Introduction to next generation microbiome data analytics

Open & reproducible microbiome data analysis spring school
Wageningen, The Netherlands, May 28-30, 2018

Gerben DA Hermes, PhD

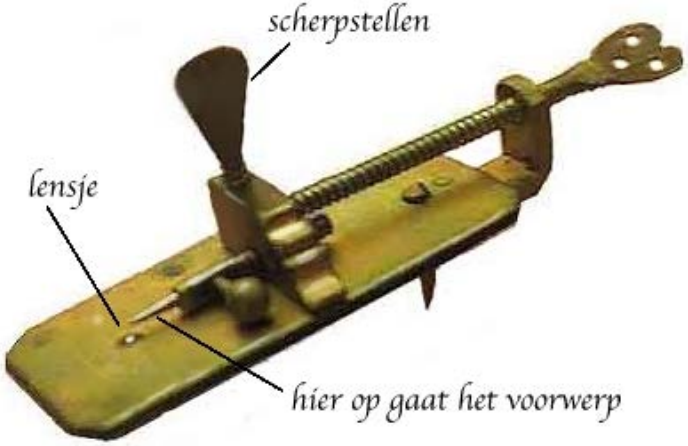
Laboratory of Microbiology,

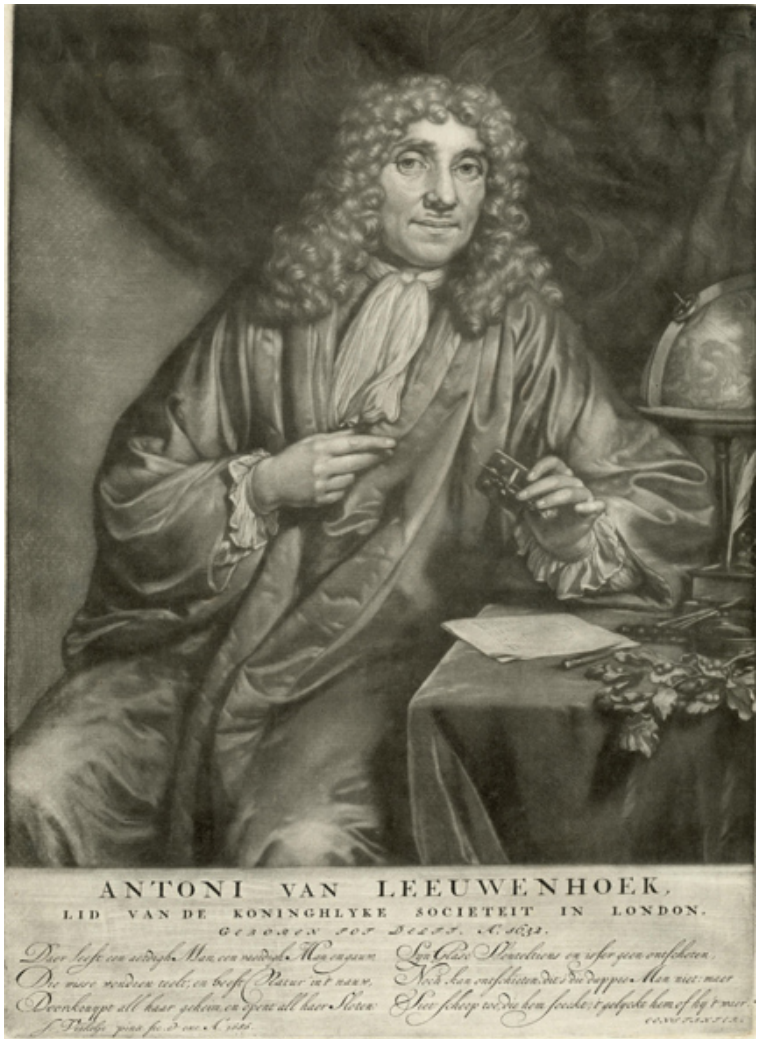
Wageningen University & Research

‘previous generation’ microbiomics

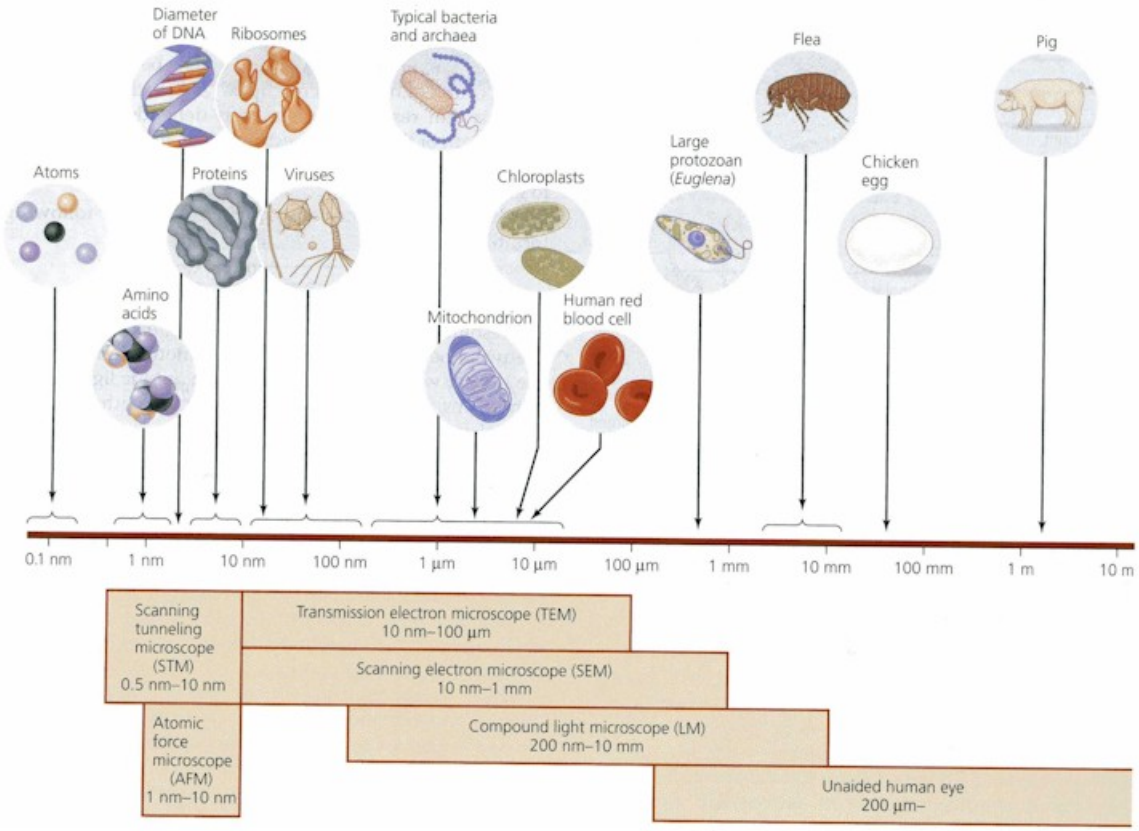
The Godfather

- 1632 - 1723
- >1674
  - Spermatozoa
  - Blood cells
  - Dental plaque





A sense of scale



Metric Measurement of Size

1 Meter (m) = 1m

10 Decimeters (dm) = 1m

100 Centimeters (cm) = 1m

1, 000 Millimeters (mm) = 1m

1,000,000 Micrometers (μm) = 1m

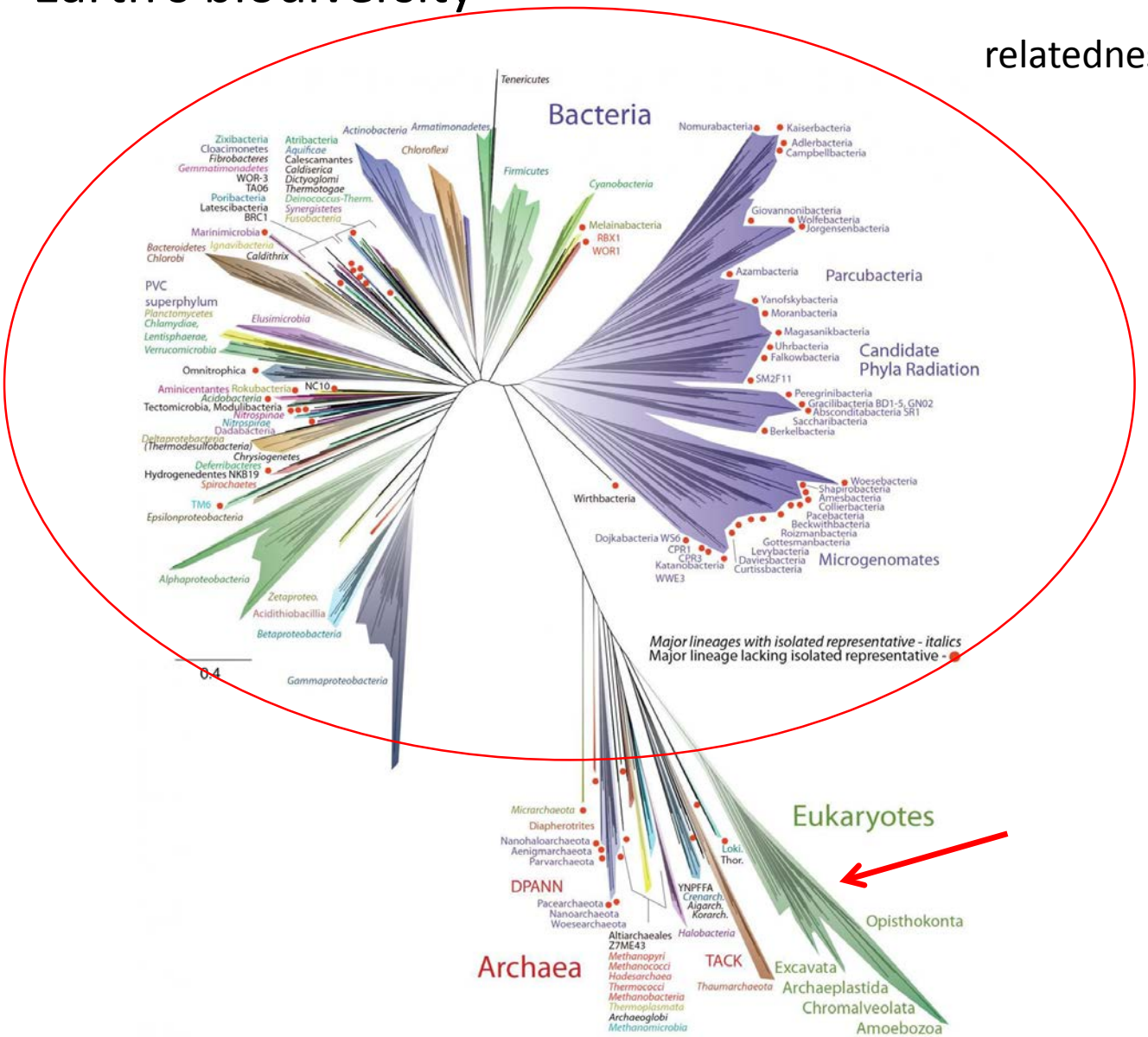
1,000,000,000 Nanometers (nm) = 1m

1,000,000,000,000 Picometers (pm) = 1m

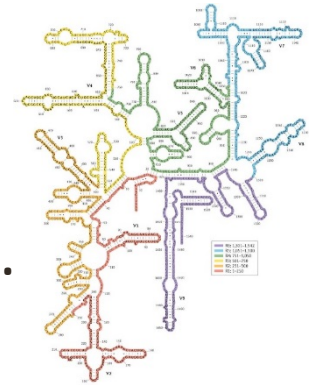
The tree of life

Earth's biodiversity

relatedness/evolution/time

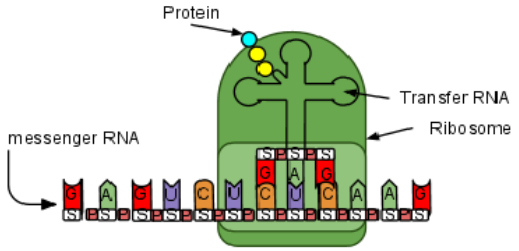


Building a prokaryotic phylogenetic tree
Target gene: Molecular clock

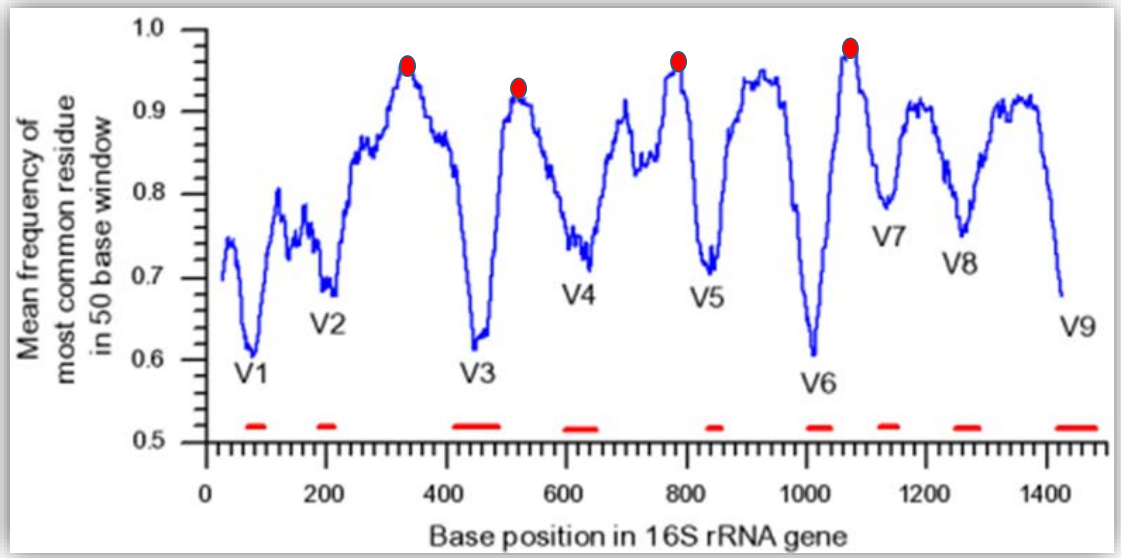


16S rRNA gene: ~1500nt

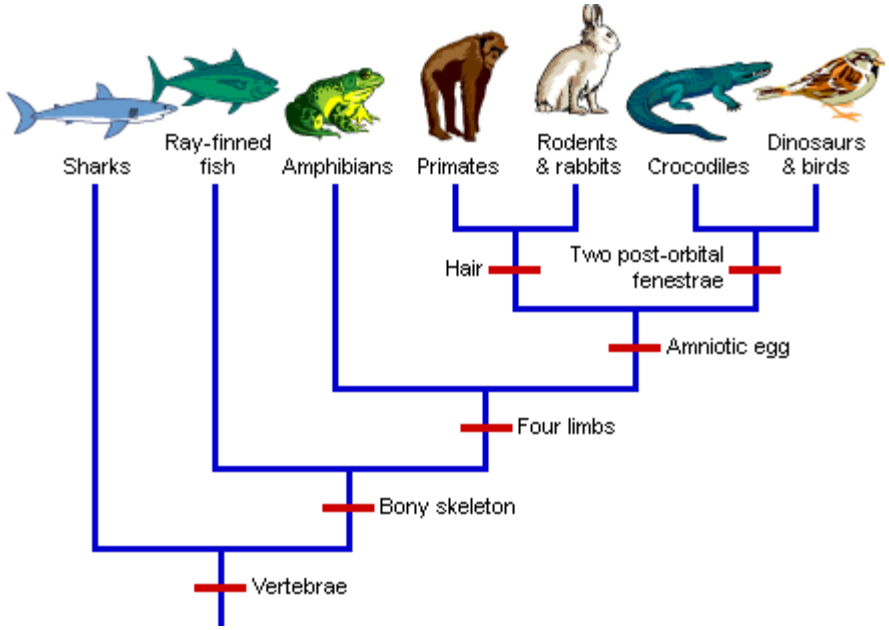
- Target for detection & identification of bacteria.
- Ideal phylogenetic marker.
  - Universally distributed, functionally constant,
