KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (오프라인)

Human microbiome studies with bioinformatics approaches

이선재 _ GIST

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 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 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 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 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 : 2월 24일 (토)

시간	강 의 (자연과학대학 28동 101호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:30	**의료빅데이터/인공지능 총론** 김헌성 교수(가톨릭대학교)
14:30-14:45	휴식
14:45-16:15	**의료영상 인공지능의 이해 및 의료영상 레이블링 실습** 백서연 교수(연석대학교)
16:15-16:30	휴식
16:30-18:00	**의료 정보처리 자동화 실습 / 독자적인 어플리케이션 만들기** 김선근 대표(원닥 주식회사), 서사도 조교

시간	강 의 (자연과학대학 28동 102호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:20	**EMR 데이터를 활용한 머신러닝 기반 예후예측:** **Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)** 고태훈 교수(가톨릭대학교)
14:20-14:40	휴식
14:40-16:00	**Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단:** **Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)** 고태훈 교수(가톨릭대학교)
16:00-16:20	휴식
16:20-17:40	**심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network +** **Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)** 고태훈 교수(가톨릭대학교)

DAY1 : 2월 26일 (월)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**DNN (이론)** 이상근 교수(고려대학교)
10:50-11:00	휴식
11:00-12:10	**CNN (이론)** 이상근 교수(고려대학교)
12:10-13:40	점심
13:40-15:10	**RNN, ChatGPT, XAI (이론)** 이상근 교수(고려대학교)
15:10-15:20	휴식
15:20-16:50	**CNN/RNN 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)** 이정현 조교, 한성민 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Best practice for single-cell data analysis** 박종은 교수(KAIST)
11:00-11:10	휴식
11:10-12:40	**Practice1: Scanpy basic workflow** 정성민 조교, 고용준 조교
12:40-14:10	점심
14:10-15:30	**Public database, data integration, reference mapping, multiomics** 박종은 교수(KAIST)
15:30-15:40	휴식
15:40-16:50	**Practice2: Advanced single-cell analysis (siVI universe)** 정성민 조교, 고용준 조교

DAY1 : 2월 27일 (화)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**AI-based protein structure prediction** **- Intro to protein structure prediction** **- Early AI-based approaches** **- AlphaFold and RoseTTAFold** 백민경 교수(서울대학교)
10:50-11:00	휴식
11:00-12:10	**단백질 구조 예측 실습** **- ColabFold를 활용한 단백질 구조 및 상호작용 예측** **- Tips &Tricks for better structure modeling** 백민경 교수(서울대학교)
12:10-13:40	점심
13:40-15:10	**AI-based protein design** **- Intro to protein design** **- Protein backbone design using RFdiffusion** **- Protein sequence design using ProteinMPNN** 백민경 교수(서울대학교)
15:10-15:20	휴식
15:20-16:50	**단백질 디자인 실습** **- RFdiffusion 및 ProteinMPNN의 활용법 실습** 백민경 교수(서울대학교)

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Single-cell biology** 최정민 교수(고려대학교)
11:00-11:10	휴식
11:10-12:40	**i. Unsupervised Spatial transcriptome analysis** **ii. Tumor Boundary Determination in Spatial Transcriptomics** 유광민 조교, 이문영 조교
12:40-14:10	점심
14:10-15:30	**i. Deconvolution Analysis Using Single-cell RNA Sequencing and Spatial Transcriptomics** **ii. Cell-Cell Interaction Analysis in Spatial Transcriptomics** 김지현 조교, 최승지 조교
15:30-15:40	휴식
15:40-16:50	**i. Open Chromatin Region Analysis and Biological Interpretation of Using scATAC-seq Dataset** **ii. Construction of Gene Regulatory Networks Based on Integrated Analysis of scATAC-seq and scRNA-seq Datasets** 천하림 조교, 이호진 조교

DAY1 : 2월 28일 (수)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Transformers (이론)** 전민지 교수 (고려대학교)
11:00-11:10	휴식
11:10-12:40	**Introduction to Transformers (실습)** 봉현수 조교, 임우택 조교
12:40-14:10	점심
14:10-15:40	**Deep learning in Bioinformatics** 노미나 교수(한양대학교)
15:40-15:50	휴식
15:50-17:20	**Deep learning model을 이용한 실습** 박예솔 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**마이크로바이옴 기본 이론** 이선재 교수(GIST)
10:50-11:00	휴식
11:00-12:10	**16S rRNA amplicon seq. - DADA2** 조준우 조교, 백재우 조교
12:10-13:40	점심
13:40-14:40	**최신 메타지놈 분석 기법의 현황** 이선재 교수(GIST)
14:40-14:50	휴식
14:50-16:50	**Shotgun metagenome 분석 (Linux)** 조준우 조교, 백재우 조교

DAY1 : 2월 29일 (목)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**화학정보학 기초(Cheminformatics) / 약물특성 및 약물다움(druglikeness)** **Molecular Notations &Descriptors / AI 신약개발을 위한 Databases** **AI 신약개발을 위한 Programming 기초** 김동섭 교수(KAIST)
10:50-11:00	휴식
11:00-12:10	**Google Colab에 RDKit 설치 / 화합물 정보 읽기 실습** **Bioactivity database 검색 및 정보 읽기 실습** **Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습** 정수재 조교, 나민주 조교
12:10-13:40	점심
13:40-15:10	**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(KAIST)
15:10-15:20	휴식
15:20-16:50	**QSAR modeling 전체 과정 실습/ 화합물의 Bioactivity 예측 모델 개발** **Virtual screening 과정을 통한 신약후보물질 발굴 실습** 정수재 조교, 나민주 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Single cell multiomics 이론 / Gene regulatory network 이론** 김준일 교수(숭실대학교)
11:00-11:10	휴식
11:10-12:40	**Seurat/Signac, ArchR, TENET+ 실습** 김현규 조교, 정회빈 조교
12:40-14:10	점심
14:10-15:40	**롱리드 시퀀싱 소개 및 유전체 조립 실습** 김준 교수(충남대학교)
15:40-15:50	휴식
15:50-17:20	**변이 분석 및 시각화 실습** 김준 교수(충남대학교)

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/식습관 등의 차이로 인한 confounding factor들이 큰 문제이다. 본 강의에서는 현재 마이크로바이옴 연구의 최근 동향과 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 (선택)

