Next-generation Sequencing

Lecture 14

Introduction of metagenomics

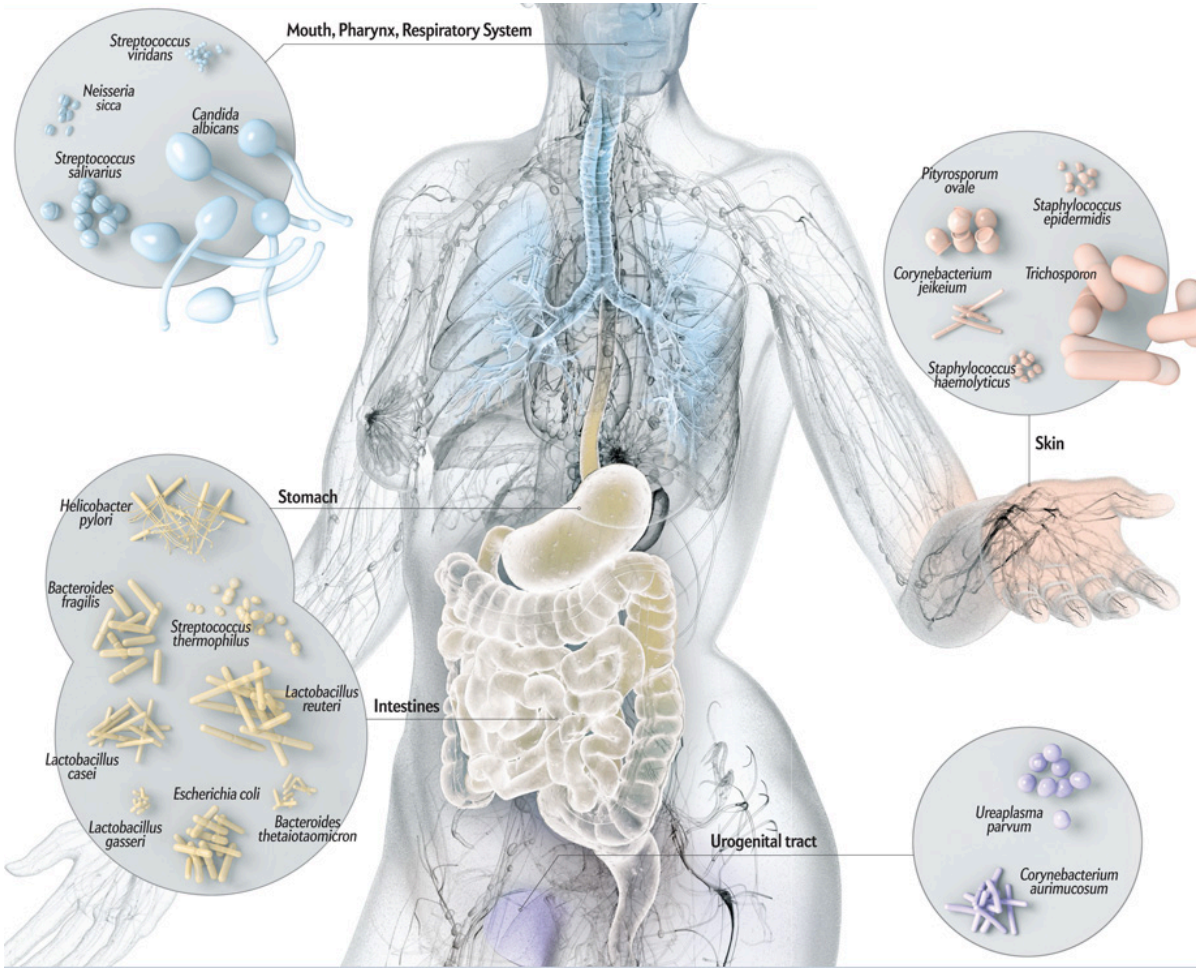
What is Metagenomics? Meta + Genomics

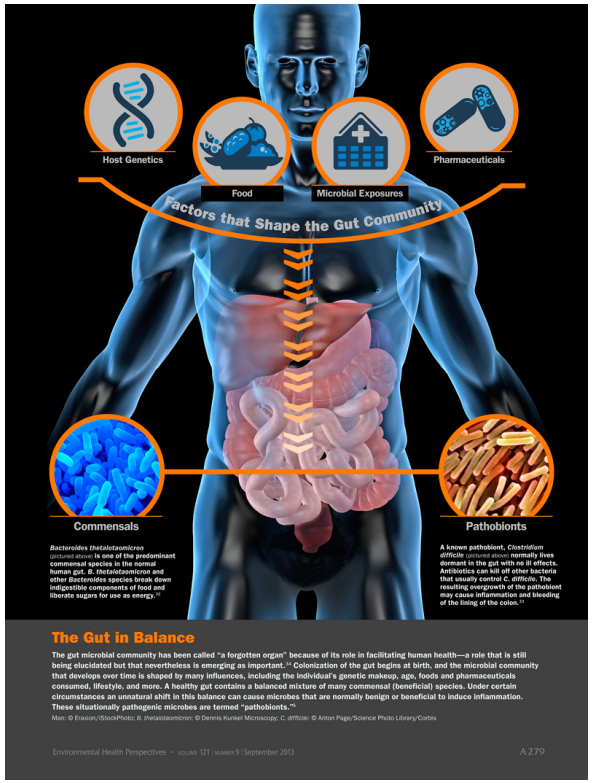
- **Metagenomics** (**Environmental Genomics, Ecogenomics** or **Community Genomics**) is the study of genomes recovered from environmental samples without the need for culturing them.

- **Key words**: Microbiology, Microbiome, Genomes, Genomics, High-throughput sequencing, Biodiversity, Microbial ecology, PCR, Cloning, 16S rRNA, 18S rRNA, Bioinformatics.

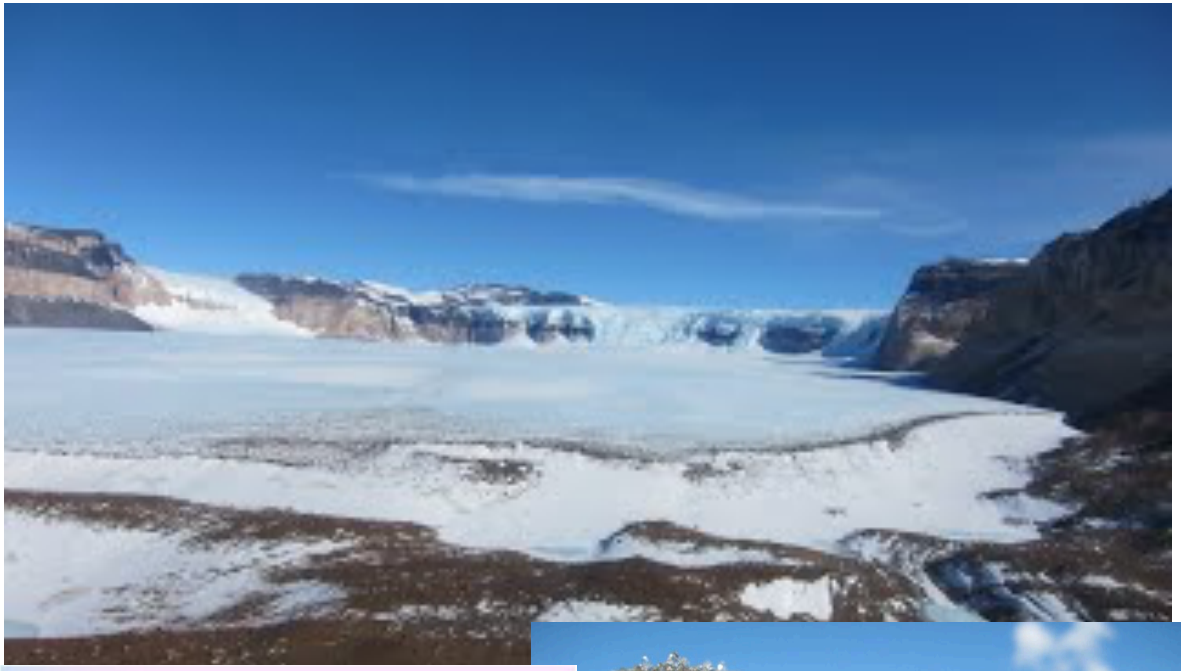
Metagenomics everywhere

Human gut, mouse, upper respiratory tract, skin, urogenital tract, and etc.



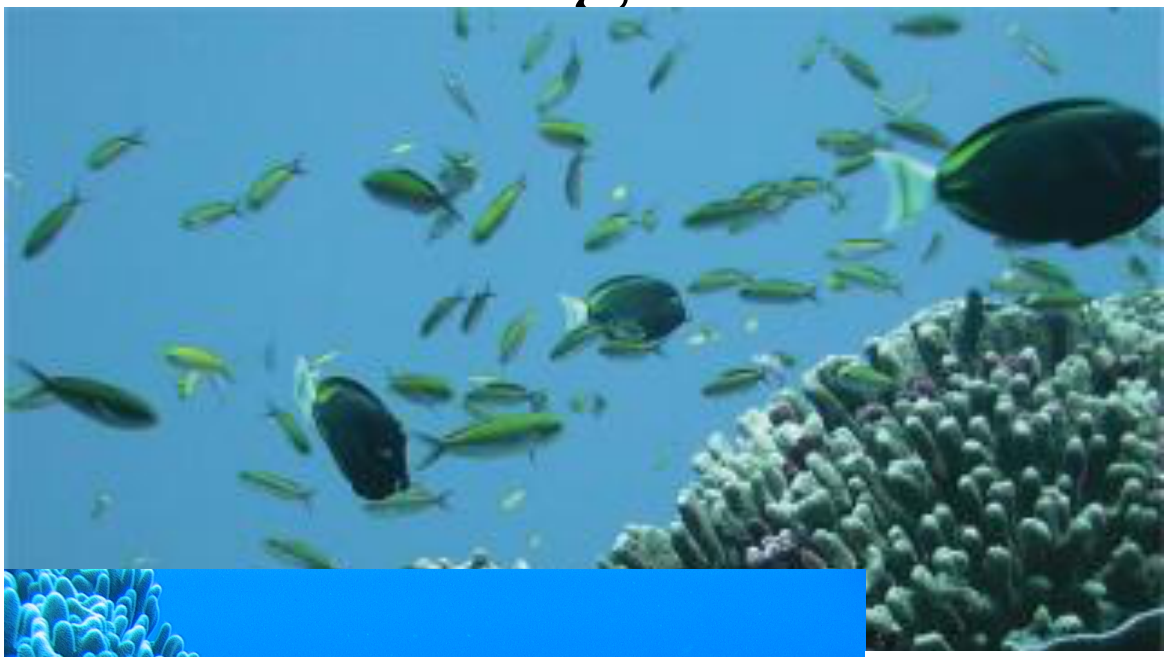


Metagenomics everywhere

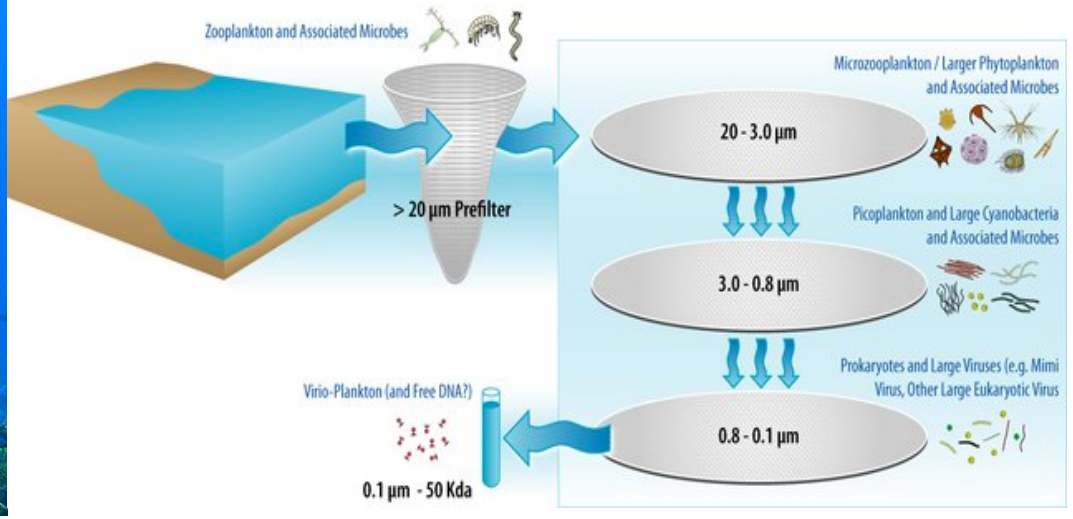


Soil, desert, ice layer, glacial, oil, and etc.

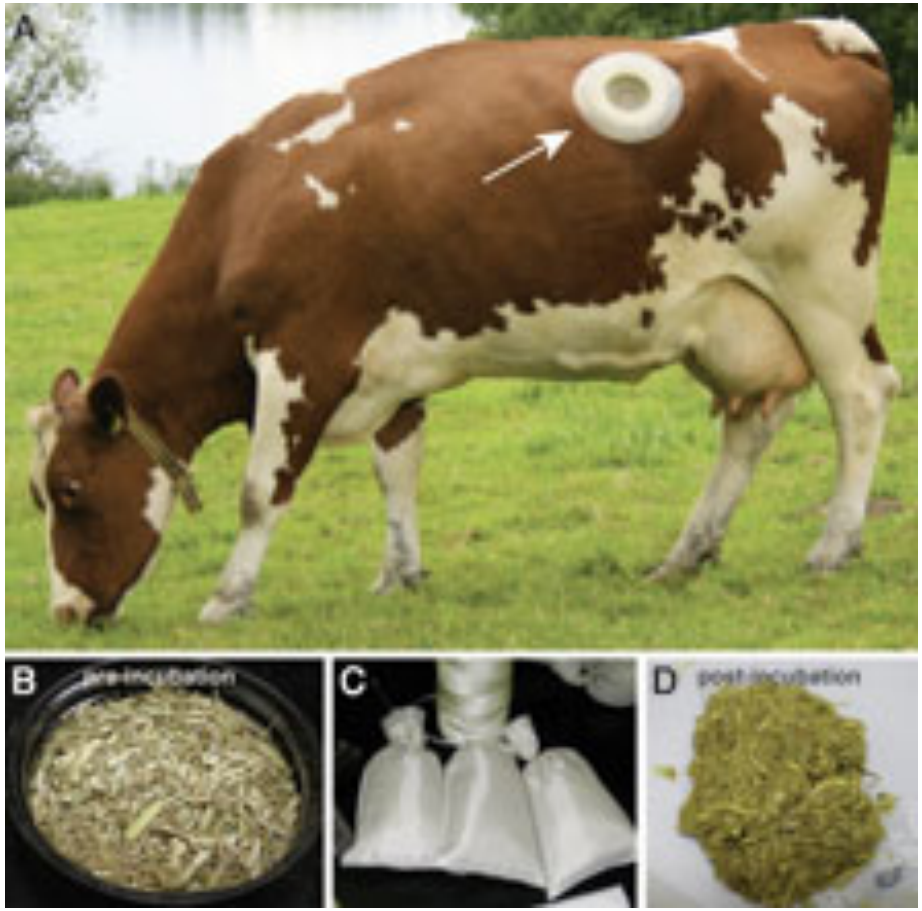
Metagenomics everywhere



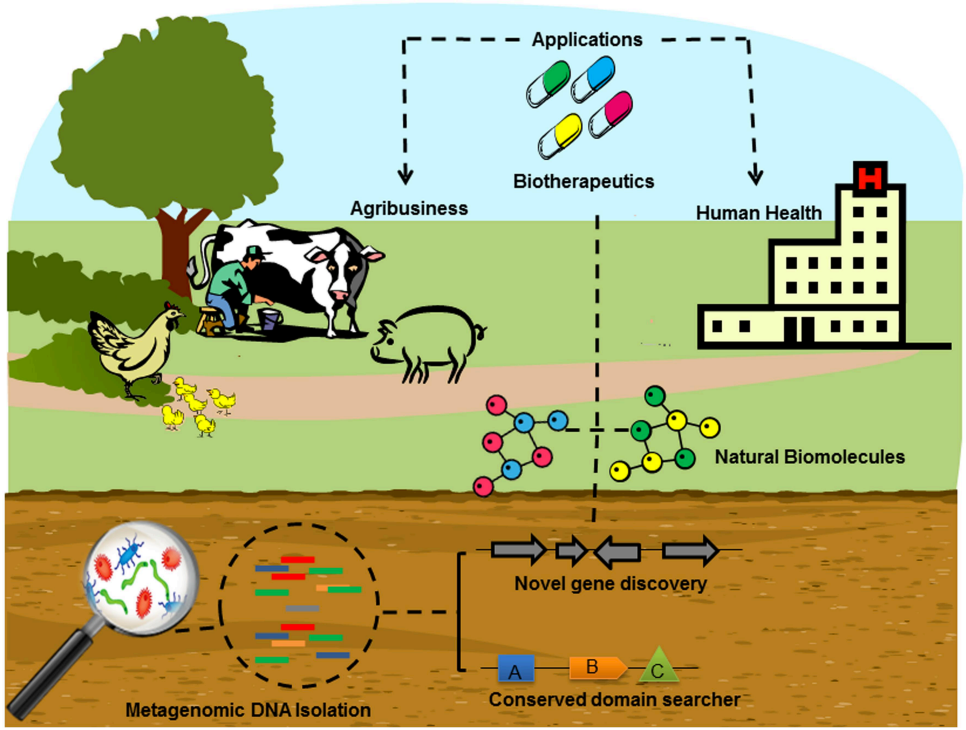
Sea surface, fish gut, Coral, Sponge, host-symbiosis, plankton, and etc.



