KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(오프라인)

Human microbiome studies with bioinformatics approaches

이선재 _ GIST

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 (2.6 월)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	개회사/공지사항전달			
09:30-10:50 (80)	**Best practice for single-cell data analysis**	**박종은 교수**	**Introduction to ML & DNN (이론)**	**이상근 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Practice1: Scanpy basic workflow**	**김우석 김성룡 조교**	**CNN (이론)**	**이상근 교수**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**Public data, batch correction, cell annotation**	**박종은 교수**	**RNN, GAN, XAI (이론)**	**이상근 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**Practice2: Advanced single-cell analysis**	**김우석 김성룡 조교**	**AI 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)**	**이정현 한성민 조교**

DAY2 (2.7 화)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**Introduction to protein structure prediction - Homology modeling - Coevolution-guided modeling Early AI-based approaches**	**백민경 교수**	**Pre-trained Models for Transfer Learning (이론)**	**전민지 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**단백질 구조 예측 실습 - MSA generation, template search - homology modeling contact prediction & modeling**	**백민경 교수**	**Pre-trained Models for Transfer Learning (실습)**	**정민수 조교**
12:10-13:40 (90)	점심			
13:40-15:10 (90)	**AI-based protein structure prediction - AlphaFold/RoseTTAFold Applications to PPI prediction & protein design**	**백민경 교수**	**Deep learning in Bioinformatics**	**노미나 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**단백질 구조 예측 실습 II AlphaFold, RoseTTAFold 실습 및 응용**	**백민경 교수**	**Deep learning model을 이용한 실습**	**곽호진 박예솔 조교**

DAY3 (2.8 수)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**화학정보학 기초(Cheminformatics) 약물특성 및 약물다움(druglikeness) Molecular Notations & Descriptors AI 신약개발을 위한 Databases AI 신약개발을 위한 Programming 기초**	**김동섭 교수**	**마이크로바이옴 기본 이론**	**이선재 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Google Colab에 RDKit 설치 화합물 정보 읽기 실습 Bioactivity database 검색 및 정보 읽기 실습 Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습**	**문채영 나민주 조교**	**16S rRNA amplicon seq. - DADA2**	**서영창 조준우 조교**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**AI 신약개발을 위한 기계학습법 기초 QSAR 모델링 기초 AI 신약개발을 위한 딥러닝 모델 Virtual screening (ligand-based, structure-based) 및 de novo design**	**김동섭 교수**	**최신 메타지놈 분석 기법의 현황**	**이선재 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**QSAR modeling 전체 과정 실습 화합물의 Bioactivity 예측 모델 개발 Virtual screening 과정을 통한 신약후보물질 발굴 실습**	**문채영 나민주 조교**	**Shotgun metagenome 분석 (Linux)**	**서영창 조준우 조교**

강의개요

Human microbiome studies with bioinformatics approaches

최근 들어 마이크로바이옴이 인체 생리작용에 끼치는 영향이 계속해서 밝혀짐에 따라서, 마이크로바이옴의 조성과 생리적인 기능을 이해하려는 연구가 크게 각광을 받고 있다. 예를 들어, 비만, 당뇨병, 간질환, 파킨슨병, 치매 등이 마이크로바이옴과 높은 관련성이 밝혀졌으며, 분변이식술 실험을 통해 숙주 인체의 표현형이 전달될 수 있고, 이를 활용하여 치료 역시 가능해짐이 밝혀지고 있다.

그러나 마이크로바이옴은 다른 오믹스 데이터와 달리 여러가지 challenge들이 남아있다. 첫번째로, 정해진 레퍼런스가 없는 "Microbial dark matter" 문제, 두번째 heterogeneous한 마이크로바이옴 데이터로 인한 분석의 어려움, 특히 각 사람마다의 생활습관/식습관등의 차이로 인한 confounding factor들이 큰 문제이다. 본 강의에서는 현재 마이크로바이옴 연구의 최근동향과 NGS 기법을 활용한 마이크로바이옴 분석에 대한 강의가 진행된다.

강의는 다음의 내용을 포함한다:

- 마이크로바이옴 이론
- Amplicon-based 16S rRNA sequencing 분석
- Shotgun metagenomics 분석

* 교육생준비물:
  노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

* 강의 난이도: 중급

* 강의: 이선재 교수 (광주과학기술원 생명과학부)

Curriculum Vitae

Speaker Name: Sunjae Lee, Ph.D.

▶ **Personal Info**

Name	Sunjae Lee
Title	Assistant Professor
Affiliation	Gwangju Institute of Science and Technology (GIST)

▶ **Contact Information**

Address	123, Chumdangwagi-Ro, Buk-Gu, Gwangju, 61005
Email	leesunjae@gist.ac.kr
Phone Number	062-715-2505

Research Interest

Systems biology, Bioinformatics, Microbiome, Metabolism

Educational Experience

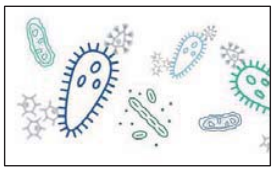
2006	B.S. in Bioinformatics, KAIST, Korea
2004	M.S. in Bioinformatics, KAIST, Korea
2007	Ph.D. in Bioinformatics, KAIST, Korea

Professional Experience

2015-2018	Post-doctoral researcher, KTH – Royal institute of technology, Sweden
2018-2020	Senior Research Associate, Centre for Host-Microbiome Interactions, King's College London, UK
2020-	Assistant professor, School of Life Sciences, Gwangju Institute of Science and Technology (GIST)

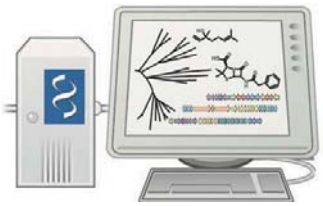
Selected Publications (5 maximum)

1. Vishal Patel*, Sunjae Lee* et al., "Rifaximin reduces gut-derived inflammation and mucin degradation in cirrhosis and encephalopathy: RIFSYS Randomised-Controlled Trial", **J Hepatology**, 2021
2. Mathias Uhlen, Cheng Zhang, Sunjae Lee et al., "A pathology atlas of the human cancer transcriptome", **Science**, 2018
3. Sunjae Lee*, Cheng Zhang*, Zhengtao Liu* et al., "Network analyses identify liver-specific targets for treating liver diseases", **Molecular Systems Biology**, 2017
4. Sunjae Lee*, Cheng Zhang*, Murat Kilicarslan* et al., "Integrated Network Analysis Reveals an Association between Plasma Mannose Levels and Insulin Resistance", **Cell Metabolism**, 2016
5. Sunjae Lee, Adil Mardinoglu, Cheng Zhang et al., "Dysregulated signaling hubs of liver lipid metabolism reveal hepatocellular carcinoma pathogenesis", **Nucleic Acids Research**, 2016



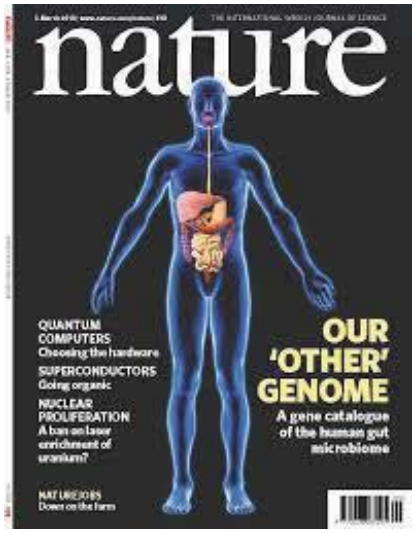
BIML-workshop
한국생명정보학회

Human microbiome studies with Bioinformatics approaches



GIST | Life Mining Lab
total 160 pages

We humans are “microbial”

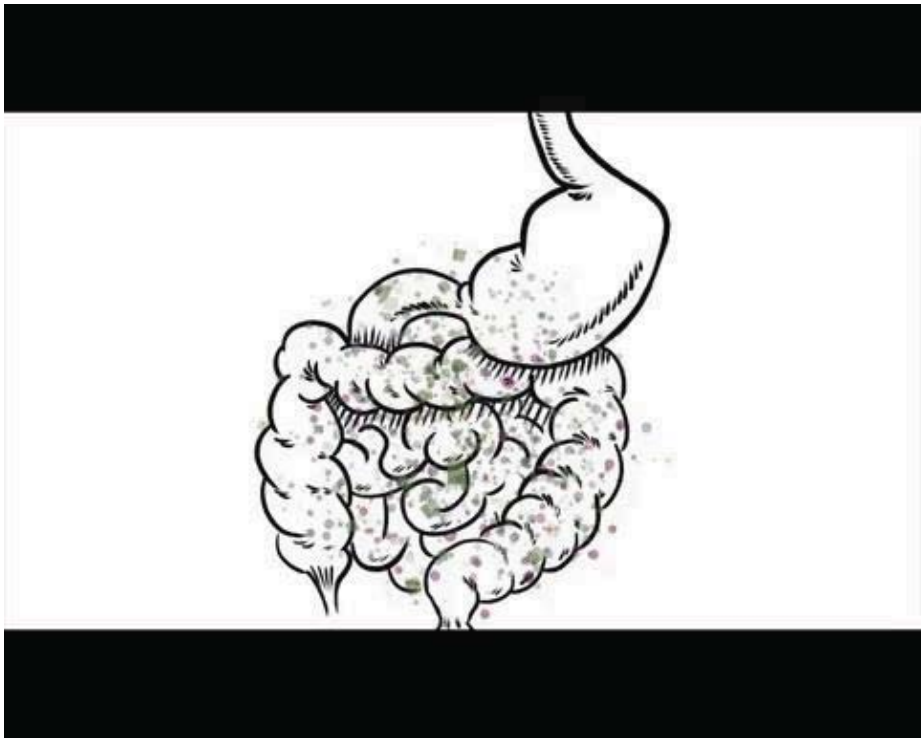


100 trillion cells **10X** **human cells**

20 million genes **100X** **human genes**

2

Individuals harbours own unique microbiome



3

Individuals harbours own unique microbiome

Only 20~30% were shared

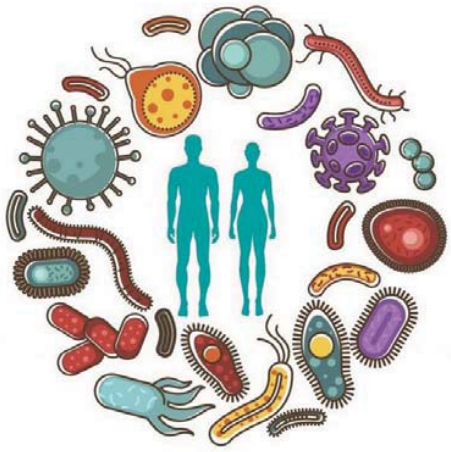
Savannah

Jungle

Tundra

REF | Rob Knight, TED talks

4



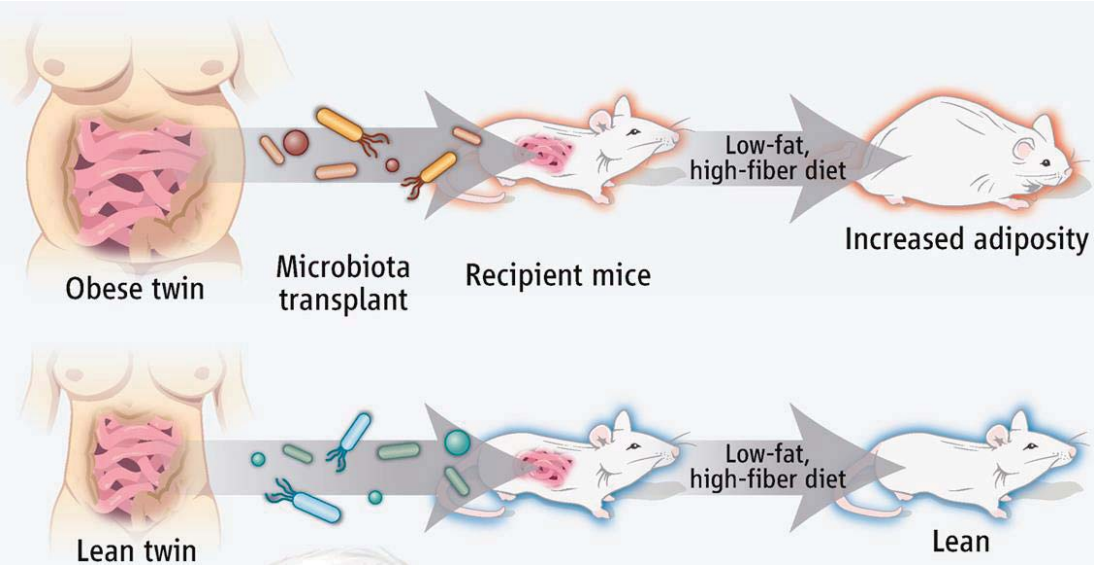
Microbiome =
Our Second Genome

5

Why microbiome is important?

6

1 Microbiome determines host phenotypes

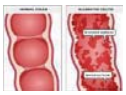


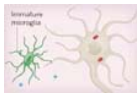
REF | Alan W. Walkter, Julian Parkhill, Science, 2013

7

1 Microbiome determines host phenotypes


Obesity


Colitis

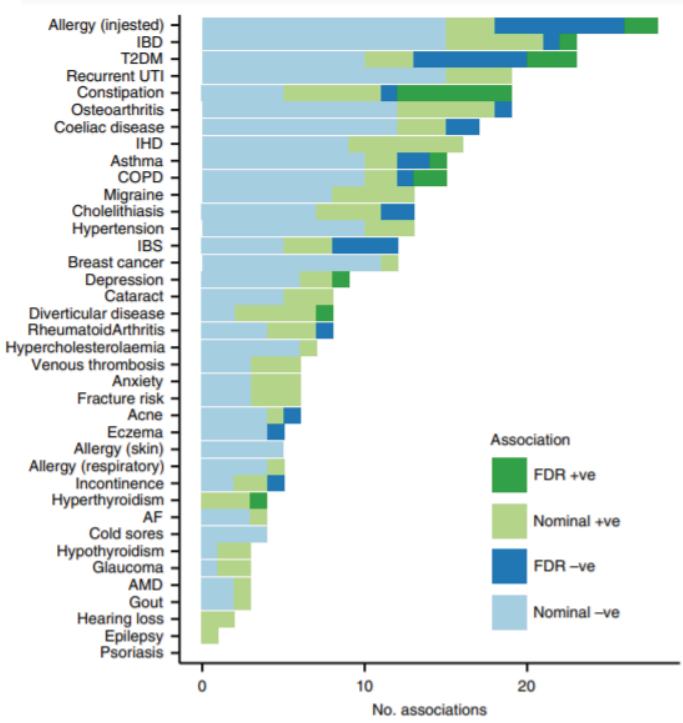

Parkinson's


Autism


Schizophrenia

8

1 Microbiome determines host phenotypes

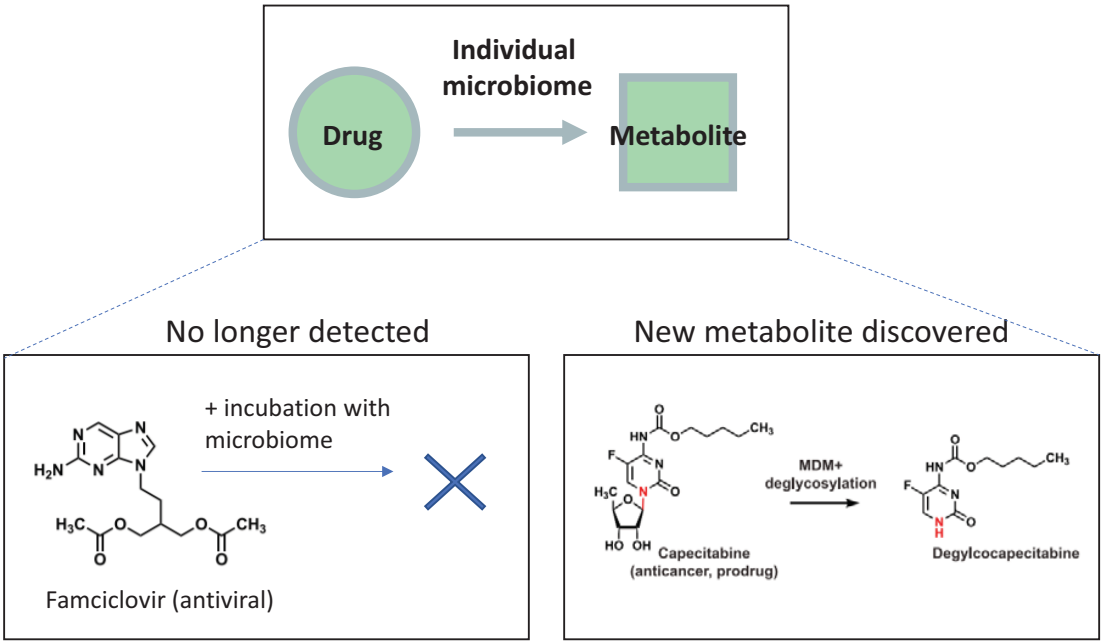


<u>Many common diseases</u> <u>associated with microbiome changes</u>

Allergy
Constipation
Migraine
T2D
Asthma
Hypertension
…

REF | Matthew A Jackson et al., Nature Communications (2018)

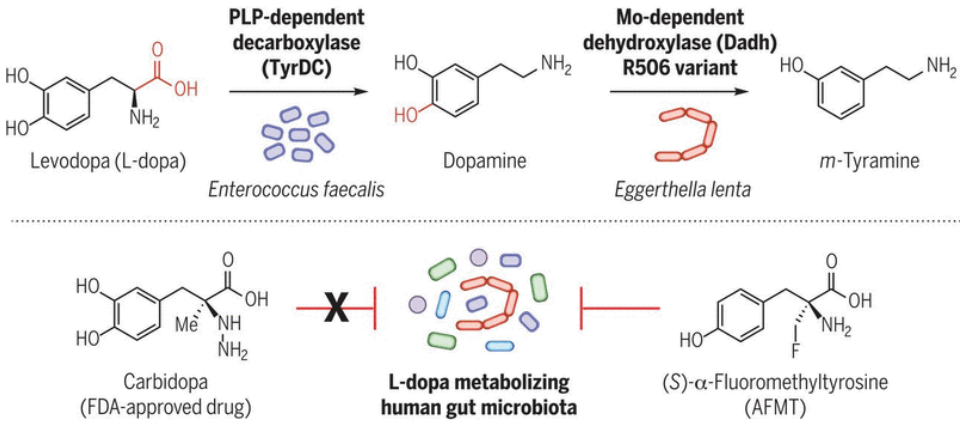
2 Microbiome affects individual drug response



REF | Bahar Javdan et al., Cell, 2020

2 Microbiome affects individual drug response

Microbiome affects Levodopa efficiency



REF | Vayu Maini Rekda et al,. Science 2019

3 Healthy microbiome can treat diseases

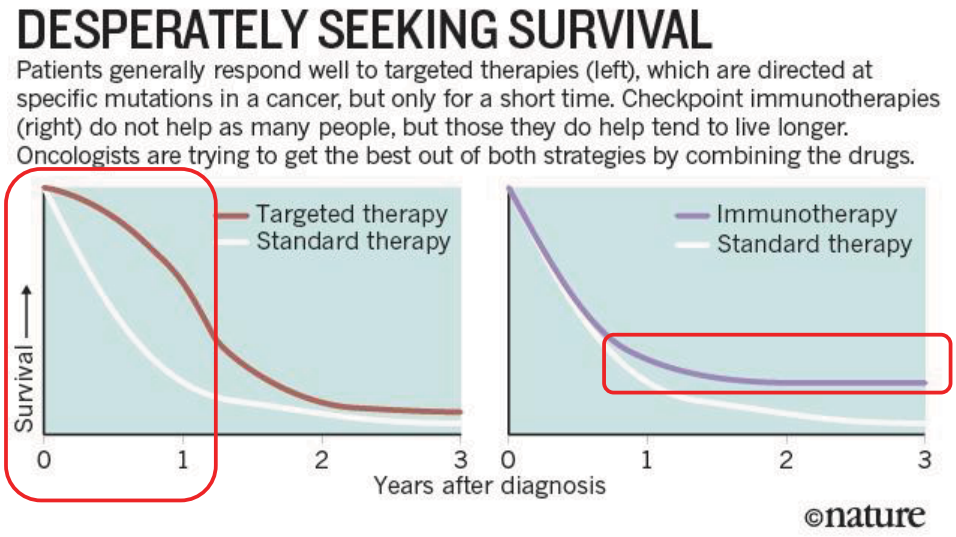
Faecal microbiota transplant

Colitis, C. difficile infection, etc

Healthy colon

REF | Matthew A Jackson et al., Nature Communications (2018)

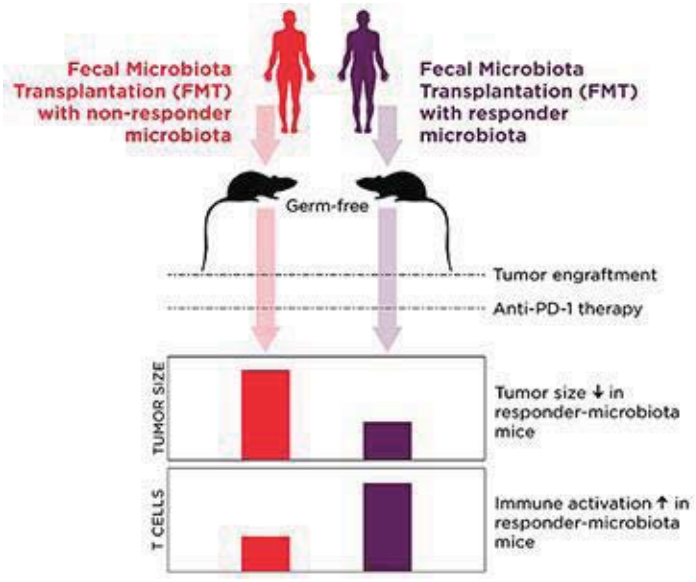
4 Microbiome affects immunotherapy efficacy



REF | Bertrand Routy et al., Science, 2018

13

4 Microbiome affects immunotherapy efficacy



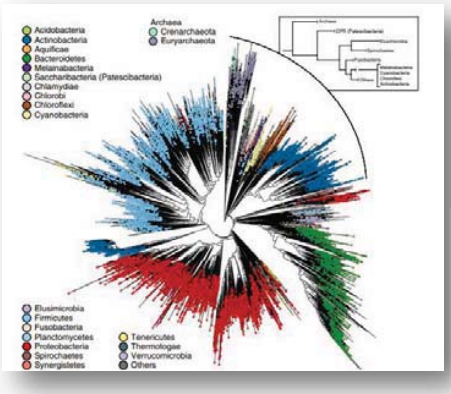
- Fecal microbiota of responders increased the efficacy of immunotherapy!

REF | Bertrand Routy et al., Science, 2018

14

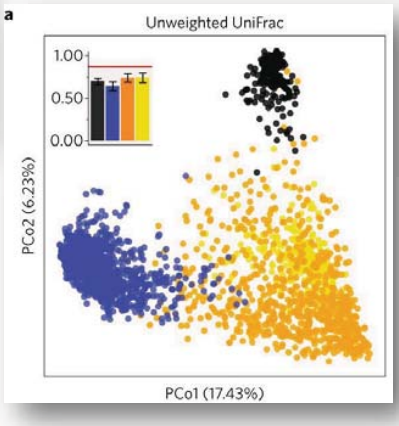
How to study microbiome?

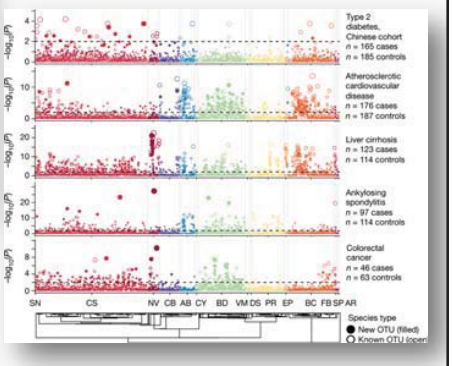
15



Phylogenetic trees

Principal coordinate analysis





Metagenome-wide association study

16

Metagenomic approaches

- 16S rRNA sequencing (ribotyping)
  - Amplicon sequencing of 16S rRNA regions
  - Economical costs
  - Taxonomic profiling
- Whole genome shotgun (WGS) methods
  - Species/strain-level in-depth analysis
  - Extensive costs
  - Taxonomic & functional profiling

17

Overview

- **Theoretical Backgrounds**
  - Key definitions
  - Taxonomy
  - Phylogeny
  - Diversity
  - Dysbiosis
- **Bioinformatics analysis**
  - 16S rRNA amplicon sequencing = DADA2
  - Shotgun metagenome analysis = MetaPhlan
- **Prerequisite**
  - R programming skills

18

Key definitions

19

Key definitions …

Microbe?

Microbiome?

Metagenome?

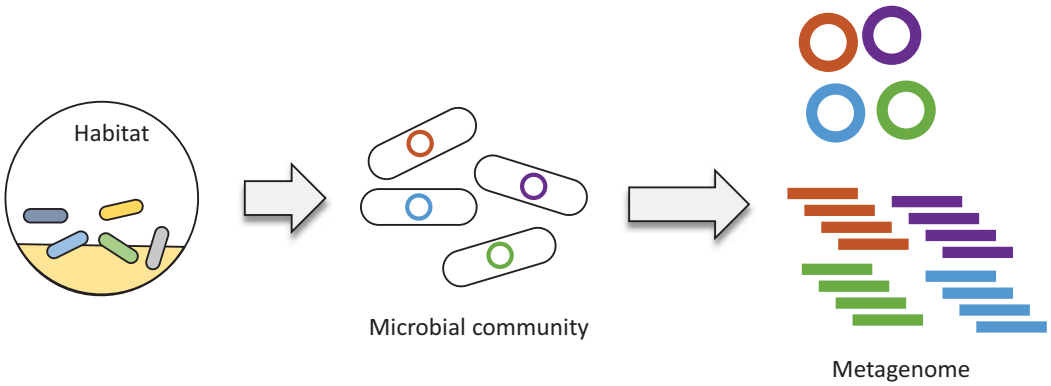
Microorganisms?

Microbiota?