* 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

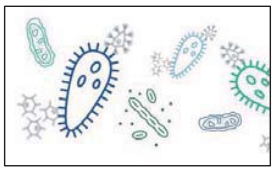
2006	B.S. in Bioinformatics, KAIST, Korea
2004	M.S. in Bioinformatics, KAIST, Korea
2007	Ph.D. in Bioinformatics, KAIST, Korea

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)

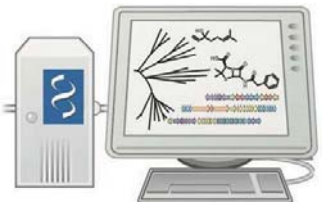
Selected Publications (5 maximum)

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



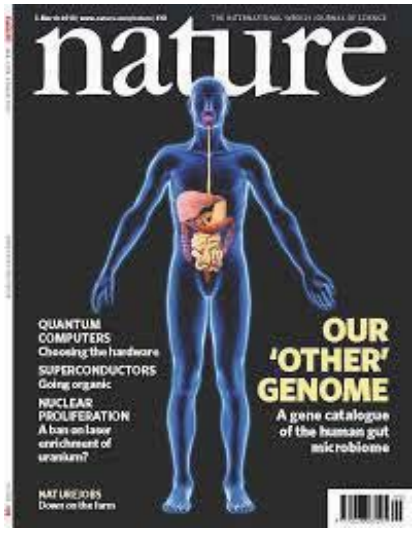
GIST | Life Mining Lab
total 149 pages

Notice

- https://sites.google.com/view/gist-life-mining-lab/home/workshop

2

We humans are “microbial”