Metagenomics everywhere



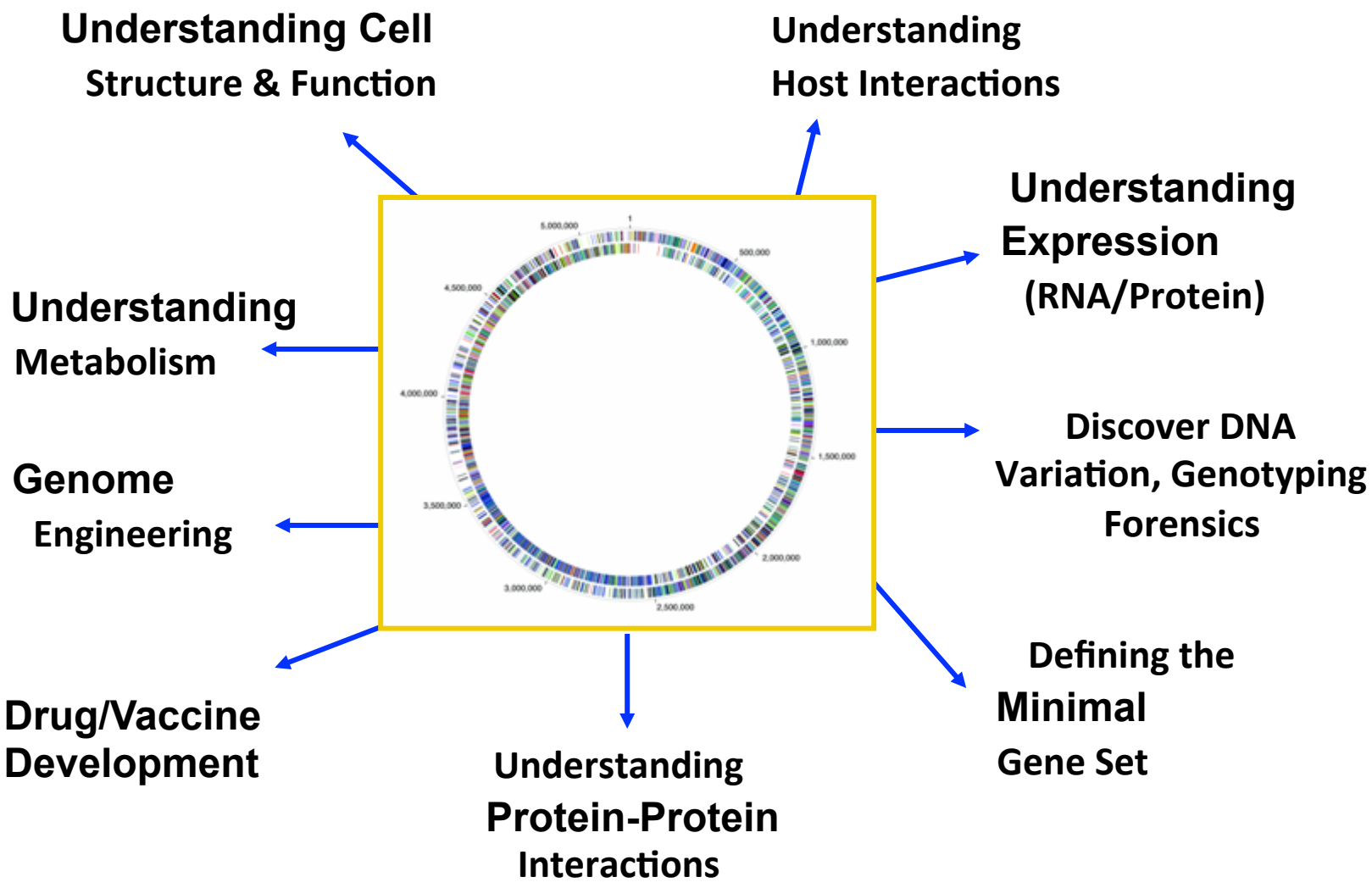
Animal gut, forest, grassland, and etc.



Why is Metagenomics Important?

- All reasons lead to more knowledge.
  - Organisms can be studied directly in their environments bypassing the need to isolate each species
  - There are significant advantages for viral metagenomics, because of difficulties cultivating the appropriate host
  - Genomic information has advanced research in a diverse array of fields, including forensic science and biomedical research

Why Do METAGENOMICS?

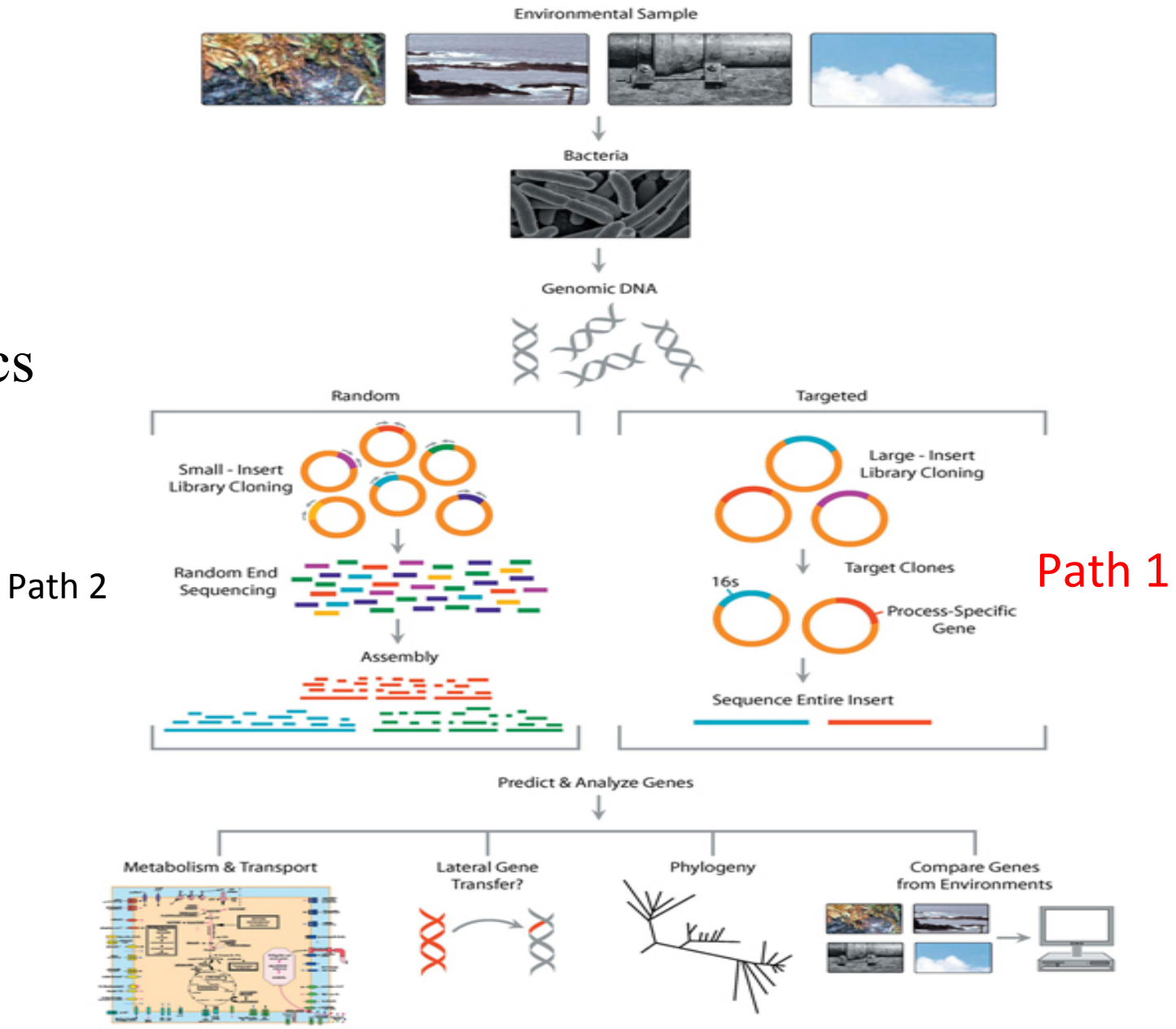


Jane Philips in University of Minnesota

Sampling in Metagenomics
some issues for library preparation

- Take a sample off of the environment
- Isolate and amplify DNA/mRNA
- Biggest challenge: environmental DNA extraction
  - Physical breaking cell wall, Beadbeater (shaking)
  - Chemical breaking cell wall, enzyme (digesting)

How to do analysis for metagenomics



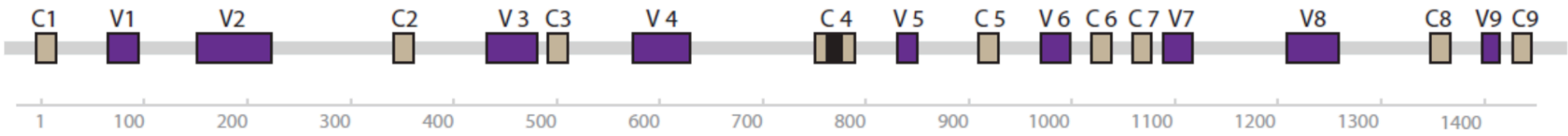
Path 2

Path 1

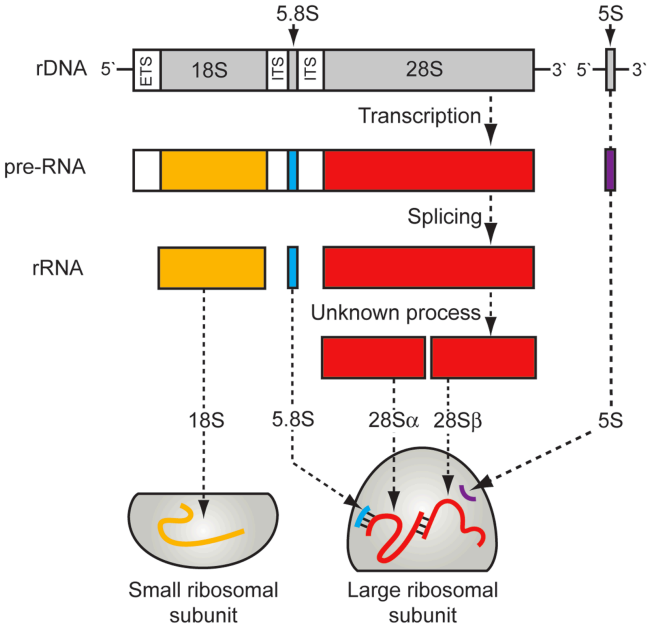
Path 1: PCR+sequencing

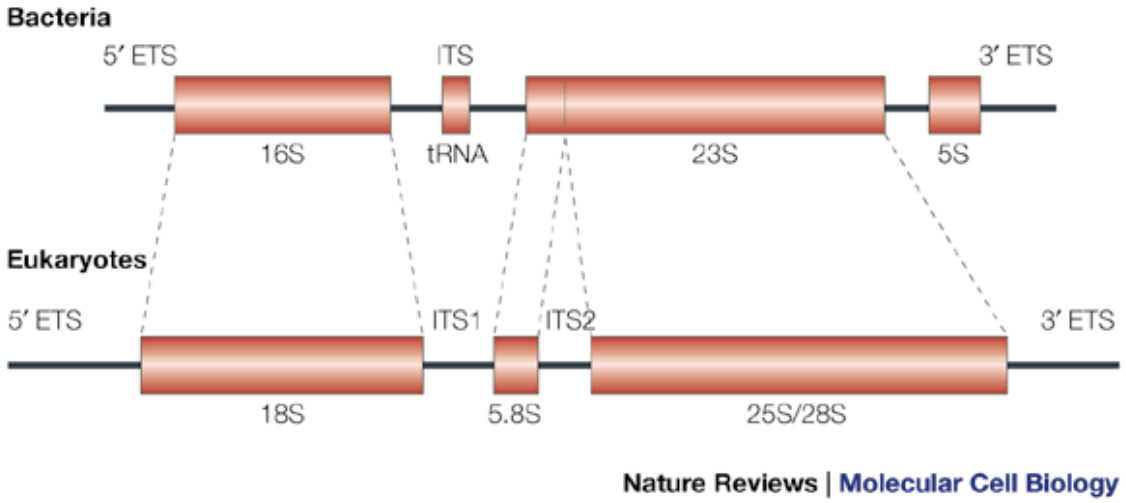
- **Goals**
  - Identify species (by identifying species-specific genes, such as 16S rRNA)
  - Richness & diversity
  - Comparison
- **Why 16S rRNA**
  - conserved within a species, and generally different between species.
  - All bacteria have 16s rRNA.
  - Database
- **Limitations**
  - PCR for the abundance
  - Only for known species
  - No gene annotation

16S & 18S rRNA



Schematic representation of the 16S rRNA gene. Location of variable (purple) and conserved (brown) regions in a canonical bacterial 16S rRNA. The black region is invariable in all bacteria.