20

Key definitions …

- **Metagenome:** study of genomes of whole biological communities from a particular habitat
- **Habitat**: specific site of organism growth



21

Key definitions …

- **Microbiota**:
  ecological communities of symbiotic and pathogenic **microorganisms** found in and on all multicellular organisms (e.g. vertebrates)

22

Key definitions …

- **Microbiome:**
  genomes of all microorganisms, symbiotic and pathogenic, **living in and on multicellular organism** (e.g. vertebrates)

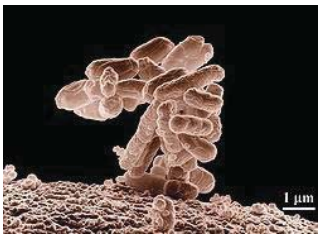
- i.e. **= metagenome of microbiota**

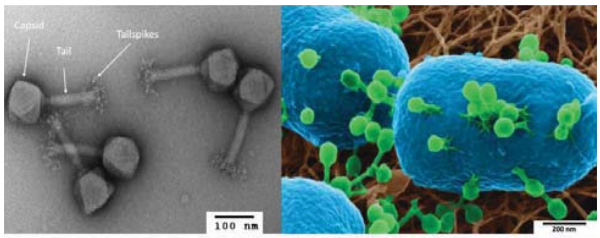
23

Microorganisms (미생물)

- Microorganisms = microbes = microscopic organisms with singled-cell form

  e.g. bacteria, archaea, fungi, virus, and protozoan

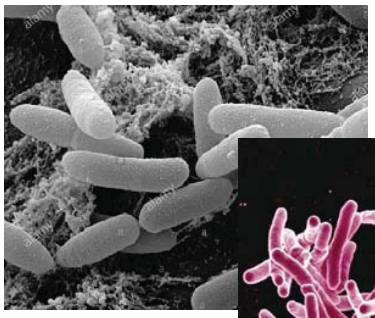




(a) (b)

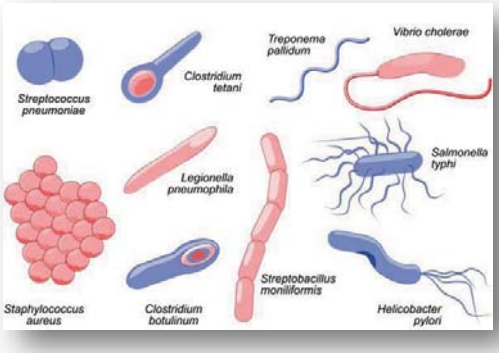
24

Bacteria (세균)



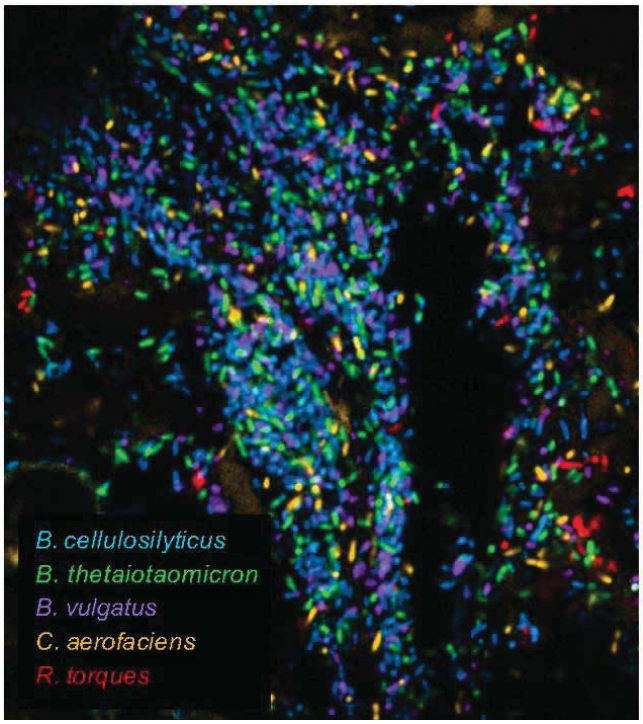
Escherichia coli

Mycobacterium tuberculosis



25

Bacteria (세균)

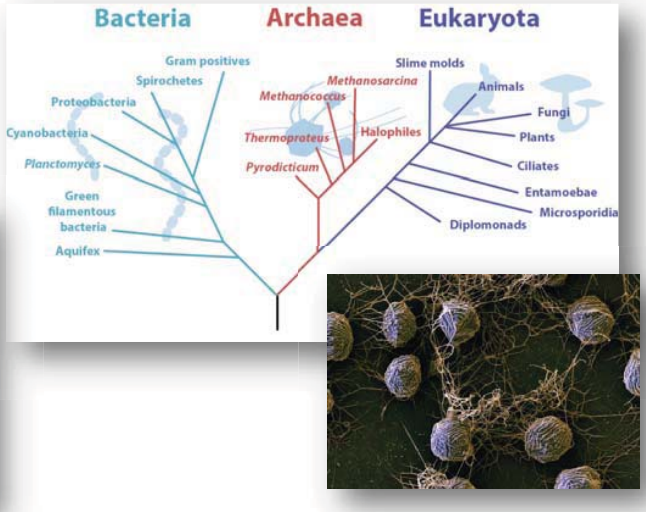


Spatial organization of human gut microbiota established in mice

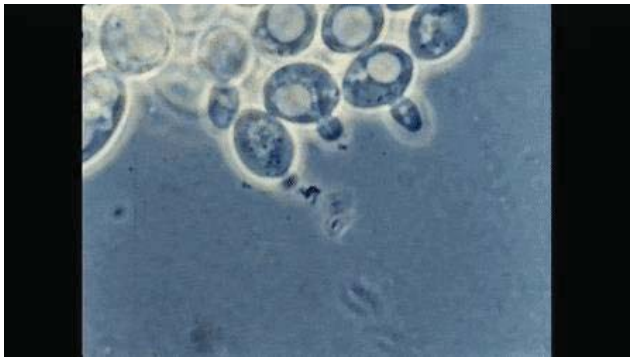
REF | Jessica L. Mark Welch et al., PNAS, 2017

26

Archaea (고균)



Fungi (진균)



Budding yeast

Candida species

Key terminology…

Taxonomy?

Phylogeny?

Diversity …?

Symbiosis…

Dysbiosis…

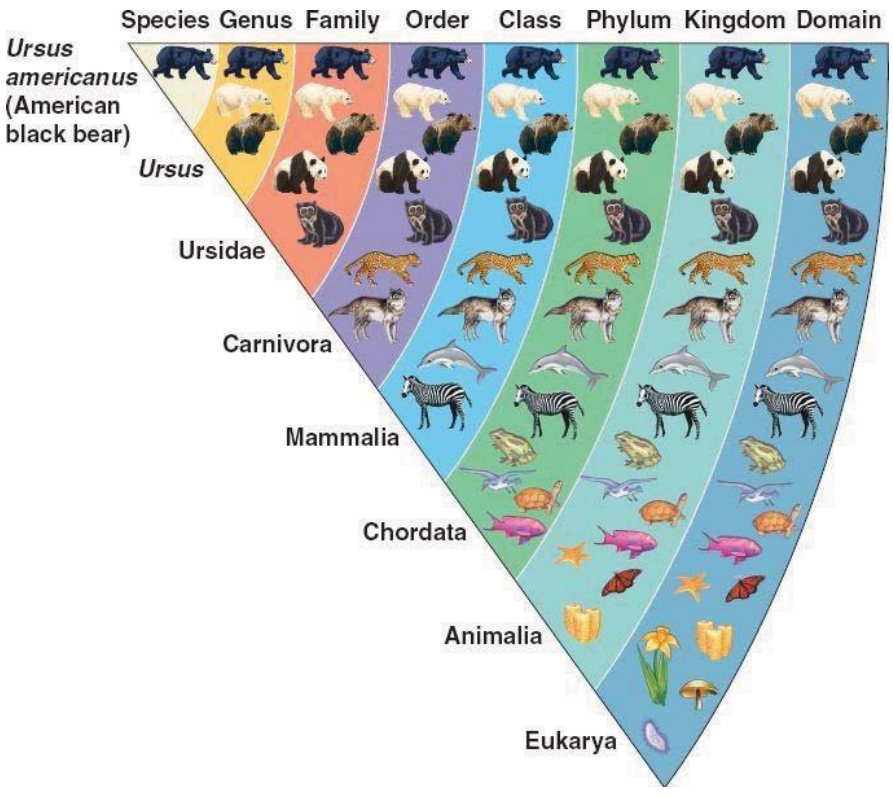
29

Taxonomy!

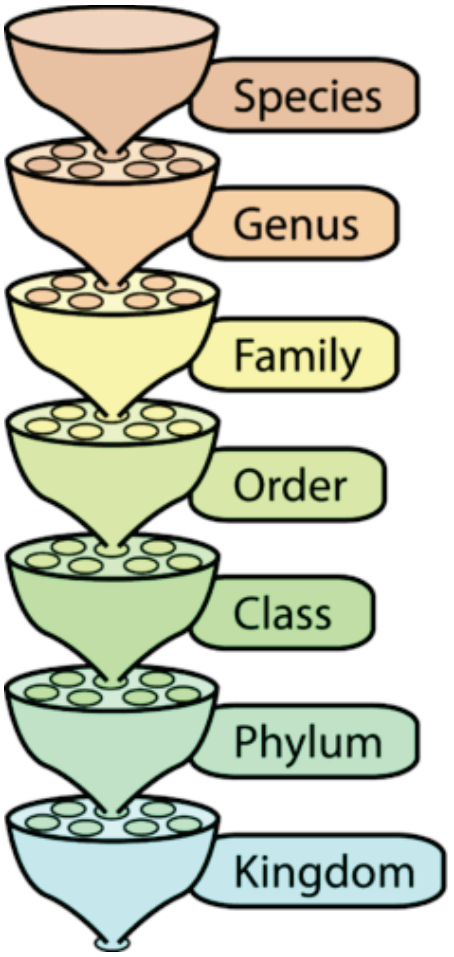
30

Taxonomy = classification = identity

31



32



Homo sapiens
Member of the genus Homo with a high forehead and thin skull bones.

Homo
Hominids with upright posture and large brains.

Hominids
Primates with relatively flat faces and three-dimensional vision.

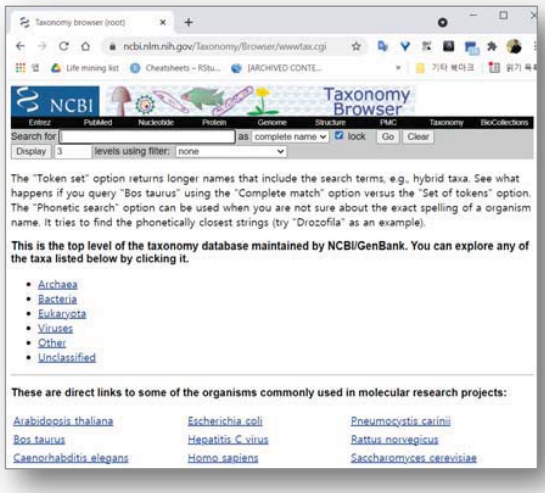
Primates
Mammals with collar bones and grasping fingers.

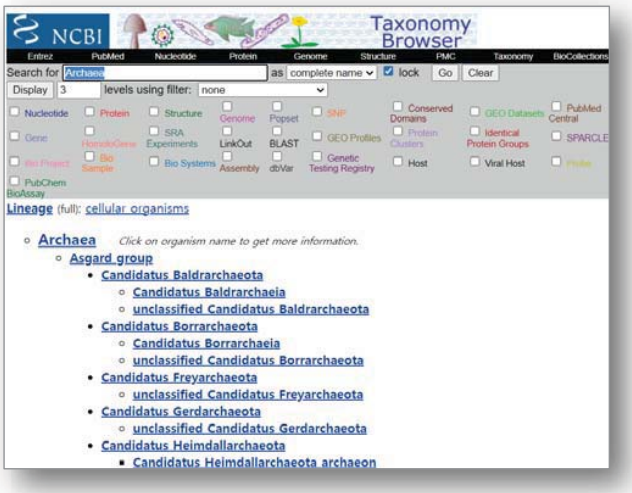
Mammals
Chordates with fur or hair and milk glands.

Chordates
Animals with a backbone.

Animals
Organisms able to move on their own.

Taxonomy database





NCBI taxonomy database:
https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

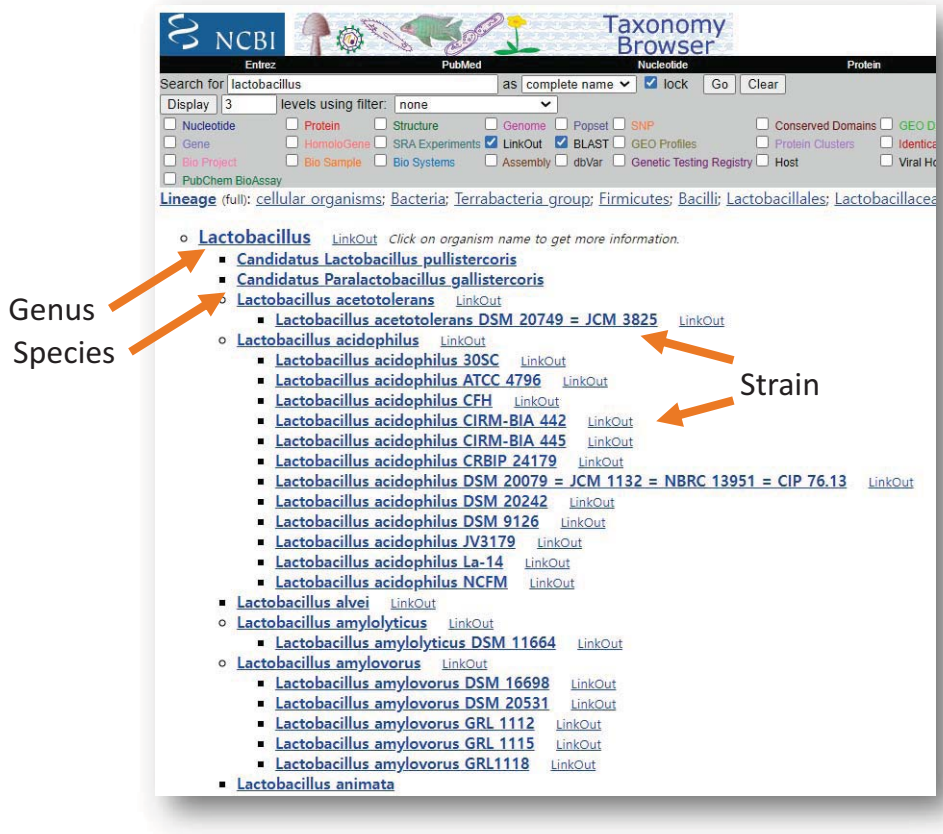
Taxonomy database

락토바실러스 유산균?

Taxonomy database

락토바실러스 (Lactobacillus) = genus 명칭

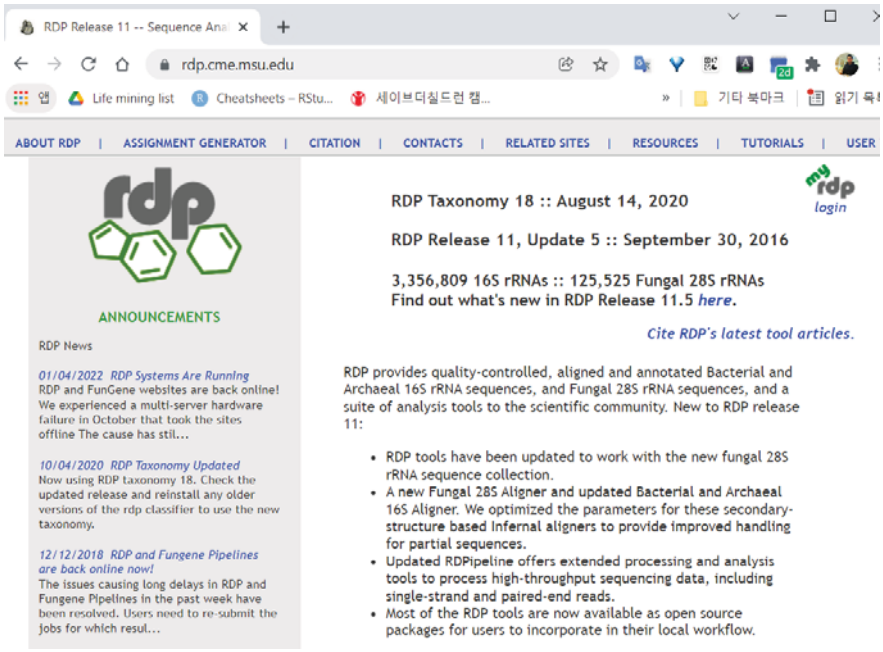
Taxonomy database



Database of 16S rRNA with taxonomy data

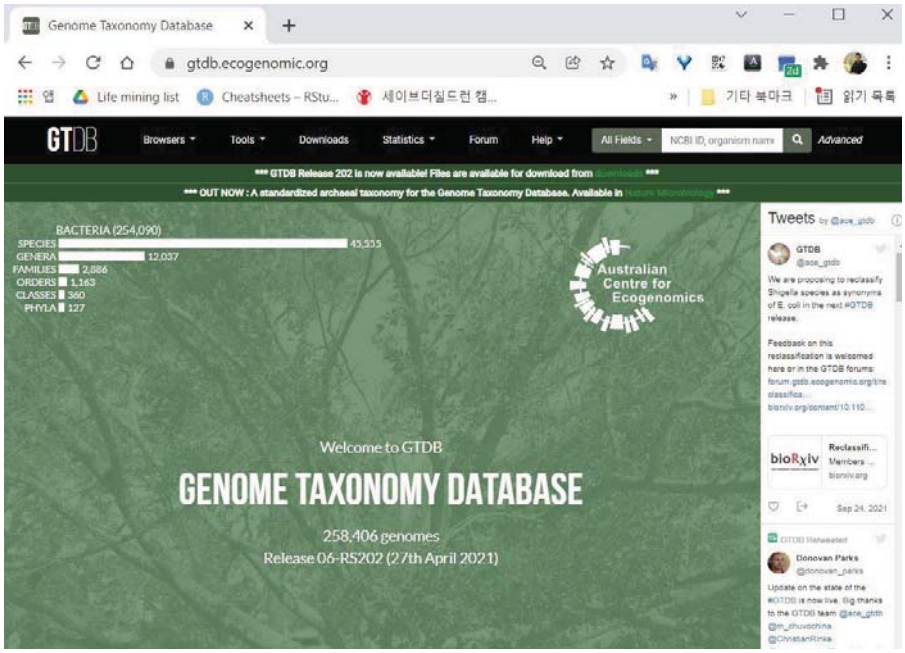
SILVA

Database of 16S rRNA with taxonomy data



RDP

Database of genomes with taxonomy data



GTDB

Phylogeny?

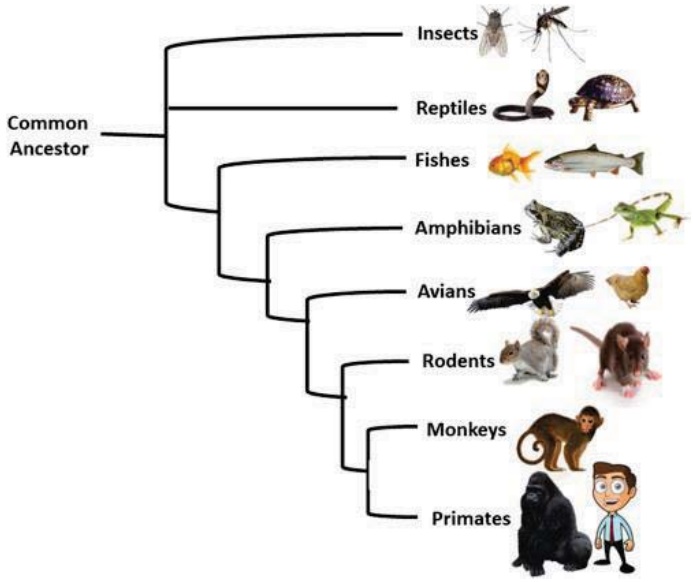
41

Phylogeny = evolutionary relationship

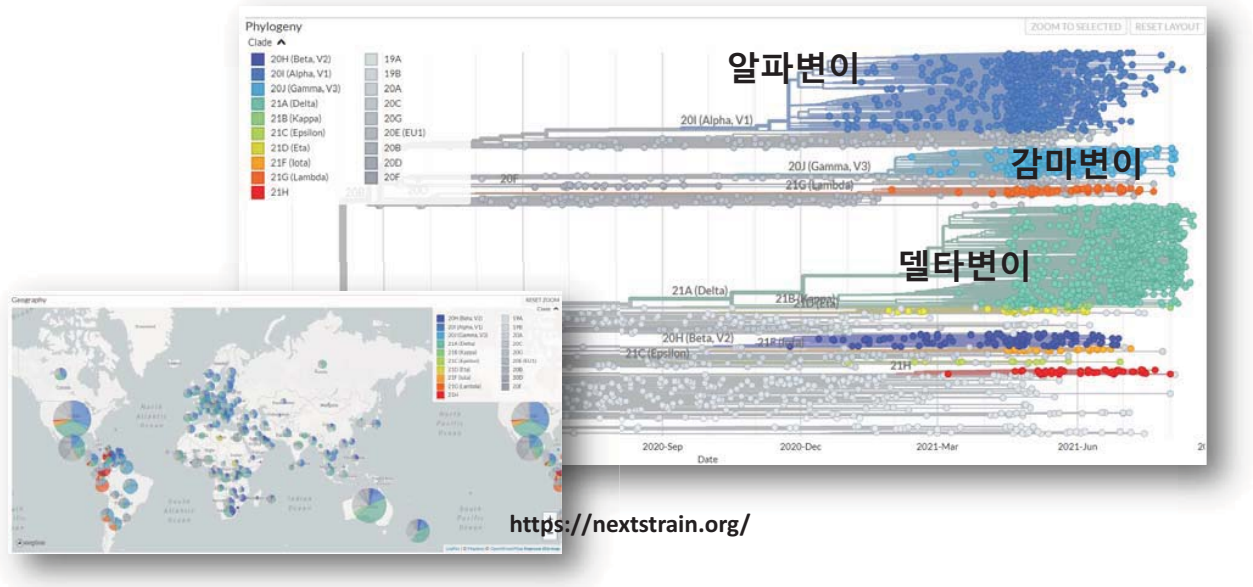
42

Phylogeny

- Evolutionary history between organisms

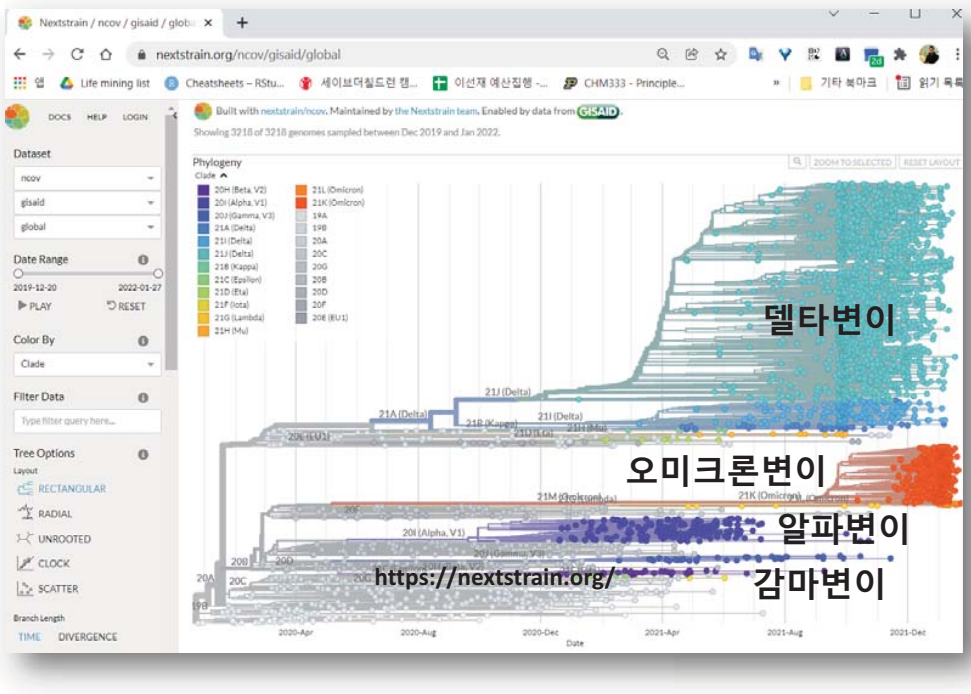


Evolutions of SARS-CoV-2 strains



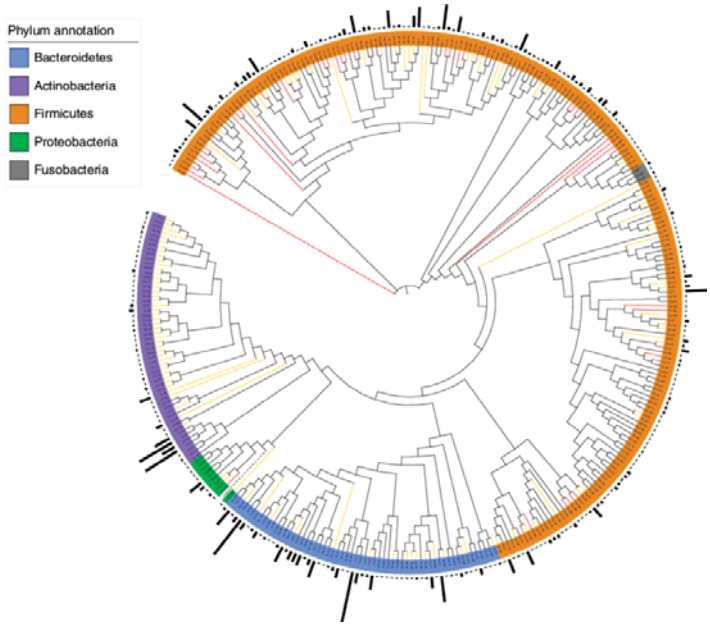
7개월전 데이터…

Evolutions of SARS-CoV-2 strains



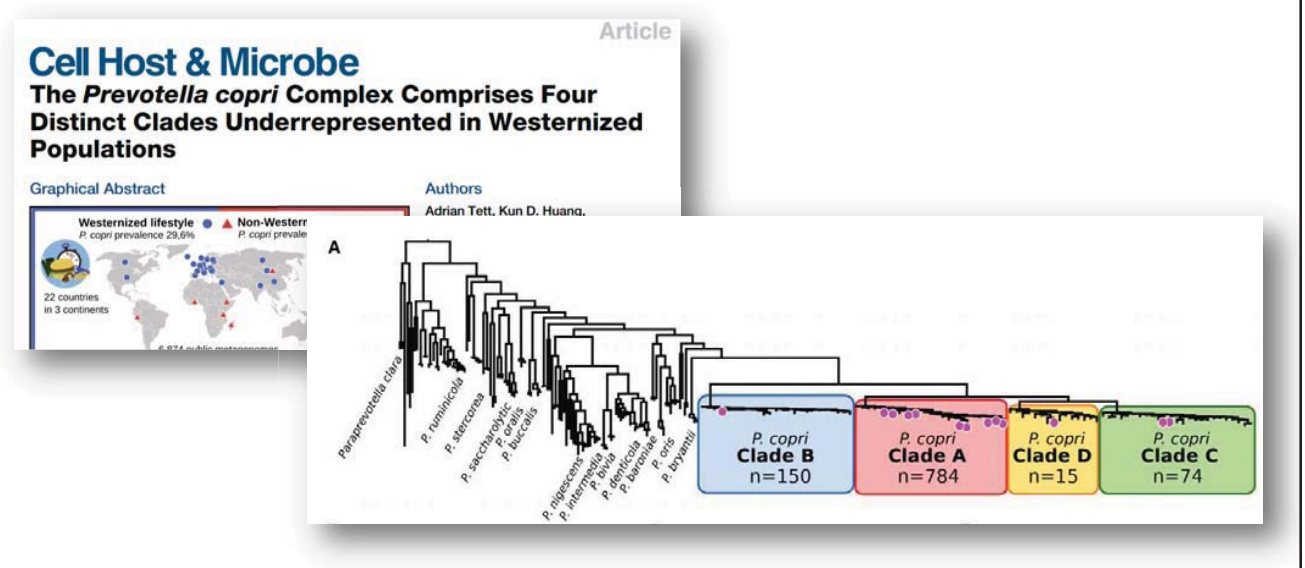
현재 데이터...