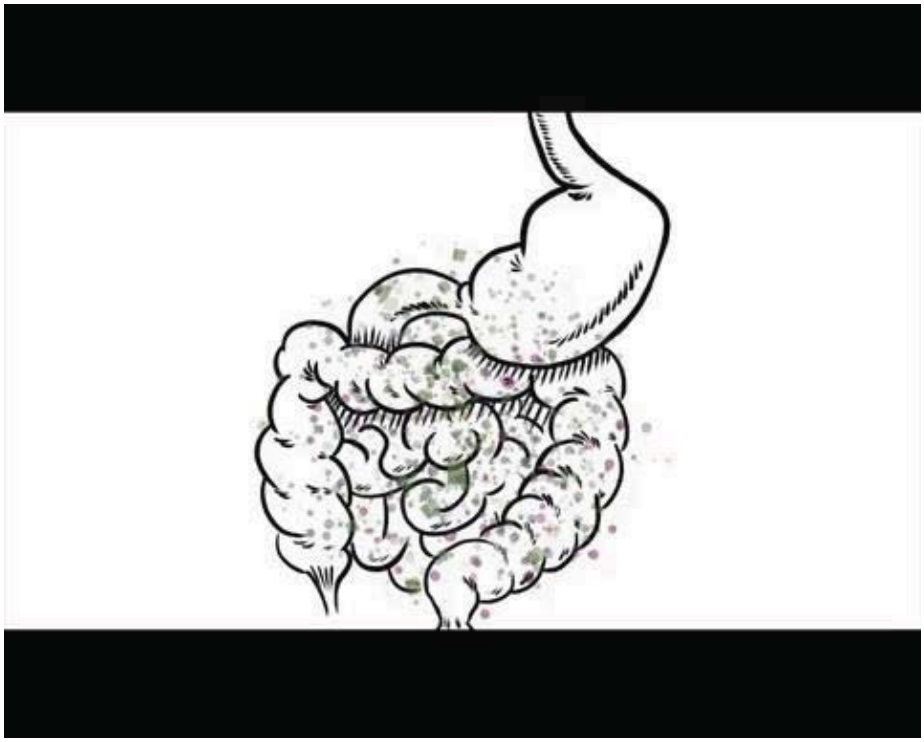


100 trillion cells **10X** human cells

20 million genes **100X** human genes

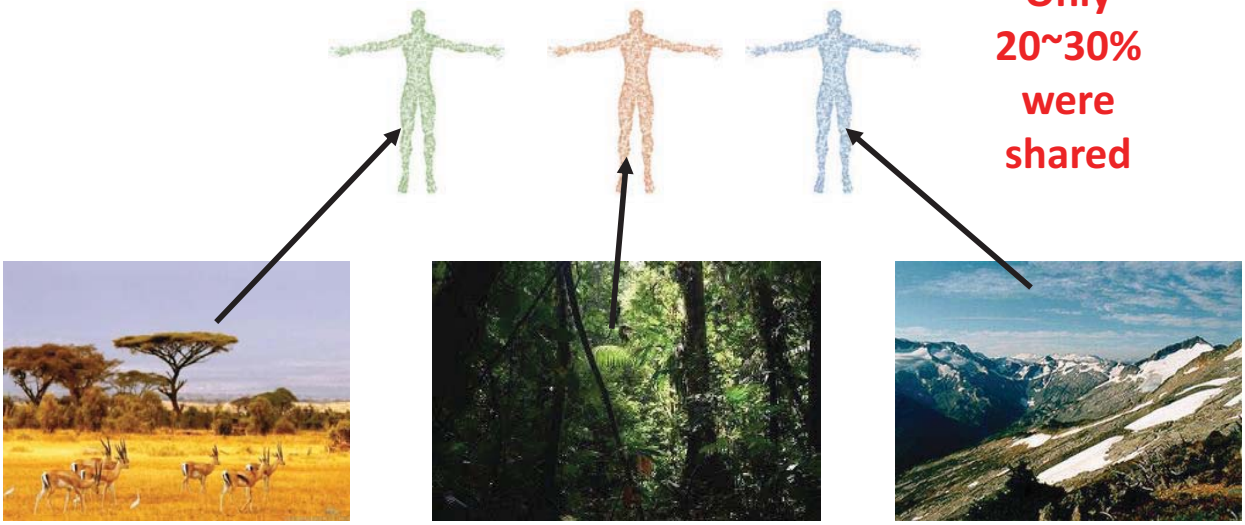
3

Individuals harbours own unique microbiome



4

Individuals harbours own unique microbiome

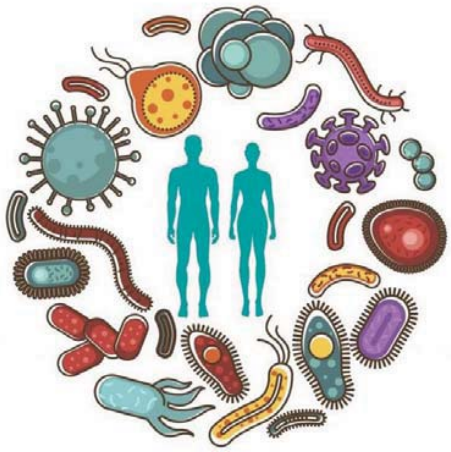


Only 20~30% were shared

Savannah **Jungle** **Tundra**

REF | Rob Knight, TED talks

5

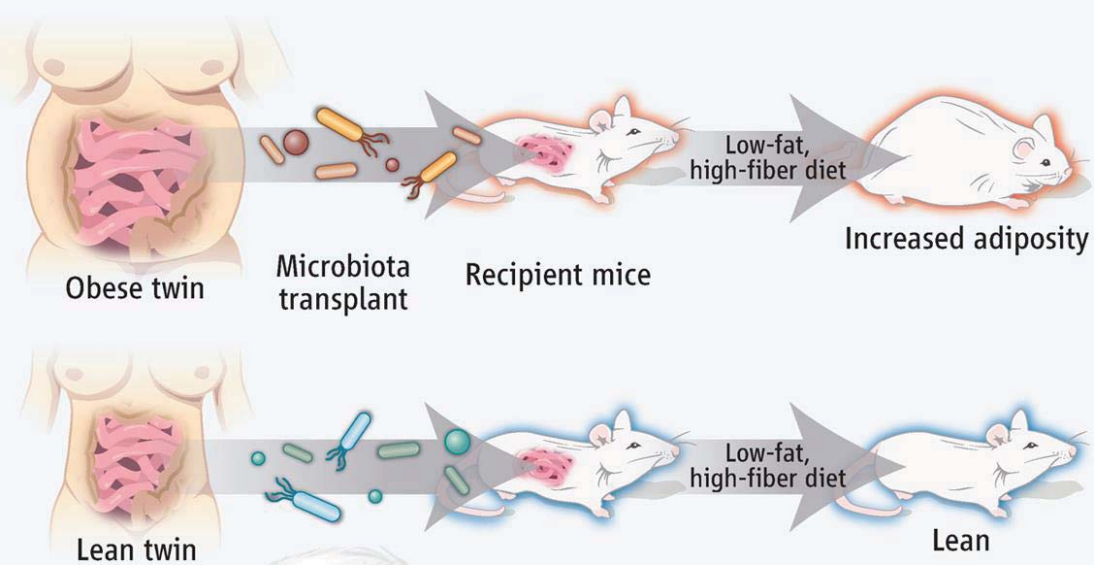


Microbiome = Our Second Genome

6

Why microbiome is important?

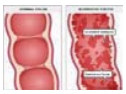
1 Microbiome determines host phenotypes



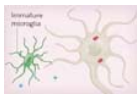
REF | Alan W. Walkter, Julian Parkhill, Science, 2013