  - Sufficiently conserved, no horizontal transfer (?)
- More than 3,000,000 sequences in databases.
- Alternating variable and conserved sequence domains.



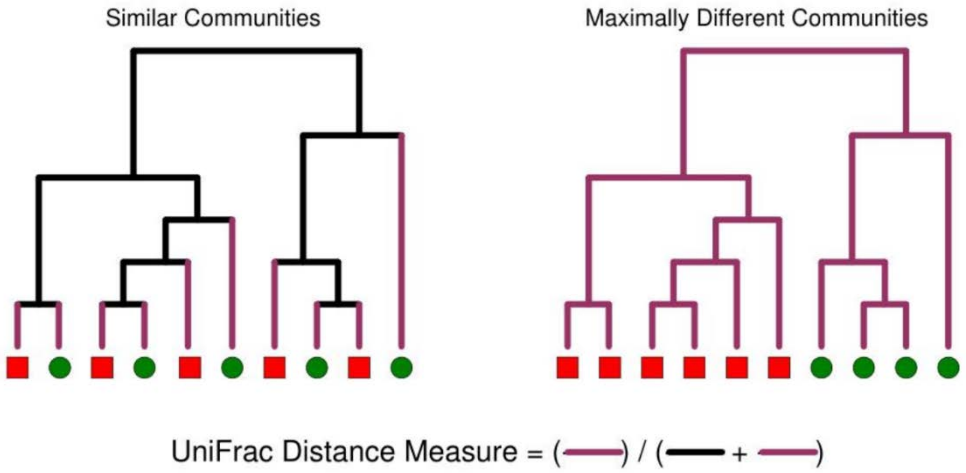
variability



base position

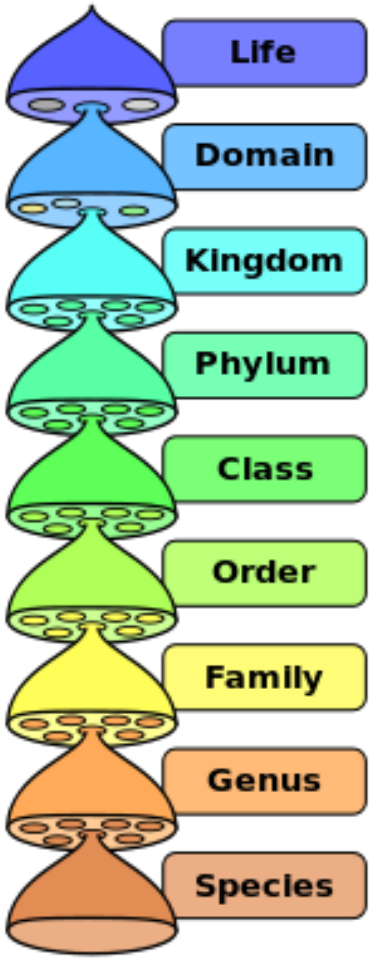


- Measures the evolutionary distance between microbial communities



Classification of all known life

prokaryotic species concept



De virussen worden hier niet behandeld, omdat deze geen organismen zijn en niet voldoen aan de belangrijke criteria voor het begrip leven.

Haeckel (1894) 3 rijken	Whittaker (1969) 5 rijken	Woese (1977) 6 rijken	Woese (1990) 3 domeinen	Cavalier-Smith (1998) 2 domeinen en 6 rijken		Keeling (2004) 3 domeinen en 5 supergroepen	
Animalia	Animalia	Animalia	Eukarya	Eukaryota	Animalia	Eukaryota	Unikonta
Plantae	Fungi	Fungi			Fungi		Excavata
	Plantae	Plantae			Plantae		Archaeplastida
	Protista	Protista			Chromista		Chromalveolata
Protista *(niet behandeld door Linnaeus)*					Protozoa		Rhizaria
	Monera	Archaebacteria	Archaea	Prokaryota	Bacteria	Archaea	
		Eubacteria	Bacteria			Bacteria	

16S rRNA gene similarity prokaryotes

~88%

~92%

~95%

~97%

- DNA-DNA hybridization (70%)
- %G+C content
- Free fatty acid analysis
- morphological, biochemical and physiological characterization