Phylogenetic trees of isolated gut bacteria



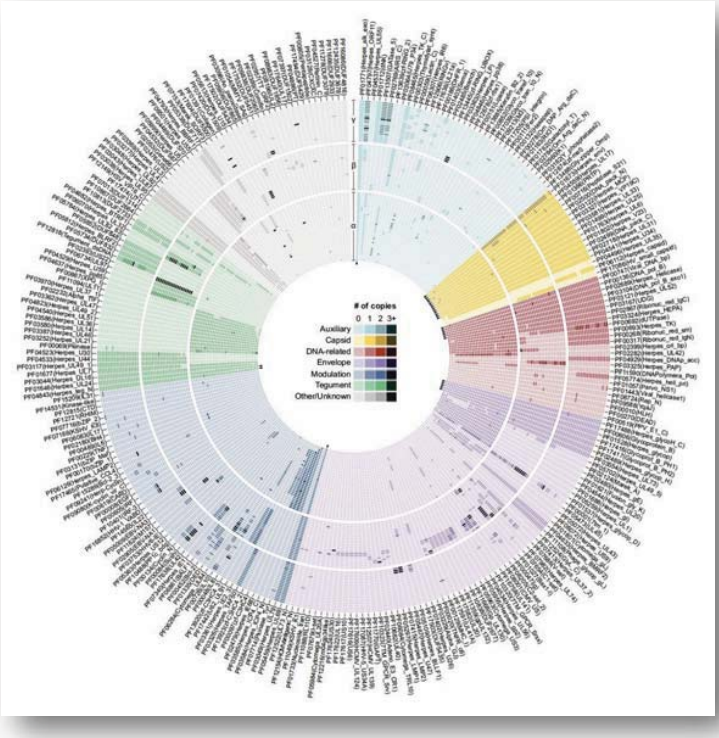
REF | Yuanqiang Zhou et al., Nature Biotechnology, 2019

Phylogenetic trees of “metagenome-assembled genomes”



REF | Yuanqiang Zhou et al., Nature Biotechnology, 2019

Website - iTOL: interactive Tree of Life

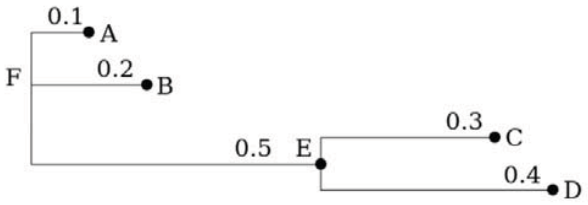


REF | https://itol.embl.de/

Tree file format

- Newick format
- Nexus format
- PhyloXML format

Newick format




```
(,,(,));                                no nodes are named
(A,B,(C,D));                            leaf nodes are named
(A,B,(C,D)E)F;                          all nodes are named
(:0.1,:0.2,(:0.3,:0.4):0.5);            all but root node have a distance to parent
(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;        all have a distance to parent
(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);        distances and leaf names (popular)
(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;      distances and all names
((B:0.2,(C:0.3,D:0.4)E:0.5)F:0.1)A;     a tree rooted on a leaf node (rare)
```

Nexus format

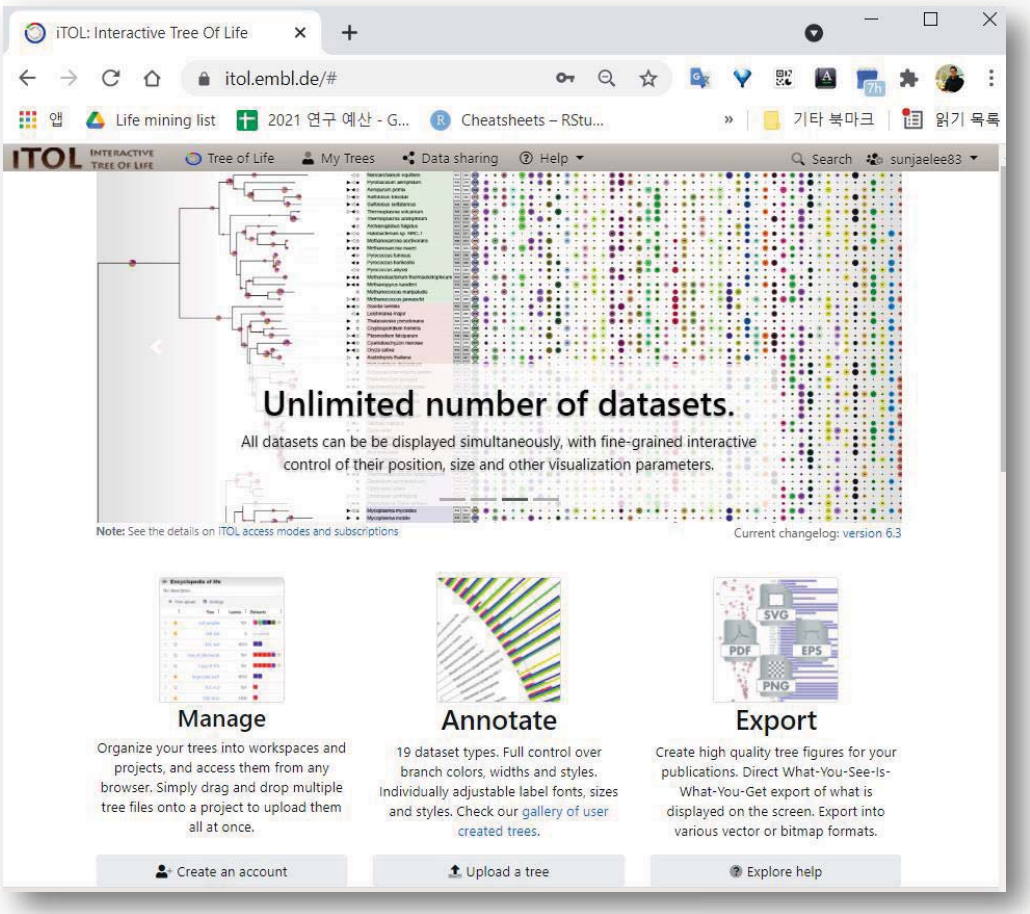
```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf
End;

Begin data;
  Dimensions nchar=15;
  Format datatype=dna missing=? gap=- matchchar=.;
  Matrix
    [ When a position is a "matchchar", it means that it is the same as the first entry at the same position. ]
    SpaceDog   atgctagctagctcg
    SpaceCat   ......??...-.a.
    SpaceOrc   ...t.......-.g. [ same as atgttagctag-tgg ]
    SpaceElf   ...t.......-.a.
  ;
End;

BEGIN TREES;
  Tree tree1 = (((SpaceDog,SpaceCat),SpaceOrc,SpaceElf));
END;
```

TAXA + DATA + TREES

PhyloXML format

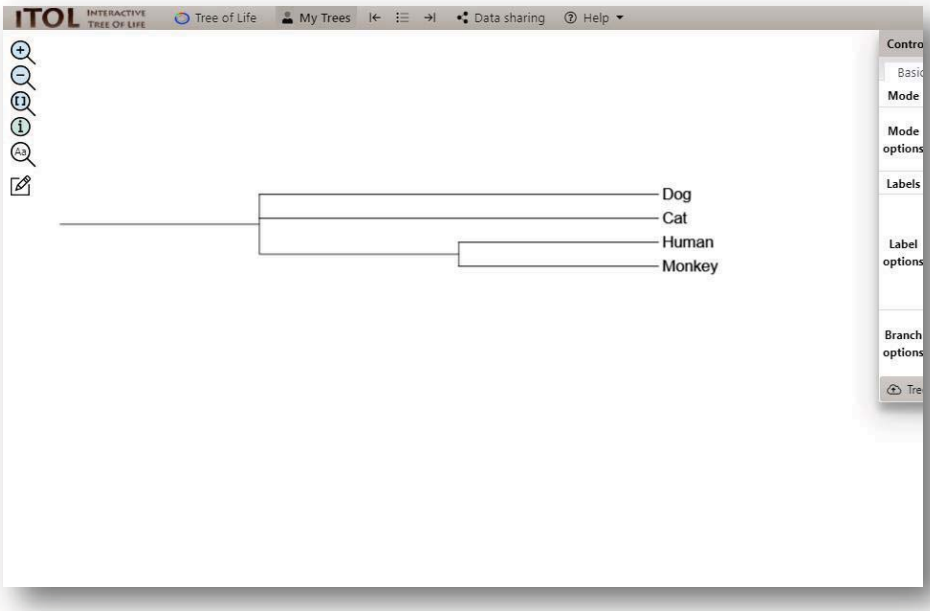
```
<phyloxml xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.phyloxml.org http://www.phyloxml.org/1.10/phyloxml.xsd"
  xmlns="http://www.phyloxml.org">
  <phylogeny rooted="true">
    <name>example from Prof. Joe Felsenstein's book "Inferring Phylogenies"</name>
    <description>MrBayes based on MAFFT alignment</description>
    <clade>
      <clade branch_length="0.06">
        <confidence type="probability">0.88</confidence>
        <clade branch_length="0.102">
          <name>A</name>
        </clade>
        <clade branch_length="0.23">
          <name>B</name>
        </clade>
      </clade>
      <clade branch_length="0.5">
        <name>C</name>
      </clade>
    </clade>
  </phylogeny>
</phyloxml>
```

Customized XML format



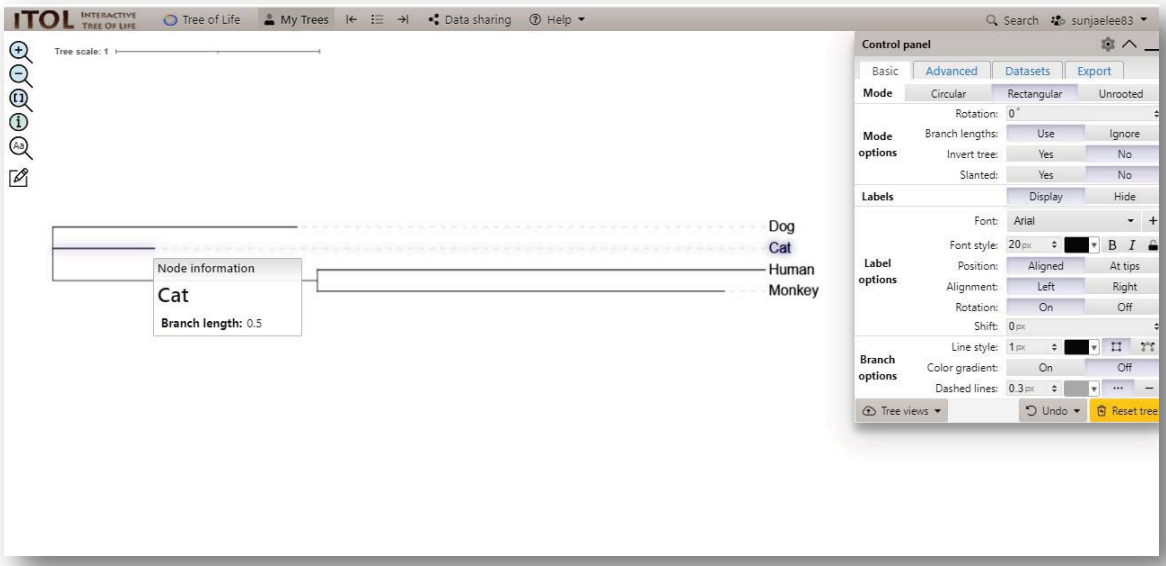
Example

(Cat,Dog,(Monkey,Human));



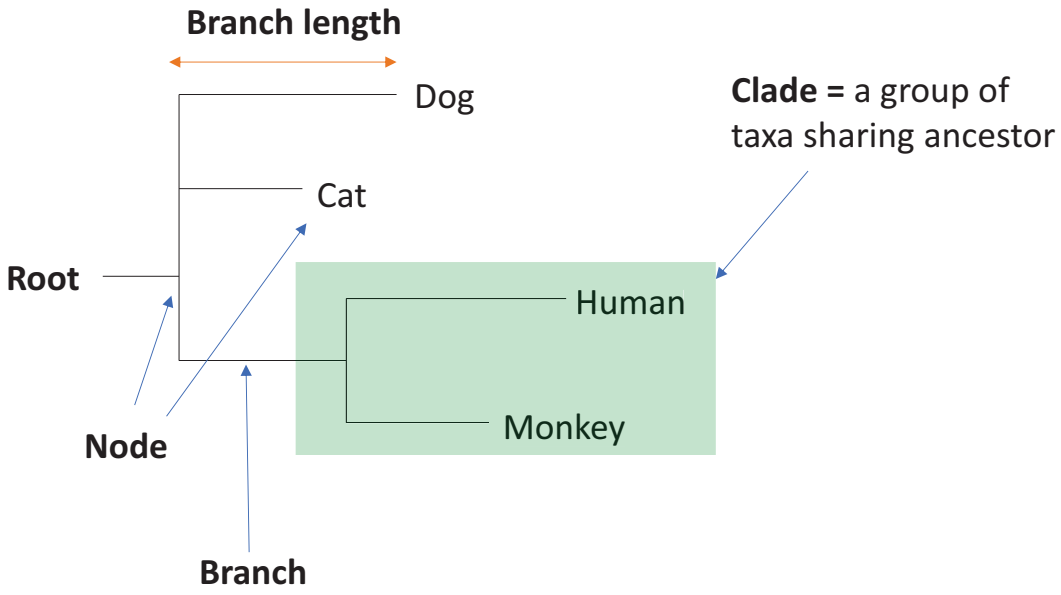
Example

(Cat:0.5,Dog:1.2,(Monkey:2,Human:2.2):1.3);



55

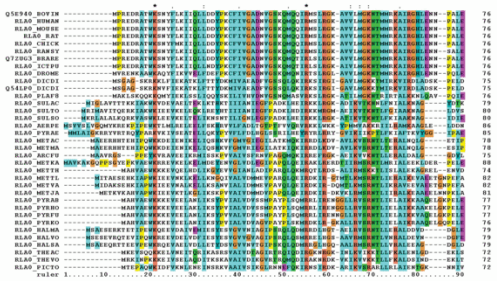
Phylogenetic tree

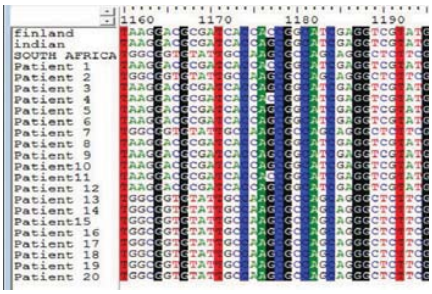


56

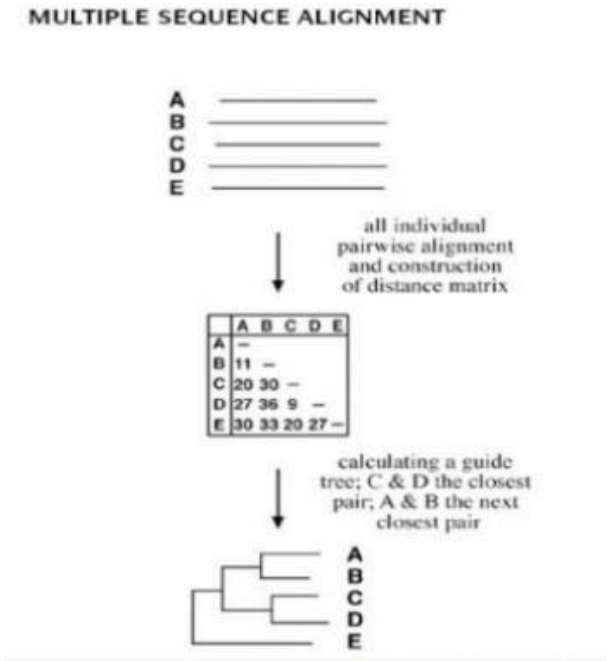
Phylogeny – sequence alignment

- Inferring regions of homology
  - 16S rRNA, house-keeping genes (multi locus sequence typing) etc
- Tools: Clustal, MUSCLE, T-Coffee, etc





Building distance matrix



REF | https://www.slideshare.net/ArghadipSamanta1/multiple-sequence-alignmentjust-glims-of-viewes-on-bioinformatics

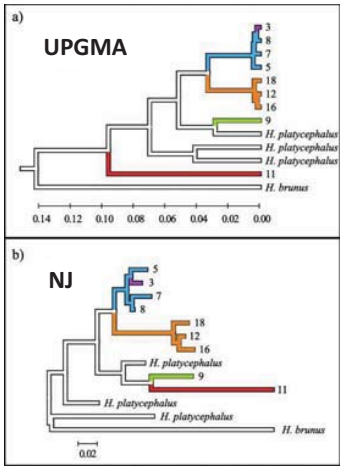
Phylogenetic tree methods

Distance-based methods

- Neighbour-joining method
- UPGMA method
- Minimum evolution method

Criterion-based methods

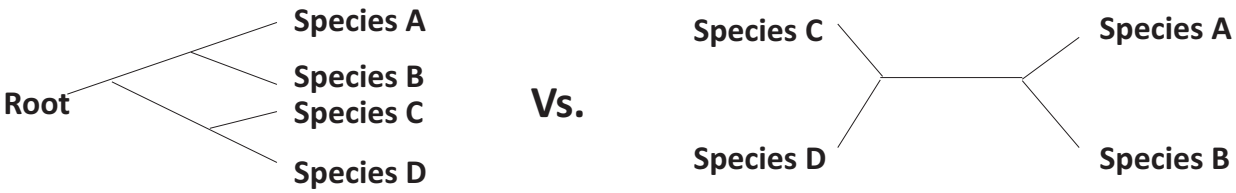
- Seeing sequences as characters
- Maximum-likelihood method
- Maximum parsimony method



REF | Robert E Bingham et al., Bulletin of the Museum of Comparative Zoology, 2018

Phylogenetic tree

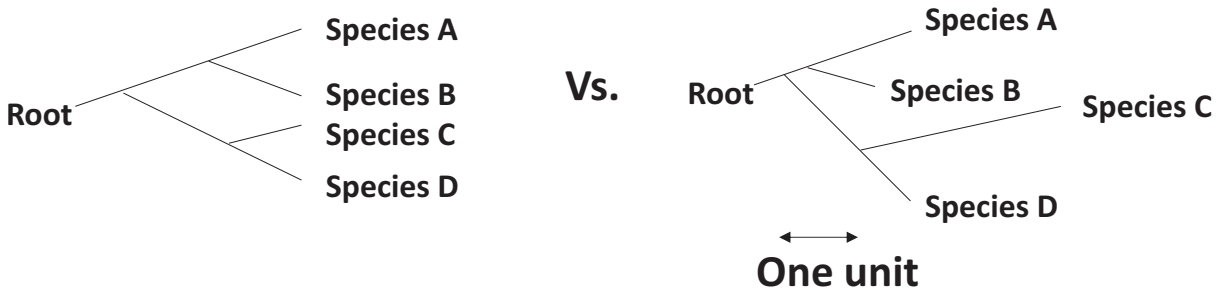
Rooted vs unrooted trees



UPGMA **NJ**

Phylogenetic tree

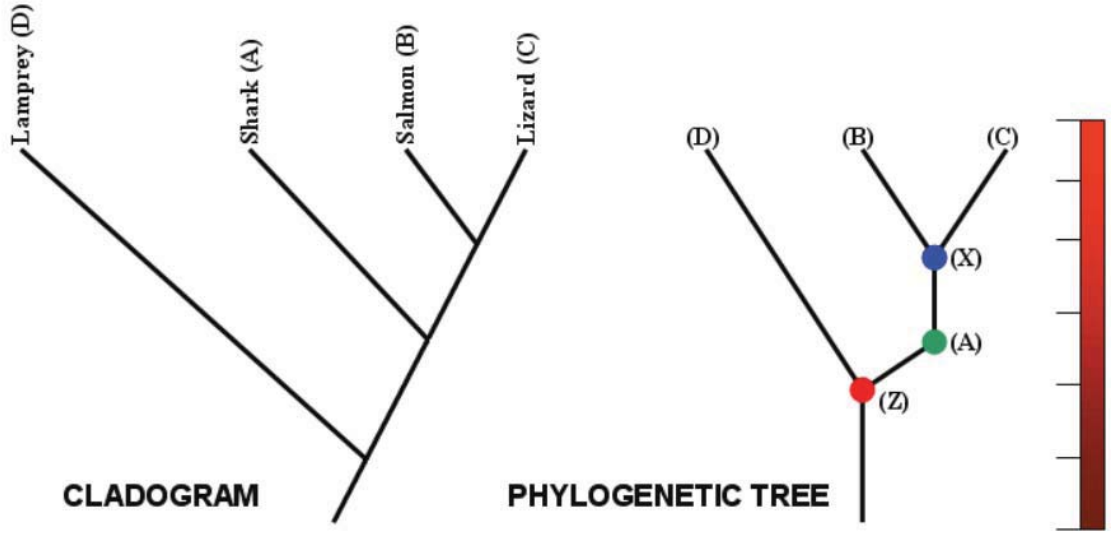
Scaled vs unscaled branches (i.e. with or without branch lengths)



61

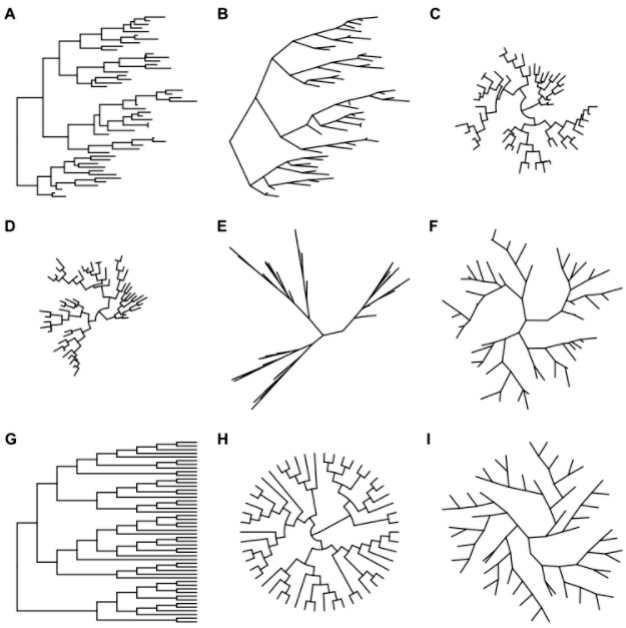
Phylogenetic tree vs cladogram

No meaning on branch lengths



62

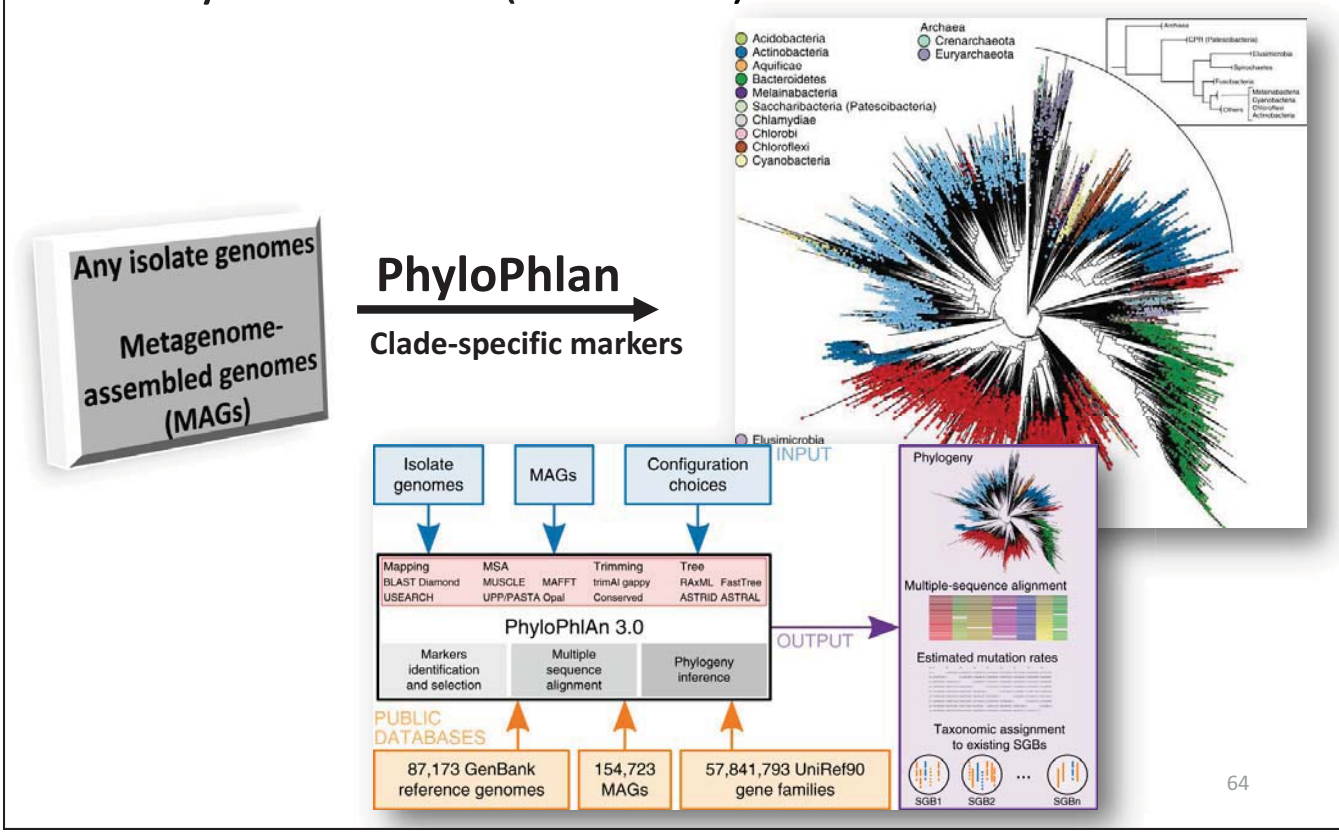
R package, ggtree




```
library(ggtree)
set.seed(2017-02-16)
tree <- rtree(50)
ggtree(tree)
ggtree(tree, layout="slanted")
ggtree(tree, layout="circular")
ggtree(tree, layout="fan", open.angle=120)
ggtree(tree, layout="equal_angle")
ggtree(tree, layout="daylight")
ggtree(tree, branch.length='none')
ggtree(tree, branch.length='none', layout='circular')
ggtree(tree, layout="daylight", branch.length = 'none')
```