1 Microbiome determines host phenotypes

Obesity



Colitis

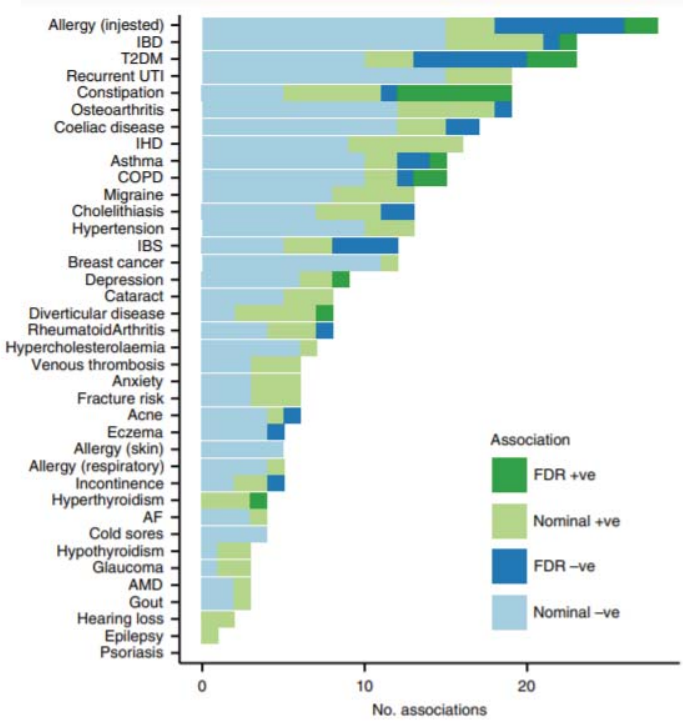


Parkinson's

Autism

Schizophrenia

1 Microbiome determines host phenotypes

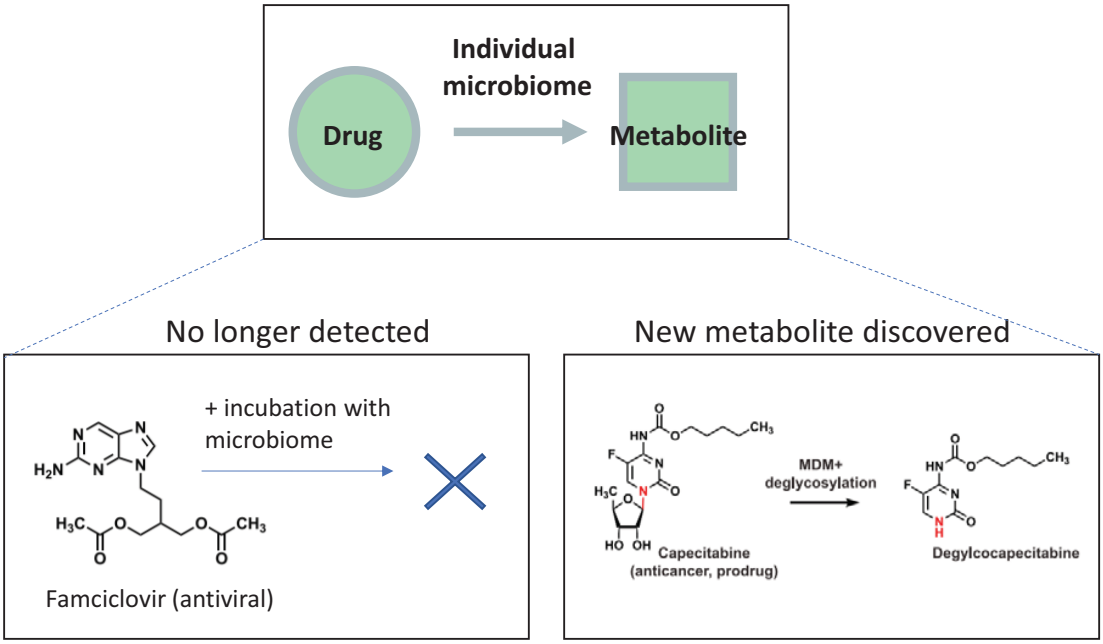


<u>Many common diseases</u> <u>associated with microbiome changes</u>

Allergy
Constipation
Migraine
T2D
Asthma
Hypertension
…

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

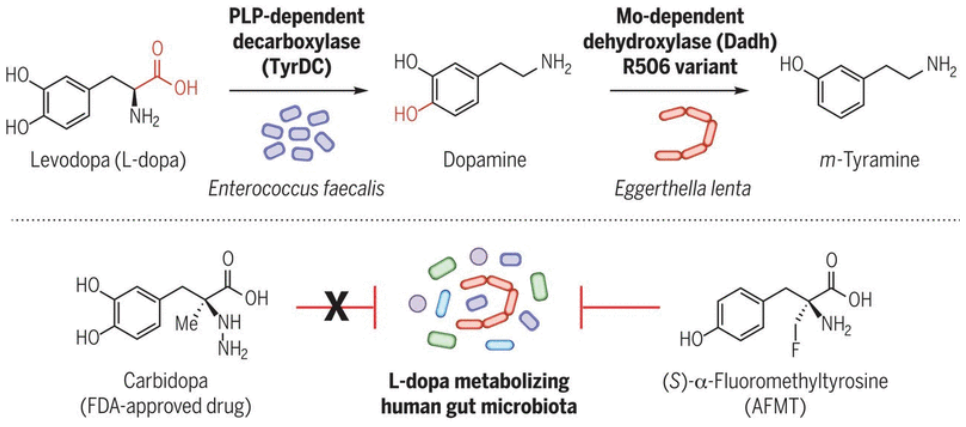
2 Microbiome affects individual drug response



REF | Bahar Javdan et al., Cell, 2020

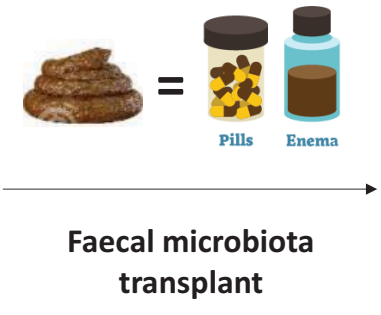
2 Microbiome affects individual drug response

Microbiome affects Levodopa efficiency



REF | Vayu Maini Rekda et al,. Science 2019

3 Healthy microbiome can treat diseases



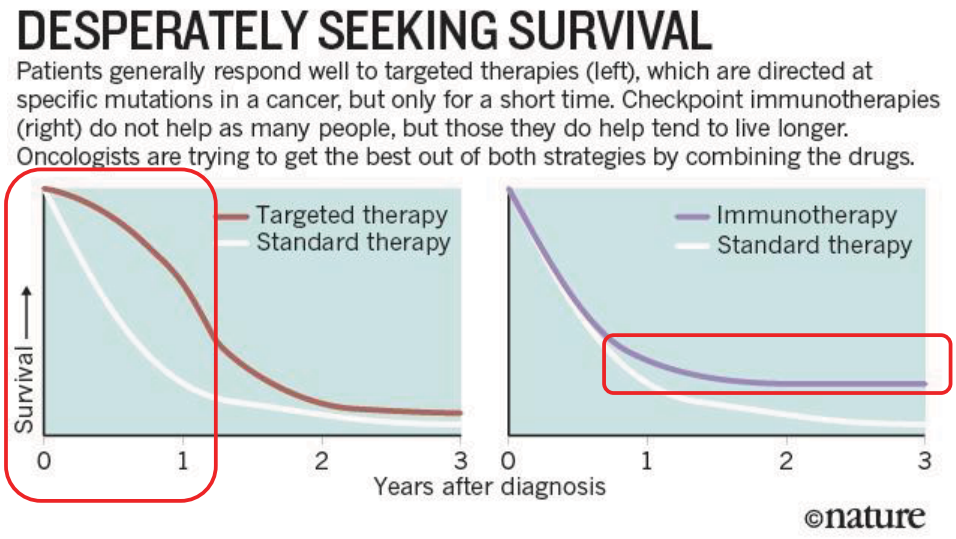
Colitis, C. difficile infection, etc

Healthy colon

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

13

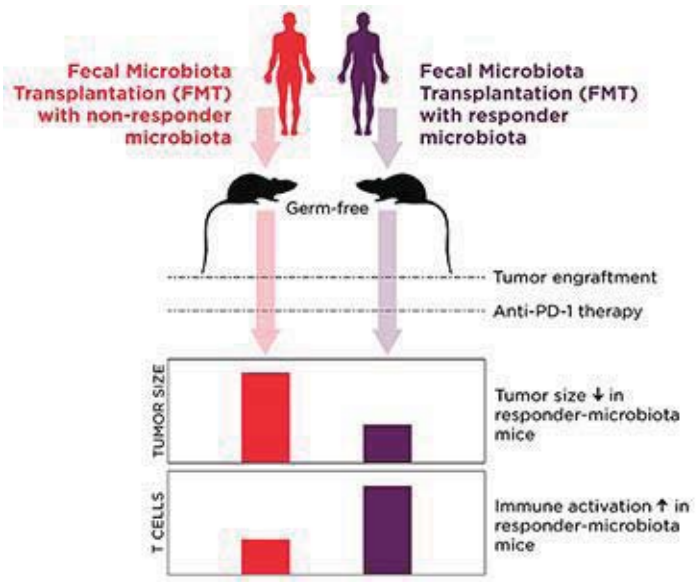
4 Microbiome affects immunotherapy efficacy



REF | Bertrand Routy et al., Science, 2018

14

4 Microbiome affects immunotherapy efficacy



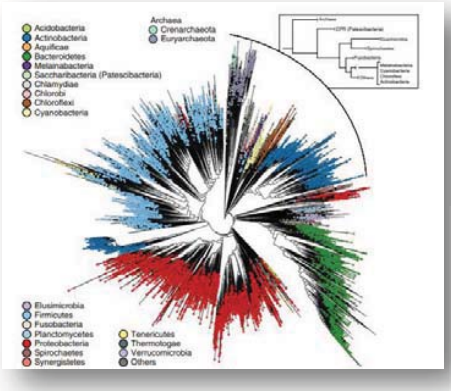
- Fecal microbiota of responders increased the efficacy of immunotherapy!