- Almost free exchange of genomic info

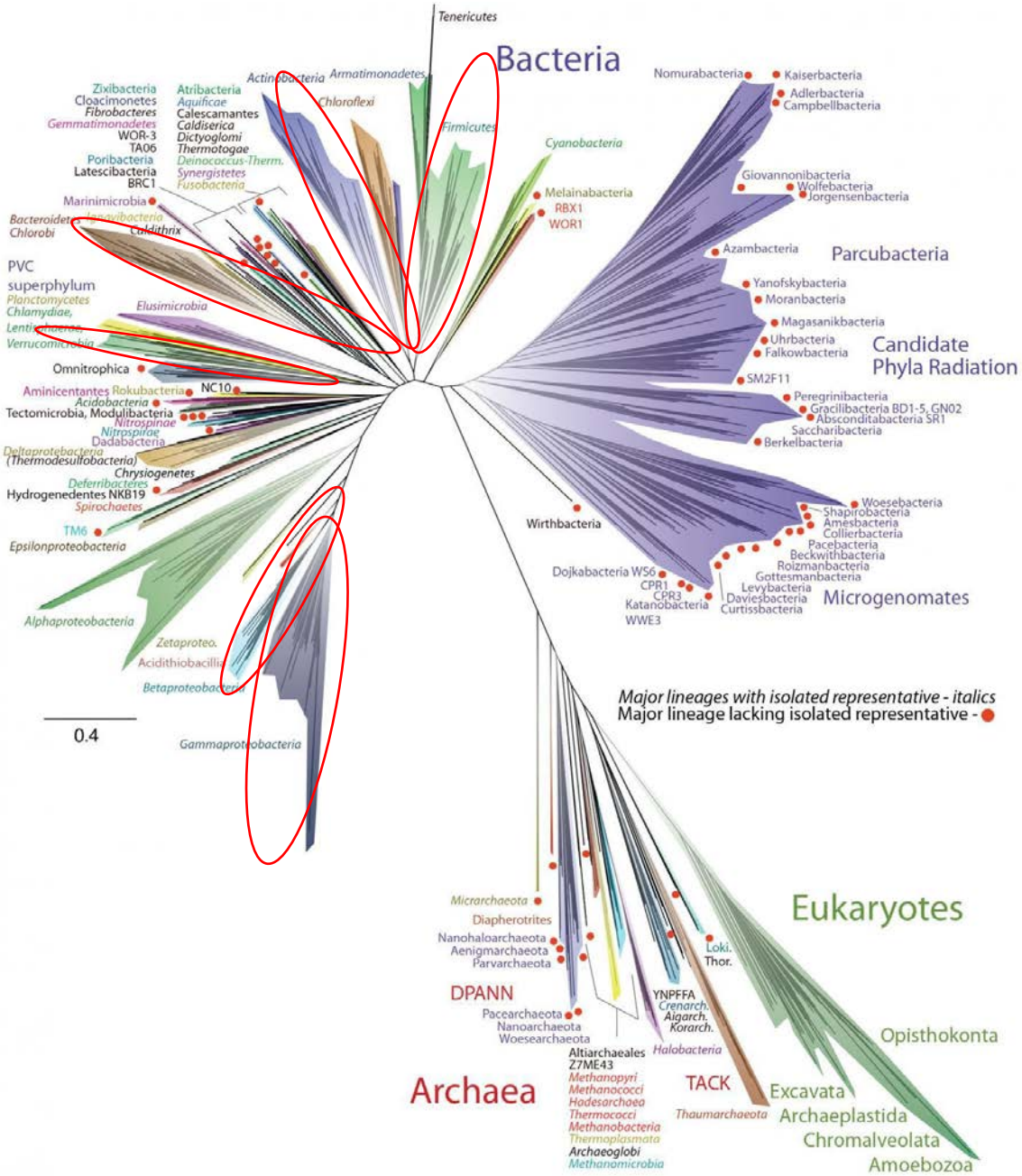
16S rRNA gene Escherichia Coli & Salmonella >98% similar

Human gastro-intestinal microbiota

Complex and diverse ecosystem

- ~ 100 +/- 50 species/ind (>200 strains), ~1000 sp. identified
- **5 dominante phyla**:
  Firmicutes, Bacteroides, Proteobacteria, Actinobacteria and Verrucomicrobia
- Associated with numerous phenotypes/diseases states

© Fuentes 2017


Bacteria
Tenericutes
Actinobacteria
Armatimonadetes
Chloroflexi
Firmicutes
Cyanobacteria
Melainabacteria
RBX1
WOR1
Zixibacteria
Cloacimonetes
Fibrobacteres
Gemmatimonadetes
WOR-3
TA06
Poribacteria
Latescibacteria
BRC1
Atribacteria
Aquificae
Calescamantes
Caldiserica
Dictyoglomi
Thermotogae
Deinococcus-Therm.
Synergistetes
Fusobacteria
Marinimicrobia
Bacteroidetes
Chlorobi
Ignavibacteria
Calditrichx
PVC
superphylum
Planctomycetes
Chlamydiae,
Lentisphaerae
Verrucomicrobia
Elusimicrobia
Omnitrophica
Aminicentantes
Rokubacteria
NC10
Acidobacteria
Tectomicrobia, Modulibacteria
Nitrospinae
Nitrospirae
Dadabacteria
Deltaprotebacteria
(Thermodesulfobacteria)
Chrysiogenetes
Deferribacteres
Hydrogenedentes NKB19
Spirochaetes
TM6
Epsilonproteobacteria
Alphaproteobacteria
Zetaproteo.
Acidithiobacillia
Betaproteobacteria
Gammaproteobacteria
0.4
Nomurabacteria
Kaiserbacteria
Adlerbacteria
Campbellbacteria
Giovannonibacteria
Wolfebacteria
Jorgensenbacteria
Azambacteria
Parcubacteria
Yanofskybacteria
Moranbacteria
Magasanikbacteria
Uhrbacteria
Falkowbacteria
SM2F11
Candidate
Phyla Radiation
Peregrinibacteria
Gracilibacteria BD1-5, GN02
Absconditabacteria SR1
Saccharibacteria
Berkelbacteria
Wirthbacteria
Woesebacteria
Shapirobacteria
Amesbacteria
Collierbacteria
Pacebacteria
Beckwithbacteria
Roizmanbacteria
Gottesmanbacteria
Levybacteria
Daviesbacteria
Curtissbacteria
Dojkabacteria WS6
CPR1
CPR3
Katanobacteria
WWE3
Microgenomates
Major lineages with isolated representative - italics
Major lineage lacking isolated representative -
Micrarchaeota
Diapherotrites
Nanohaloarchaeota
Aenigmarchaeota
Parvarchaeota
DPANN
Pacearchaeota
Nanoarchaeota
Woesearchaeota
Altiarchaeales
Z7ME43
Methanopyri
Methanococci
Hadesarchaea
Thermococci
Methanobacteria
Thermoplasmata
Archaeoglobi
Methanomicrobia
Archaea
YNPFFA
Crenarch.
Aigarch.
Korarch.
Halobacteria
Loki.
Thor.
TACK
Thaumarchaeota
Eukaryotes
Opisthokonta
Excavata
Archaeplastida
Chromalveolata
Amoebozoa

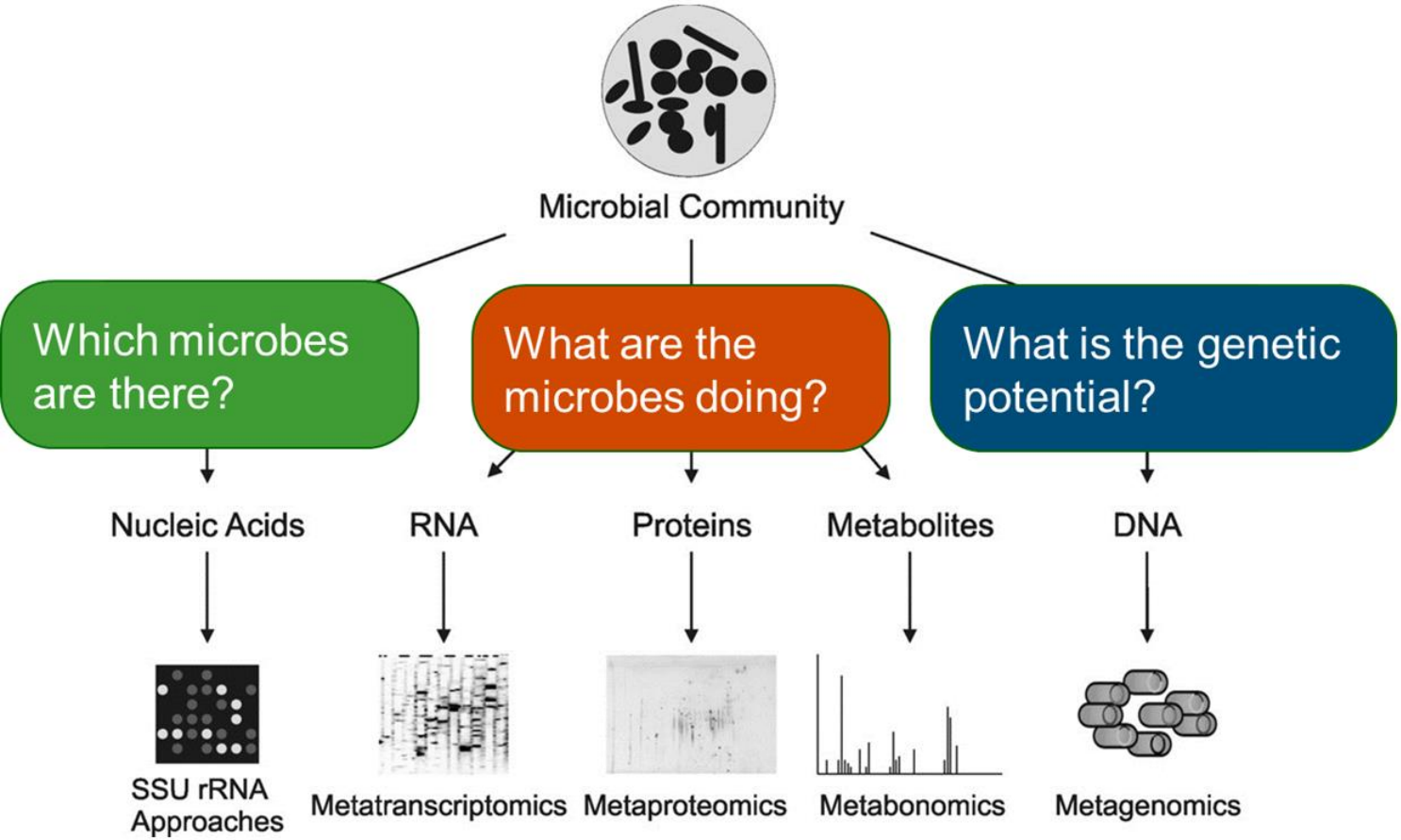
Which method do you use to study the microbiota?

Depends on your research question!!!