REF | https://guangchuangyu.github.io/ggtree-book/chapter-ggtree.html

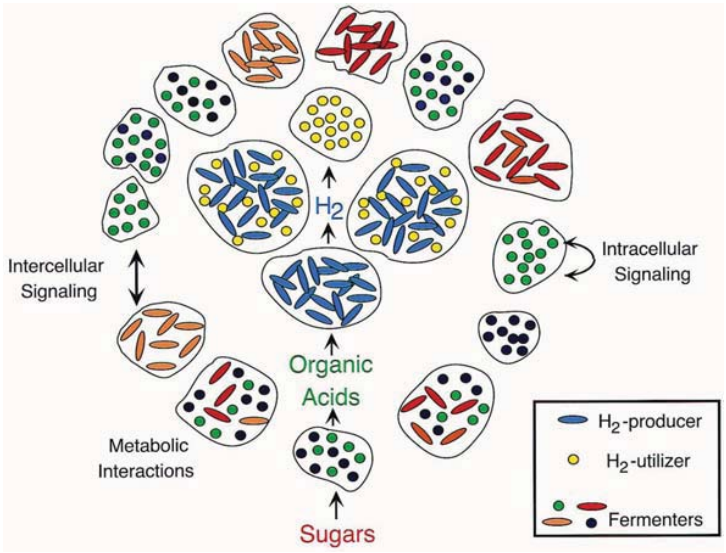
PhyloPhlan (ver3.0)



Diversity

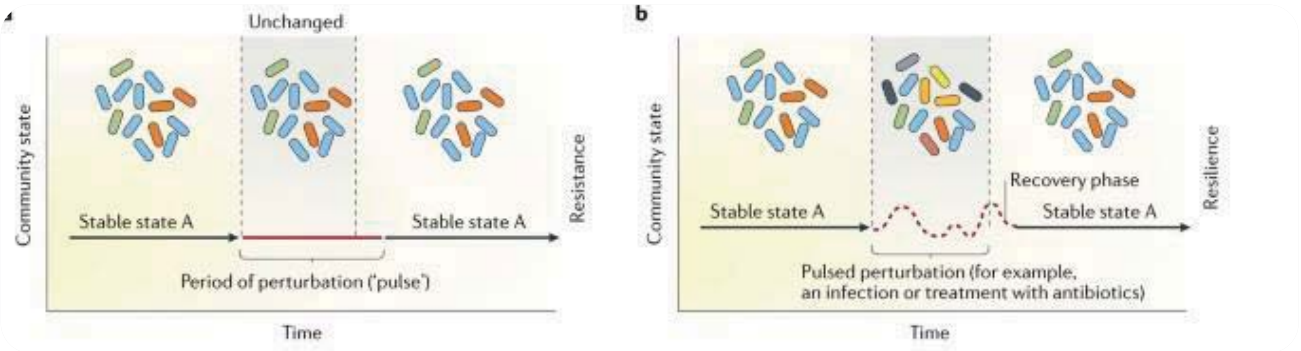
Microbes are living as a community

Microbial biofilm



Why microbial diversity?

- **Insurance hypothesis**
  - Biodiversity ensures ecosystems against decreases in their functionality
  - In gut, rich diversity is also crucial and protective for sustaining a microbial equilibrium



67

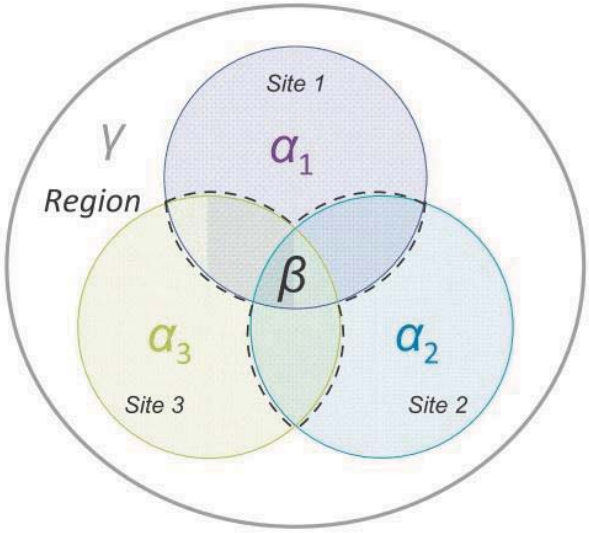
Key terminology

- Diversity
- Richness + Evenness
- Coverage

68

Diversity

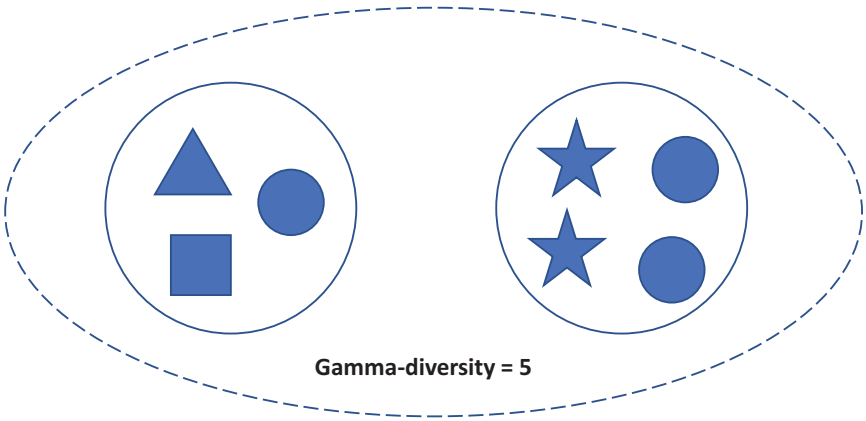
- Alpha-diversity
- Gamma-diversity
- Beta-diversity



69

Diversity

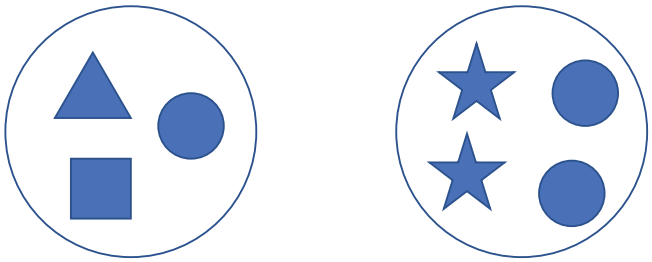
- Gamma-diversity = total diversity in a landscape
  = alpha + beta diversity



70

Diversity

- Alpha-diversity = diversity within ecological units or habitats



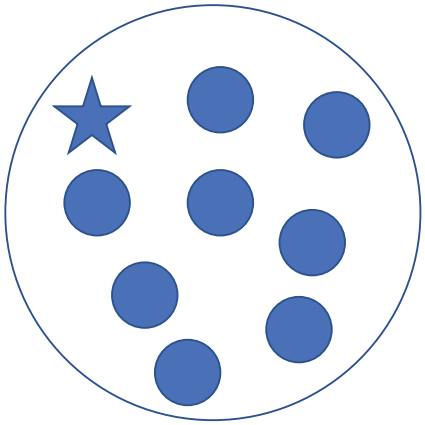
Alpha-diversity = 3

Alpha-diversity = 2

71

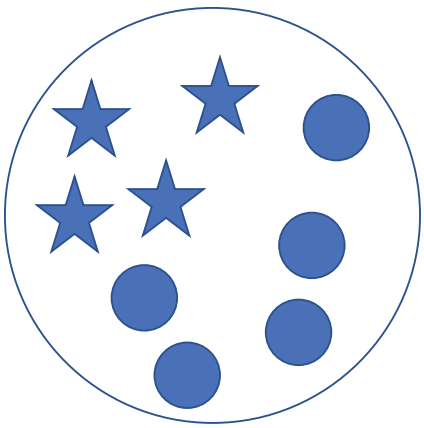
Diversity

Unit1



Vs.

Unit2



Which unit has a higher diversity?

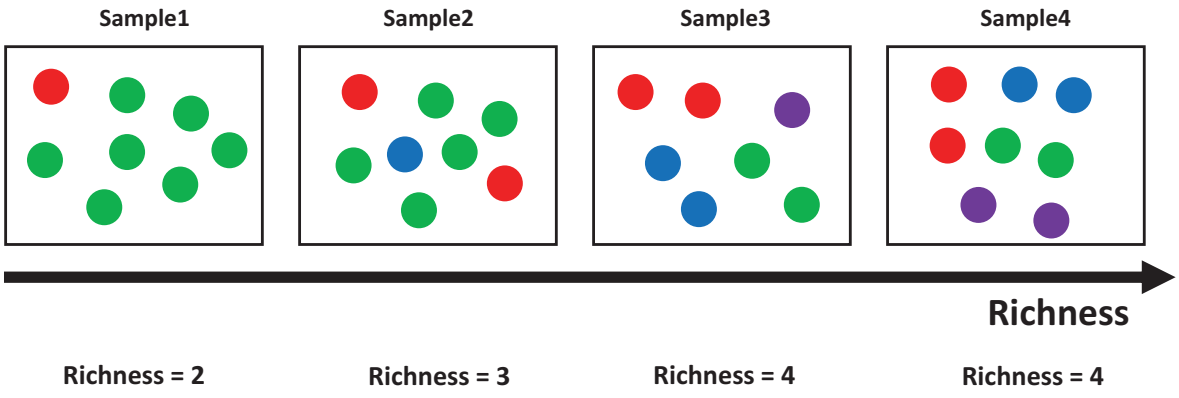
72

Diversity

- **Diversity** = measure of **richness** + **evenness**

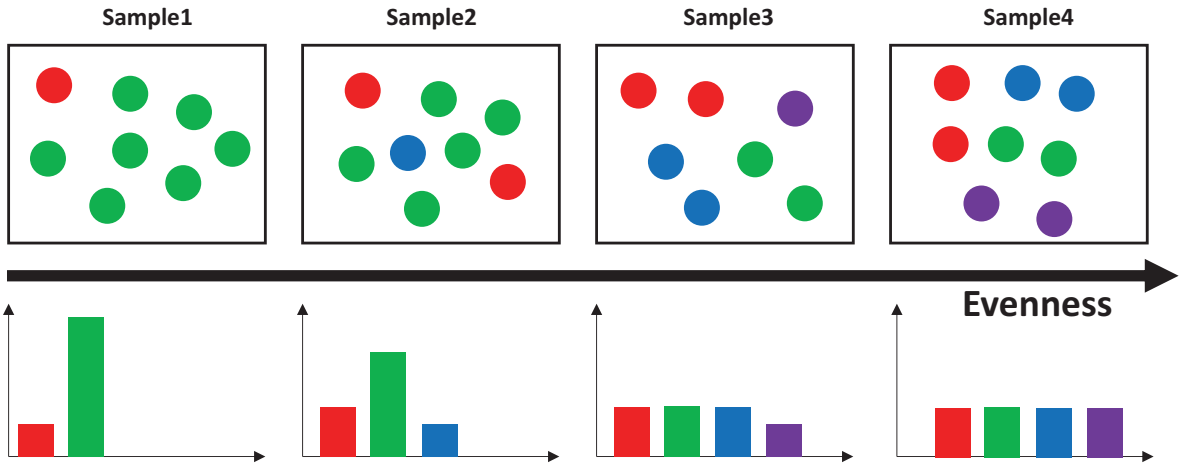
- Richness = number of species present
- Evenness = measuring how different species in community are similar in numbers

Diversity



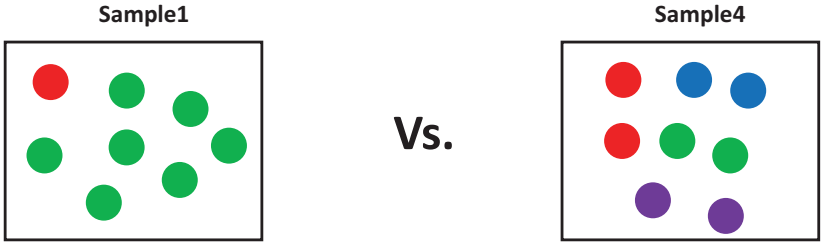
REF | http://www.evolution.unibas.ch/walser/bacteria_community_analysis

Diversity



75

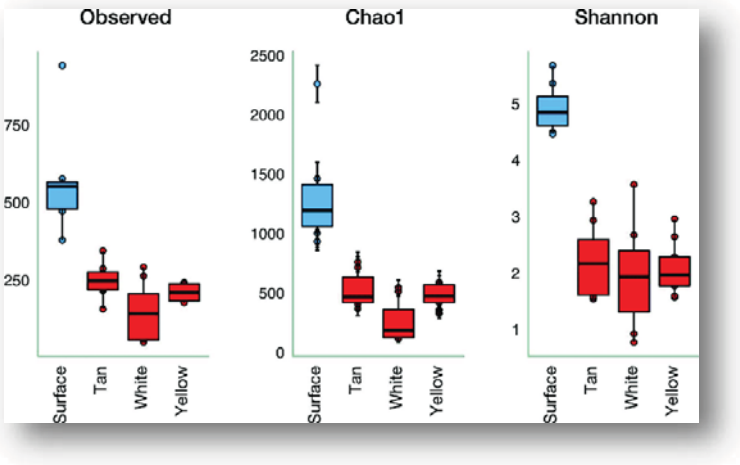
Diversity



76

Diversity

- **Alpha-diversity comparison →** normally, mean alpha diversity measures are compared



77

Diversity

- Well-known alpha diversity indices
  - Shannon & Inverse Simpson

Shannon

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

Inverse Simpson

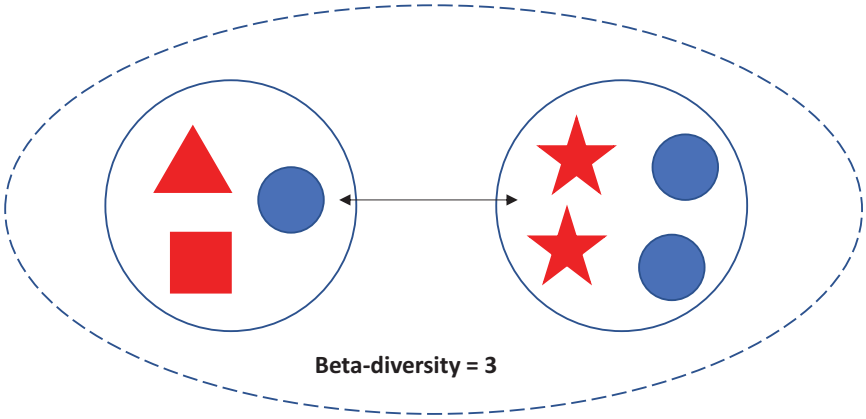
$$\frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{R} p_i^2} = {}^2D$$

P_i = the proportion of samples belonging to i^{th} species in the dataset
R = a number of species, i.e. richness

78

Diversity

- Beta-diversity = differences in diversity between habitats



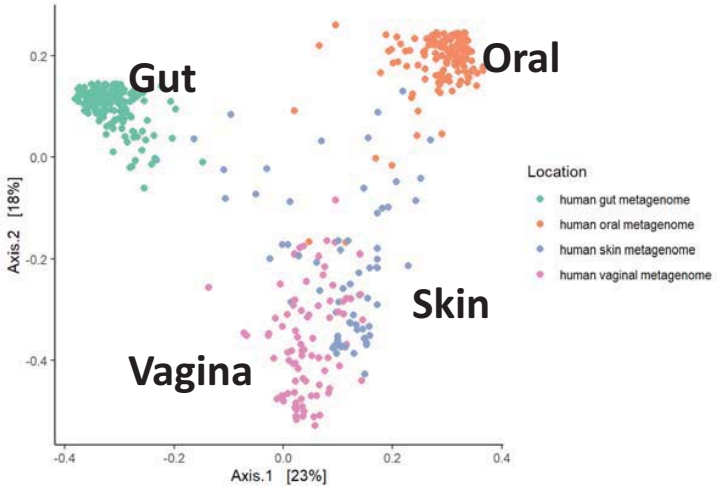
Diversity

Beta-diversity comparison

- Comparing compositional differences

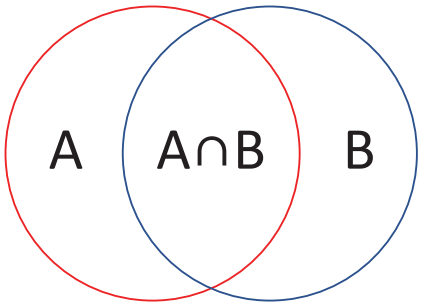
Popular Beta-diversity measure

	Categorical	Phylogenetic
Presence/ absence	Jaccard	Unifrac
Abundance	Bray-Curtis	Weighted unifrac



Diversity

Jaccard distance
= fraction of shared types
= 1 – (A ∩ B)/(A U B)

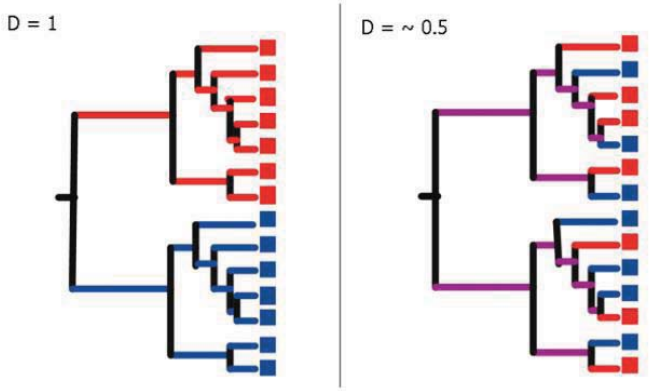


Diversity

Bray-Curtis distance
= sum of absolute differences over total abundance
= ∑|xi - yi | / (∑xi + ∑yi)
= ■ / (■ + ■)

Diversity

Unifrac distance
= fraction of unshared branch lengths over tree
= (■ + ■) / (■ + ■ + ■)

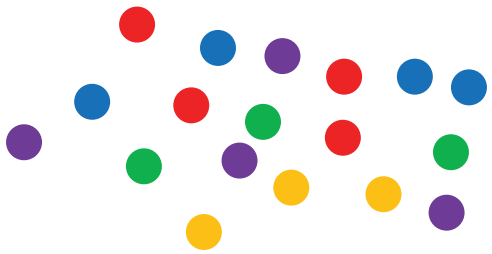


REF | Benjamin Callahan lecture

***weighted unifrac considers abundance changes**

Coverage

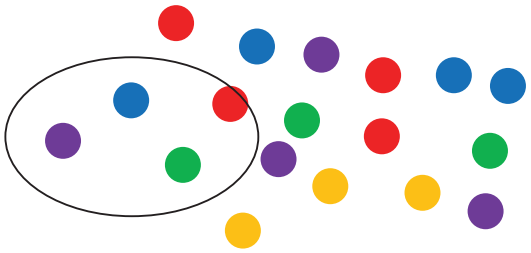
- Sampling is inevitable



Total richness = 5

Coverage

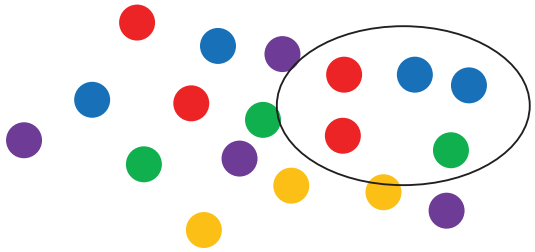
- Sampling is inevitable



Richness = 4

Coverage

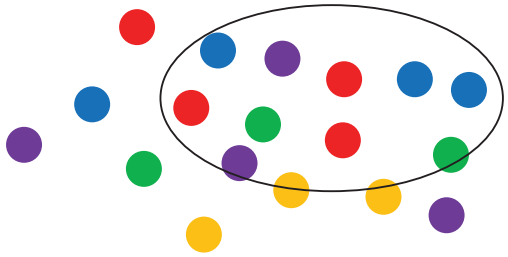
- Sampling is inevitable



Richness = 3

Coverage

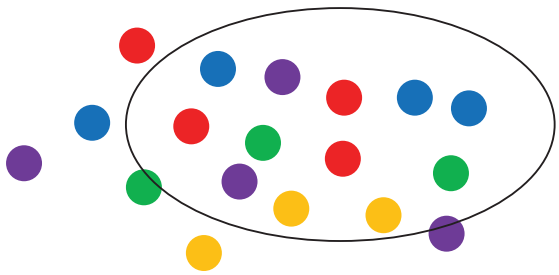
- Sampling is inevitable



Richness = 4

Coverage

- Sampling is inevitable

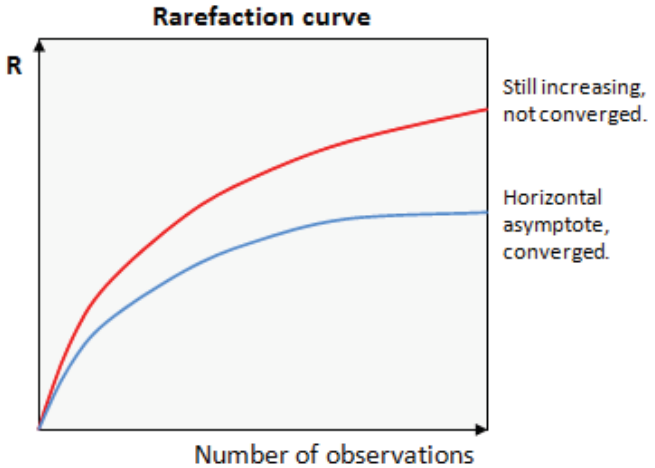


Richness = 5

Coverage

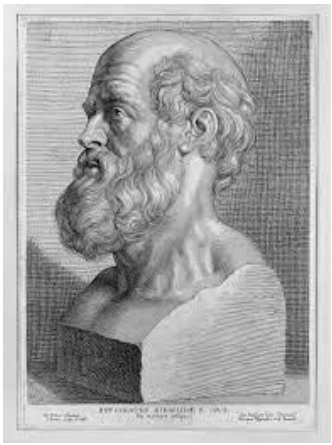
- Sampling is inevitable
- Coverage
  - proportion of community revealed by sampling
  - Let's say 100 species in the community
  - 80 marker gene sequences might give 80% or less coverage
  - 20 marker gene sequences might give 20% or less coverage
- Can be checked with rarefaction curves

Rarefaction curve



Dysbiosis

91



Ancient Greek physician
Hippocrates

"All disease begins in the gut"

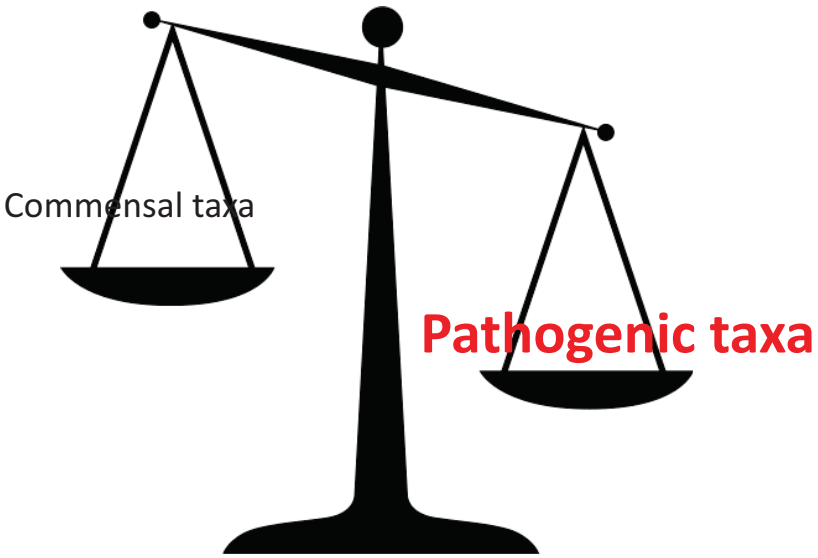
92

Humans are enriched with symbiotic bacteria

- Microbes **outnumbers** human cells
- All the human microbiota roughly **weighs 1-2 kg**
  (= 1-3% of total body mass)
  (= same weight to “liver”)
- Generally non-pathogenic
- **Many are symbiotic**
  - **Commensal**
  - **Opportunistic pathogens**

93

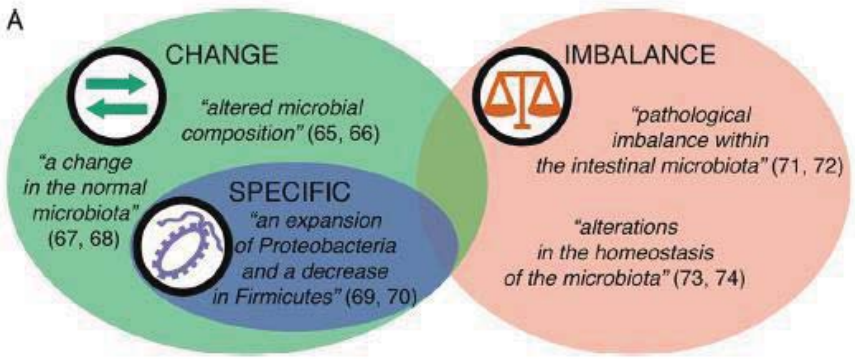
Dysbiosis?