REF | Bertrand Routy et al., Science, 2018

15

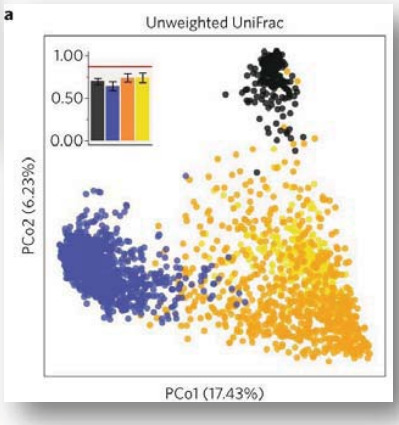
How to study microbiome?

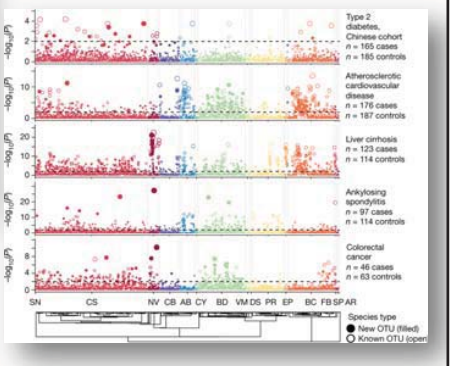
16



Phylogenetic trees

Principal coordinate analysis





Metagenome-wide association study

17

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

18

Overview

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

19

Key definitions

20

Key definitions …

Microbe?

Microbiome?

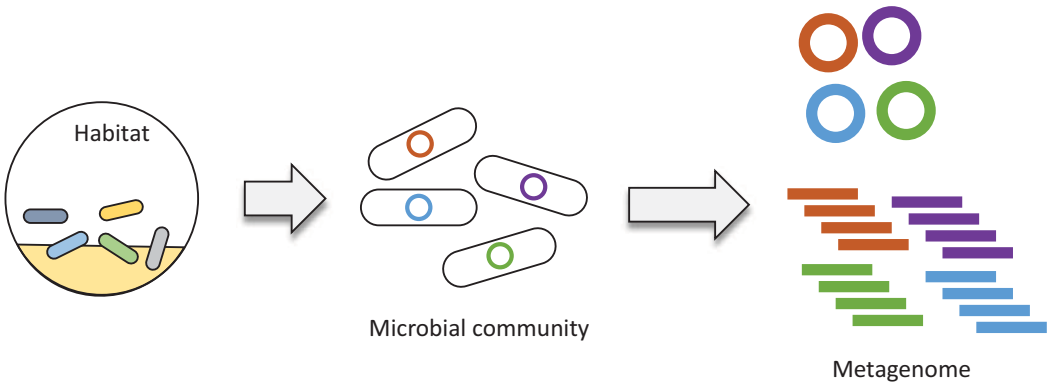
Metagenome?

Microorganisms?

Microbiota?

Key definitions …

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



Key definitions …

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

23

Key definitions …

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

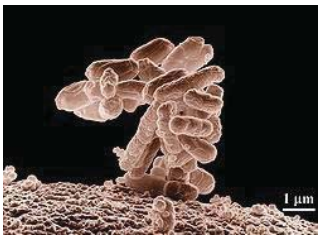
- i.e. **= metagenome of microbiota**

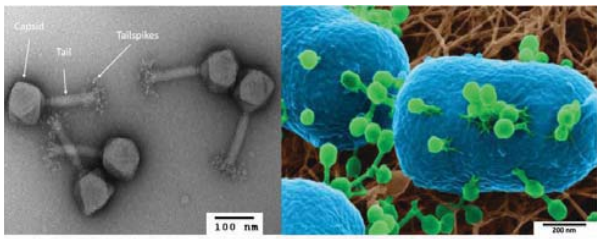
24

Microorganisms (미생물)

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

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

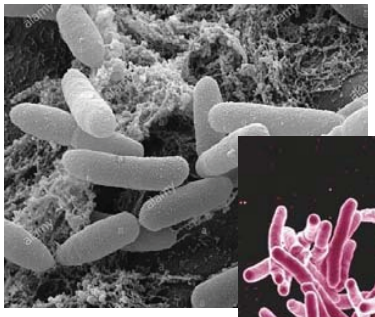




(a) (b)

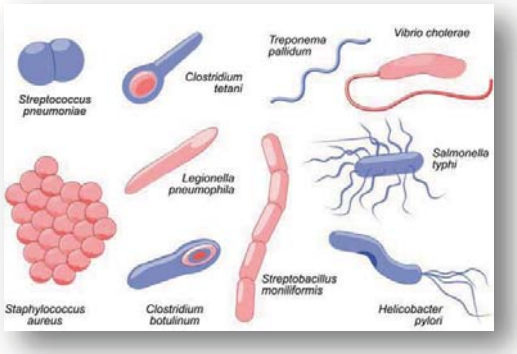
25

Bacteria (세균)



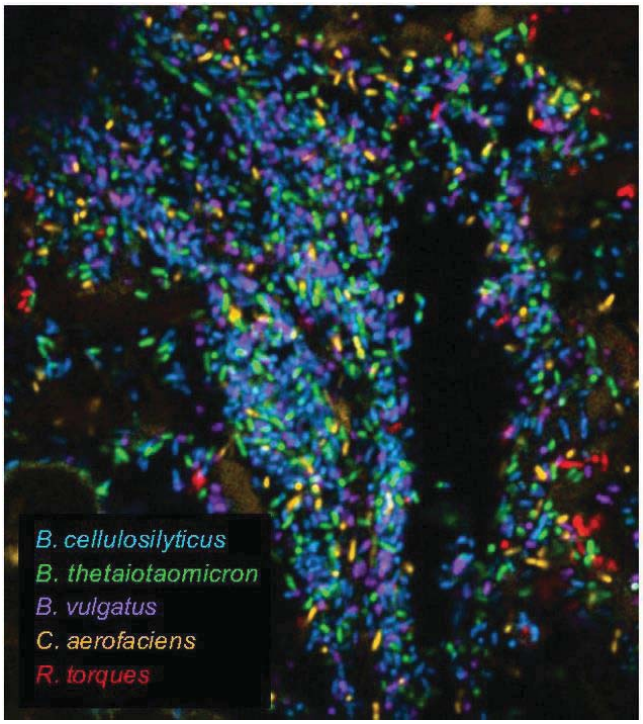
Escherichia coli

Mycobacterium tuberculosis



26

Bacteria (세균)