© Zoetendal

Studying Microbiota Composition and Function

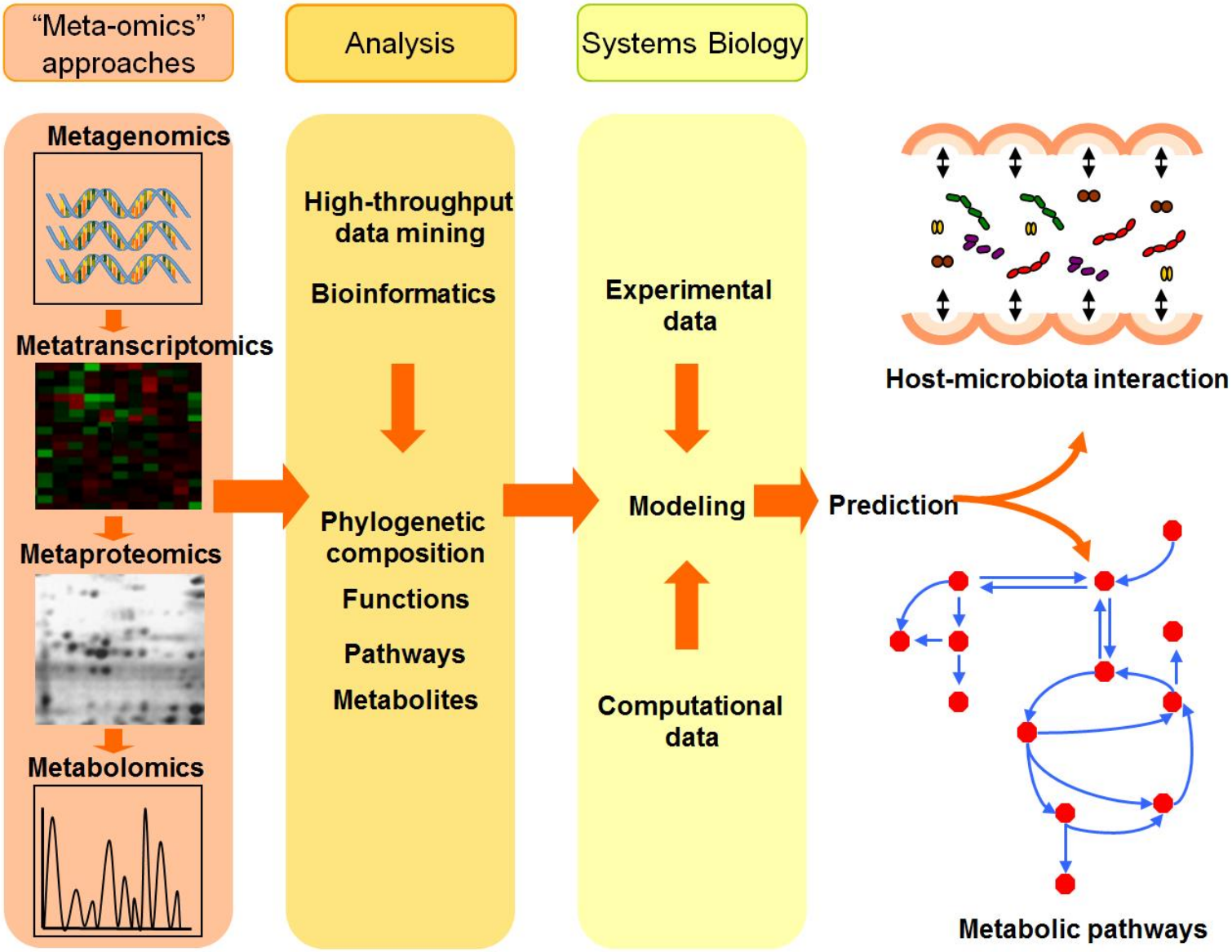
No universal method that provide all answers



Cultivation

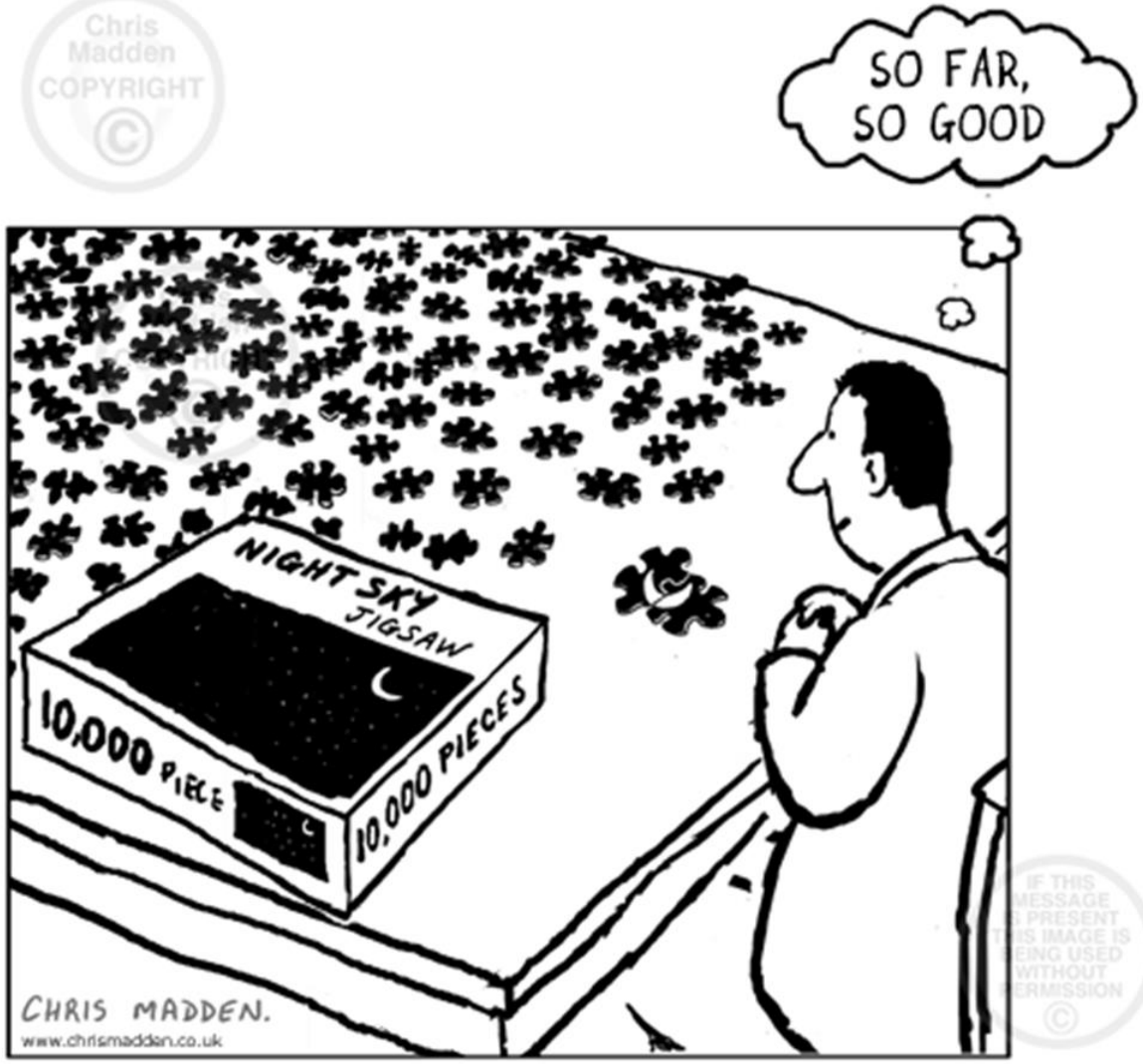
Zoetendal et al, 2008, Gut.

Integrated ‘meta-omics’ approach to understand microbiological concepts



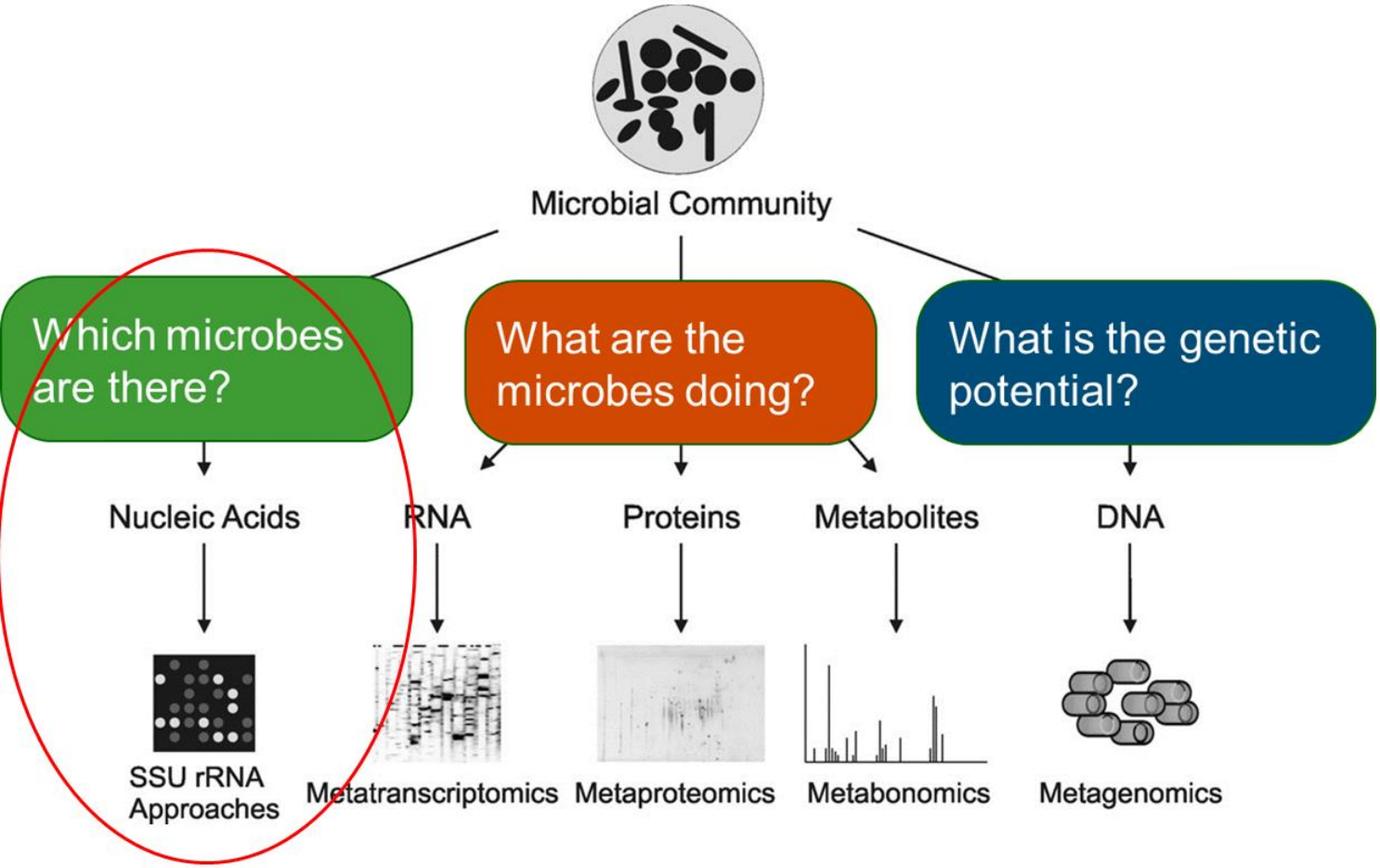
Van den Bogert and Leimena et al. 2011. Handbook of Mol. Micr. Ecol. Chapter 18