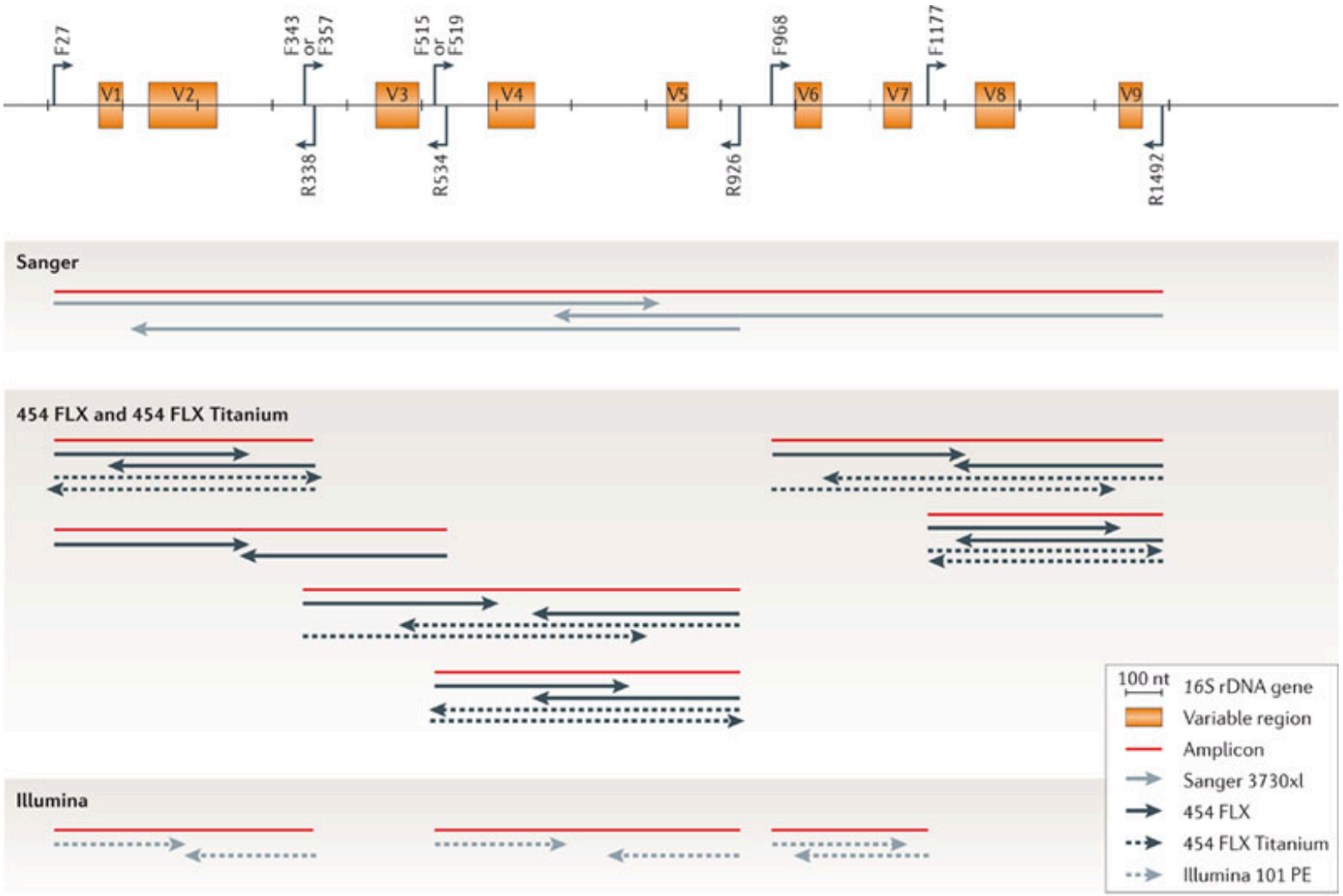




Sequencing driven metagenomics

- Sanger sequencing
- 454 pyrosequencing
- Illumina/Miseq sequencing
- Single cell sequencing: Multiple displacement amplification (MDA) of genomic DNA

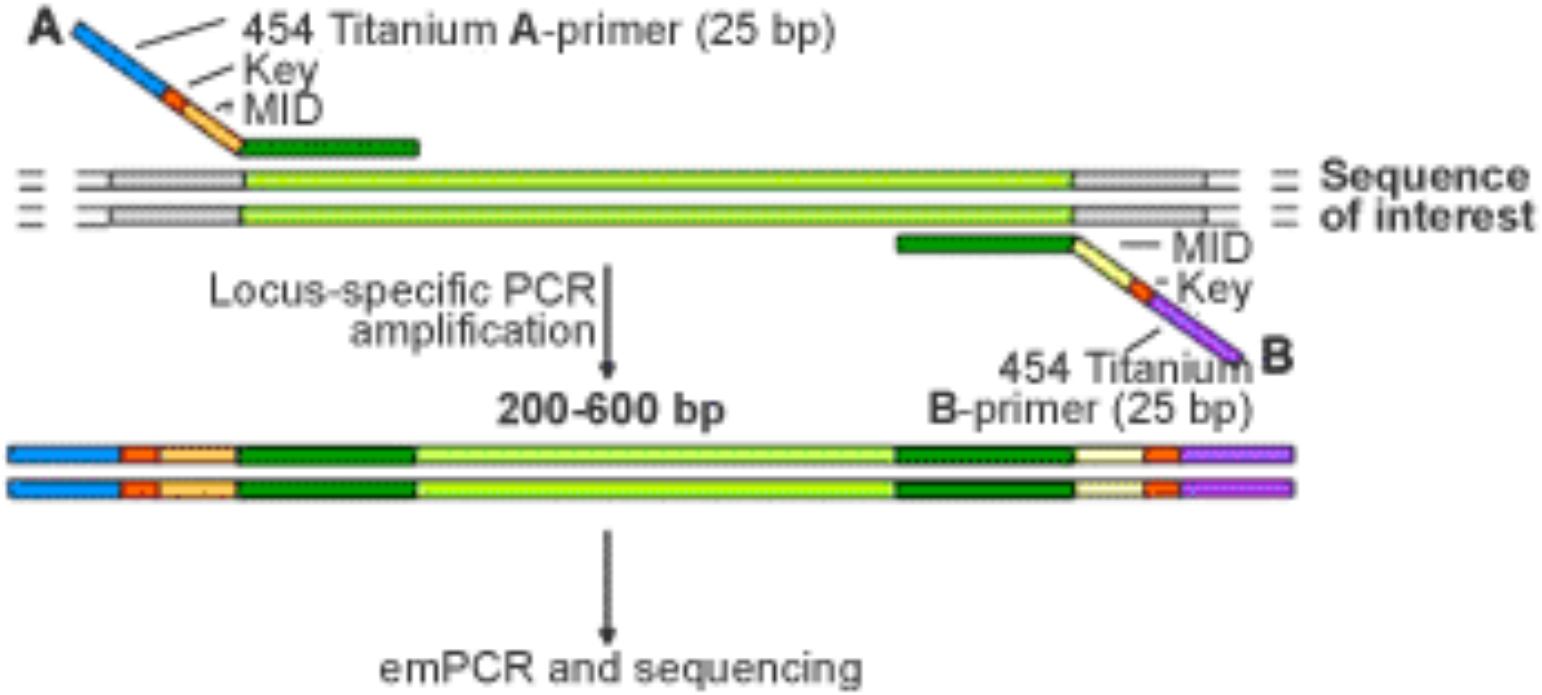
PCR for 16S rRNA amplicon



Justin et al., 2012

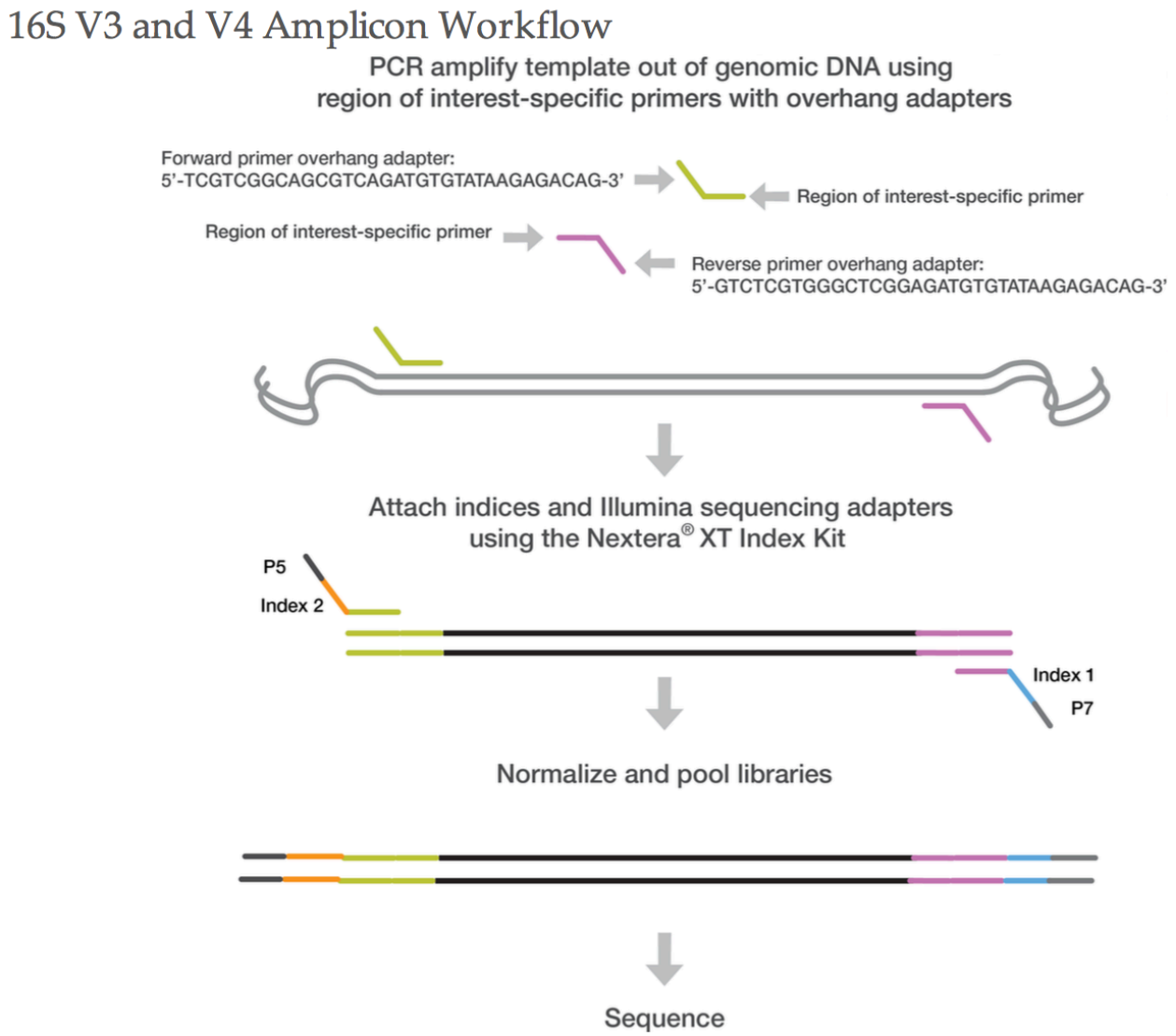
Nature Reviews | Genetics

Barcoded PCR & sequencing for amplicon in 454



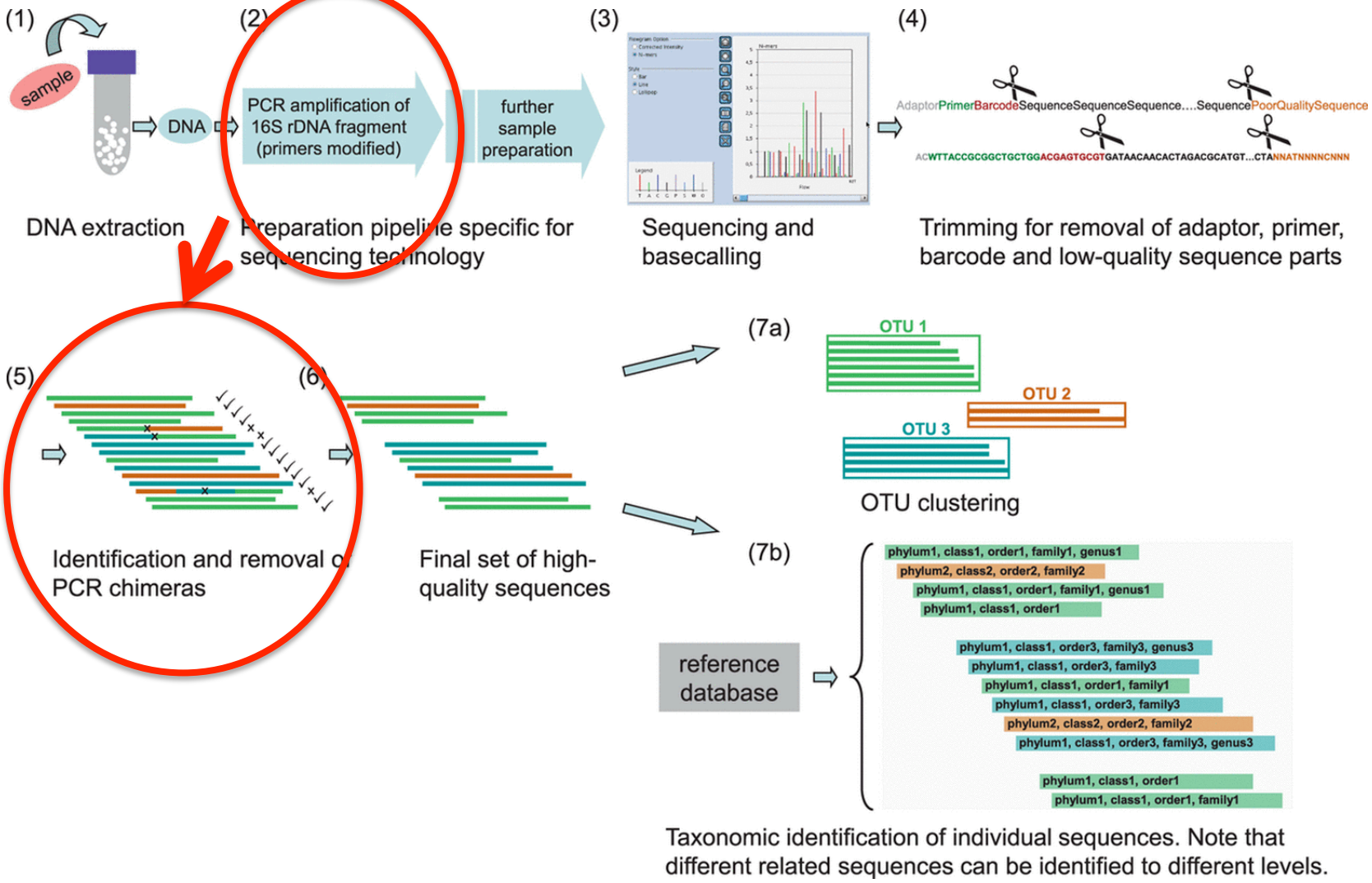
MID: Multiplex Identifiers, DNA barcode.
Key: sequencing key "TCAG" for bidirectional or unidirectional Sequencing

Barcoded PCR & sequencing for amplicon in Illumina



What is next after PCR？

Step1： Preprocessing



Yang et al., 2013

Fichot et al., 2013

What is chimera

aborted amplification

next cycle's "primer"