94

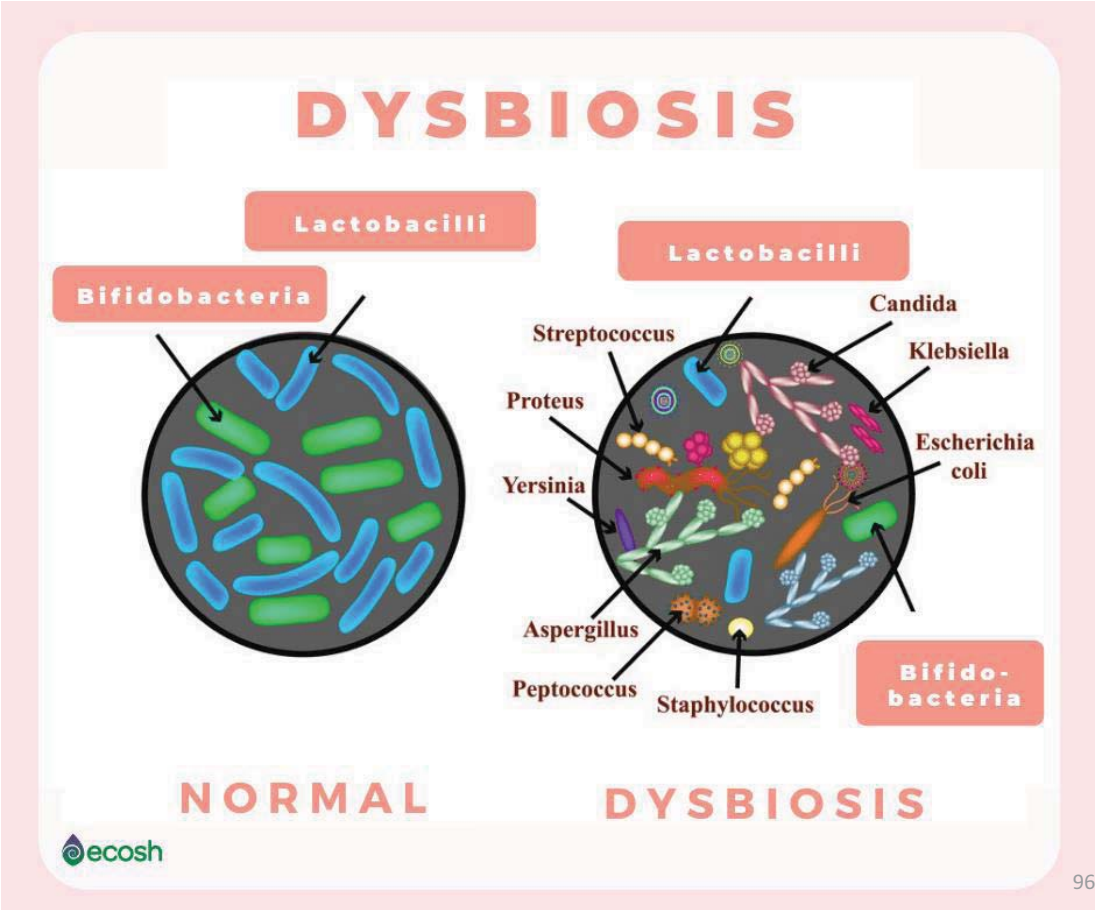
Dysbiosis

- “changes” in the microbiome
- “imbalance” in the microbiome
- “specific “ alteration in the microbiome



REF | Katarzyna B Hooks & Maureen A O’Malley, mBio, 2017

95



96

Terminology

- **Symbiont** = an organism living in symbiosis with another

- **Pathobiont** = a symbiont that is able to promote pathology only wen specific genetic or environmental conditions are altered in the host

REF | Lara Jochum & Barbel Stecher, Trends in Microbiology, 2020

Terminology

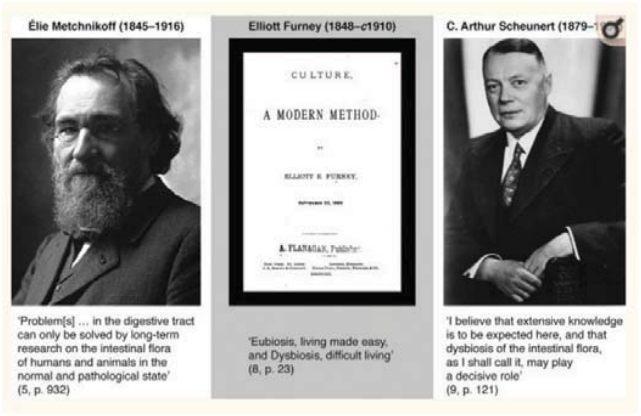
- **Dysbiosis** = condition of having imbalance in microbial community

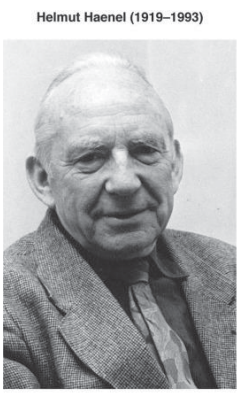
- **Eubiosis** = microbial balance within the body

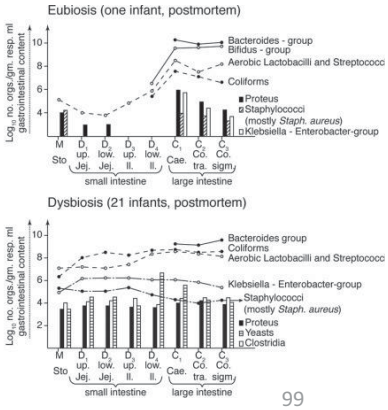
- **Symbiosis** = living together

Terminology

- Elie Metchnikoff (메치니코프 박사)
  - Pointing out resident microbes that could be “normal” or “pathological”
- Elliot Furney
  - coined “**eubiosis**” and “**dysbiosis**” in a science fiction novel
  - Not the context of microbiology
- Helmut Haenel
  - The first to promote dysbiosis and Eubiosis as we see it today
- C Arthur Scheunert
  - First claimed associations between gut dysbiosis and diseases



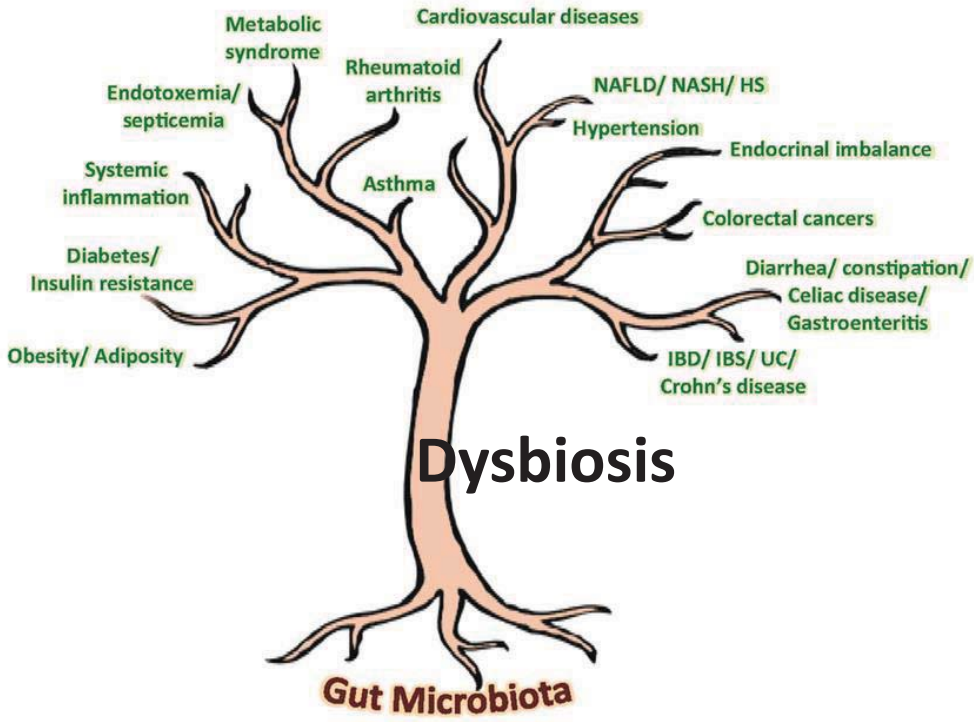




REF | Katarzyna B Hooks & Maureen A O’Malley, mBio, 2017

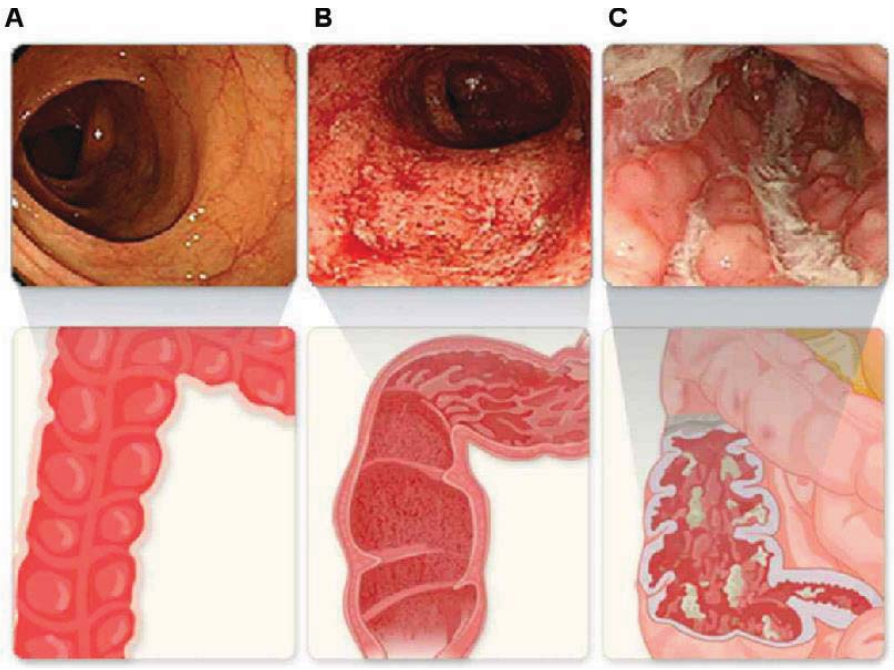
99

Dysbiosis → disease



100

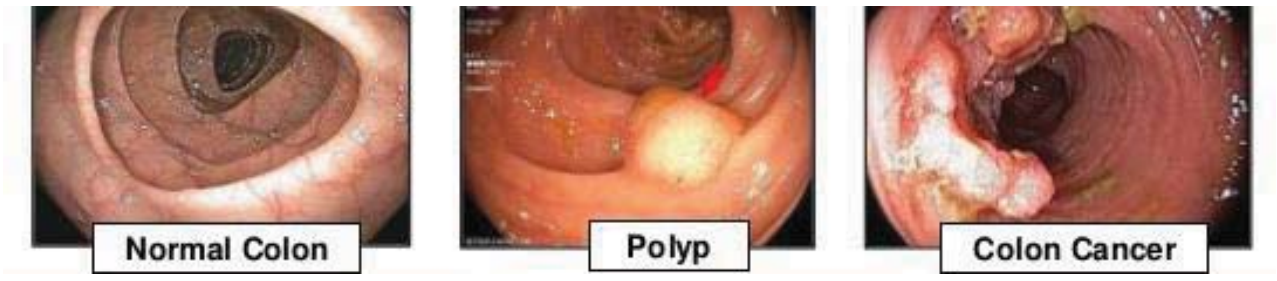
Dysbiosis → disease



Healthy | Ulcerative colitis | Chron's disease

101

Dysbiosis → disease



102

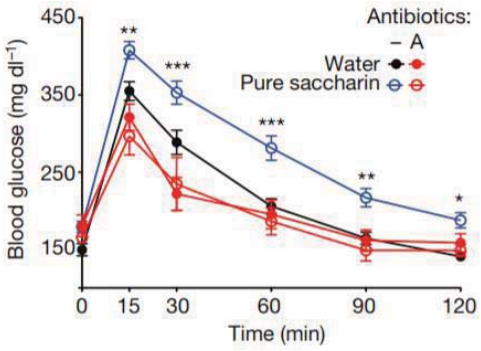
Dysbiosis → disease

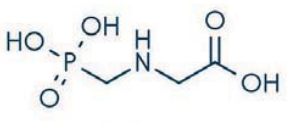


Atopic dermatitis

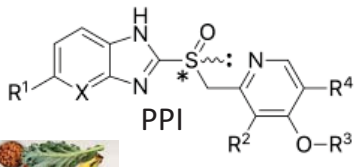
Cause of dysbiosis?

- Dietary changes
- Decreased in fermentable fibre
- Antibiotics
- Glyphosate (herbicide)
- Sweetener: e.g. saccharin
- Medications: e.g. PPIs, steroids, chemotherapy



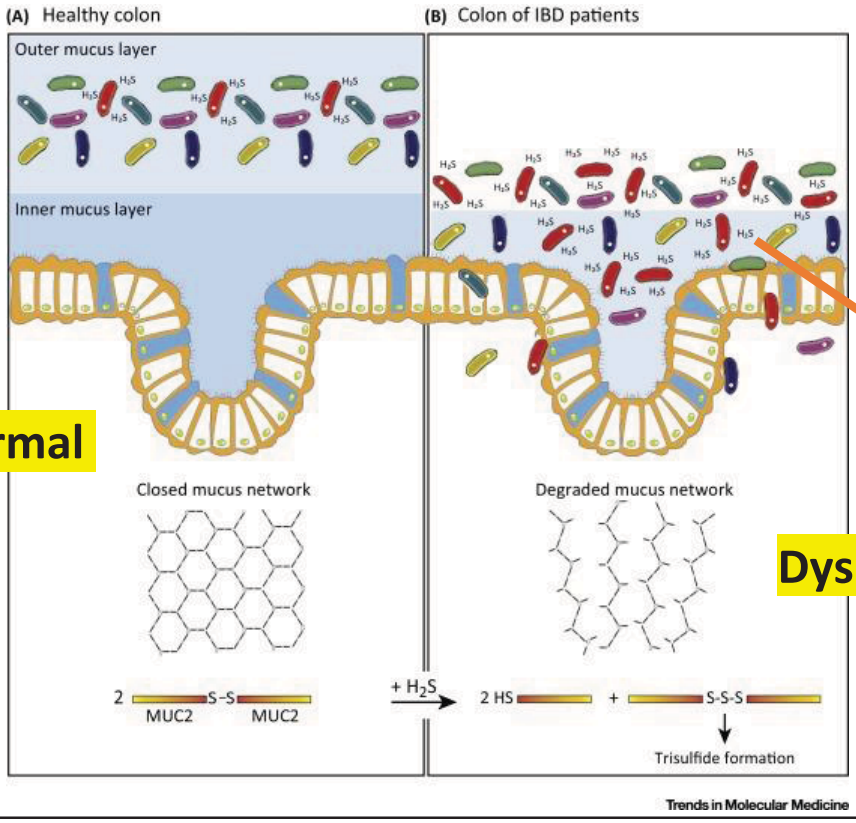


glyphosate



PPI

Low fibre diet → dysbiosis → inflammation

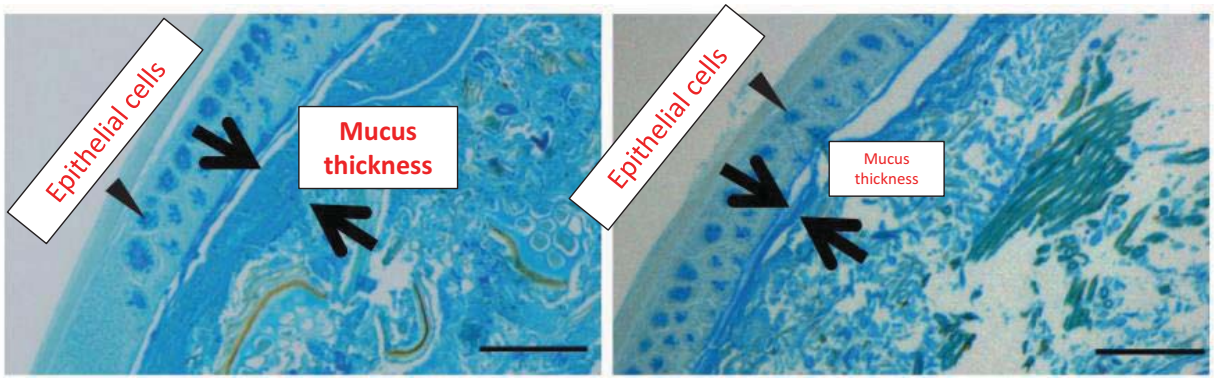


Normal

Degrading mucus by H_2S reducing bacteria

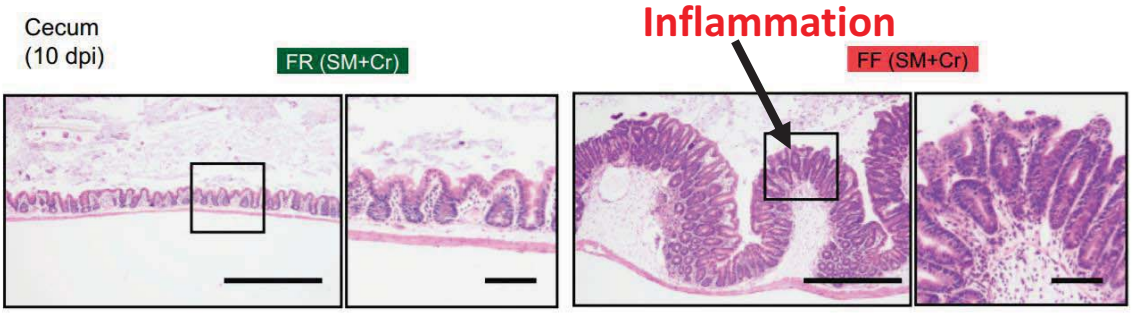
Dysbiosis

Low fibre diet → dysbiosis → inflammation



High-fibre diet **Fibre-free diet**

Low fibre diet → dysbiosis → inflammation

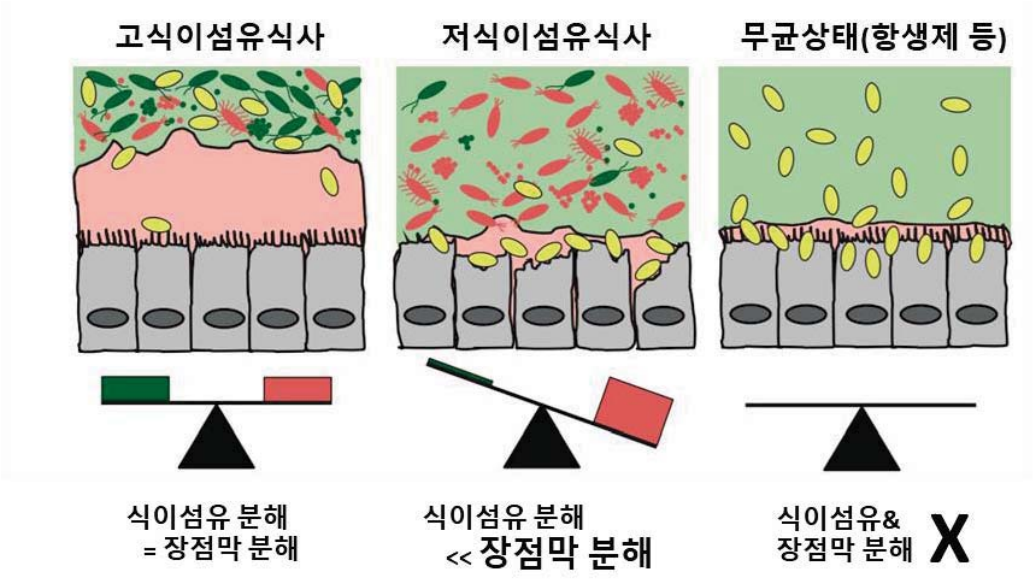


High fibre diet + pathogens

Low fibre diet + pathogens

107

Low fibre diet → dysbiosis → inflammation



108

Rebalancing the gut microbiome?

- Administration of probiotic bacteria
- Administration of prebiotics to favour the overgrowth of probiotic bacteria
- Administration of probiotics & prebiotics (called synbiotics)
- Phage therapy
- Fecal microbiota transplant

109

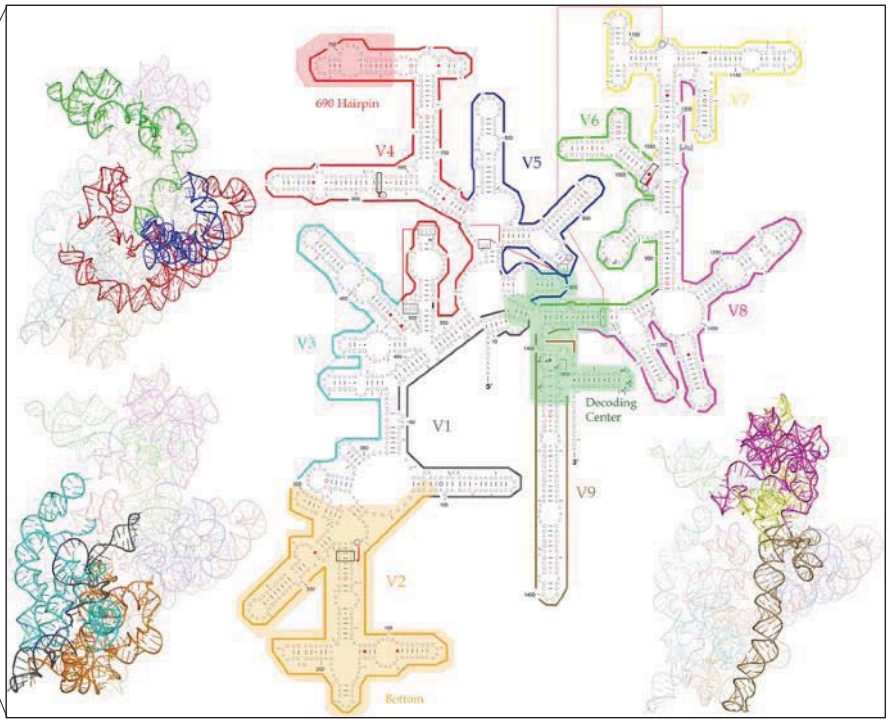
Bioinformatics analysis

110

16S rRNA amplicon sequencing

111

16S rRNA = universal phylogenetic marker

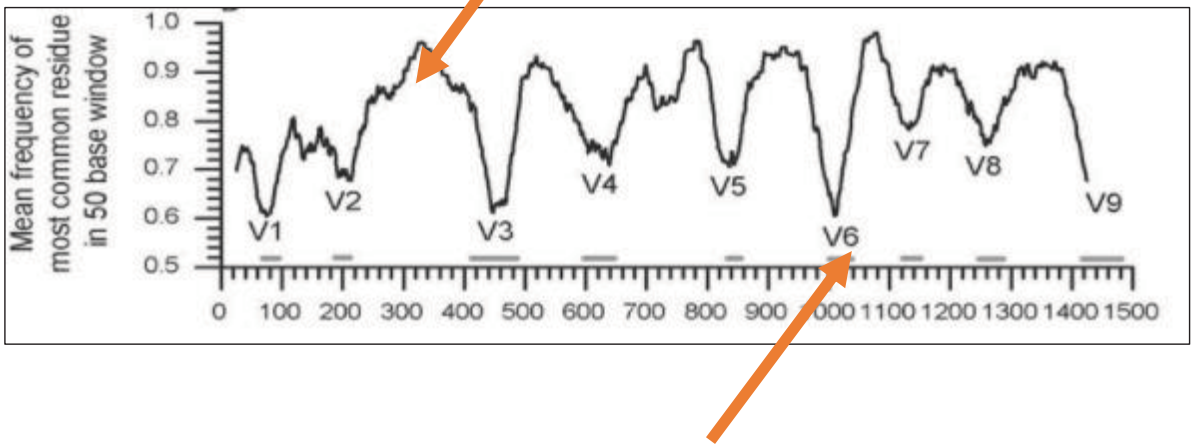


- **Present in all species**
- **Ubiquitous**
- **Extreme sequence conservation**
- **Single copy**
- Well-annotated references

112

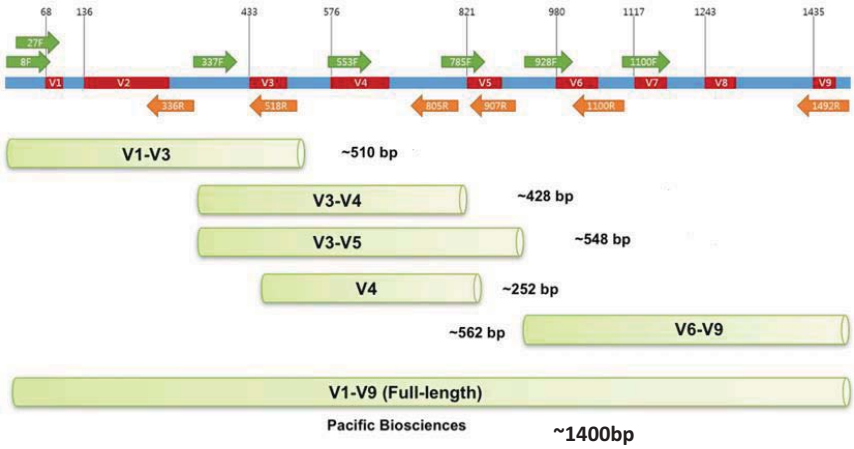
16S rRNA = universal phylogenetic marker