Integrated ‘meta-omics’ approach



Studying Microbiota Composition and Function

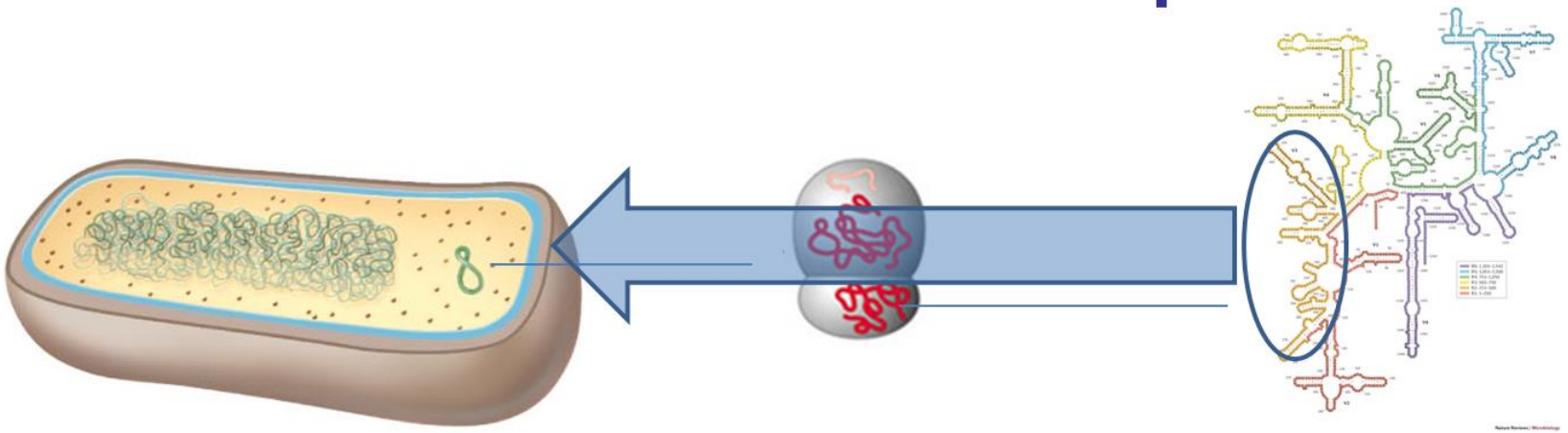
No universal method that provide all answers

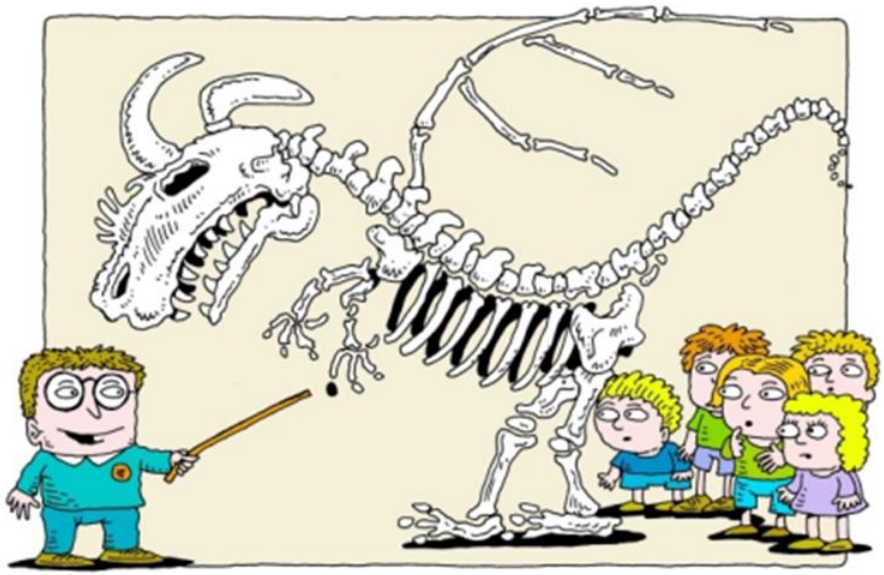


Cultivation

Zoetendal et al, 2008, Gut.

Be careful with functional interpretations!

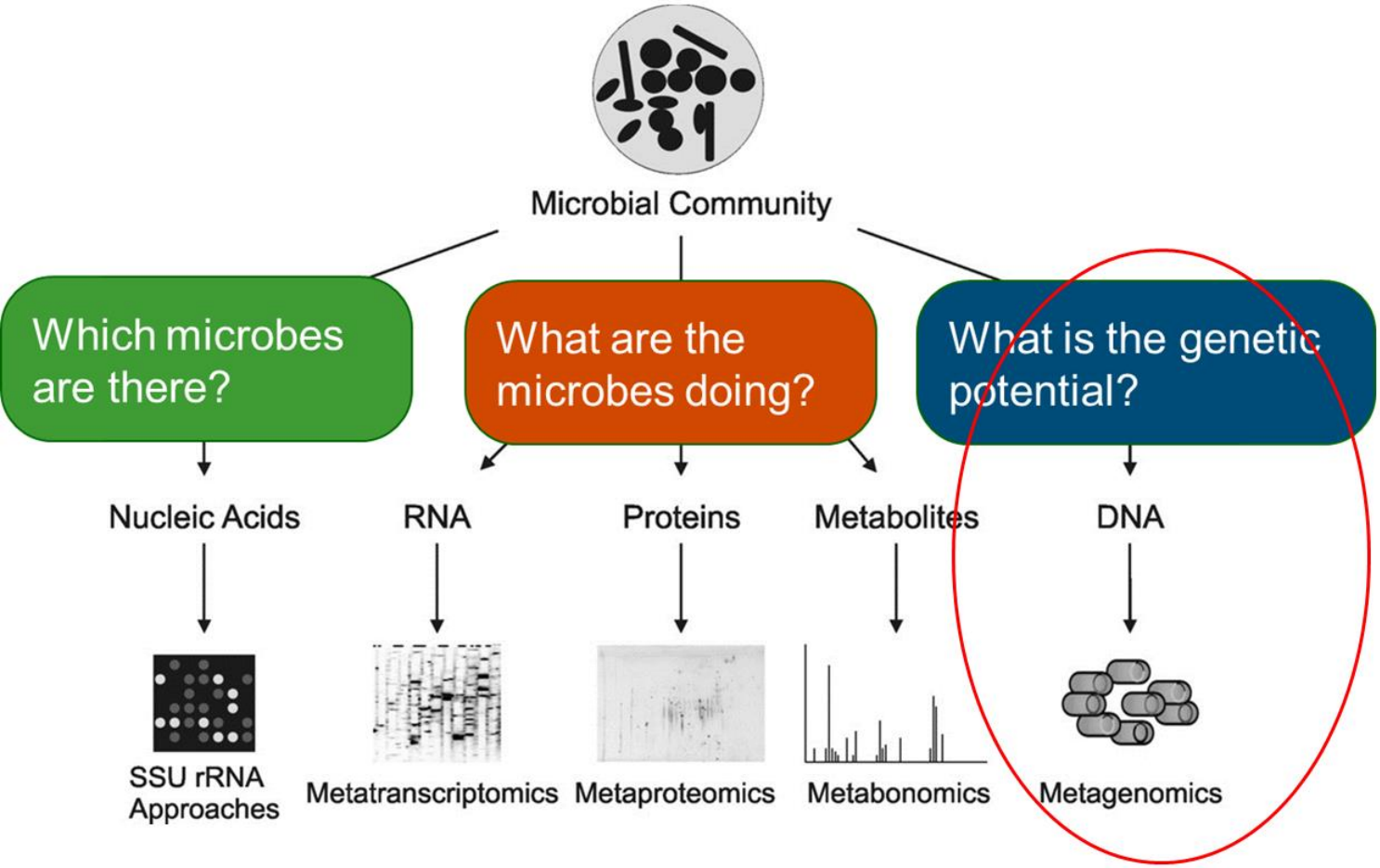




"Luckily we found THIS bone, so we were able to reconstruct the whole creature..."

Studying Microbiota Composition and Function

No universal method that provide all answers

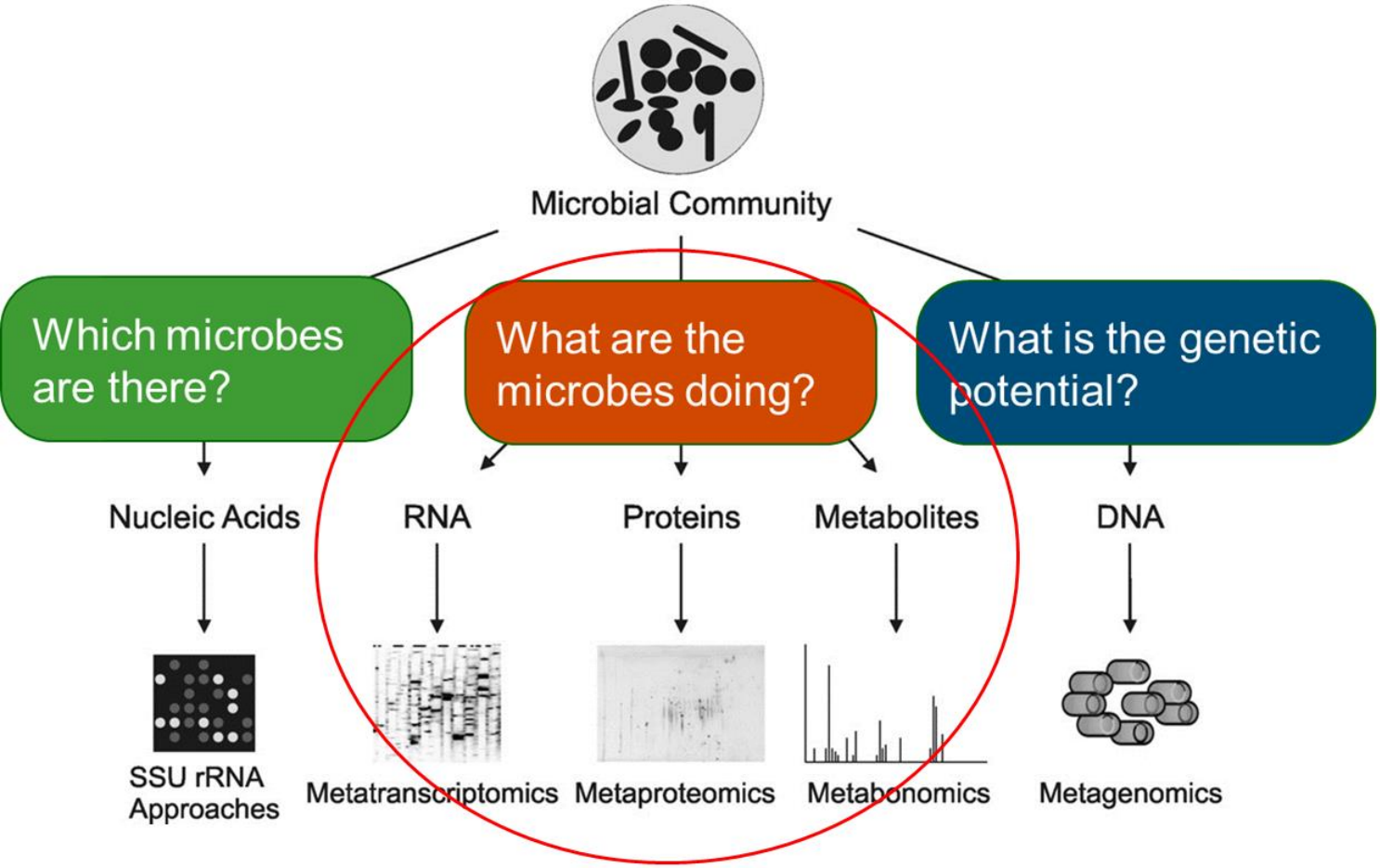


Cultivation

Zoetendal et al, 2008, Gut.