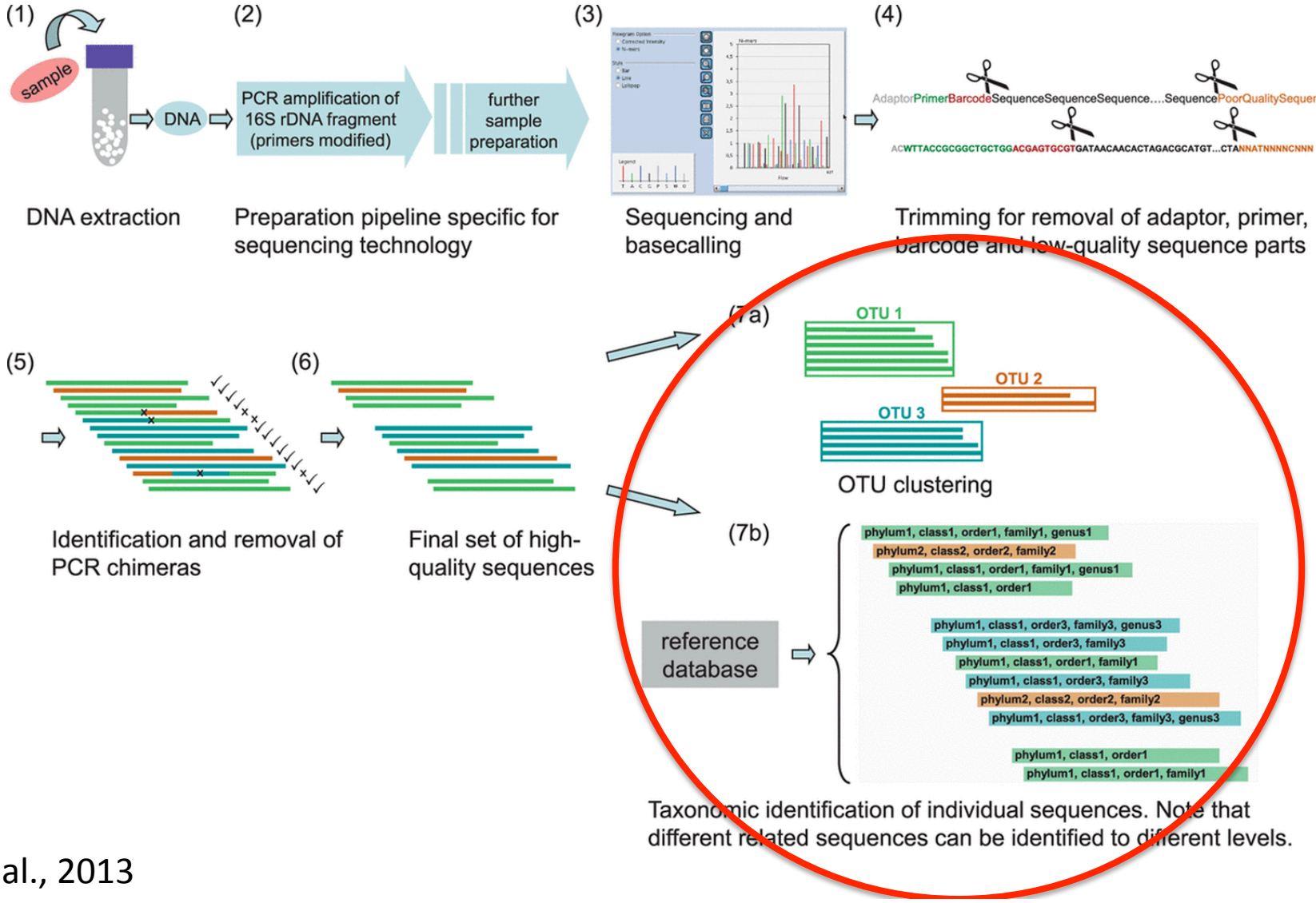
chimeric sequence

Chimera filtering tool

- ChimeraSlayer
  - Reference based.
  - Two ends' alignment and scoring.
- USEARCH (UCHIME)
  - Reference based (uchime_ref).
  - De novo chimera checking (uchime_denovo). 3-way alignment of a query sequence with two parent sequences.
- Perseus
- DECIPHER

What is next after PCR?

Step2： Data analysis



Yang et al., 2013

Data analysis

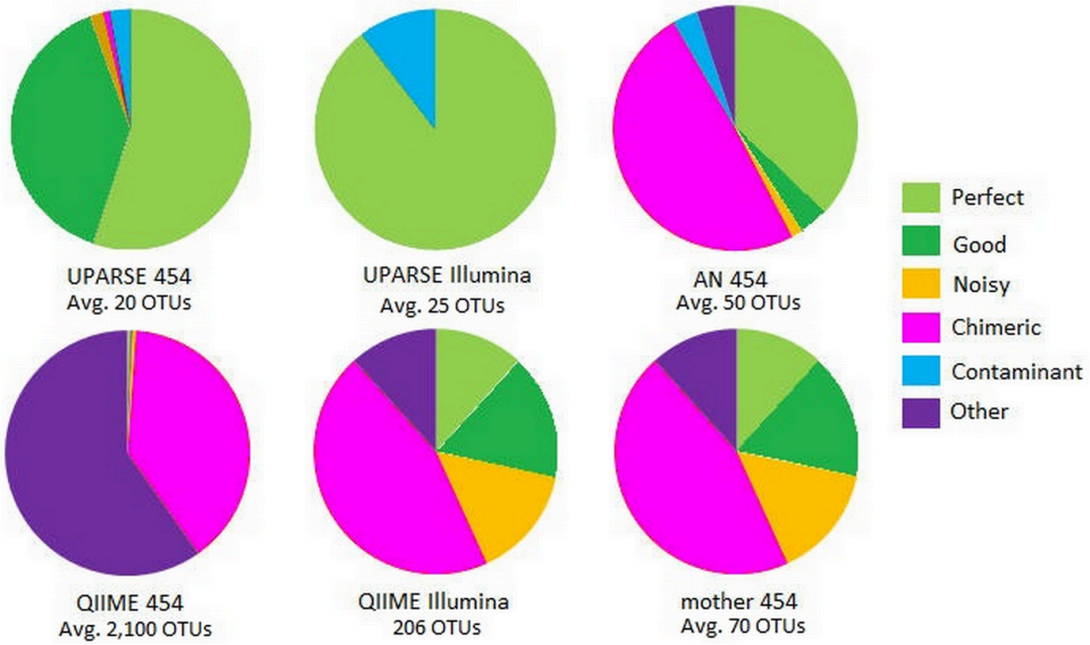
1. OTU table
2. Biodiversity
3. Phylogenetics
4. Taxonomy
5. PCA, etc.

Terms you should know

- **OTU**: operational taxonomic unit. Taxonomic level of sampling selected by the user to be used in a study, such as individuals, populations, species, genera, or bacterial strains.
- **RDP**: Ribosomal Database Project. RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community.

OTU clustering

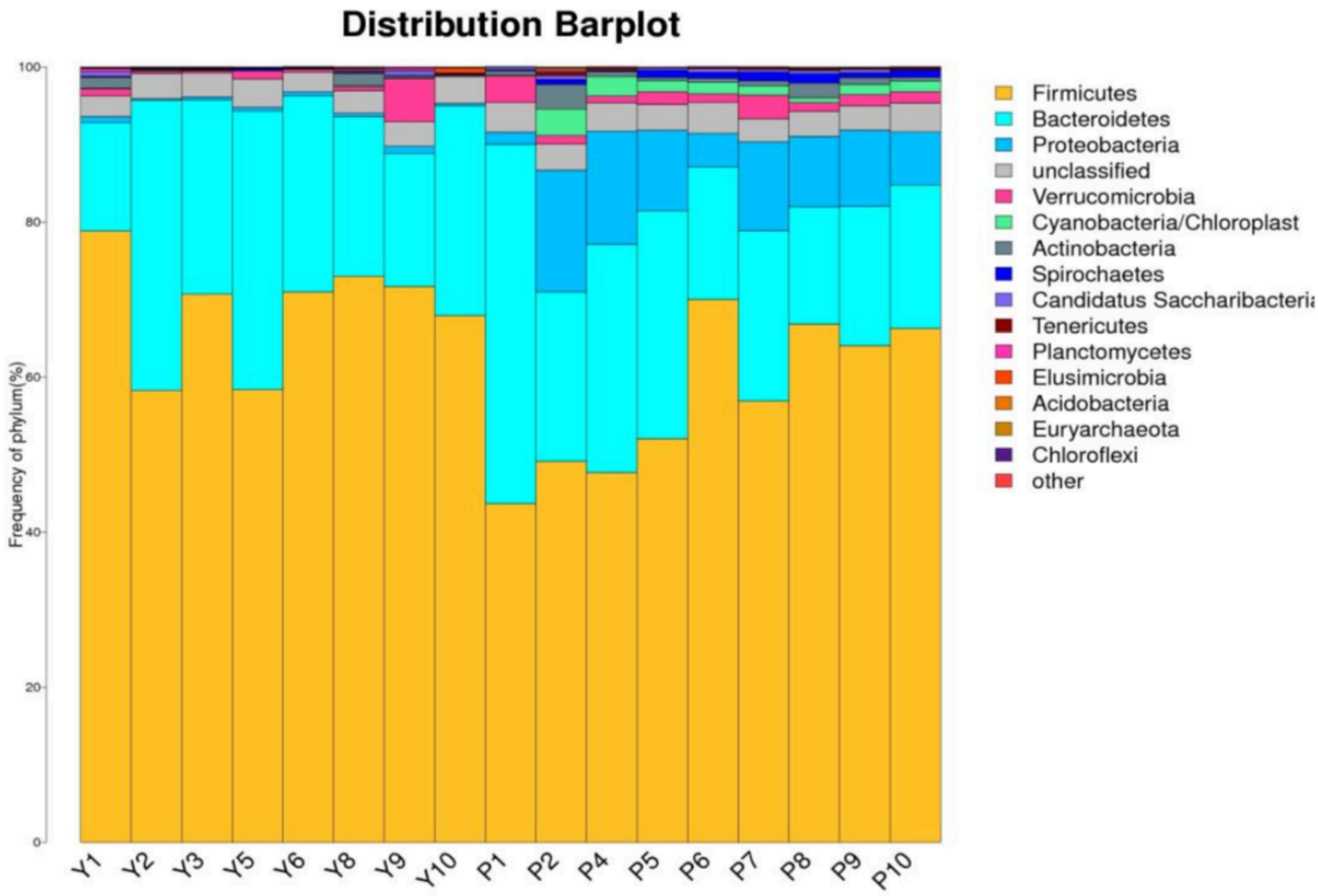
- **QIIME**: Next-generation reads are clustered into OTUs. This requires quality filtering, dereplication, discarding singletons (optional), and finally clustering into OTUs, typically at a 97% identity threshold.
- **Mothur**: Alignment based OTUs. align.seqs().



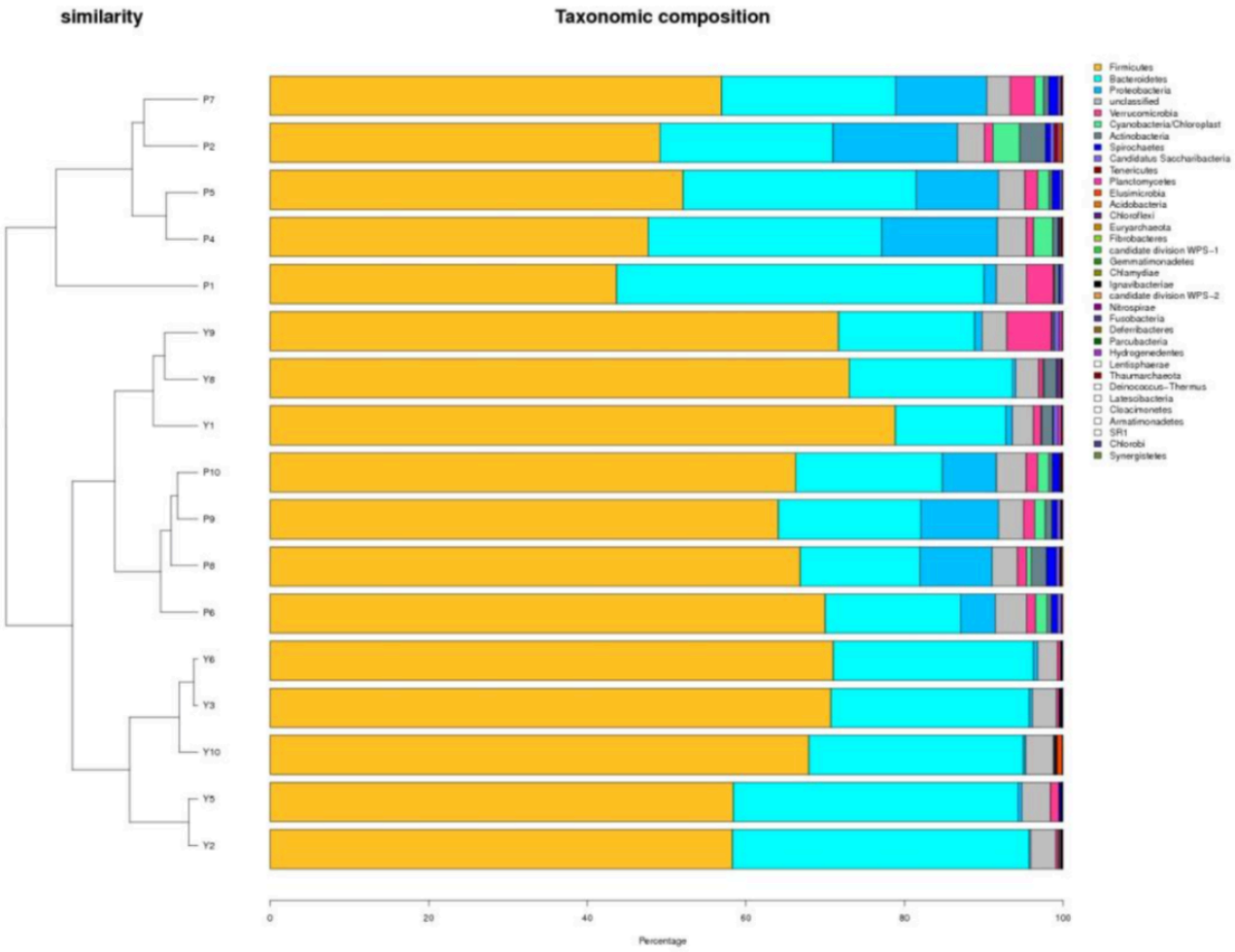
OTU identity assessed by USEARCH

Accuracy measure	Summary
Sequence quality Are OTUs accurate reconstructions of biological sequences?	Most USEARCH OTUs are >=99% identical to a biological sequence. Most QIIME, mothur and AmpliconNoise OTUs are >3% diverged from a biological sequence. Roughly half are chimeric.
Diversity Does the number of OTUs correspond to the number of species?	USEARCH generated from 0.8 to 1.0 OTUs per detectable species. Mothur and AmpliconNoised produced 2.3x to 6.7x more OTUs than species. QIIME produced thousands of OTUs, far more than the number of species.

Distribution of Phylum for yak and pika

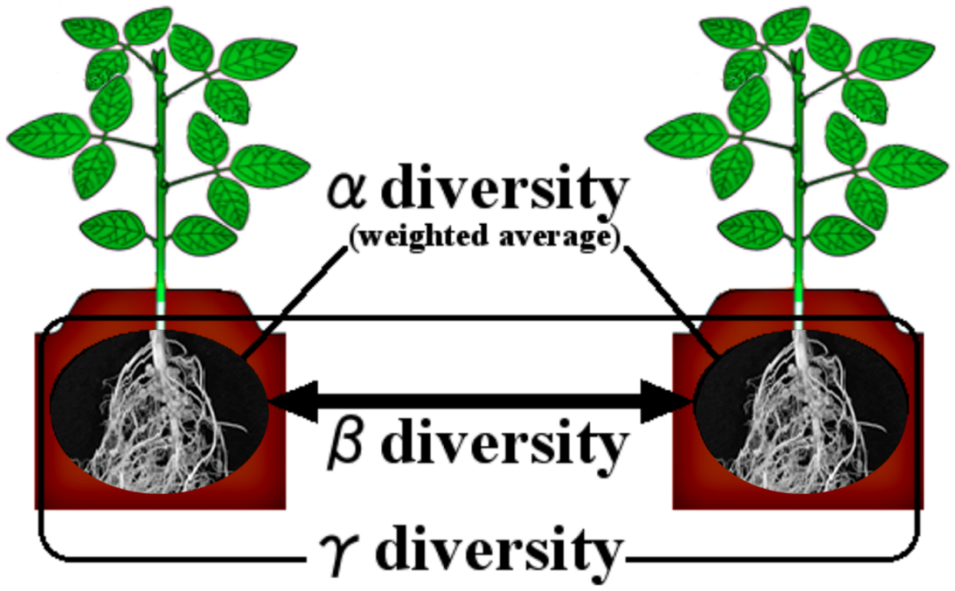


Clustering results for yak and pika



Biodiversity

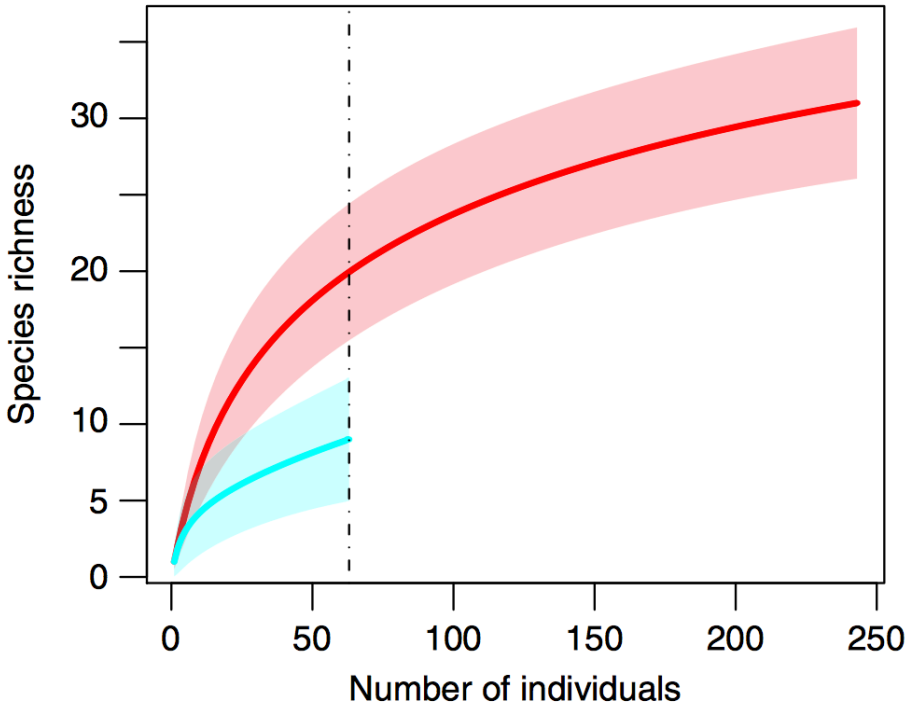
- **Alpha & Beta biodiversity**: the total species diversity in a landscape (gamma diversity) is determined by two different things, the mean species diversity in sites or habitats at a more local scale (alpha diversity) and the differentiation among those habitats (beta diversity).

