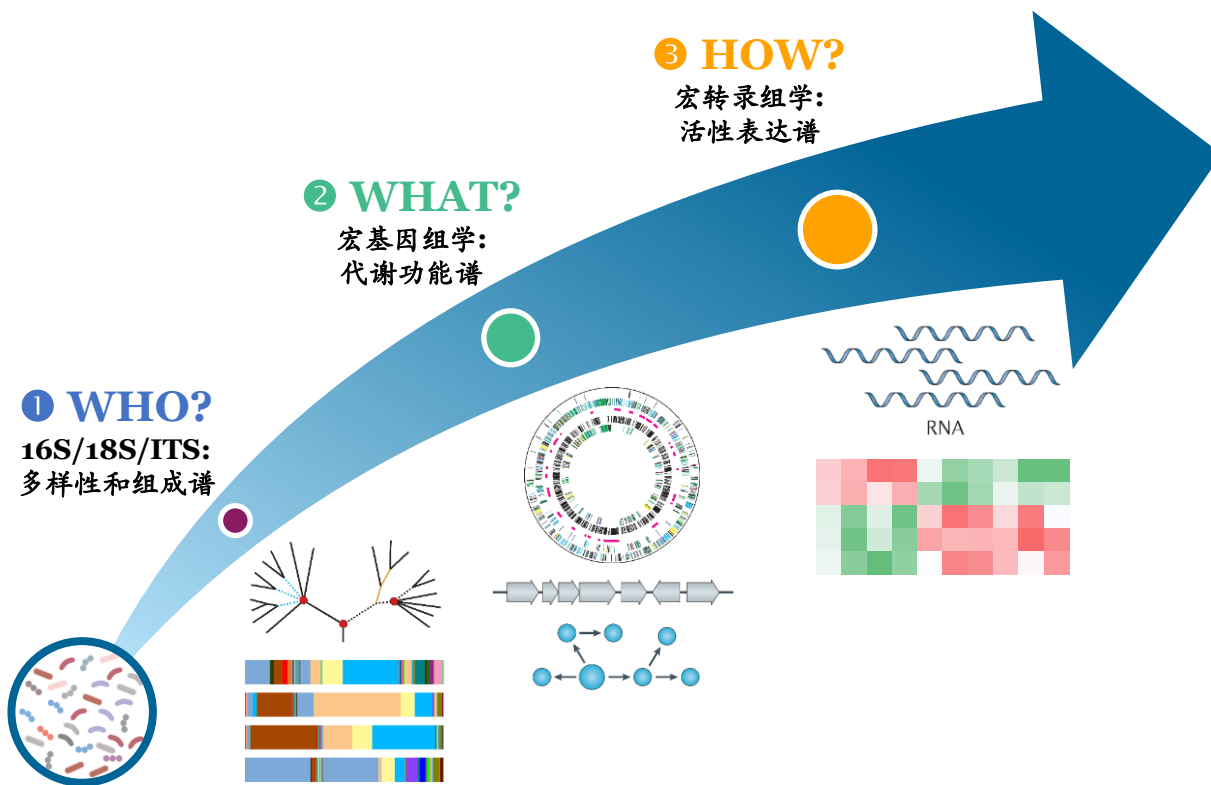
物种分类柱状堆积图

微生物组软件（USEARCH/QIIME）

- USEARCH是好用的扩增子分析软件，但是代码不开源，用于分析较大数据的64位版本收费。VSEARCH是USEARCH的免费、开源代替品。VSEARCH主要功能有：嵌合体检测、聚类、去冗余、两两比对、排序、抽样、物种分类等。
- QIIME(Quantitative Insights Into Microbial Ecology)是一个用于比较和分析微生物基因组的开源软件，其开发者是美国科罗拉多大学的Rob Knight团队。QIIME能够处理各种测序平台上扩增子高通量测序结果。



基于高通量测序技术的微生物组学研究



微生物组仿真实验 操作练习

扫描右边二维码查看操作说明👉



超星在线课程《微生物育种实验》