Highly conserved



Hypervariable regions

16S rRNA region has 9 hypervariable regions

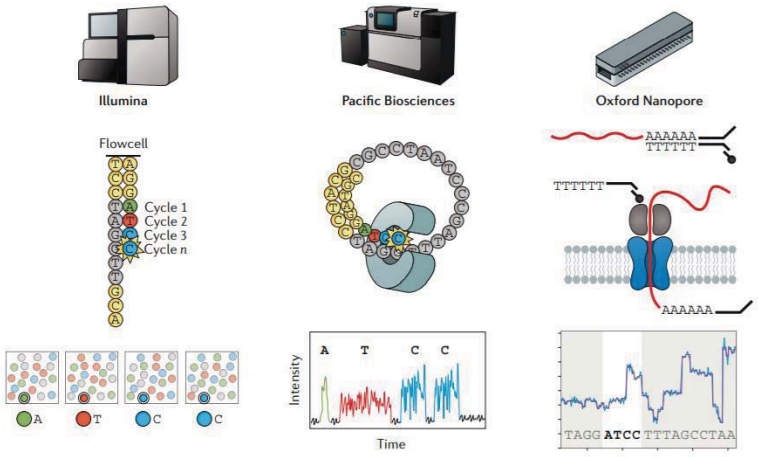


Primers

Primer Name	Sequence (5'-3')	SEQ ID NO:
V1/2	AGAGTTTGATCCTGGCTCAG	SEQ ID NO: 18
V5/4	CCGTCAATTYYTTTRAGTTT	SEQ ID NO: 19
U1492R	GGTTACCTTGTTACGACTT	SEQ ID NO: 20
928F	TAAAACTYAAAKGAATTGACGGG	SEQ ID NO: 21
336R	ACTGCTGCSYCCCGTAGGAGTCT	SEQ ID NO: 22
1100F	YAACGAGCGCAACCC	SEQ ID NO: 23
1100R	GGGTTGCGCTCGTTG	SEQ ID NO: 24
337F	GACTCCTACGGGAGGCWGCAG	SEQ ID NO: 25
907R	CCGTCAATTCCTTTRAGTTT	SEQ ID NO: 26
785F	GGATTAGATACCCTGGTA	SEQ ID NO: 27
805R	GACTACCAGGGTATCTAATC	SEQ ID NO: 28
533F	GTGCCAGCMGCCGCGGTAA	SEQ ID NO: 29
518R	GTATTACCGCGGCTGCTGG	SEQ ID NO: 30

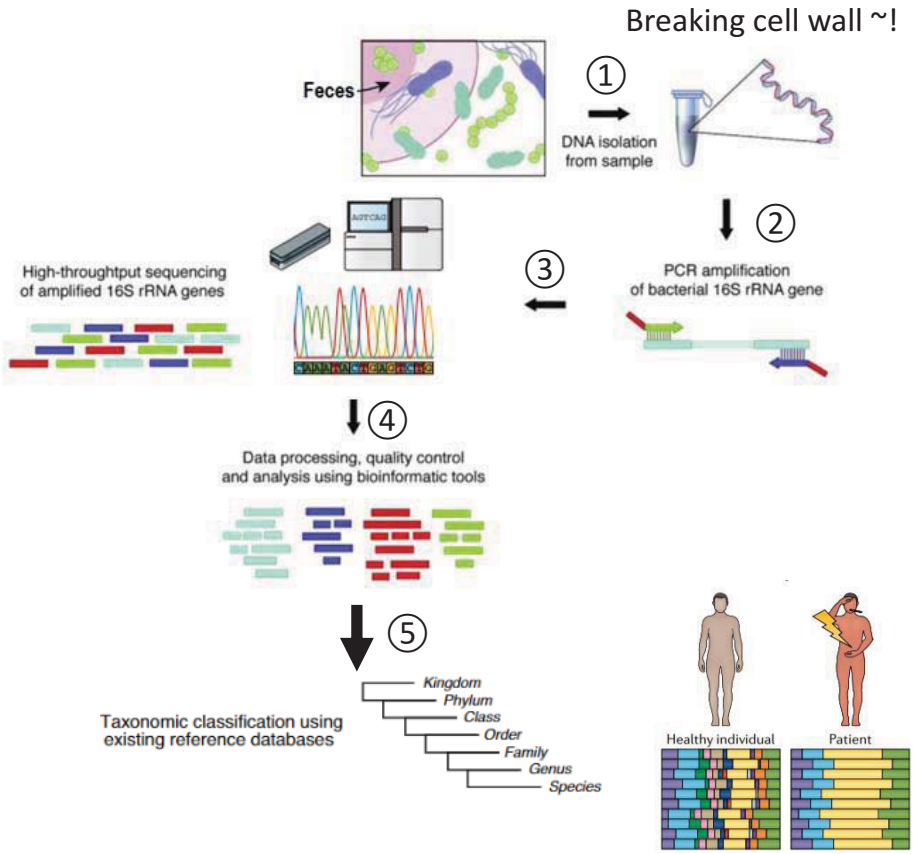
Useful for taxonomical classifications

Next-generation sequencing for 16S rRNA amplicon sequencing



Illumina MiSeq used commonly (2 X 300bp):
it can cover 500~600bp 16S rRNA amplicons mostly

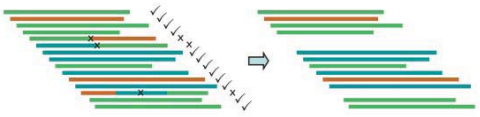
16S rRNA sequencing workflow



16S rRNA preprocessing

Quality assessment & trimming

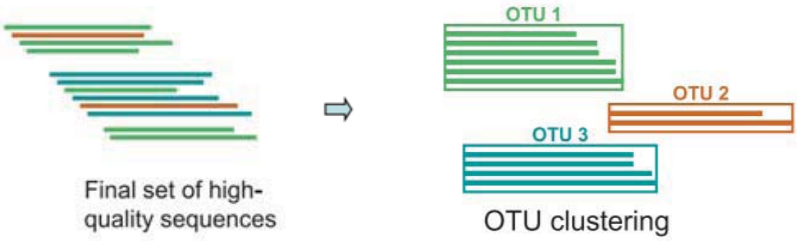
- Removing adapters, PCR primers & low-quality bases
- Removing PCR chimeric sequences
  - Common when closely related sequences are amplified



Identification and removal of PCR chimeras

Final set of high-quality sequences

Binning sequences into operational taxonomy unit (OTU)



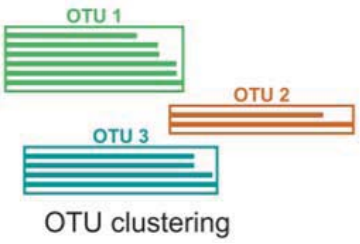
Final set of high-quality sequences

OTU clustering

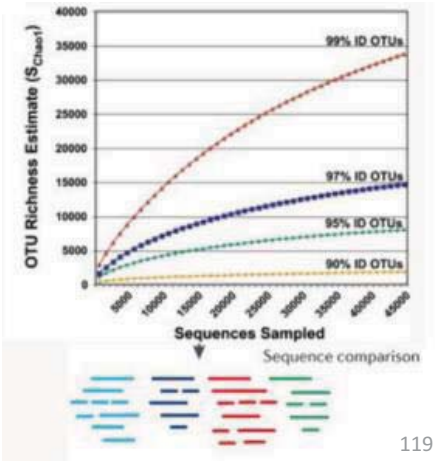
Binning sequences into operational taxonomy unit (OTU)

- Operational taxonomy unit (OTU)
  - A group of sequences grouped together based on sequence similarity
  - 97% identity threshold used frequently
  - Not necessarily equivalent to taxonomic entities

- Two ways of OTU clustering/picking
  - Reference-based (closed & open)
  - *De novo*



OTU clustering



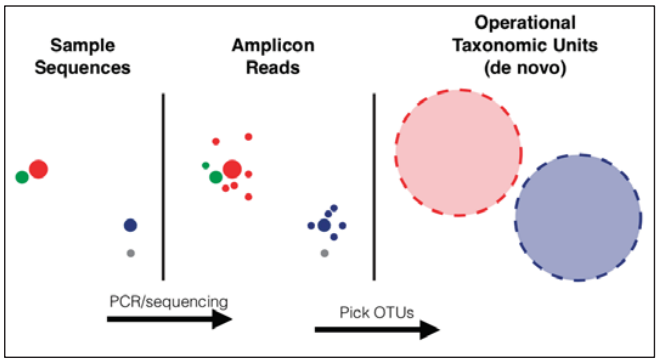
Amplicon sequence variants (ASV)

Amplicon sequence variant

- a single DNA sequence recovered from a high-throughput marker gene analysis

- Basically, it infers true sequences from sequencing reads

- It allows sequence variations by a single nucleotide change

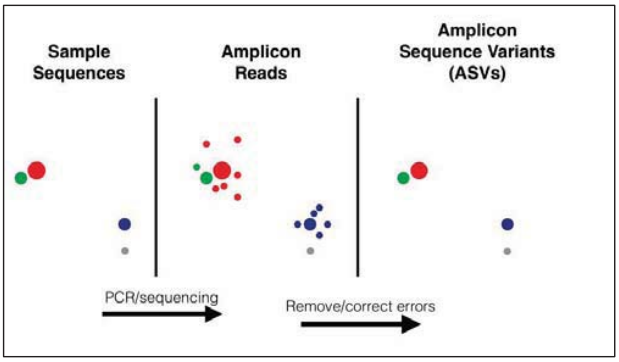


OTU methods

Amplicon sequence variants (ASV)

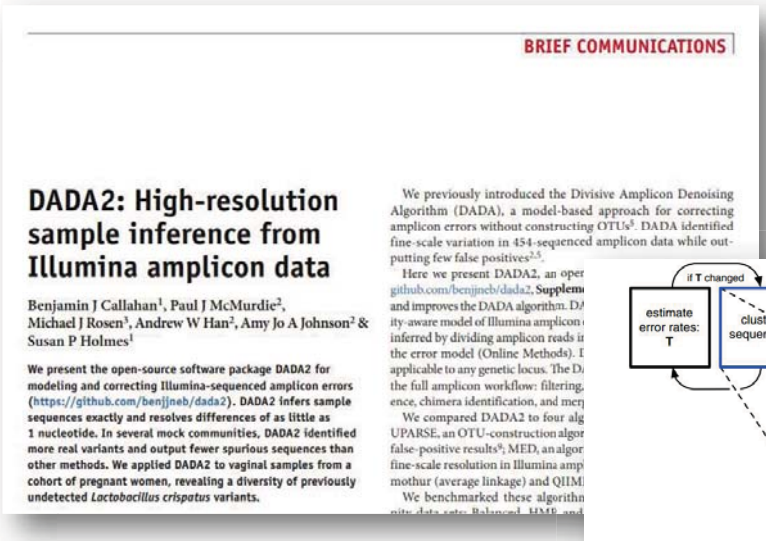
Amplicon sequence variant

- a single DNA sequence recovered from a high-throughput marker gene analysis
- Basically, it infers true sequences from sequencing reads
- It allows sequence variations by a single nucleotide change



ASV methods

ASV pipeline: DADA2 R package



BRIEF COMMUNICATIONS

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

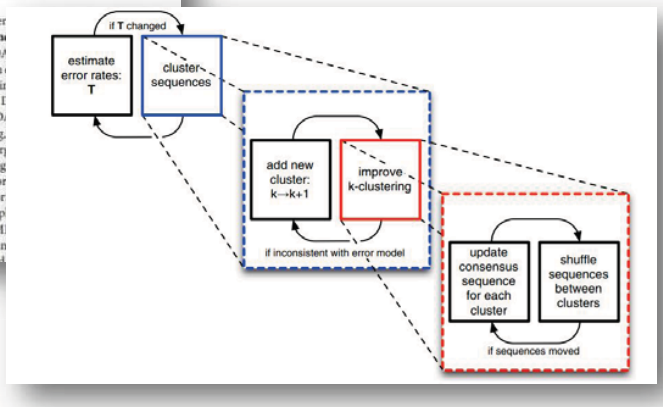
We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as 1 nucleotide. In several mock communities, DADA2 identified more real variants and output fewer spurious sequences than other methods. We applied DADA2 to vaginal samples from a cohort of pregnant women, revealing a diversity of previously undetected *Lactobacillus crispatus* variants.

We previously introduced the Divisive Amplicon Denoising Algorithm (DADA), a model-based approach for correcting amplicon errors without constructing OTUs[5]. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives[2,5].

Here we present DADA2, an oper
github.com/benjjneb/dada2, Supplem
and improves the DADA algorithm. DA
ity-aware model of Illumina amplicon
inferred by dividing amplicon reads i
the error model (Online Methods). I
applicable to any genetic locus. The DA
the full amplicon workflow: filtering,
ence, chimera identification, and merg

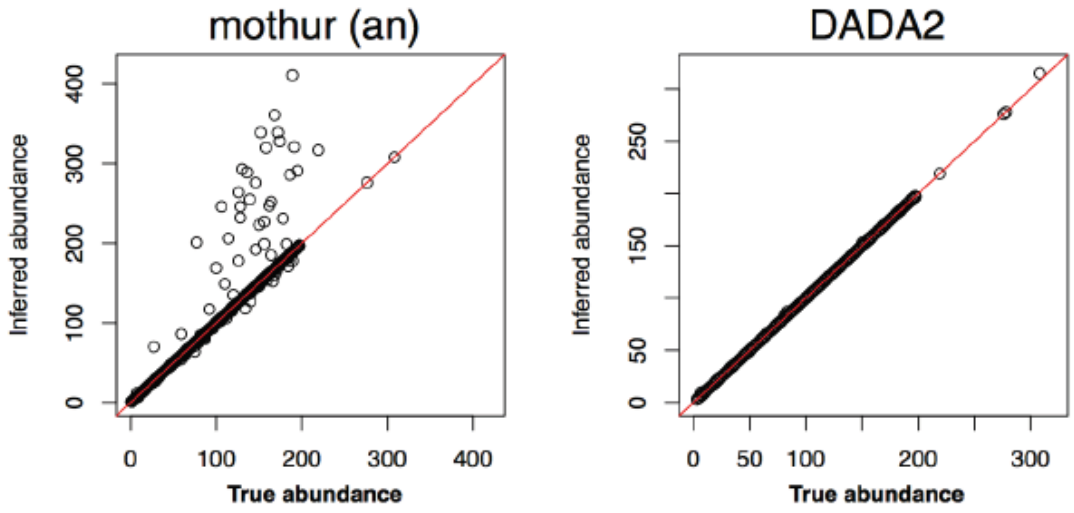
We compared DADA2 to four alg
UPARSE, an OTU-construction algor
false-positive results[9]; MED, an algor
fine-scale resolution in Illumina amp
mothur (average linkage) and QIIM

We benchmarked these algorithm
nity data sets: Balanced, HMP and



Nature Methods (2016)

Simulated dataset validation

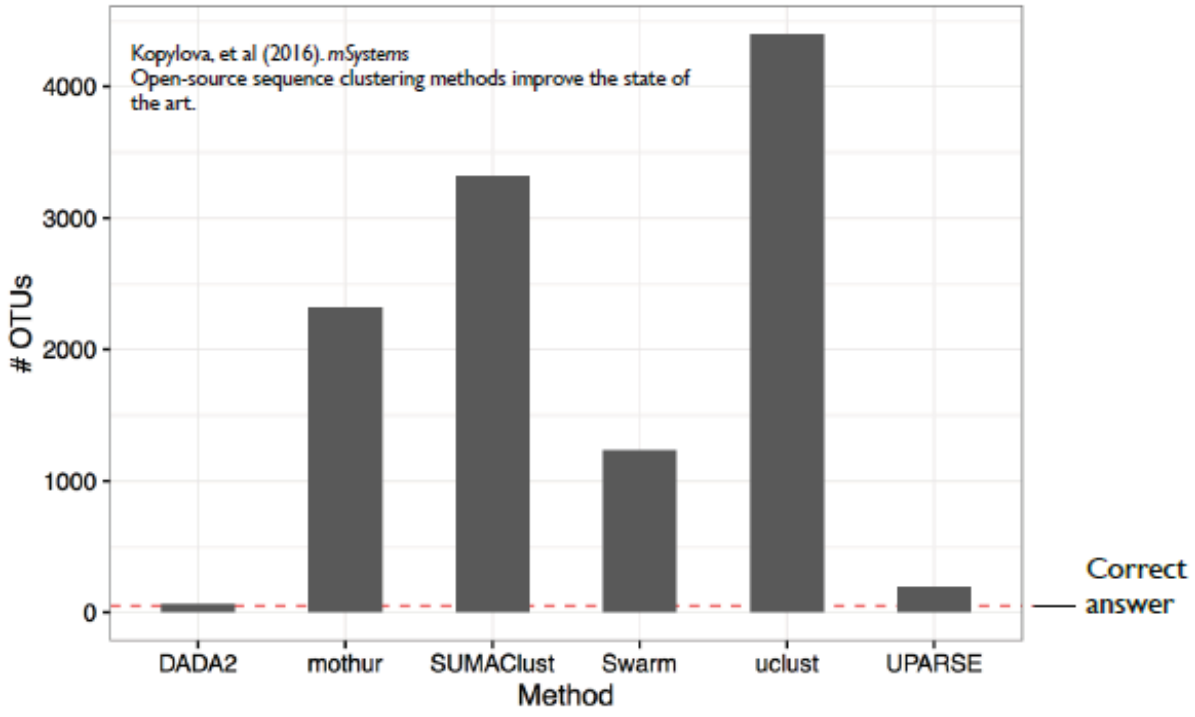


TP:	978
FP:	272
FN:	77

TP:	1042
FP:	0
FN:	13

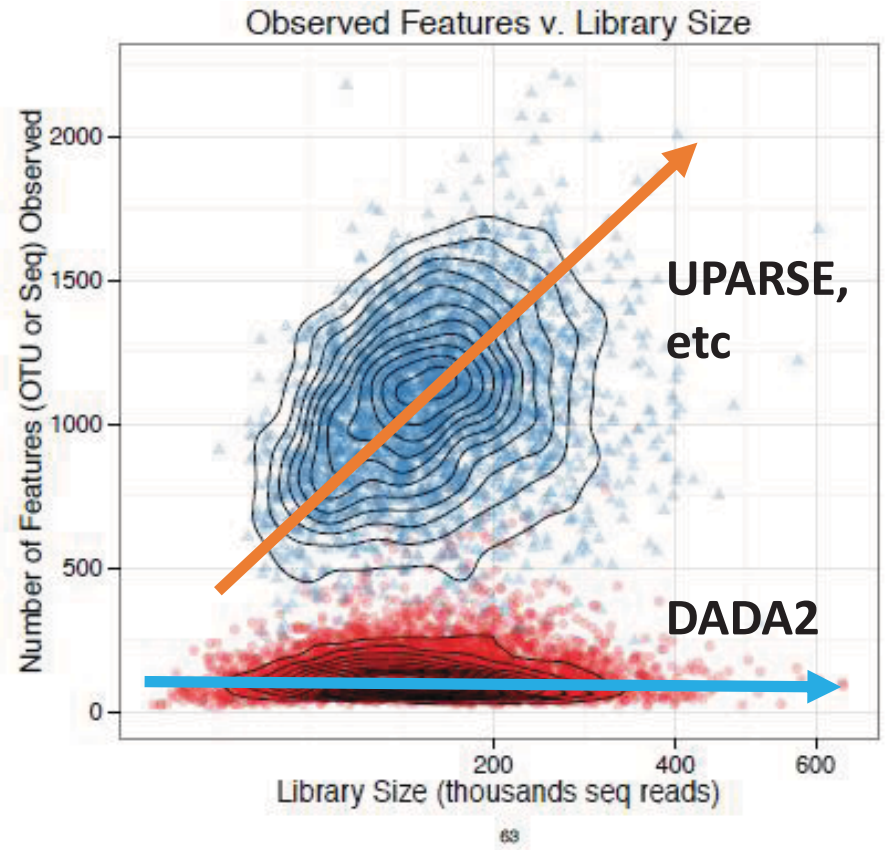
123

Mock community validation (Bokulich data)



124

Library sizes less affect ASV detection



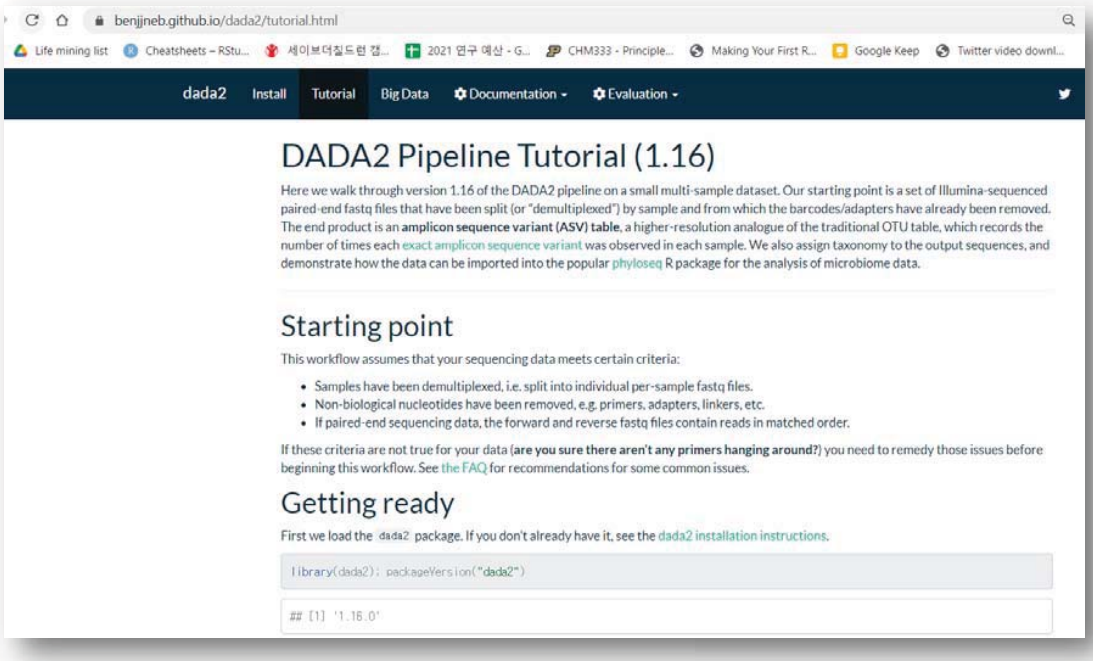
125

ASV vs OTU

	ASVs	OTUs: De novo	OTUs: Closed-ref
Precise	✓	~	~
Tractable	✓	~	✓
Reproducible	✓	✗	✓
Comprehensive	✓	✓	✗

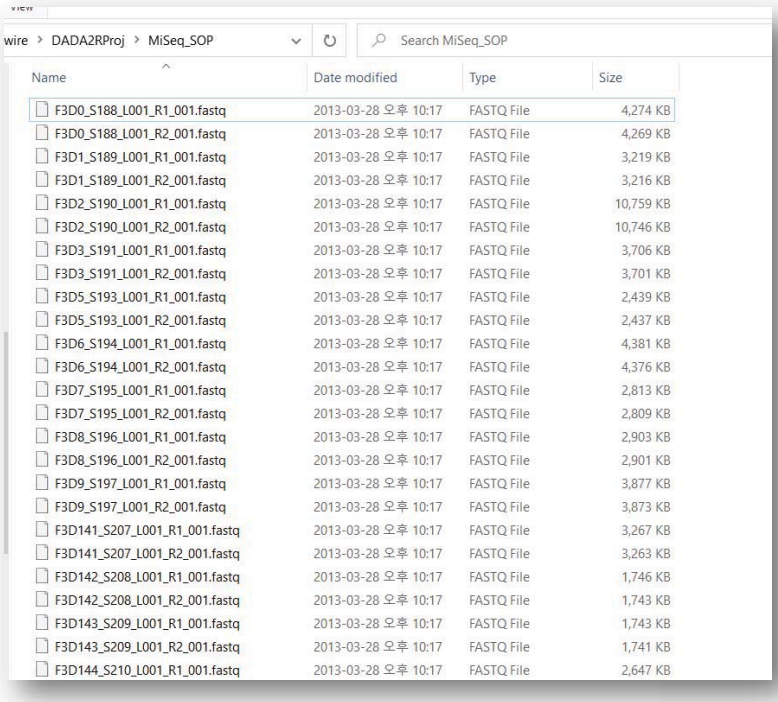
126

DADA2 pipeline tutorial



https://benjjneb.github.io/dada2/tutorial.html

Example dataset



Total 40 files
Total 163MB

Library install & setting paths

```
#### installing libraries ####
BiocManager::install("dada2", force = T)
library(dada2); packageVersion("dada2")

#### checking path ####
path <- "J:\\MiSeq_SOP/"
list.files(path)
```

129

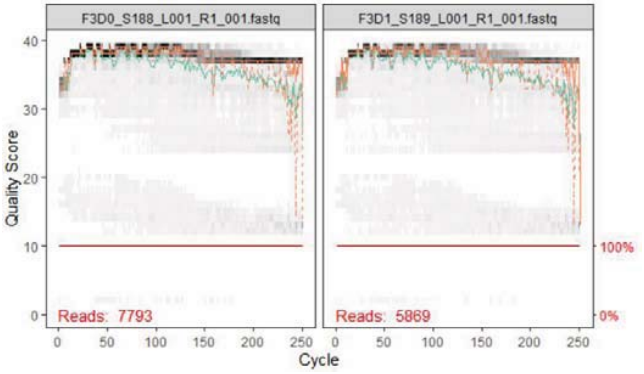
Loading file names & checking sequence quality

```
#### taking samples, forward and reverse reads ####
# Forward and reverse fastq filenames have format: SAMPLENAME_R1_001.fastq and SAMPLENAME_R2_001.fastq
fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))

# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)

#### quality check ####
plotQualityProfile(fnFs[1:2])
plotQualityProfile(fnRs[1:2])
```