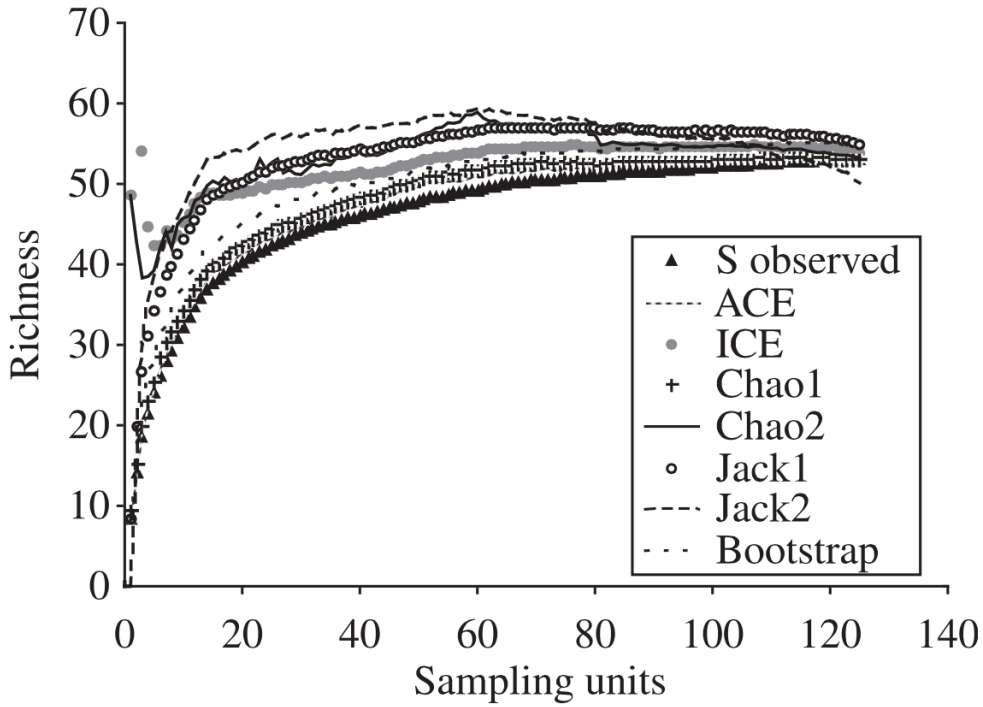


Yuichi et al., 2014

Richness and rarefaction curve

- Richness *R* simply quantifies how many different types the dataset of interest contains.
- Rarefaction curves are used to determine whether sampling depth was sufficient to accurately characterize the bacterial community being studied





Higuti et al., 2009

Database available

- **RDP** 16S database
- **SILVA** rRNA database
- **Greengenes** 16S database
- **EzTaxon-e** 16S database
- **UNITE** ITS database

Tools

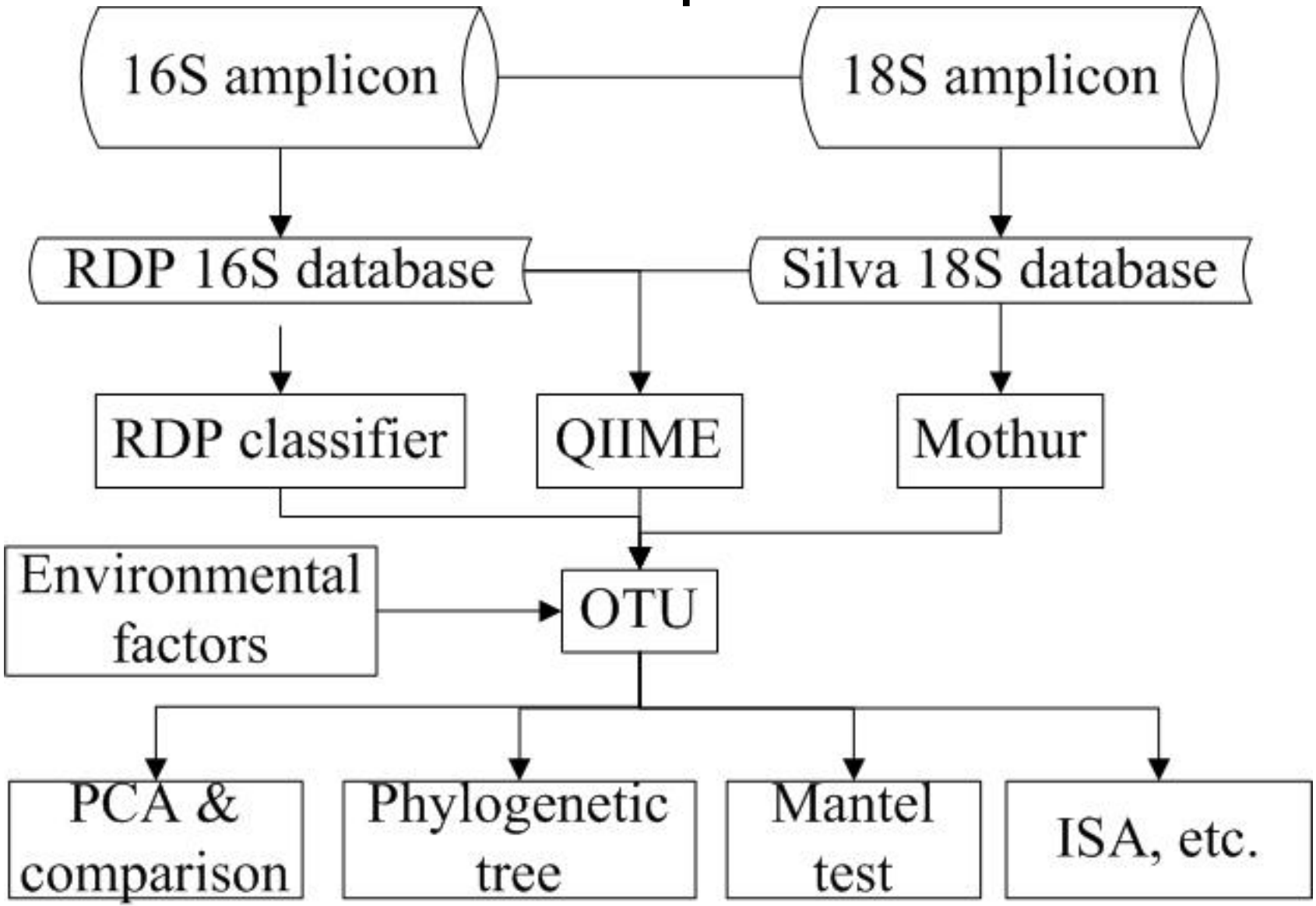
- **QIIME** http://qiime.org/install/virtual_box.html
- **Mothur** http://www.mothur.org/
- **RDP classifier** https://rdp.cme.msu.edu/classifier/
- **MEGAN** http://ab.inf.uni-tuebingen.de/software/megan/
- **MG-RAST** http://metagenomics.anl.gov/
- **CARMA** http://www.cebitec.uni-bielefeld.de/brf/carma/carma.html
- **IMG/M** http://img.jgi.doe.gov/cgi-bin/m/main.cgi

QIIME pipeline

- ✓ Demultiplexing and quality filtering sequences
- ✓ Chimera filtering.
- ✓ OTU picking
  - Pick OTUs based on sequence similarity within the reads
  - Pick a representative sequence for each OTU
  - Assign taxonomy to OTU representative sequences
  - Align OTU representative sequences
  - Make the OTU table

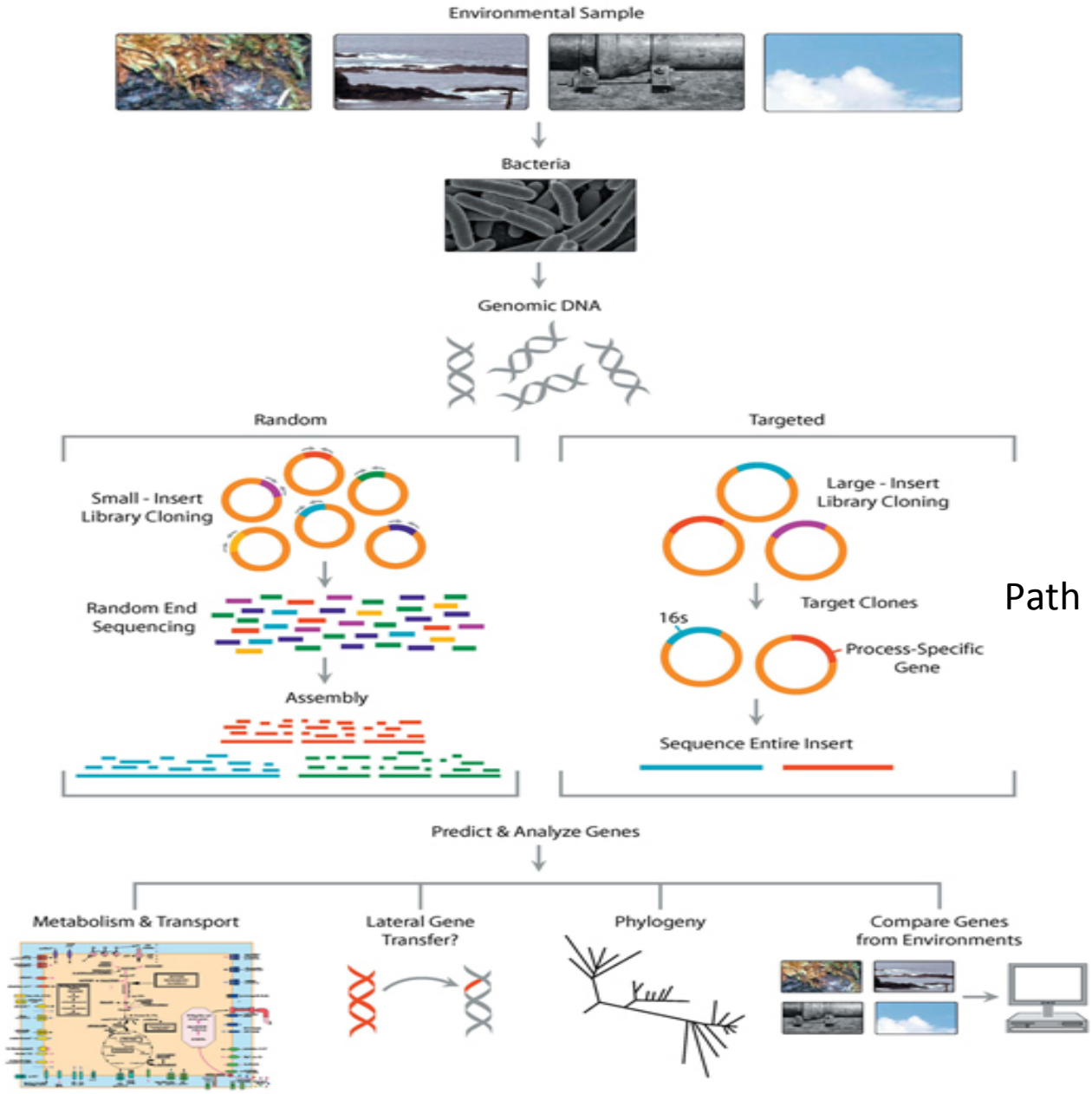
- ✓ Run diversity analyses for alpha and beta diversity

Amplicon data analysis pipeline example



How to do metagenomics

Path 2

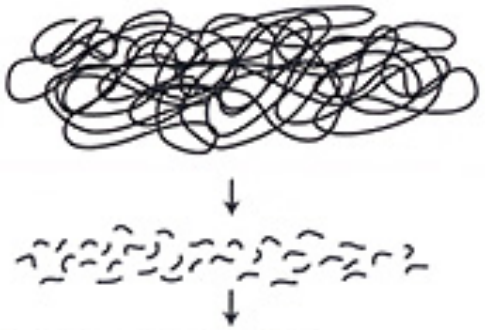


Path 1

Whole Genome Shotgun Sequencing for Metagenomics

One genome

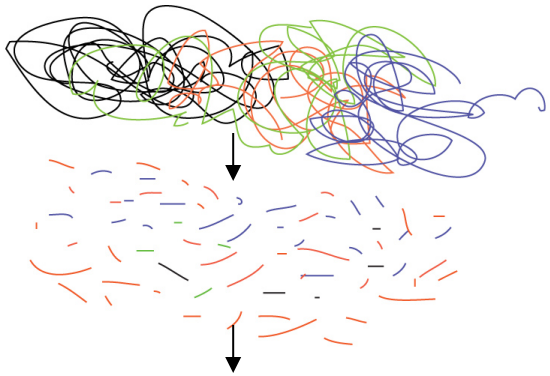
Random genome fragmentation



Genome assembly using overlaps

...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG..
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG..

Multiple genomes



Random genomes fragmentation

...ACACATATACATACAGAGATAGCCCAGATG
AGCCCAGATGGCGCTGCTGCTGGGCGCG.....
...ACATACATACAGAGATAAAAGATG
AAAAGATGCCAGATGGCGCTGCTGCTGGGCGCG.....
...ACACCATACAGAGATAGGGGTGGG
GGGTGTGGAGCCCAGATGGCGCTGCTGCTGG.....