Studying Microbiota Composition and Function

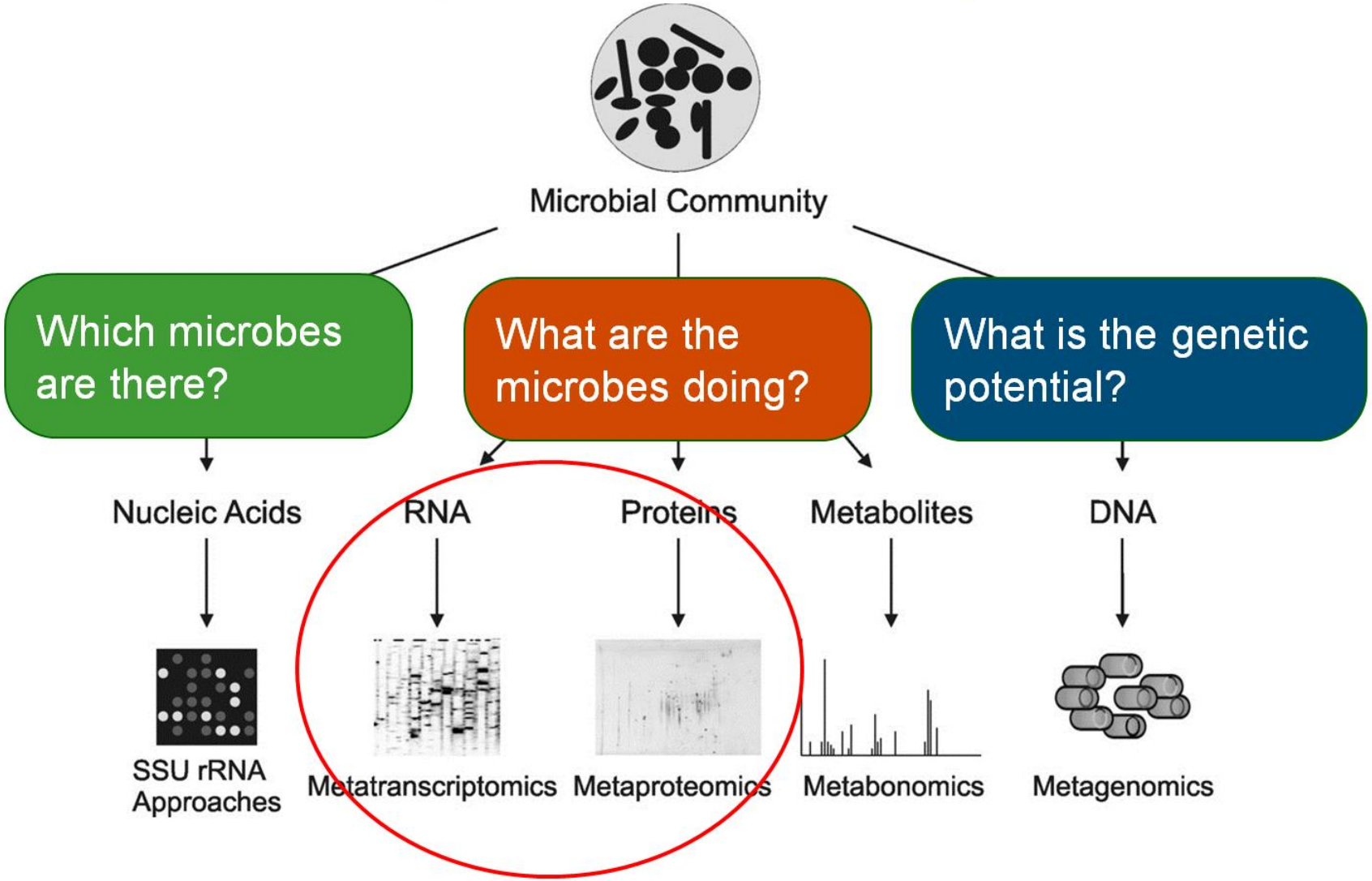
No universal method that provide all answers



Cultivation

Zoetendal et al, 2008, Gut.

Who is doing what in the intestine? Metatranscriptomics and metaproteomics



-Metatranscriptomics: Community gene expression

-Metaproteomics: Community protein production

Zoetendal et al, 2008, Gut.

Example: composition vs activity

Reciprocal diet exchange between Africans and Americans

Colon cancer rates:
-Africans:<10:100,000
-African Americans: >65:100,000

African Americans → Western diet → 2 weeks → African diet → 2 weeks

Africans → African diet → Western diet

nature COMMUNICATIONS

ARTICLE

Received 23 May 2014 | Accepted 20 Jan 2015 | Published 28 Apr 2015 DOI: 10.1038/ncomms7342

Fat, fibre and cancer risk in African Americans and rural Africans

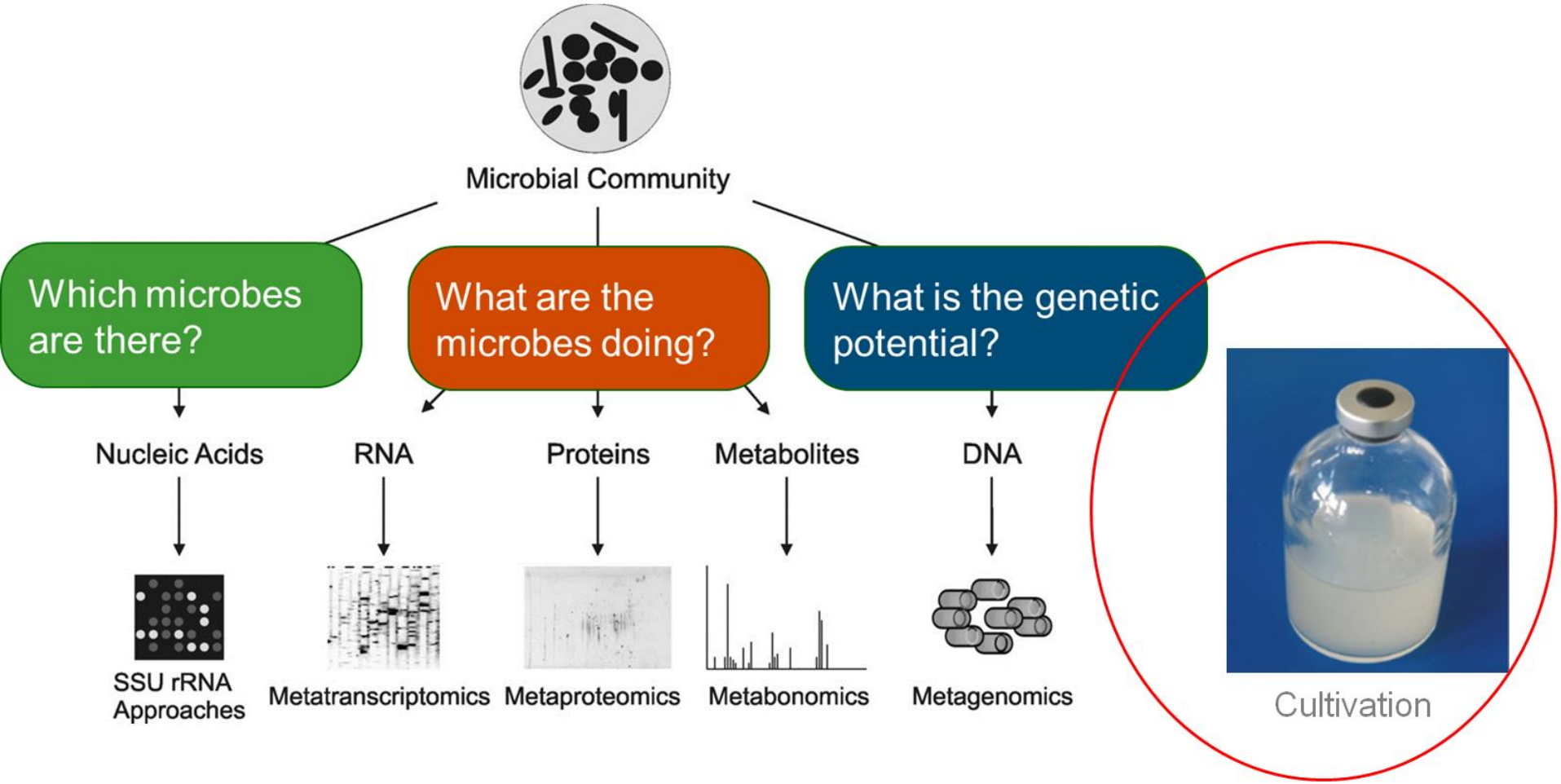
Stephen J.D. O'Keefe[1], Jia V. Li[2], Leo Lahti[3,4], Junhai Ou[1], Franck Carbonero[5,†], Khaled Mohammed[1], Joram M. Posma[2], James Kinross[2], Elaine Wahl[1], Elizabeth Ruder[6], Kishore Vipperla[1], Vasudevan Naidoo[7], Lungile Mtshali[7], Sebastian Tims[3], Philippe G.B. Puylaert[3], James DeLany[8], Alyssa Krasinskas[9], Ann C. Benefiel[5], Hatem O. Kaseb[1], Keith Newton[7], Jeremy K. Nicholson[2], Willem M. de Vos[3,4,10], H. Rex Gaskins[5] & Erwin G. Zoetendal[3]

Microbiota-mediated reduction of colon cancer risk after consumption of African diet

Limited diet effect on microbiota composition
Huge diet impact on microbiota activity

Studying Microbiota Composition and Function

No universal method that provide all answers

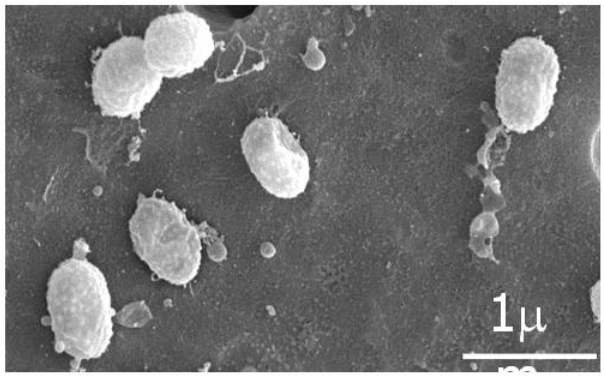


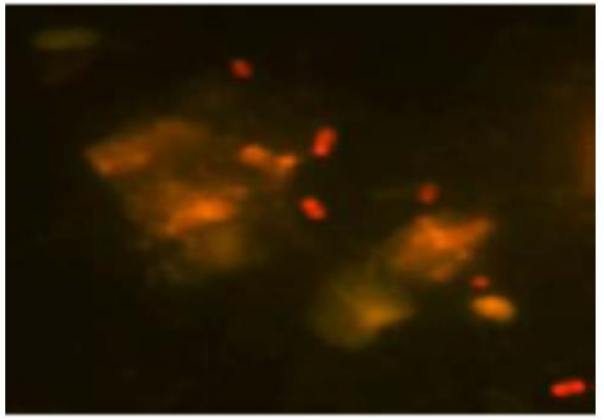
Zoetendal et al, 2008, Gut.