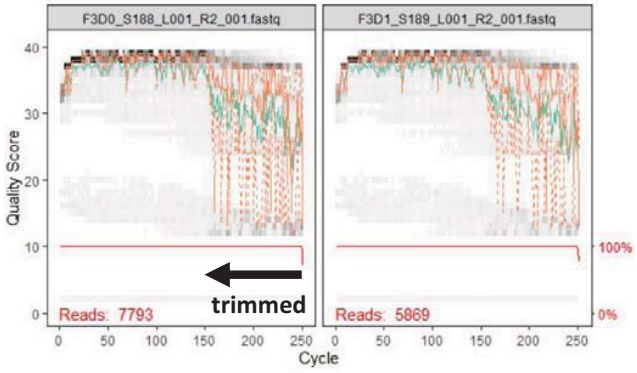
130

Loading file names & checking sequence quality

Forward reads



Reverse reads



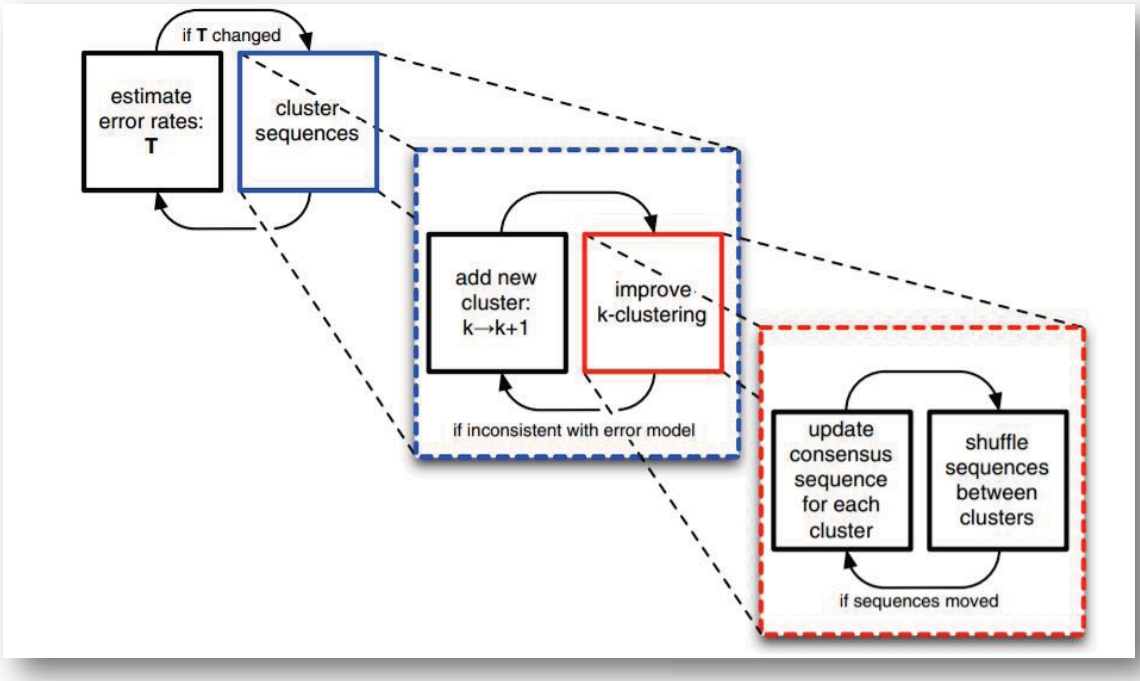
Filtering and trimming

```
#### filtering and trimming ####

# Place filtered files in filtered/ subdirectory
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
names(filtFs) <- sample.names
names(filtRs) <- sample.names

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(240,160),
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=FALSE) # On Windows set multithread=FALSE
head(out)
```

```
> head(out)
                              reads.in reads.out
F3D0_S188_L001_R1_001.fastq       7793      7113
F3D1_S189_L001_R1_001.fastq       5869      5299
F3D141_S207_L001_R1_001.fastq     5958      5463
F3D142_S208_L001_R1_001.fastq     3183      2914
F3D143_S209_L001_R1_001.fastq     3178      2941
F3D144_S210_L001_R1_001.fastq     4827      4312
```

Forward and reverse sequences will be discarded when its length less than 240bp and 160bp, respectively

Inferring error models & amplicon sequence variants (ASVs)



133

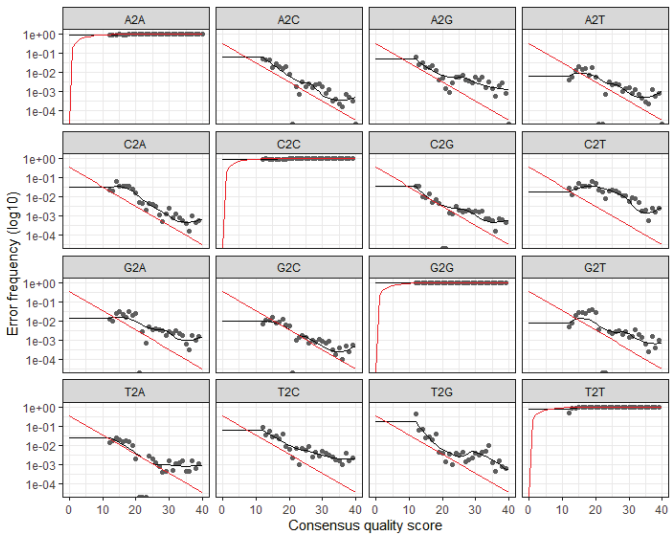
Inferring error models

learning error rate models

errF <- learnErrors(filtFs, multithread=FALSE)

errR <- learnErrors(filtRs, multithread=FALSE)

plotErrors(errF, nominalQ=TRUE)



134

Inferring amplicon sequence variants (ASVs)

```
#### running core sample inference ####
dadaFs <- dada(filtFs, err=errF, multithread=FALSE)
dadaRs <- dada(filtRs, err=errR, multithread=FALSE)

#### merging paired reads ####
mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
# Inspect the merger data.frame from the first sample
View(mergers[[1]])
```

	sequence	abundance	forward	reverse	nmatch	nmismatch	nindel	prefer	accept
1	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	582	1	4	168	0	0	1	TRUE
2	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	476	2	1	168	0	0	1	TRUE
3	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	442	3	7	168	0	0	1	TRUE
4	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	427	4	2	168	0	0	2	TRUE
5	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	340	5	5	168	0	0	1	TRUE
6	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	281	6	6	168	0	0	1	TRUE
7	TACGTAGGGGGCAAGCGTTATCCGGAATTACTGGGTGTAAAGG...	280	10	3	167	0	0	2	TRUE
8	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	219	7	43	168	0	0	1	TRUE
9	TACGGAGGATTCAAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	183	8	9	167	0	0	2	TRUE
10	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	164	11	10	168	0	0	1	TRUE
11	TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	154	9	8	167	0	0	2	TRUE
12	TACGTAGGTGGCGAGCGTTGTCCGGATTTACTGGGCGTAAAGG...	128	12	11	167	0	0	2	TRUE
13	TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGG...	103	82	83	168	0	0	1	TRUE
15	TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGG...	92	17	12	167	0	0	2	TRUE
16	TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGG...	91	16	15	167	0	0	2	TRUE

135

Generating ASV tables with removing chimera and low sequencing depth samples

```
#### constructing sequence table ####
seqtab <- makeSequenceTable(mergers)
dim(seqtab) #This table contains 292 ASVs
table(nchar(getSequences(seqtab)))

#### removing low depth samples ####
rowSums(seqtab)
seqtab = seqtab[rowSums(seqtab) > 1000,]
View(t(seqtab))

#### removing chimeras ####
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus",
multithread=FALSE, verbose=TRUE)
dim(seqtab.nochim)
dim(seqtab.nochim)[2]/dim(seqtab)[2]
sum(seqtab.nochim)/sum(seqtab)
```

```
> dim(seqtab.nochim)[2]/dim(seqtab)[2]
[1] 0.8287671
> sum(seqtab.nochim)/sum(seqtab)
[1] 0.9679396
```

136

ASV table example

	F3D0	F3D1	F3D141	F3D142	F3D143	F3D144	F3D145	F3D146	F3D147	F3D148	F3D149	F3D150
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	582	401	442	296	222	421	641	322	1485	854	874	312
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	340	350	366	304	174	267	491	233	1194	720	765	225
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	442	226	344	157	210	297	508	259	898	579	725	395
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	427	70	498	166	232	352	581	387	1083	844	892	469
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAA...	154	139	188	175	129	105	302	177	451	441	416	169
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	476	40	327	189	245	351	470	273	1169	861	634	210
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	281	99	244	162	150	237	377	219	845	567	551	236
TACGGAGGATTCAAGCGTTATCCGGATTTATTGGGTTTAAA...	183	184	321	87	81	39	124	71	74	515	513	117
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	164	106	141	80	74	155	242	99	268	196	295	148
TACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAA...	17	102	168	42	78	269	316	177	446	411	478	61
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	219	40	145	96	115	145	254	144	559	425	301	98
TACGTAGGTGGCAAGCGTTATCCGGATTTATTGGGCGTAAA...	52	128	12	102	43	16	22	4	146	18	87	64
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	103	0	66	64	59	75	119	57	278	192	161	72
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAA...	92	325	33	11	9	11	15	25	73	56	44	18
TACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAA...	80	0	103	52	40	113	126	35	305	271	175	30
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAA...	69	31	43	30	20	44	107	35	147	120	119	49

137

Assigning taxa and generating phyloseq object

```
#### assigning taxonomy ####
taxa <- assignTaxonomy(seqtab.nochim,
"silva_nr99_v138.1_wSpecies_train_set.fa", multithread=F)
taxa <- addSpecies(taxa, "silva_species_assignment_v138.1.fa")


taxa.print <- taxa # Removing sequence rownames for display only
rownames(taxa.print) <- NULL
head(taxa.print)
dim(taxa.print)
```

assignTaxonomy

- Based on naïve Bayes classifier

addSpecies

- Exact matching

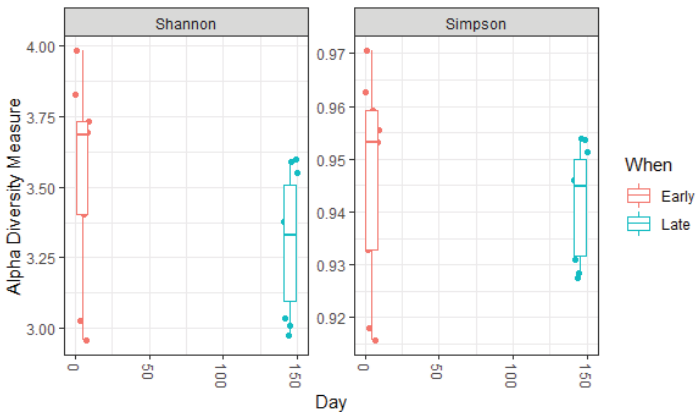
138

Downstream microbiome analysis with libraries

```
#### downstream analysis with phyloseq ####
if (!requireNamespace("phyloseq", quietly = T))
  BiocManager::install("phyloseq")
if (!requireNamespace("Biostrings", quietly = T))
  BiocManager::install("Biostrings")
if (!requireNamespace("ggplot2", quietly = T))
  install.packages("ggplot2")

library(phyloseq); packageVersion("phyloseq")
library(Biostrings); packageVersion("Biostrings")
library(ggplot2); packageVersion("ggplot2")
```

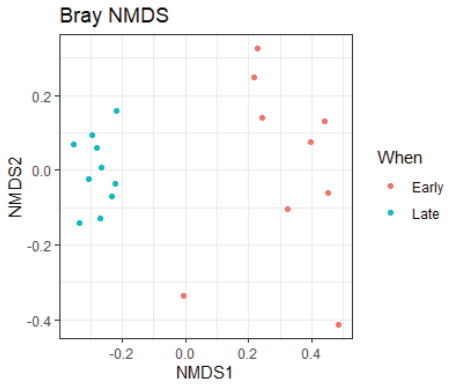

Alpha-diversity analysis

```
#### generating phyloseq object ####

ps <- phyloseq(otu_table(seqtab.nochim,
taxa_are_rows=FALSE),
               sample_data(samdf),
               tax_table(taxa))

ps <- prune_samples(sample_names(ps) != "Mock", ps) #
Remove mock sample


#### alpha diversity ####

plot_richness(ps, x="Day", measures=c("Shannon",
"Simpson"), color="When") + geom_boxplot()
```



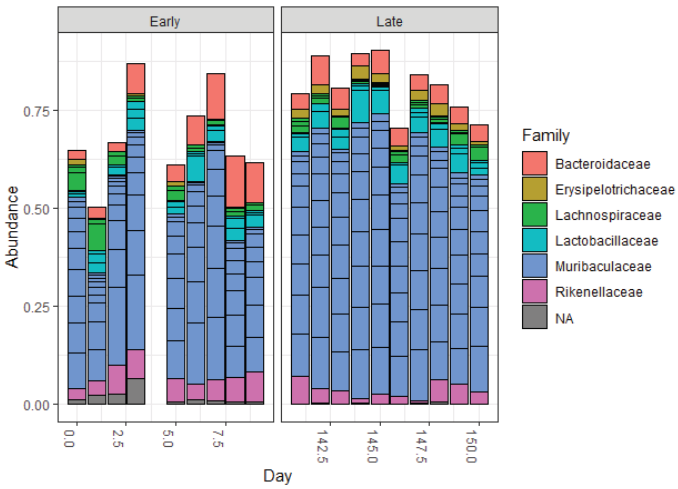
Ordination plot (beta-diversity)

```
#### ordination plots (beta-diversity) ####

# Transform data to proportions as appropriate for Bray-Curtis distances

ps.prop <- transform_sample_counts(ps, function(otu) otu/sum(otu))

ord.nmds.bray <- ordinate(ps.prop, method="NMDS", distance="bray")

plot_ordination(ps.prop, ord.nmds.bray, color="When", title="Bray NMDS")

plot_ordination(ps.prop, ord.nmds.bray, color="When", title="Bray NMDS") +
geom_point(size=5)
```

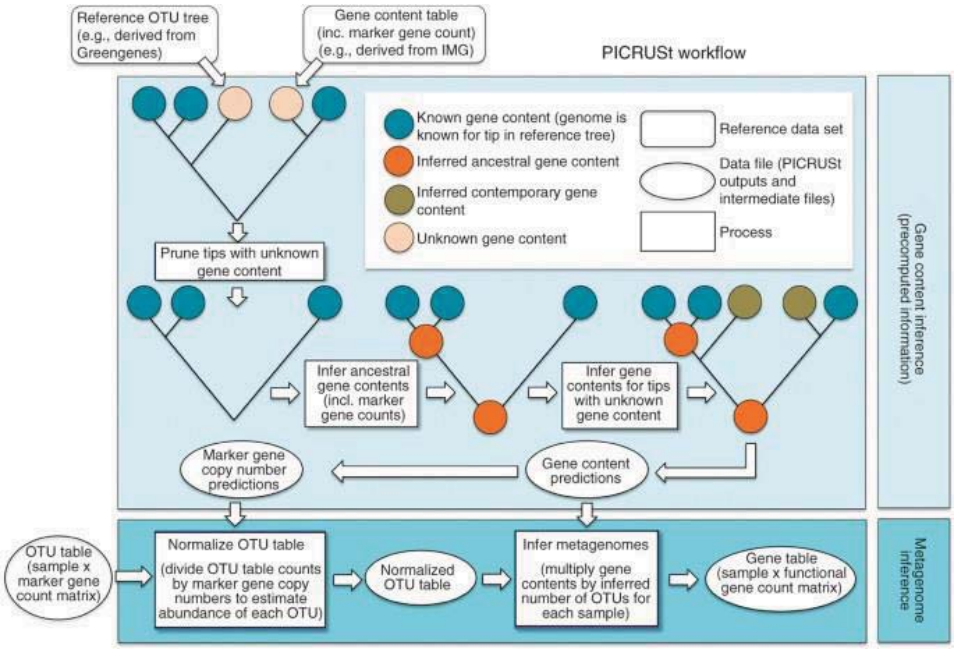


Top-20 taxa barplot

```
#### barplots of top-20 taxa ####

top20 <- names(sort(taxa_sums(ps),
decreasing=TRUE))[1:20]

ps.top20 <- transform_sample_counts(ps, function(OTU)
OTU/sum(OTU))

ps.top20 <- prune_taxa(top20, ps.top20)

plot_bar(ps.top20, x="Day", fill="Family") +
facet_wrap(~When, scales="free_x")

plot_bar(ps.top20, x="Day", fill="Genus") +
facet_wrap(~When, scales="free_x")
```

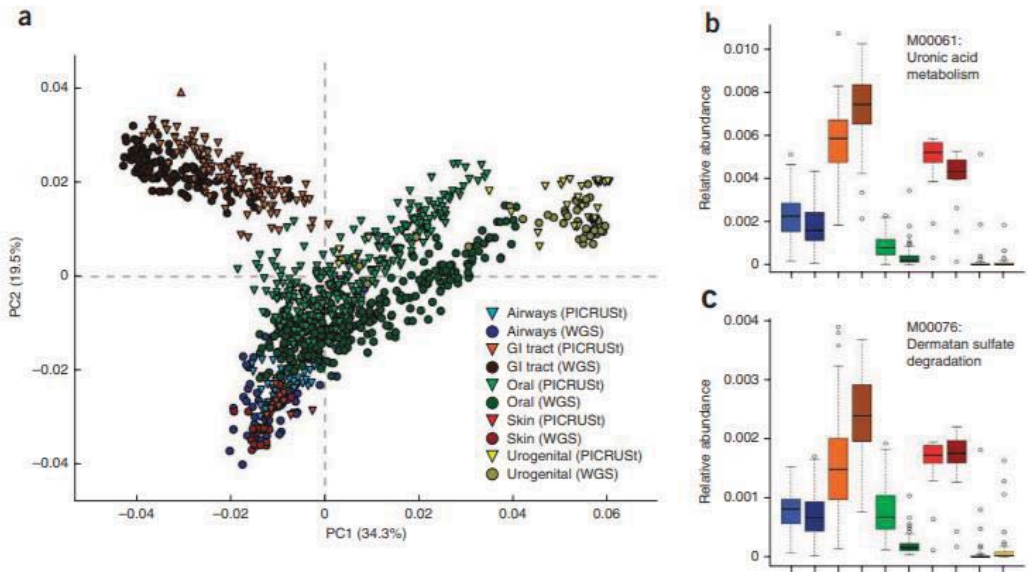


Function prediction - PICRUSt



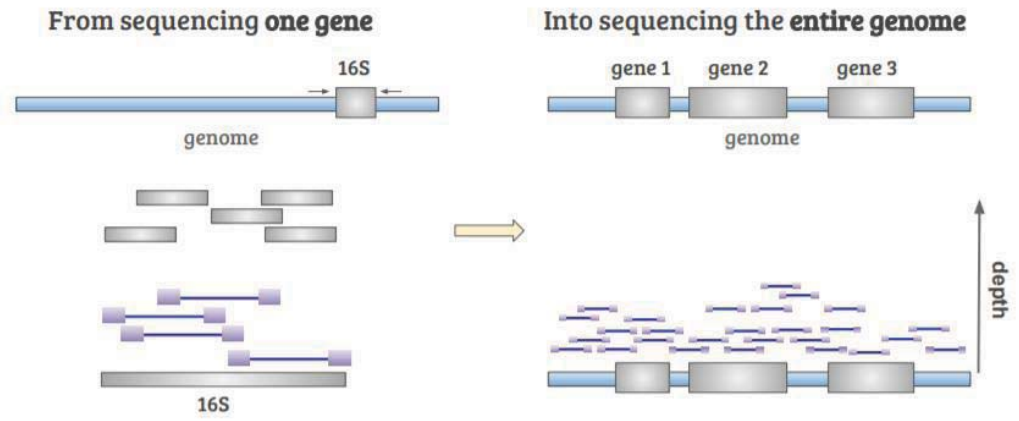
143

Function prediction - PICRUSt



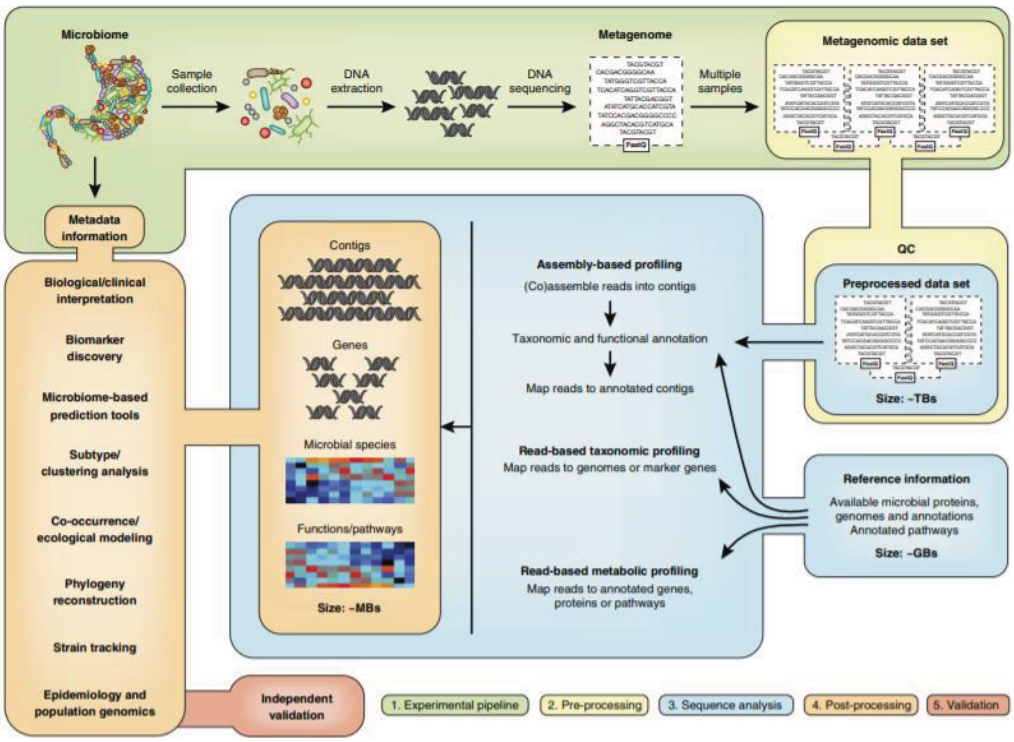
144

Shotgun metagenome analysis



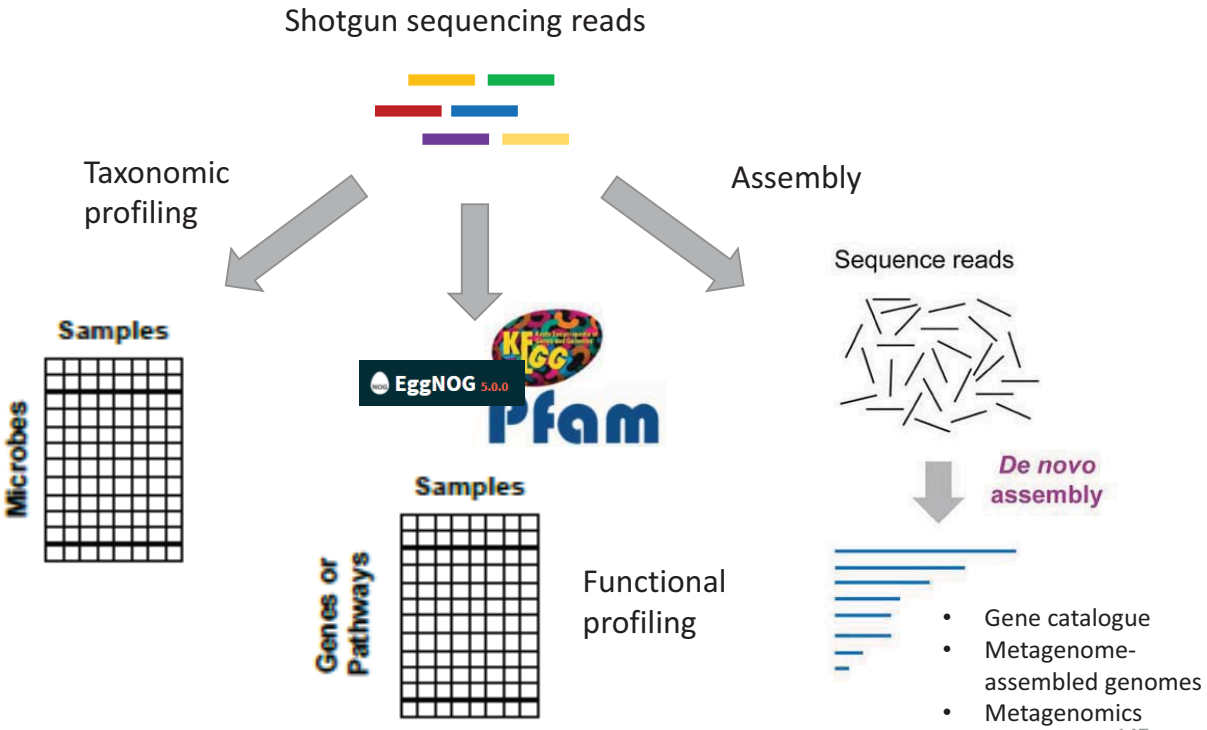
REF | BioBakery

Shotgun metagenomics enables complete overview of a complex microbiome



REF| Christopher Quince et al., Nature Biotechnology, 2017

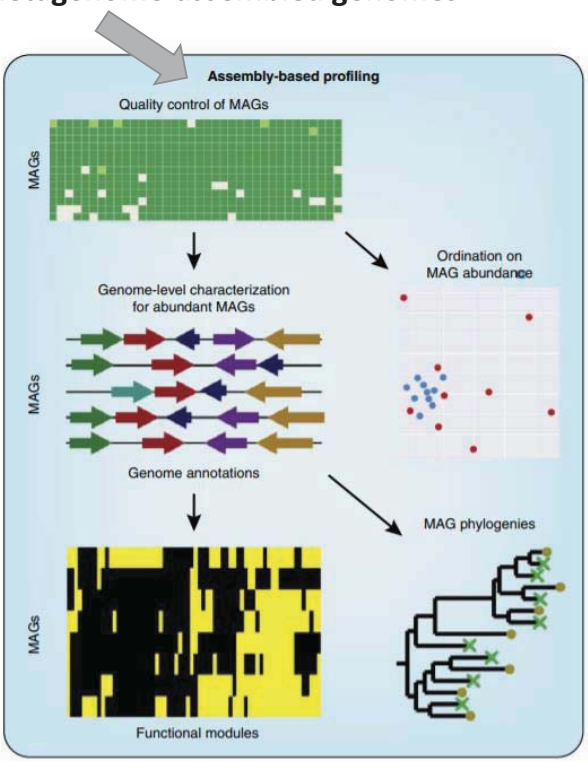
Overview of shotgun metagenome analysis



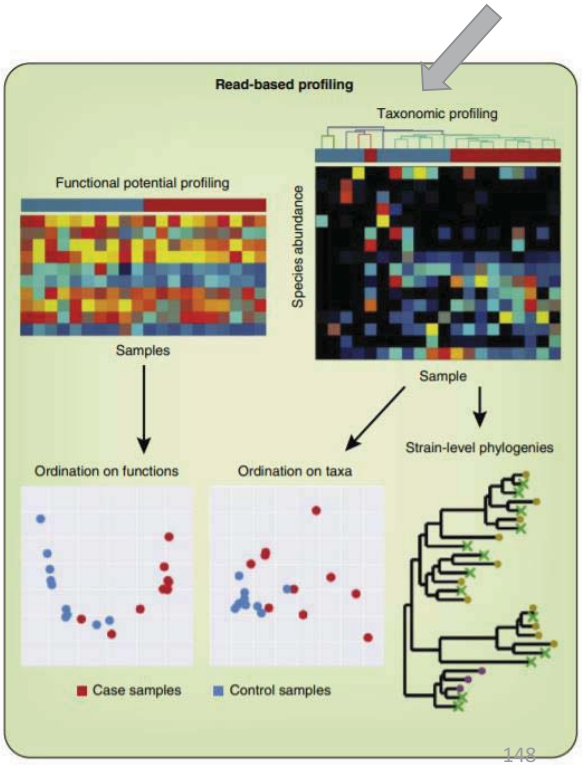
147

Taxonomic profiling (species/strain-level)