Genomes assembly using overlaps

Shotgun Sequencing for Metagenomics

- **Goals**
  - Key genes for nutrition metabolism, etc.
  - Novel genes
  - Novel species
- **Difficulties**
  - Genome assembly
  - Gene annotation
- **Solutions**
  - PE
- **Limitations**
  - Overestimation
  - Species with low abundance
  - Unknown species
  - Cost

Metagenome assembly

- **Velvet** http://www.ebi.ac.uk/~zerbino/velvet/
- **SOAP denovo** http://soap.genomics.org.cn/soapdenovo.html
- **Celera** http://www.cbcb.umd.edu/research/assembly.shtml#software
- **Metasim** http://ab.inf.uni-tuebingen.de/software/metasim/welcome.html#Download
- **Euler** http://nbcr.sdsc.edu/euler/JAZZ

Metagenome annotation tools

- **MG-RAST**
- **IMG/M**
- **CAMERA**
- **MEGAN**
- **RAMMCAP** http://weizhong-lab.ucsd.edu/rammcap/cgi-bin/rammcap.cgi
- **METAGENE** http://metagene.cb.k.u-tokyo.ac.jp/

MG-RAST

metagenomics analysis server

# of metagenomes	210,768
# base pairs	85.24 Tbp
# of sequences	678.81 billion
# of public metagenomes	29,927

MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

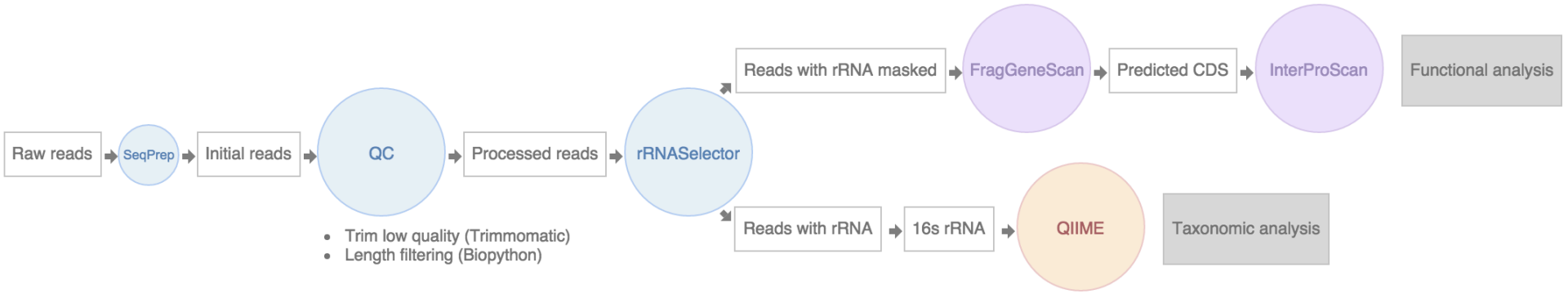
PUBLIC METAGENOMES

group by project

Current table counts

metagenomes	projects	biomes	features	materials	altitudes	depths	locations	ph's	countries	temperatures	pi's
29927	1093	185	258	176	5	169	429	1243	298	115	8

EBI Metagenomics



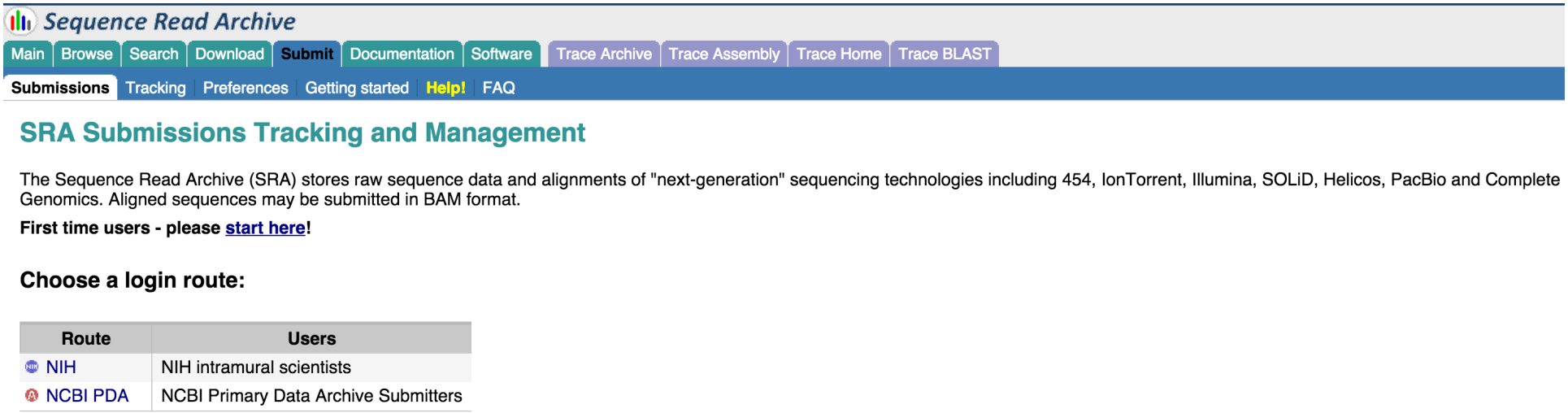
The EBI Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data that aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample.

Public	5578	runs
	4876	samples
	136	projects

Private	2097	runs
	2035	samples
	91	projects

NCBI Sequence Read Archive

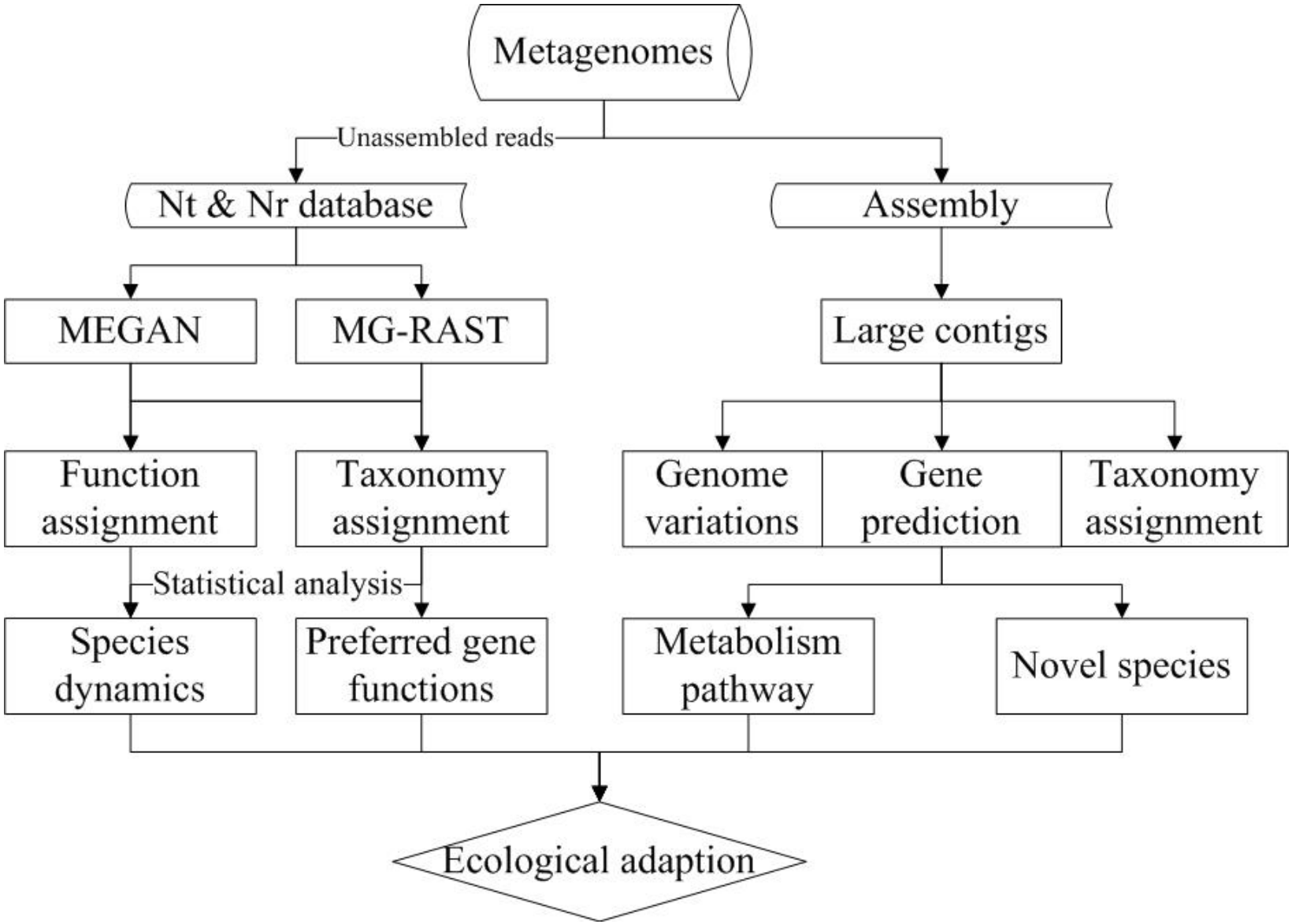
The Sequence Read Archive accepts all Next-Generation sequencing data.



Submitting assembled metagenomic contigs to WGS

Contigs that have been assembled from raw reads can be submitted as a WGS project.

Metagenome data analysis pipeline example



Case study (1)

- Metagenomics in red palm weevil.
- Analysis pipeline.
- Species distribution.
- Genome assembly.
- Nutrition metabolism.

Shangang et al., 2013

Metagenomics in red palm weevil

RPW is living inside the trunk of date palm.

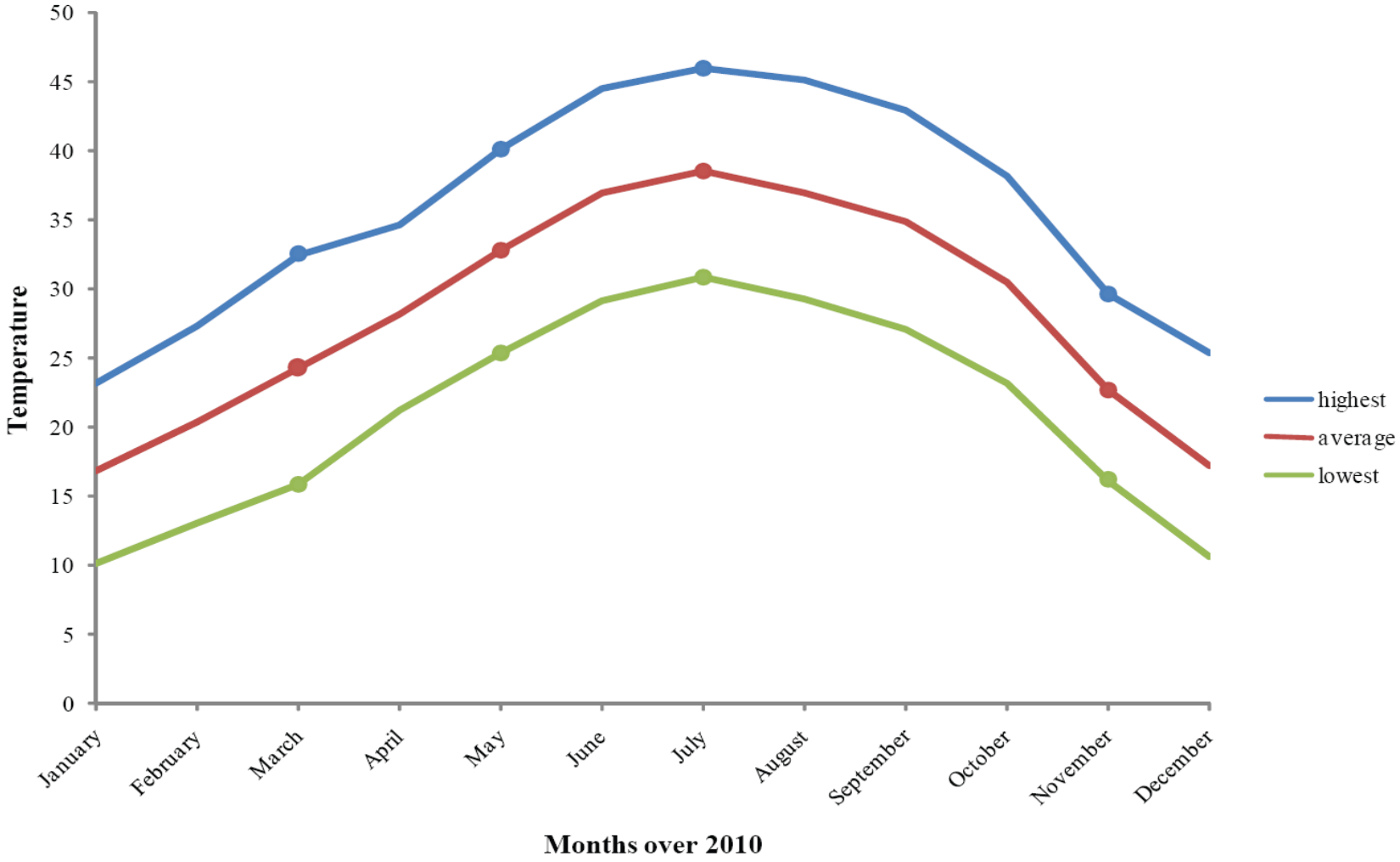
Egg, larva, pupa, and adult.