No cultured microbes, no microbiology

Previous generation microbiomics

All microbiological concepts based on cultured representatives





Example: *Akkermansia muciniphila*

-Isolated in 2004

-Gram-negative

-specialized in mucin degradation

Subsequent studies:

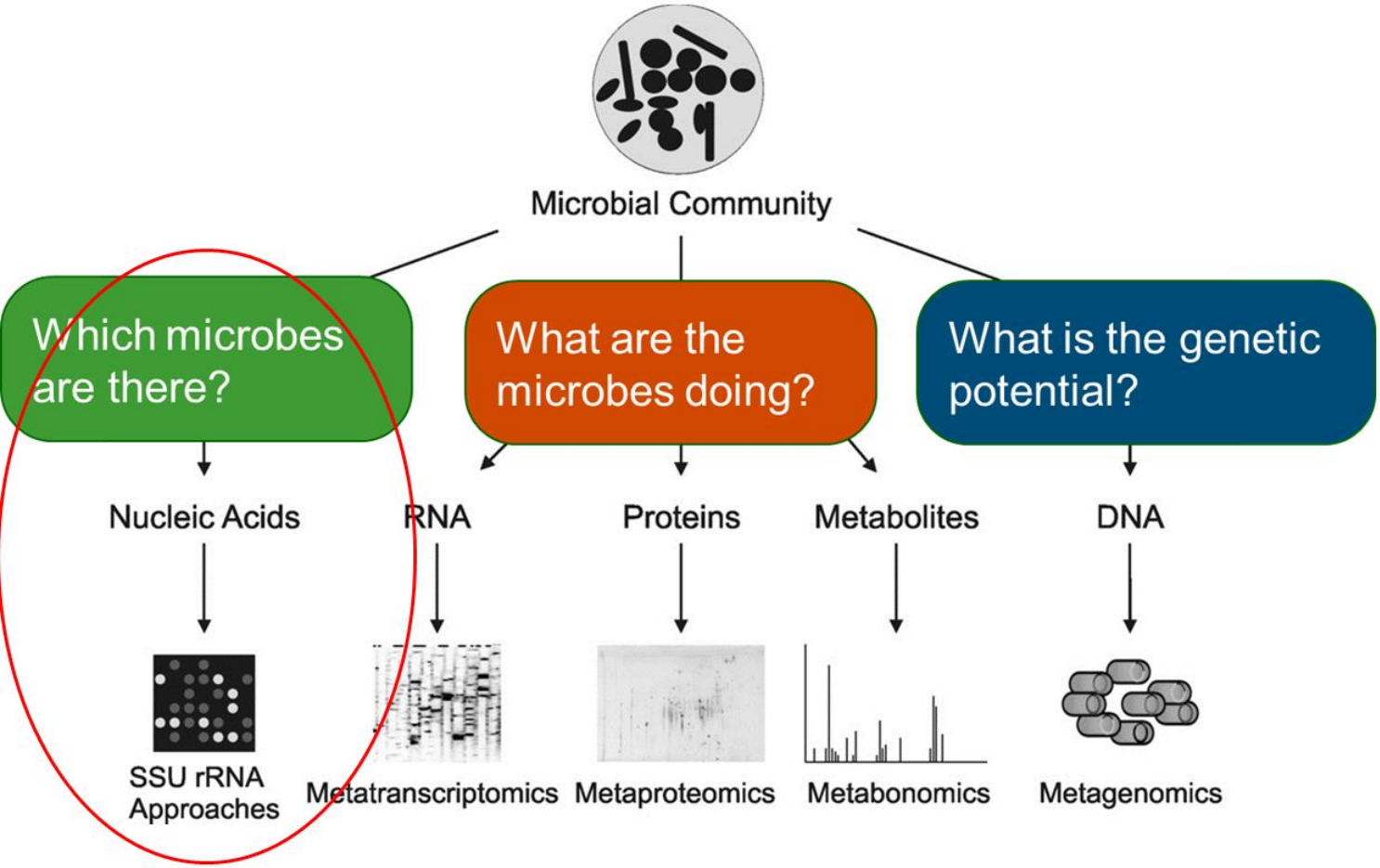
-Frequently associated to healthy controls in comparative analyses

-Counteracts diet-induced obesity in mice

Derrien et al., 2004

Studying Microbiota Composition and Function

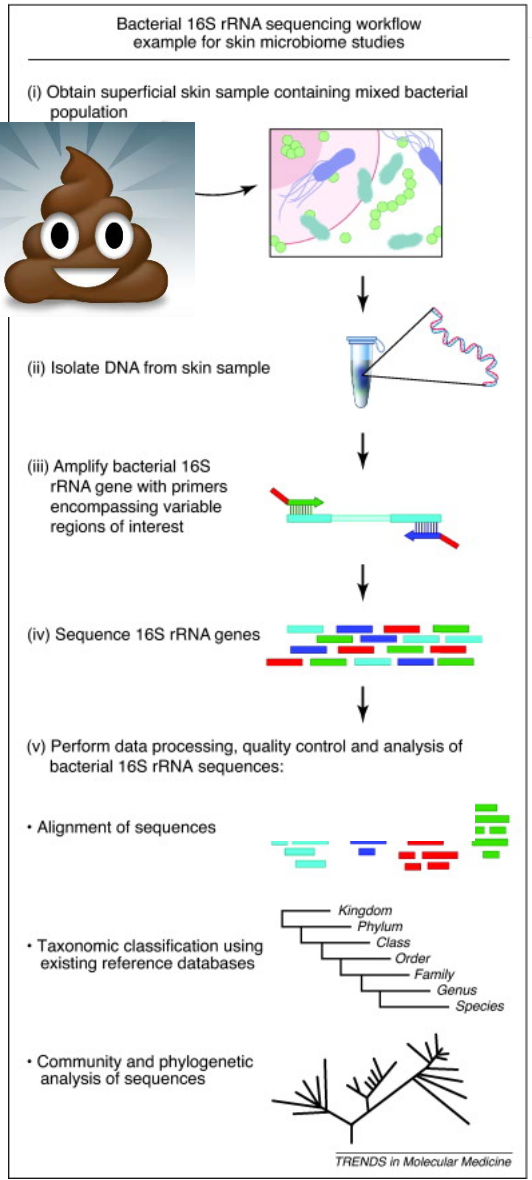
No universal method that provide all answers

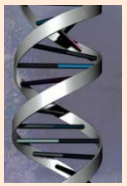


Cultivation

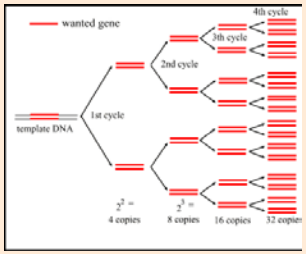
Zoetendal et al, 2008, Gut.

16S rRNA gene amplicon sequencing workflow





DNA extraction



PCR

Community profiling with NGS

‘Next Generation’ Sequencers

Table 1 | **Comparison of sequencing technologies**

	Read length	Maximum insert size	Run time (hours (h) or days (d))	Reads per run	Relative cost factor (per Mb)	Scale of reads per sample	Scale of samples per run	Raw error rate (%)			
								Total	Insertions	Deletions	Mismatches
ABI 3730	800 b	>1 Kb	2 h	96	100	10^2	10^1	0.001	<<0.1	<<0.1	<<0.1
454 FLX Titanium	300–400 b	800 b	9 h	10^6	1	10^3	10^2	1	< 1	<< 0.1	<< 1
454 FLX+	500–600 b	1200 b	23 h	10^6	0.7	10^3	10^2				
Illumina GAIIx	76–101 b	500 b	6–9 d	4×10^8	0.1	10^5–10^6	10^3–10^4	<1	<<1	<<1	<1
Illumina HiSeq 2000	101–151 b	500 b	9–15 d	3×10^9	0.002	10^5–10^6	10^3–10^4				
Illumina MiSeq	36–151 b	500 b	4h–27 h	10^7	0.06	10^4	10^2				
PacBio	1100 b	>1 Kb	1.5 h	3.5×10^7	1.5	10^3	10^1	15	13	1	1
IonTorrent	200 b	400 b	2–3 h	1.5×10^6–3×10^6	0.4	10^3	10^2	2	1	1	<1

Kuczynski et al., Nature rev. 2011

Why Next Generation Sequencing?

Resolution

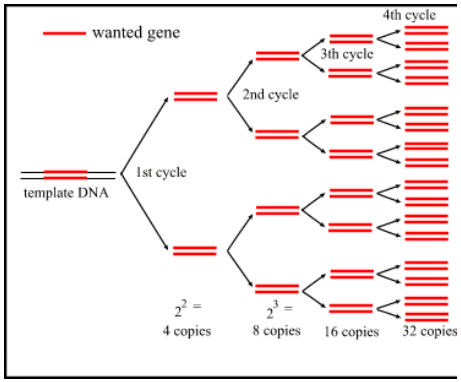
- Sanger: long sequences > more taxonomic information 450 – 1200 bp
- NGS shorter sequences > less taxonomic information 50 – 450 bp

Data

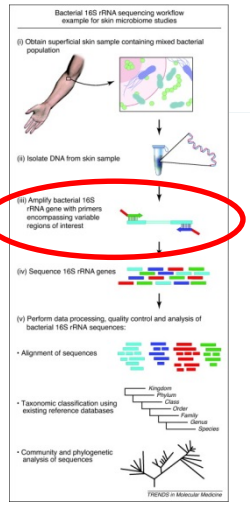
- Sanger: 10, 100, 200 sequences … (time & $)
- NGS 200.000 to > 1.000.000.000 sequences (much cheaper/sequence)

WAGENINGEN
UNIVERSITY & RESEARCH

Amplicon sequencing



20.000 pe Depends on your research question!!!



PCR

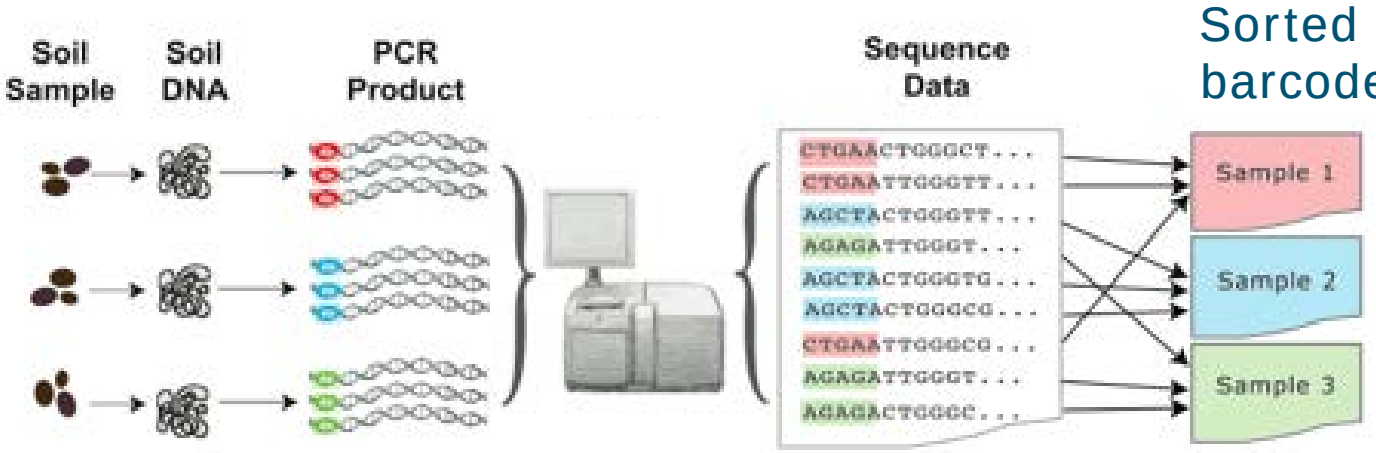
what to do with 10.000.000 sequences?