***De novo* assembly – metagenome-assembled genomes**



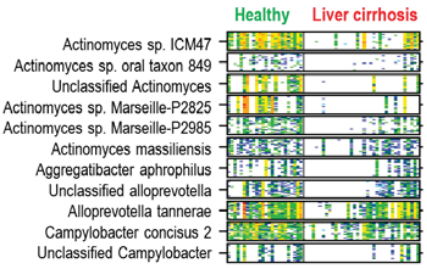
Using reference genomes



148

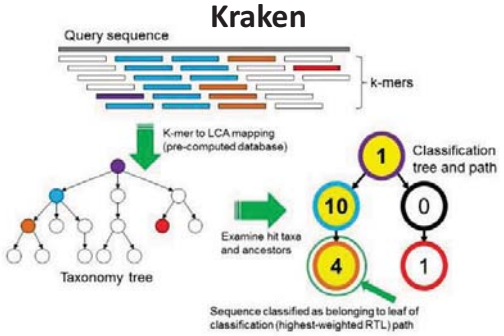
Read-based profiling

- **Index-based profiling**
  - Kraken, centrifuge
- **Marker gene-based profiling**
  - mOTU, metaPhlan
- **Metagenomic species (MGS)-based profiling**
  - Meteor/MOMR

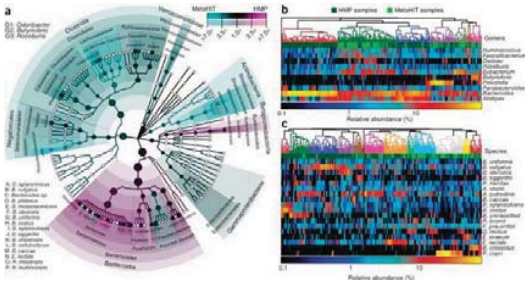


Metagenomic species (MGS)

Kraken

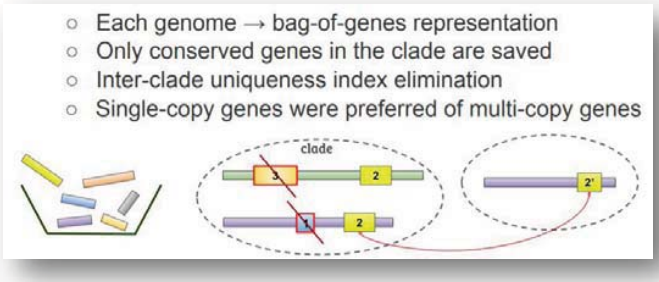


metaPhlan

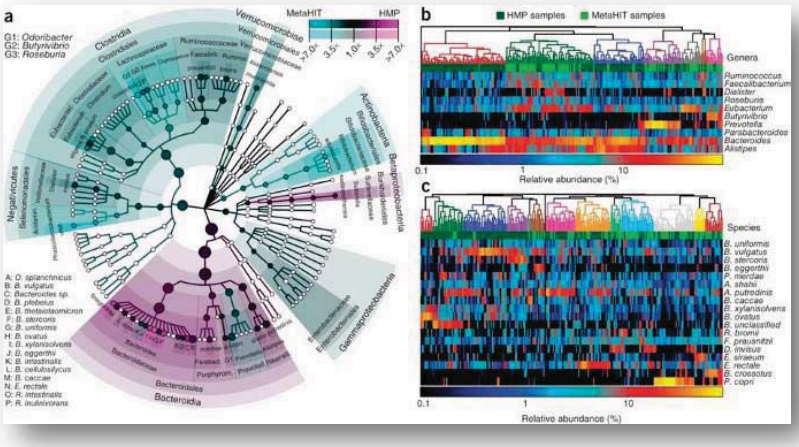


149

MetaPhlan: clade-specific marker gene based profiling

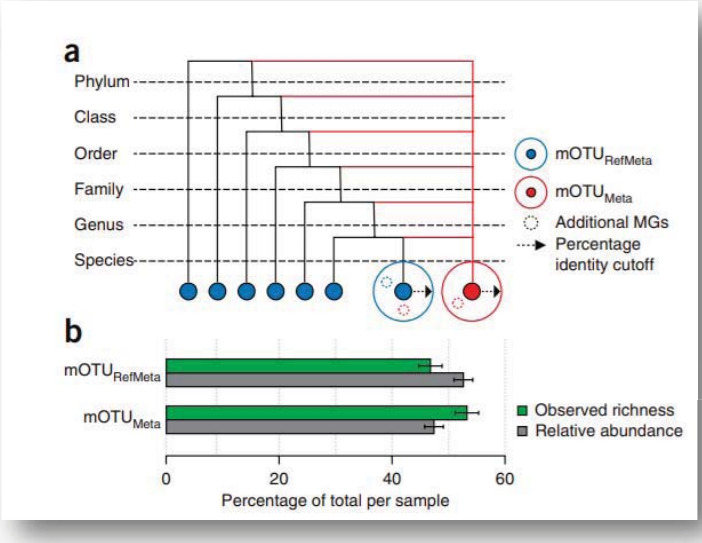


Clade-specific marker discovery



150

mOTU: universal phylogenetic marker - based profiling

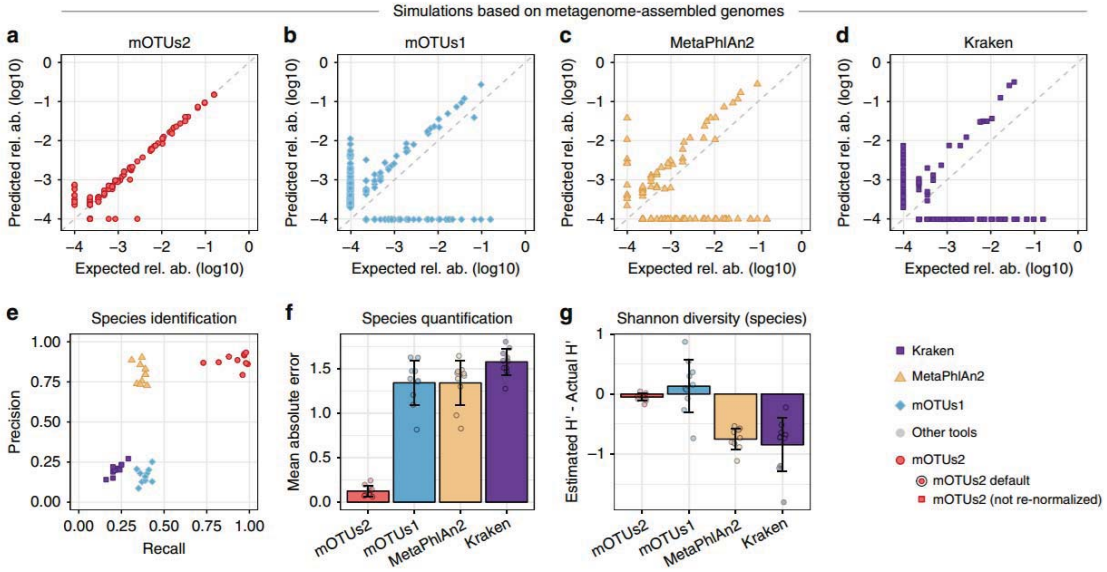


By COG	
COG ID	COG name
COG0012	Predicted GTPase, probable translation factor
COG0016	Phenylalanyl-tRNA synthetase alpha subunit
COG0018	Arginyl-tRNA synthetase
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0090	Ribosomal protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8

40 universal markers were selected

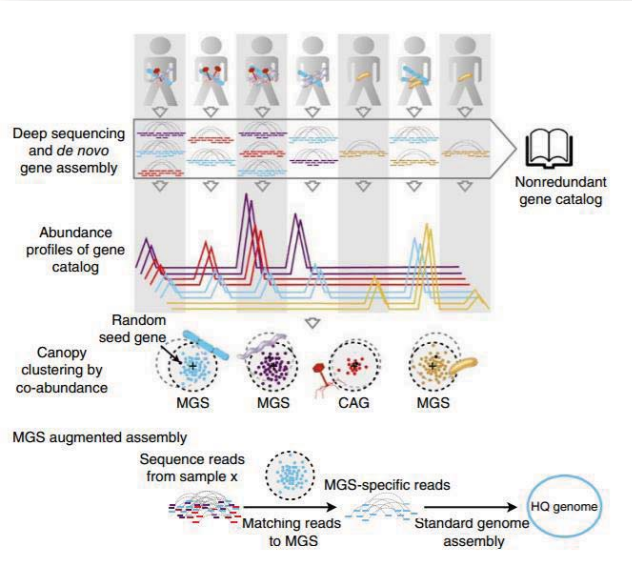
https://motu-tool.org/

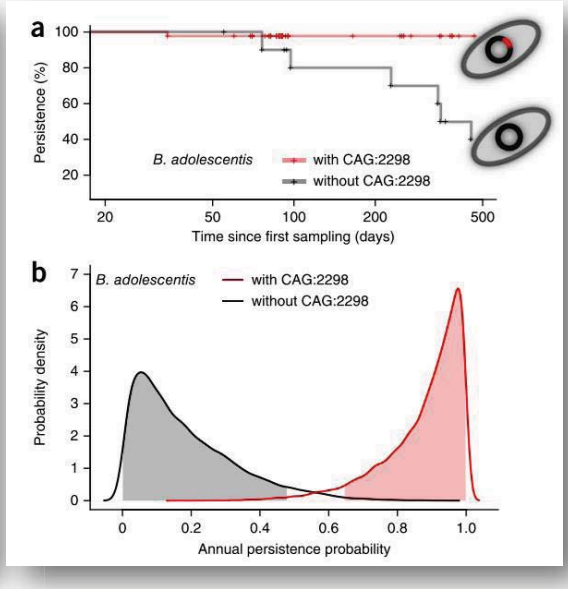
mOTU: universal phylogenetic marker - based profiling



https://motu-tool.org/

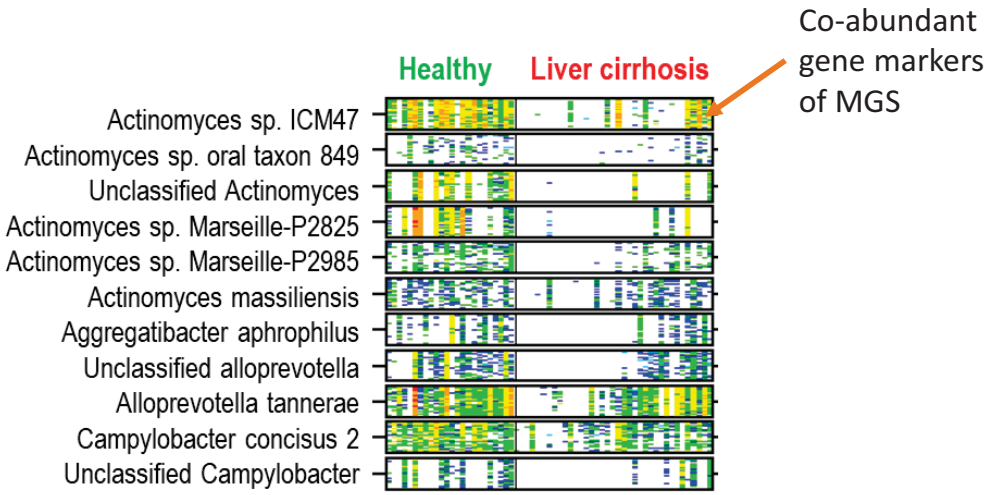
Metagenomic species-based profiling





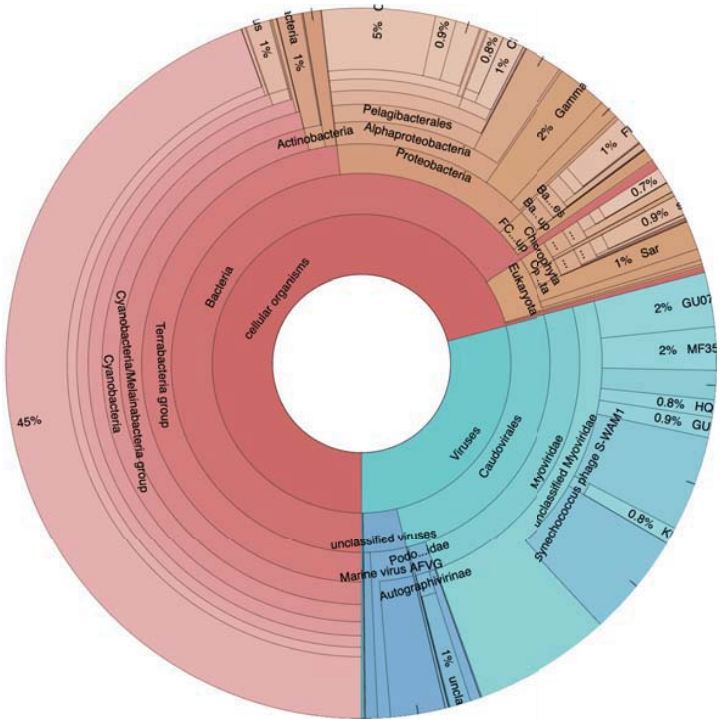
153

Metagenomic species-based profiling



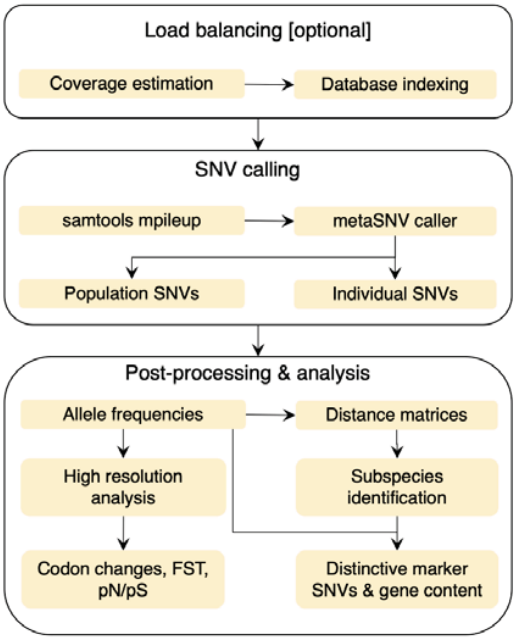
154

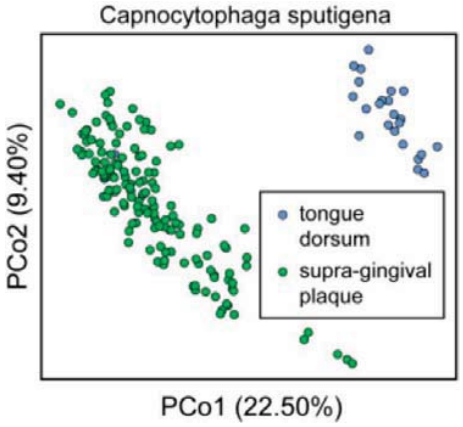
Plotting microbiome composition - Krona plot



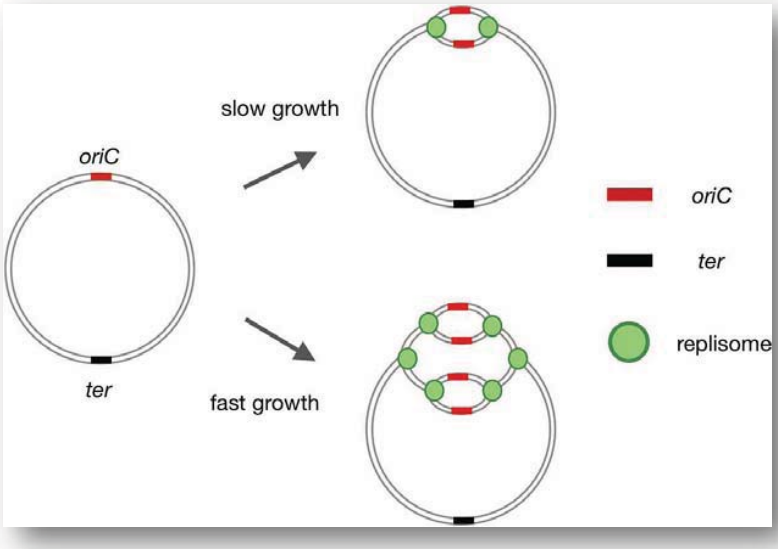
https://github.com/marbl/Krona/wiki/KronaTools

Subspecies analysis - meta-SNV





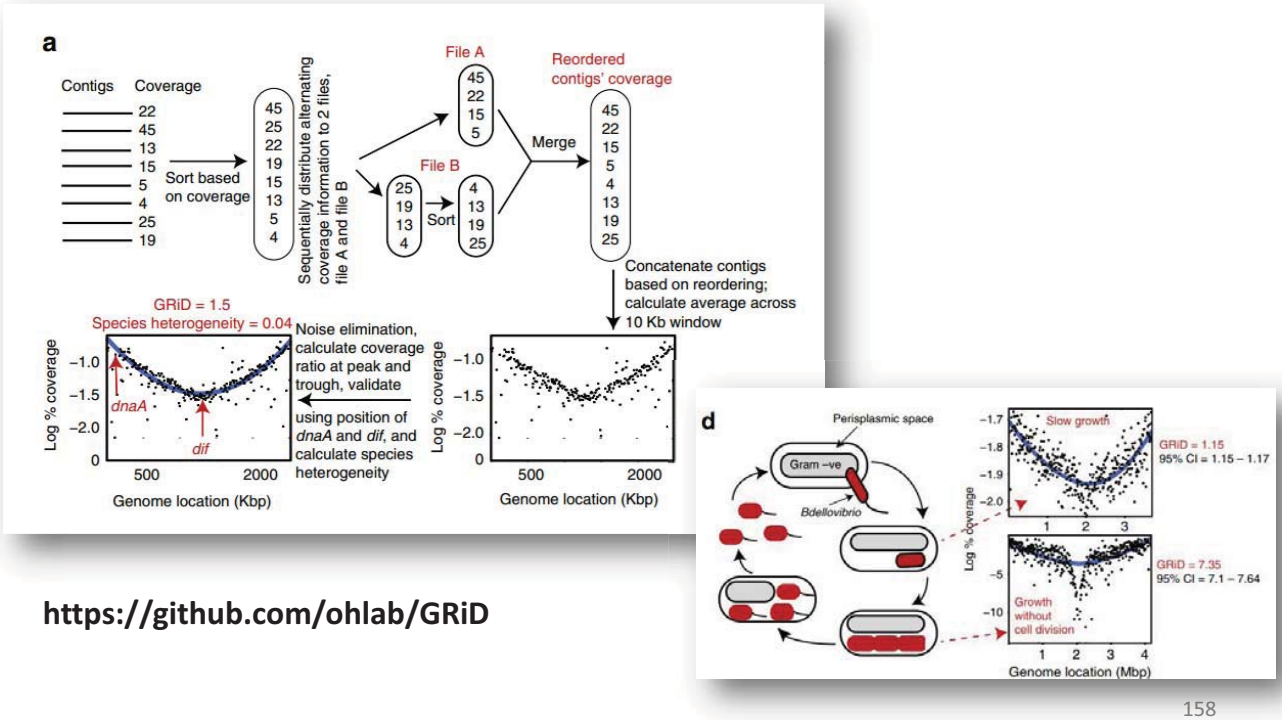
Growth-rate estimation: GRiD



Different sequencing depths from replication of origins

157

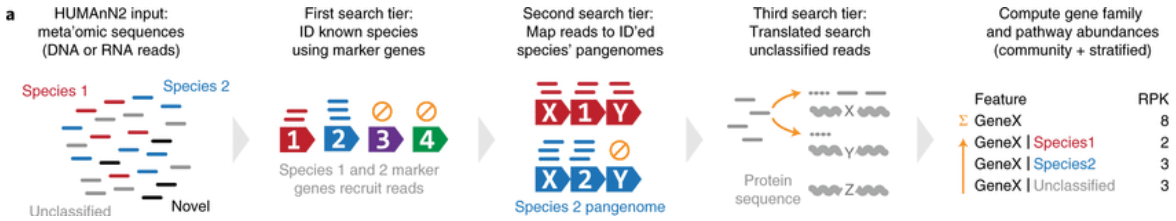
Growth-rate estimation: GRiD



https://github.com/ohlab/GRiD

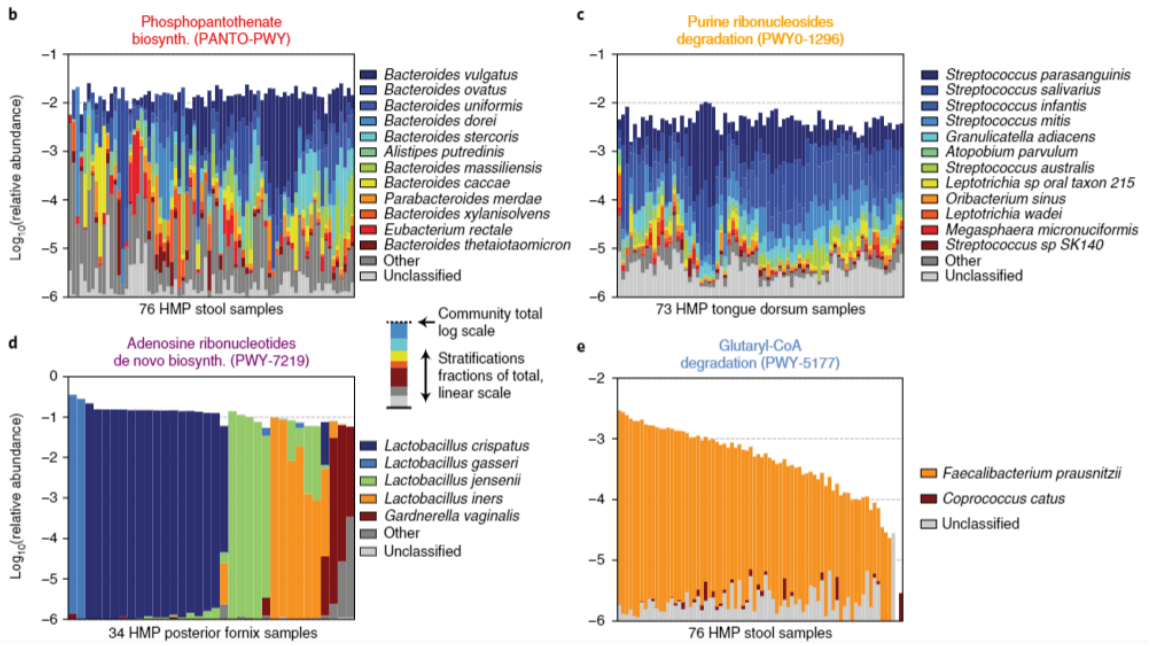
158

HUMAnN2 - Functional profiling



https://huttenhower.sph.harvard.edu/humann2/

HUMAnN2 - Functional profiling



End of lecture

- Thank you for all students and collaborators

GIST
All lab members …

King's College London

Saeed Shoaie
Gholamreza Bidkhori
David Gomez
Elizabeth Witherden
David Moyes
Gordon Proctor

Institute of
Liver Studies
and Transplantation

…

INRAe

Emmanuelle
Lechatelier
Nicolas Pons
Mathieu Almeida
Florian
Dusko Ehrlich
…

161

Appendix

162

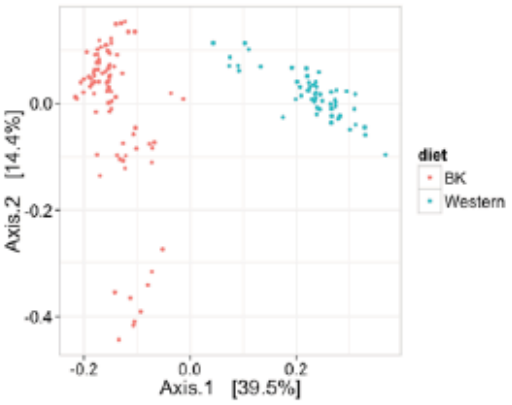
Ordination Methods

Project high-dimensional data onto lower dimensions

P taxa

N samples

0,1,5,1,0,1,2,1,0,0,9,…
7,2,0,0,0,0,0,0,1,0,0,…
0,0,0,0,0,0,8,0,0,0,1,…
0,0,0,1,0,1,2,0,0,0,5,…
0,1,0,2,0,0,0,1,0,0,4,…
0,0,0,1,9,1,2,5,2,0,1,…
0,0,0,0,0,1,2,1,8,0,0,…
0,0,0,0,9,4,0,0,0,0,1,…
.
.

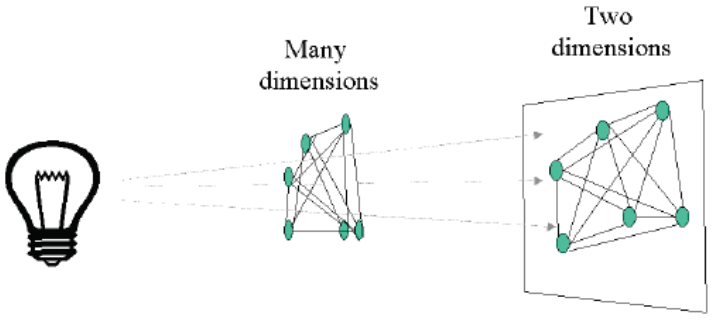


P-dimensions

2-dimensions

Multi-dimensional Scaling

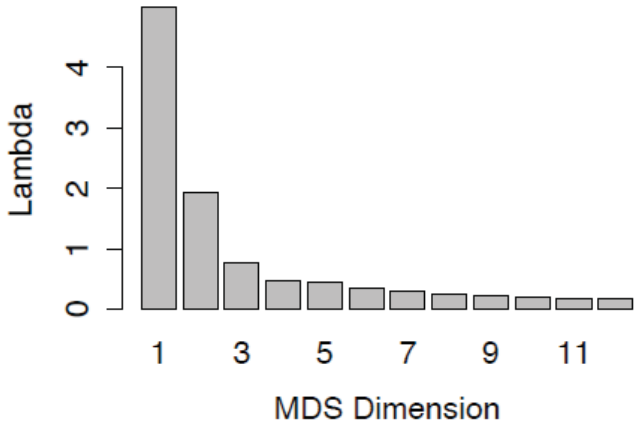
Why MDS? It works with any distance!



Input distance matrix can by Bray-Curtis, Unifrac, …

MDS Scree Plot

These values are the relative quantity of variability represented in each new dimension



165