World distribution of RPW

From Red Palm Weevil Home

Gut of RPW

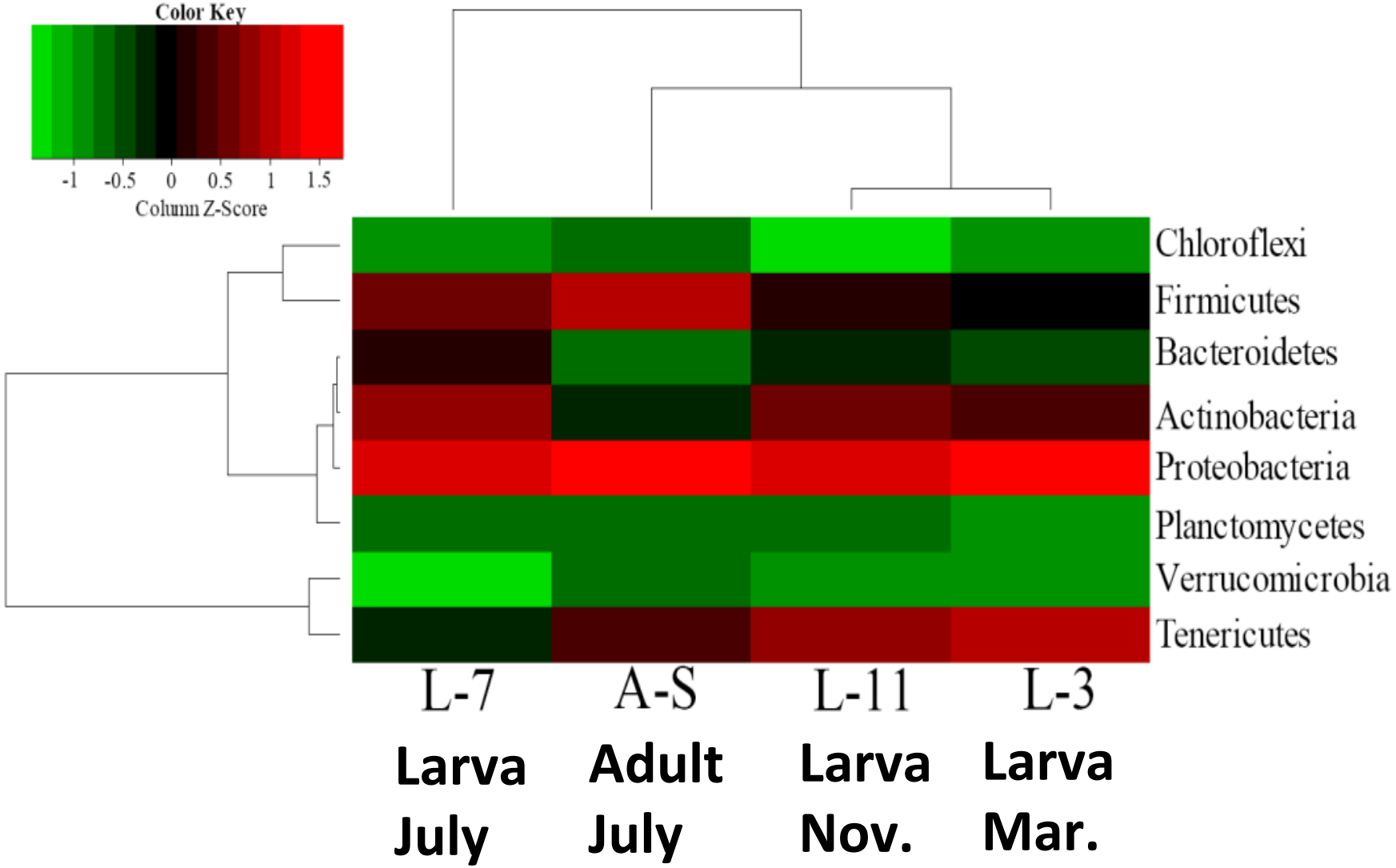
Temperature in Saudi Arabia



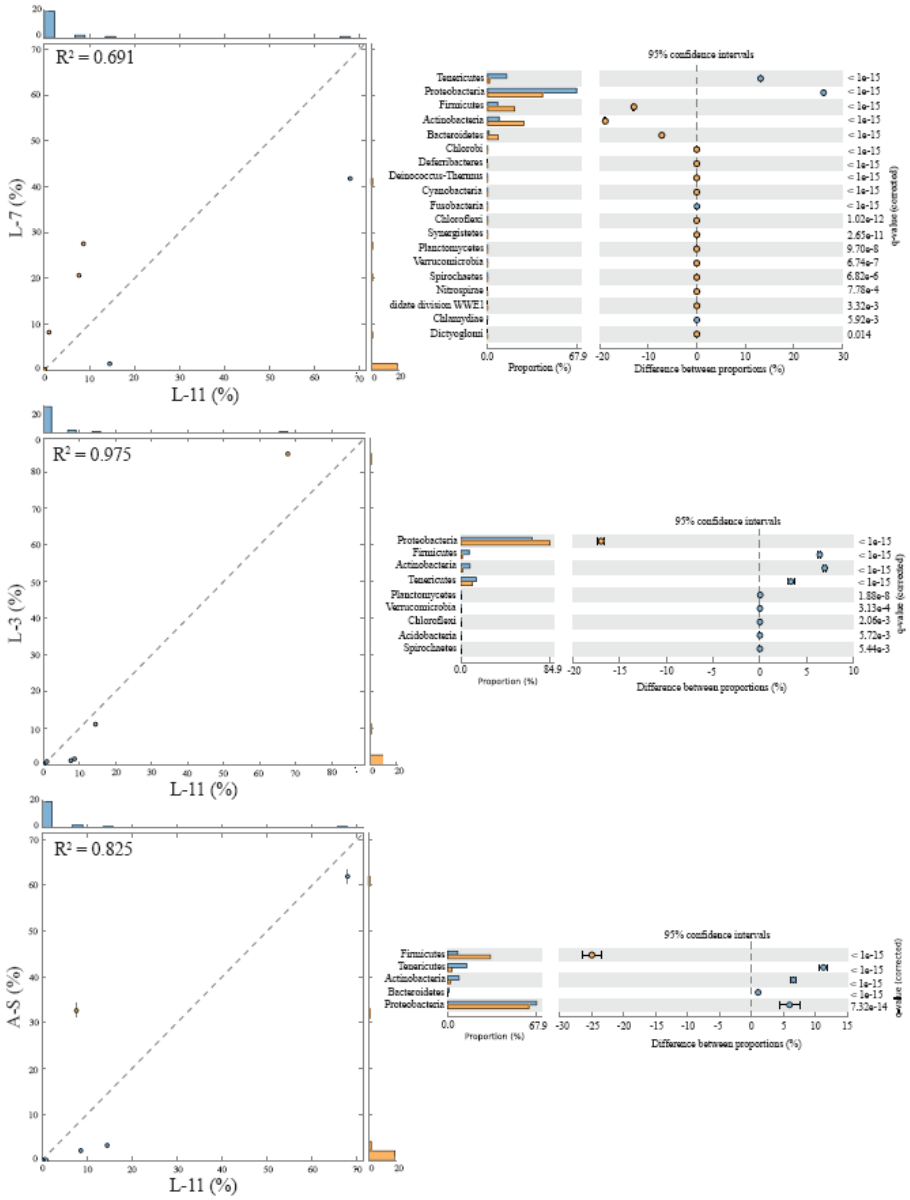
Summary of micro-organism

	Larva Mar.	**Larva July**	**Larva Nov.**	**Adult July**
	L-3	L-7	L-11	A-S
Bacteria	**72.69%**	**97.06%**	**94.86%**	**15.18%**
Archaea	0.00%	0.04%	0.05%	0.00%
Eukaryota	27.11%	2.89%	5.08%	84.52%
Viruses	0.20%	0.02%	0.01%	0.30%

Clustering in phylum level



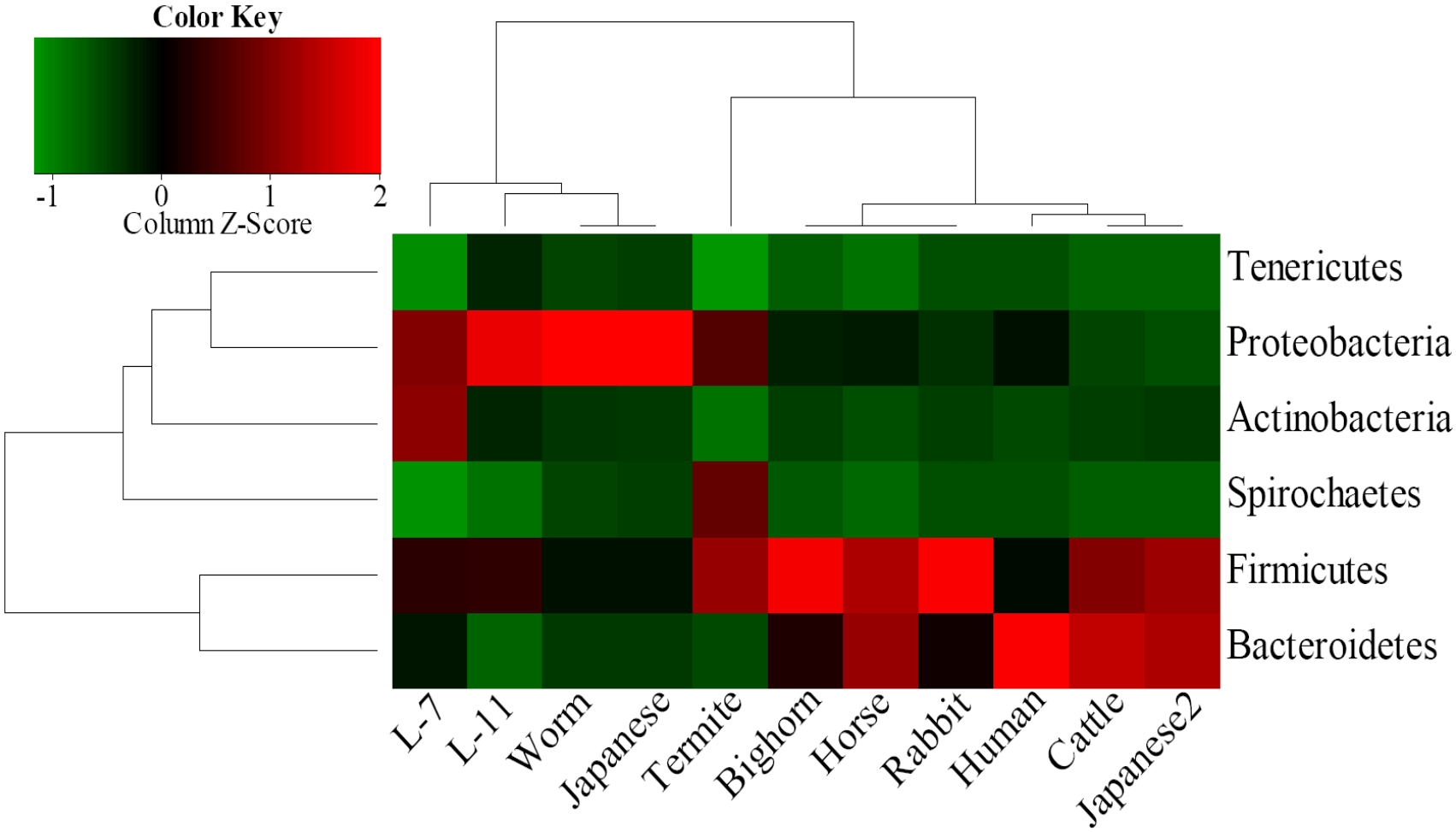
Correlation



Top 10 species

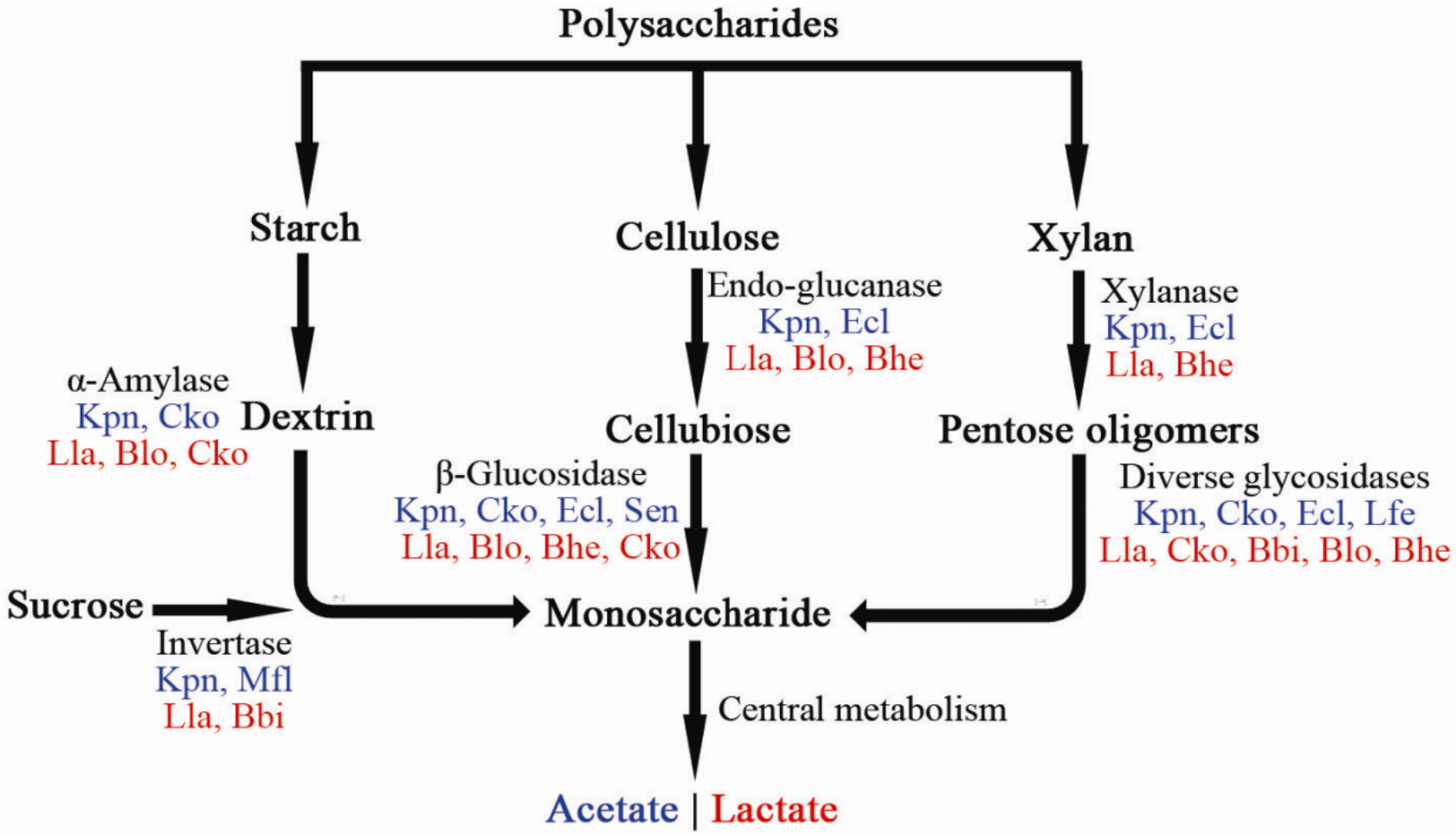
L-11		L-7		L-3		A-S	
Species	%	Species	%	Species	%	Species	%
Klebsiella pneumoniae	0.15	***Lactococcus lactis***	0.26	***Klebsiella pneumoniae***	0.18	***Lactococcus lactis***	0.3
Mesoplasma florum	0.1	***Desulfovibrio desulfuricans***	0.11	*Citrobacter koseri*	0.15	*Acinetobacter baumannii*	0.1
Citrobacter koseri	0.08	***Bifidobacterium bifidum***	0.04	*Mesoplasma florum*	0.1	*Acidovorax avenae*	0.07
Enterobacter cloacae	0.06	***Bifidobacterium longum***	0.03	*Enterobacter cloacae*	0.07	*Buchnera aphidicola*	0.05
Lactobacillus fermentum	0.04	*Rhodobacter capsulatus*	0.03	*Salmonella enterica*	0.05	*Pseudomonas mendocina*	0.04
Mycoplasma mycoides	0.04	*Bacteroides helcogenes*	0.02	*Mycoplasma mycoides*	0.04	*Pseudomonas aeruginosa*	0.03
Salmonella enterica	0.03	*Paracoccus denitrificans*	0.02	*Pantoea sp. At-9b*	0.03	*Enterobacter cloacae*	0.03
Lactobacillus salivarius	0.03	***Bifidobacterium adolescentis***	0.02	*Citrobacter rodentium*	0.03	*Citrobacter koseri*	0.03
Pantoea sp. At-9b	0.02	*Bacteroides xylanisolvens*	0.02	*Escherichia coli*	0.03	*Herbaspirillum seropedicae*	0.03
Escherichia coli	0.02	*Citrobacter koseri*	0.02	*Klebsiella oxytoca*	0.02	*Enterococcus faecalis*	0.02

Compared to others



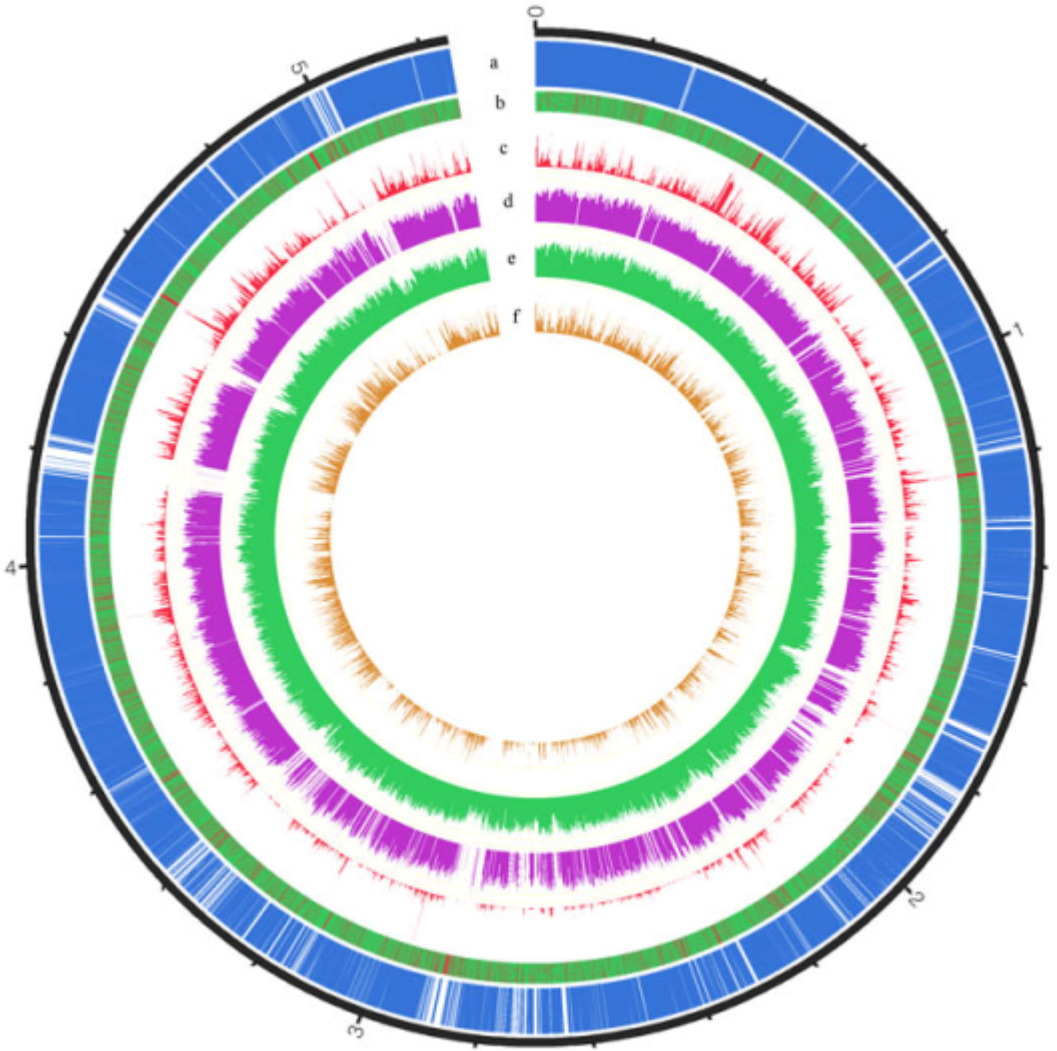
Nutrition metabolism

CAZY: Carbohydrate-Active enZYmes Database



Whole genome assembly

Circle a, the regions with or without mapped reads are colored blue or white, respectively.
Circle b, genes or intergenic regions are colored green or red, respectively.
Circles c, d, e, and f: the distribution of mapped read, GC content of the mapped reads, GC content of the references, and SNPs distribution.
The window size in Circles c, d, e, and f is1kb.