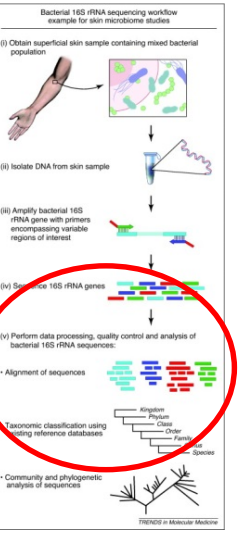
1 sample; pick 10.000.000 colonies

Or

100.000 colonies from 100 samples

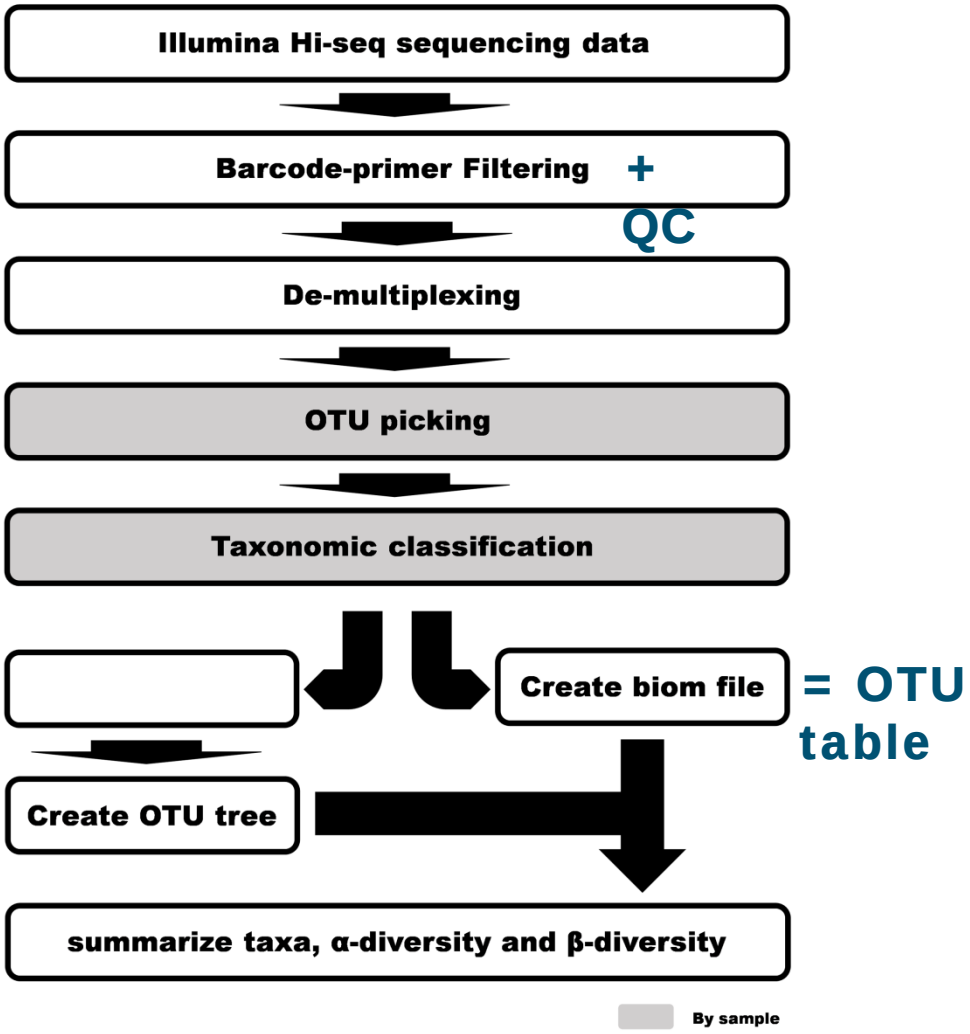
Barcoding & multiplexing



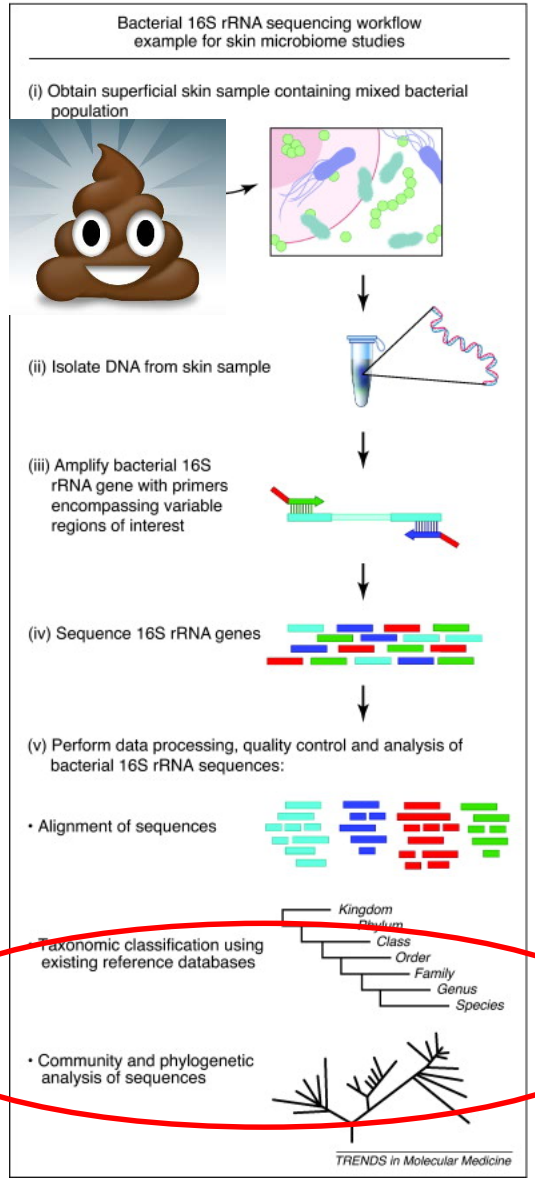


Analysis pipeline/workflow

OTU -> Operational Taxonomic Unit (nt sequence -> biological entity)



WAGENINGEN
UNIVERSITY & RESEARCH



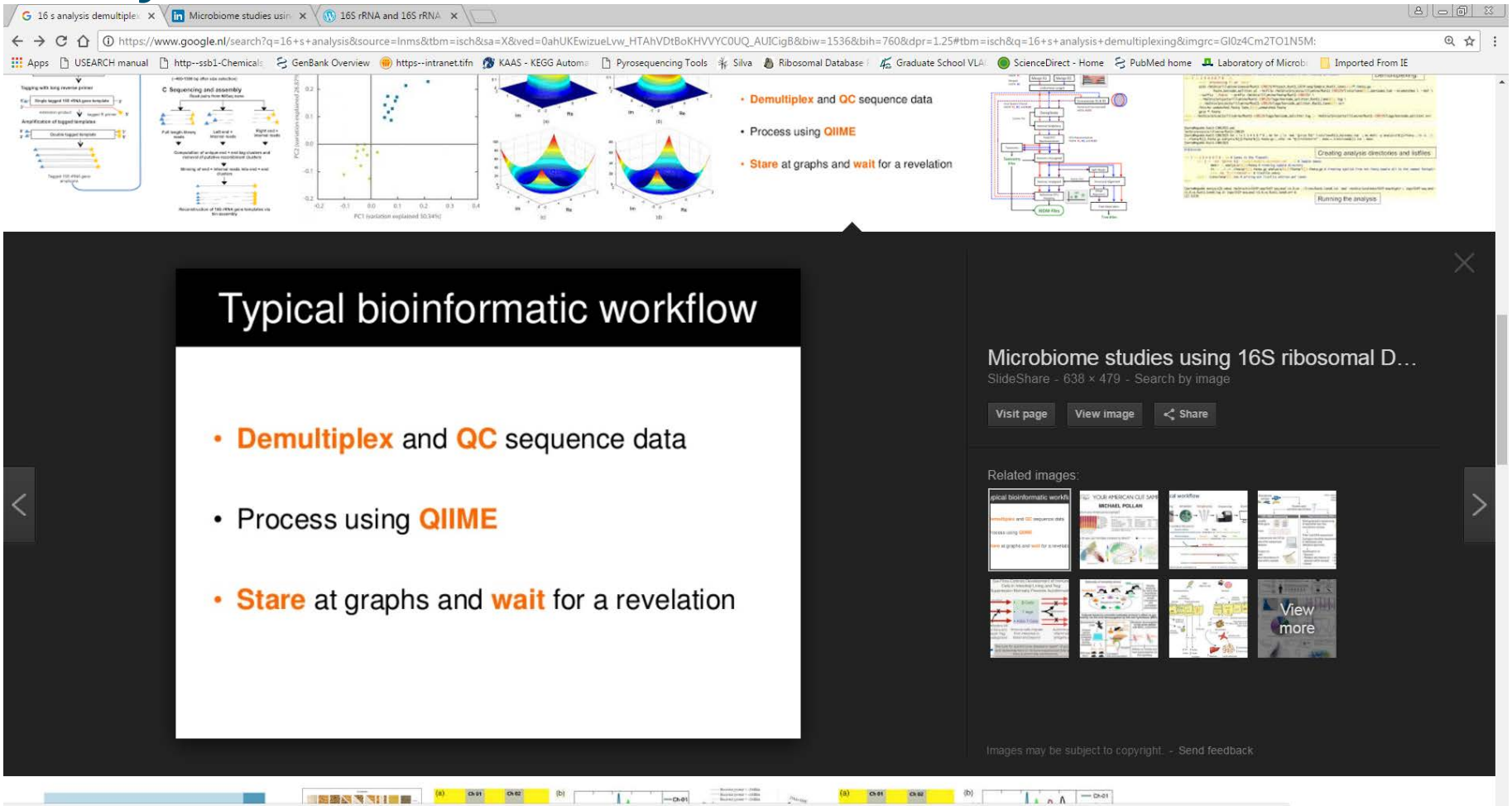
Weeks/months

- Biological interpretation
- Hypothesis testing
- Months/year?

Day 2/3

The hard part....

Day 2/3



Recap