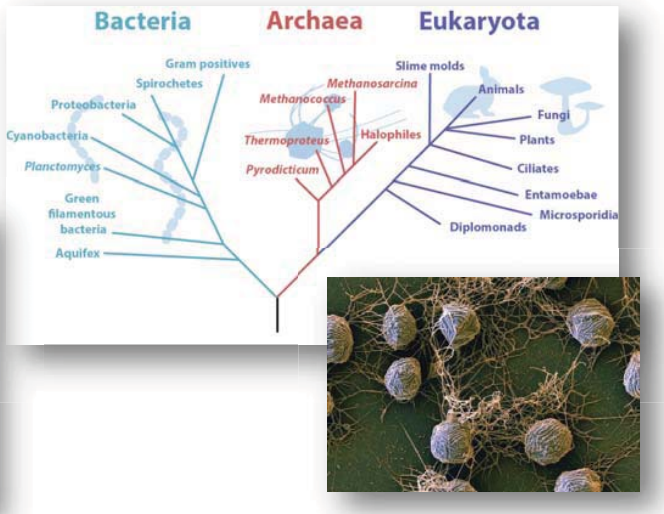


Spatial organization of human gut microbiota established in mice

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

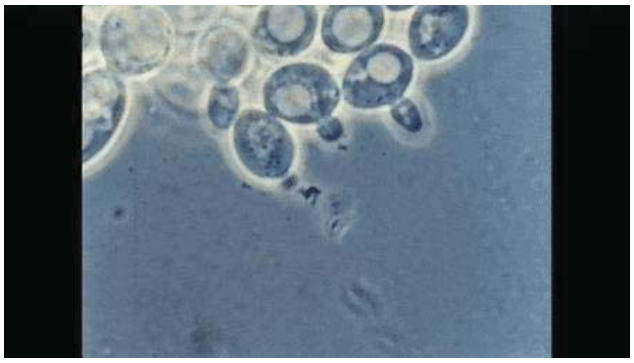
27

Archaea (고균)



28

Fungi (진균)



Budding yeast

Candida species

29

Key terminology…

Taxonomy?

Phylogeny?

Diversity …?

Symbiosis…

Dysbiosis…

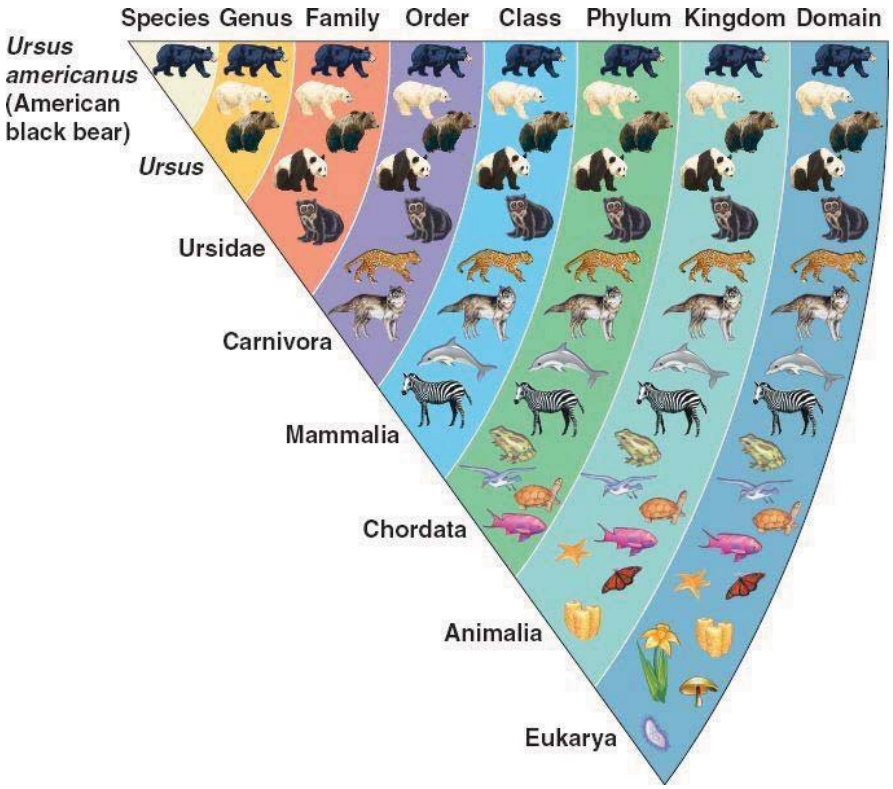
30

Taxonomy!

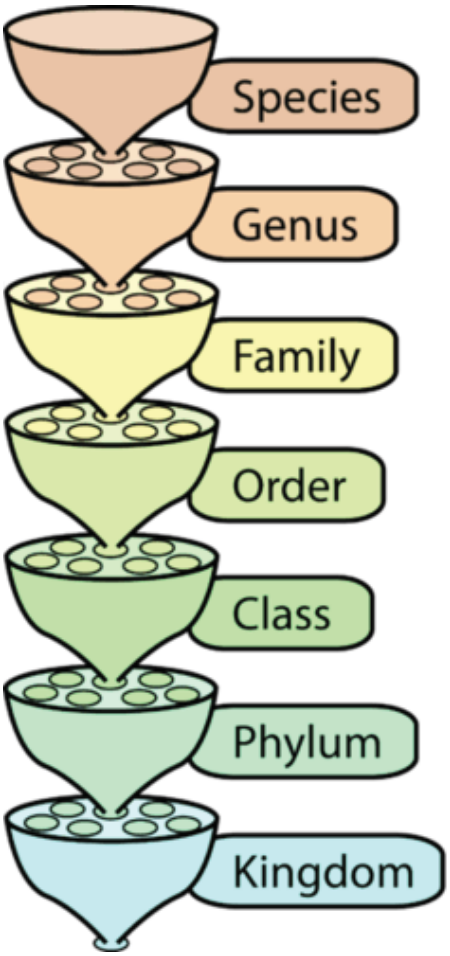
31

Taxonomy = classification = identity

32



33



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.