Klebsiella pneumoniae 342

Case study (2)

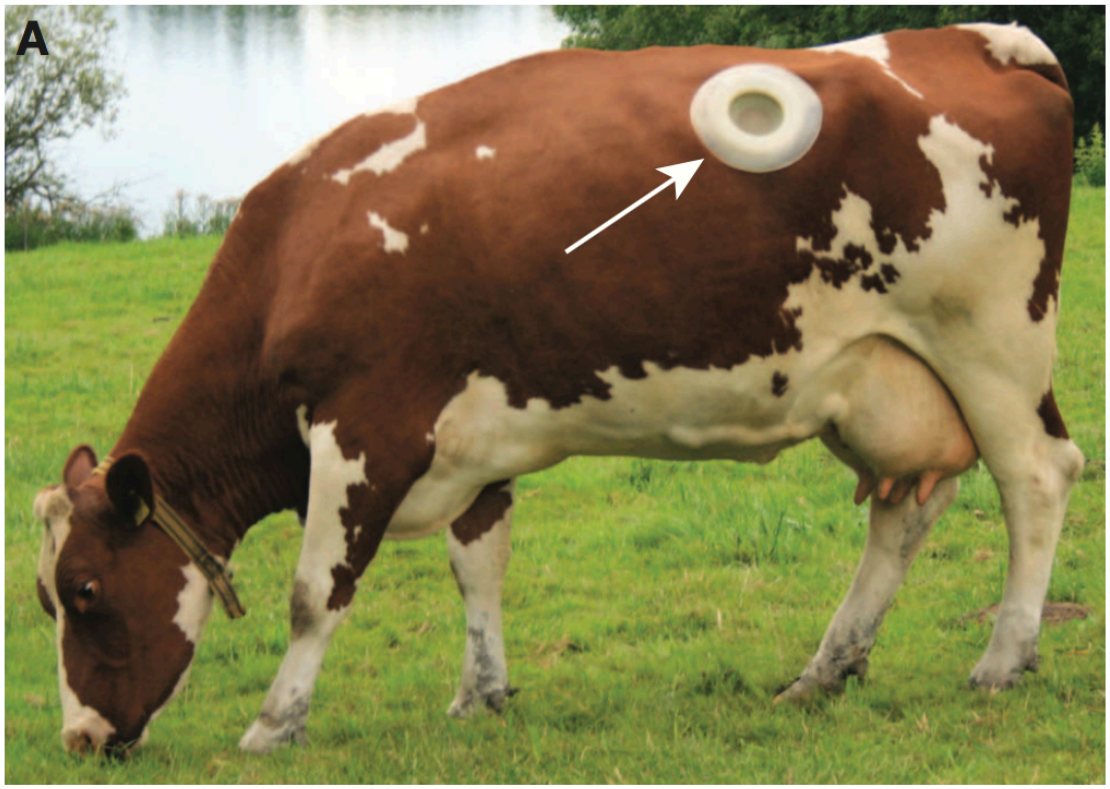


Fig. 1. (**A**) A surgically created fistula (arrow) sealed with a flexible cannula was used to study the degradation of switchgrass within the rumen. (**B**) Switchgrass before rumen incubation. (**C**) Nylon bags filled with switchgrass before insertion into the rumen. (**D**) Switchgrass after 72 hours of rumen incubation.

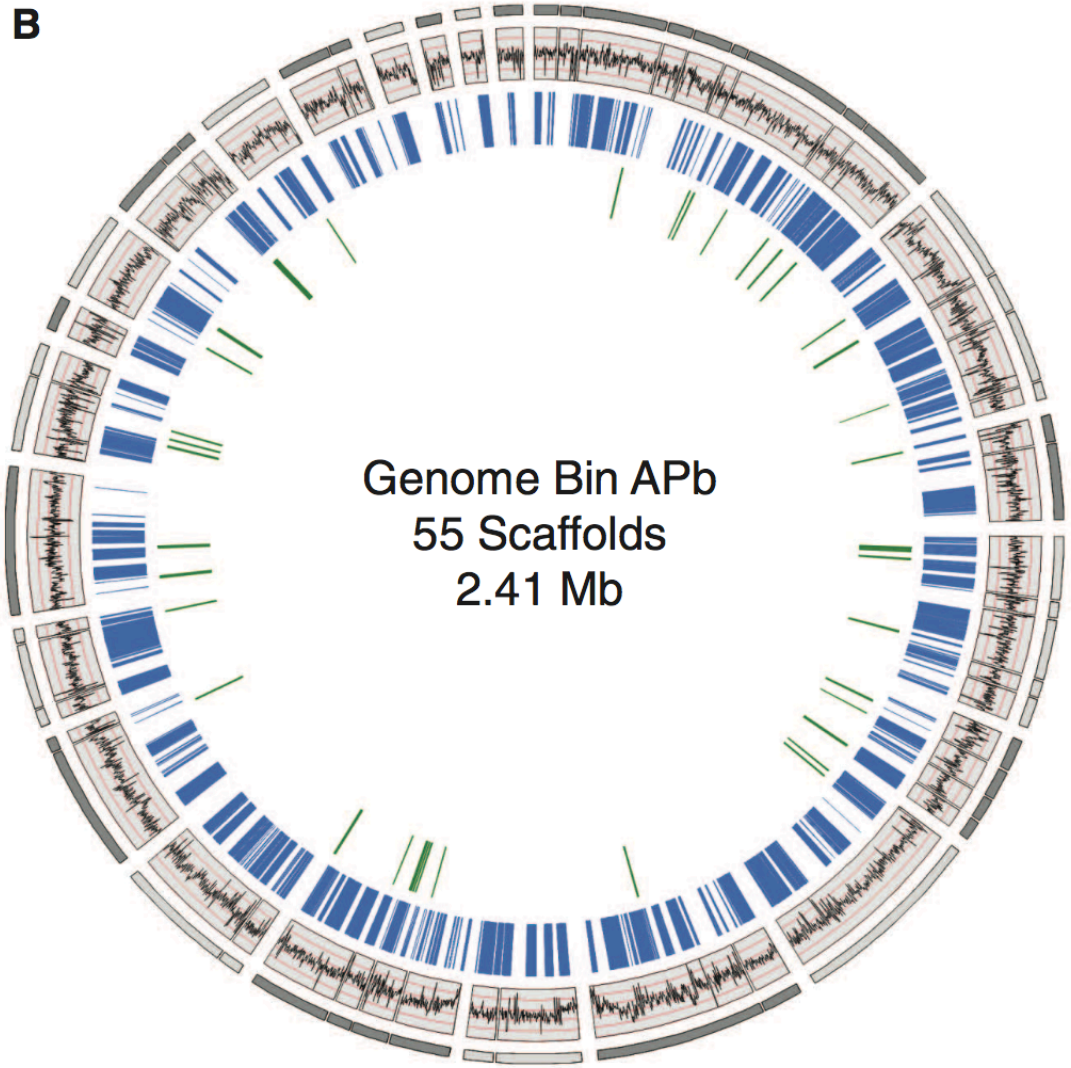
Matthias et al., 2011

Draft genomes

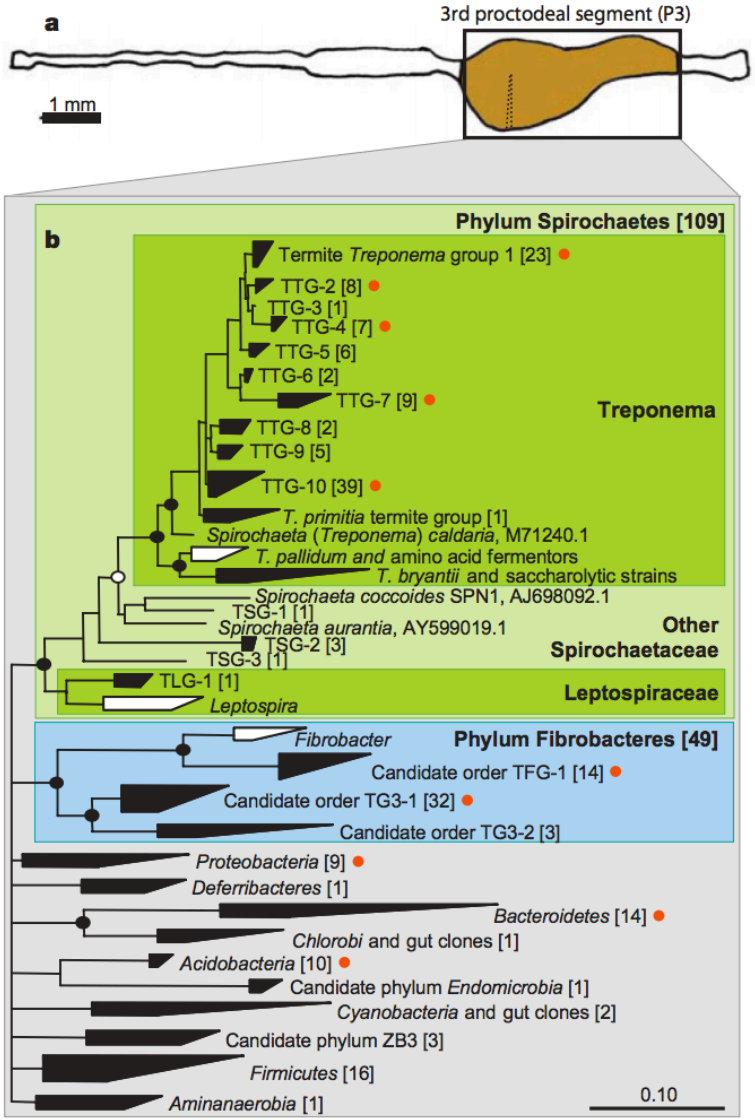
A

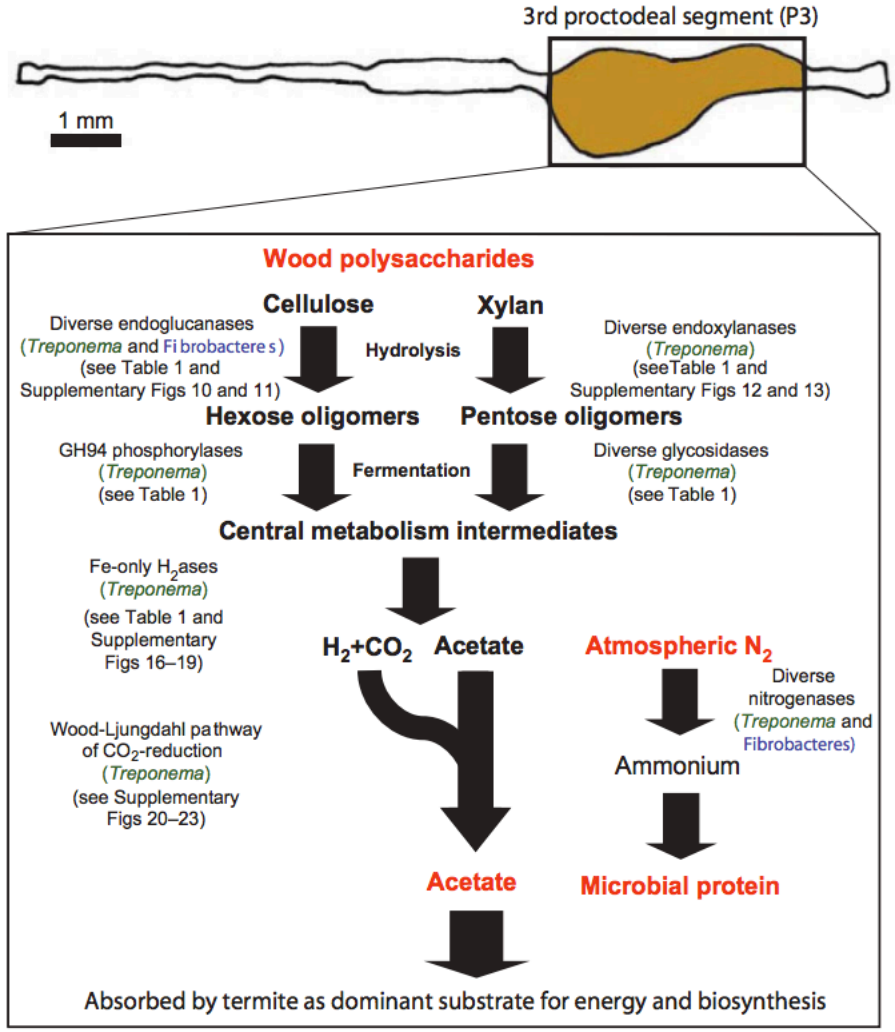
Genome Bin	Genome Size (Mb)	Phylogenetic Order	Estimated Complete-ness
AFa	2.87	Spirochaetales	92.98%
AMa	2.21	Spirochaetales	91.23%
Ala	2.53	Clostridiales	90.10%
AGa	3.08	Bacteroidales	89.77%
AN	2.02	Clostridiales	78.50%
AJ	2.24	Bacteroidales	75.96%
AC2a	2.07	Bacteroidales	75.96%
AWa	2.02	Clostridiales	75.77%
AH	2.52	Bacteroidales	75.45%
AQ	1.91	Bacteroidales	71.36%
AS1a	1.75	Clostridiales	70.99%
APb	2.41	Clostridiales	64.85%
BOa	1.67	Clostridiales	64.16%
ADa	2.99	Myxococcales	62.13%
ATa	1.87	Clostridiales	60.41%

B



Case Study (3): termite gut





Falk et al., 2007

Metatranscriptomics

- The extraction and analysis of metagenomic mRNA (the metatranscriptome) provides information on the regulation and expression profiles of complex communities.
- Metatranscriptomcs studies have made use of direct high-throughput cDNA sequencing to provide whole-genome expression and quantification of a microbial community.

Characteristic	Application
No prior knowledge of the genome sequence is required	• Discover novel transcripts and genetic features (gene mining). • Annotate functional domains in the genome.
Accurate mapping	• Mapping of sequences with an aligner is more precise than hybridization in solution. • Transcription can be studied at a much higher resolution and specificity without interference from non-specific cross-hybridization.
Dynamic range	• Greater dynamic range than fluorescence-based measurements. • Better discrimination at high and low levels of expression.

Paleomics

- **Paleome**: the genome of an extinct species.
- A sedimentary genetic record of past microbial communities.

Thank you