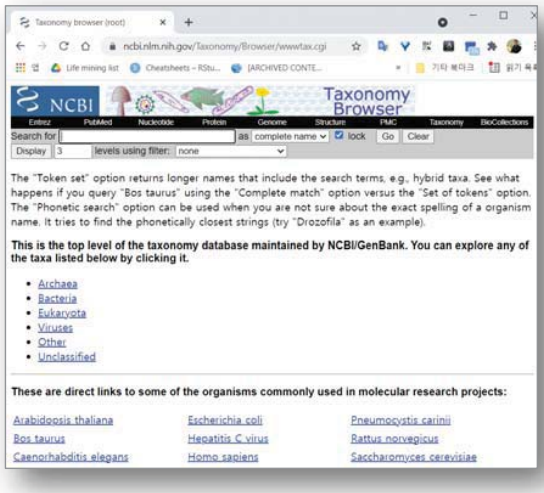
Mammals
Chordates with fur or hair and milk glands.

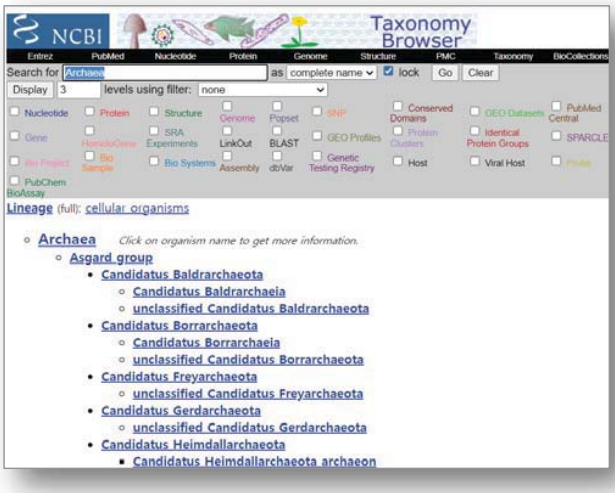
Chordates
Animals with a backbone.

Animals
Organisms able to move on their own.

34

Taxonomy database





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

35

[Taxonomy]

Taxonomy database

락토바실러스 유산균?

36

Taxonomy database

락토바실러스 (Lactobacillus) = genus 명칭

Taxonomy database

Genus

Species

Strain

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein

Search for lactobacillus as complete name lock Go Clear

Display 3 levels using filter: none

Nucleotide Protein Structure Genome Popset SNP Conserved Domains GEO D
Gene HomoloGene SRA Experiments LinkOut BLAST GEO Profiles Protein Clusters Identica
Bio Project Bio Sample Bio Systems Assembly dbVar Genetic Testing Registry Host Viral Ho
PubChem BioAssay

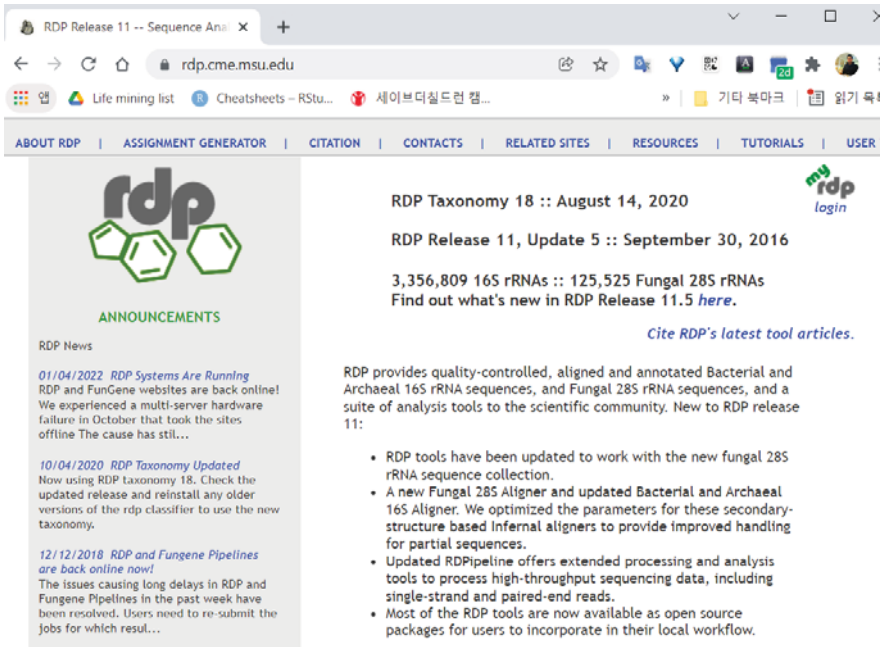
Lineage (full): cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Lactobacillales; Lactobacillacea

- Lactobacillus LinkOut *Click on organism name to get more information.*
  - Candidatus Lactobacillus pullistercoris
  - Candidatus Paralactobacillus gallistercoris
  - Lactobacillus acetotolerans LinkOut
    - Lactobacillus acetotolerans DSM 20749 = JCM 3825 LinkOut
  - Lactobacillus acidophilus LinkOut
    - Lactobacillus acidophilus 30SC LinkOut
    - Lactobacillus acidophilus ATCC 4796 LinkOut
    - Lactobacillus acidophilus CFH LinkOut
    - Lactobacillus acidophilus CIRM-BIA 442 LinkOut
    - Lactobacillus acidophilus CIRM-BIA 445 LinkOut
    - Lactobacillus acidophilus CRBIP 24179 LinkOut
    - Lactobacillus acidophilus DSM 20079 = JCM 1132 = NBRC 13951 = CIP 76.13 LinkOut
    - Lactobacillus acidophilus DSM 20242 LinkOut
    - Lactobacillus acidophilus DSM 9126 LinkOut
    - Lactobacillus acidophilus JV3179 LinkOut
    - Lactobacillus acidophilus La-14 LinkOut
    - Lactobacillus acidophilus NCFM LinkOut
  - Lactobacillus alvei LinkOut
  - Lactobacillus amylolyticus LinkOut
    - Lactobacillus amylolyticus DSM 11664 LinkOut
  - Lactobacillus amylovorus LinkOut
    - Lactobacillus amylovorus DSM 16698 LinkOut
    - Lactobacillus amylovorus DSM 20531 LinkOut
    - Lactobacillus amylovorus GRL 1112 LinkOut
    - Lactobacillus amylovorus GRL 1115 LinkOut
    - Lactobacillus amylovorus GRL1118 LinkOut
  - Lactobacillus animata

Database of 16S rRNA with taxonomy data

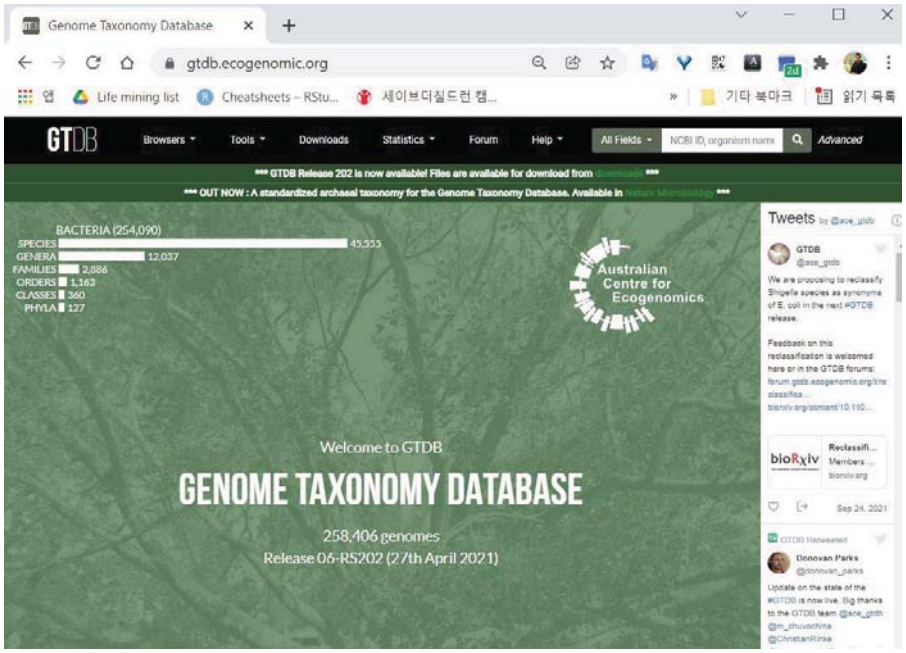
SILVA

Database of 16S rRNA with taxonomy data



RDP

Database of genomes with taxonomy data



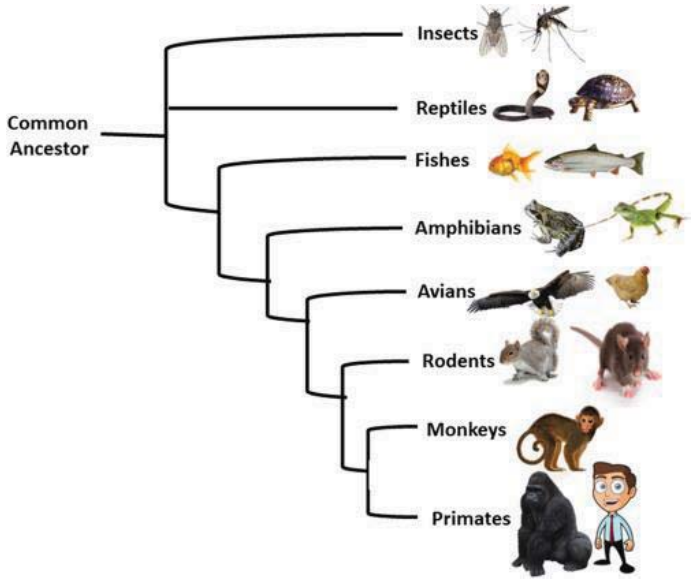
GTDB

Phylogeny?

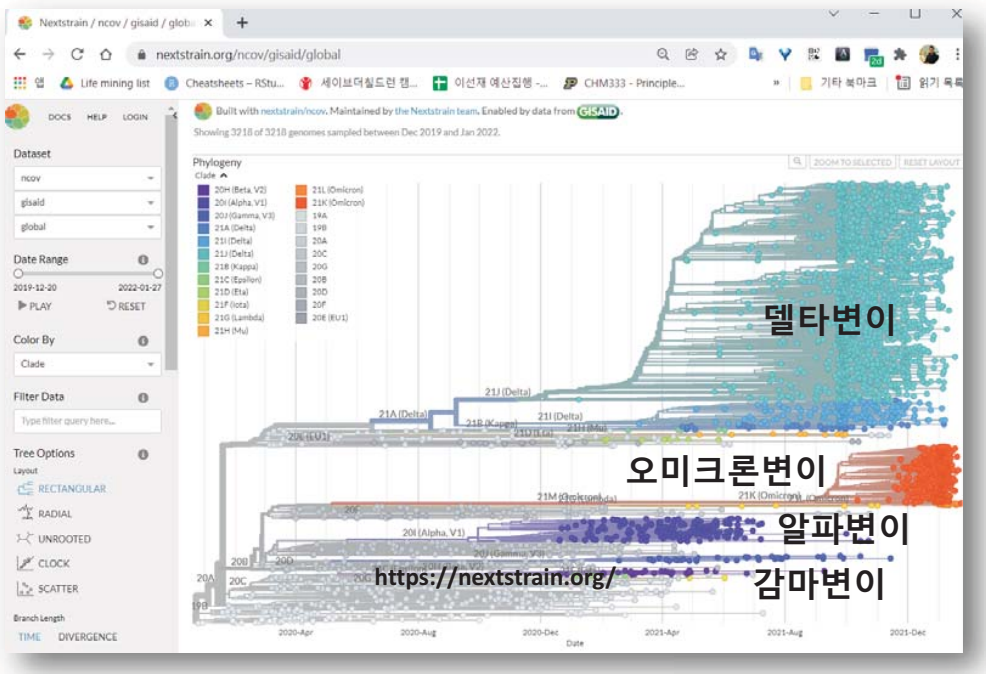
Phylogeny = evolutionary relationship

Phylogeny

- Evolutionary history between organisms

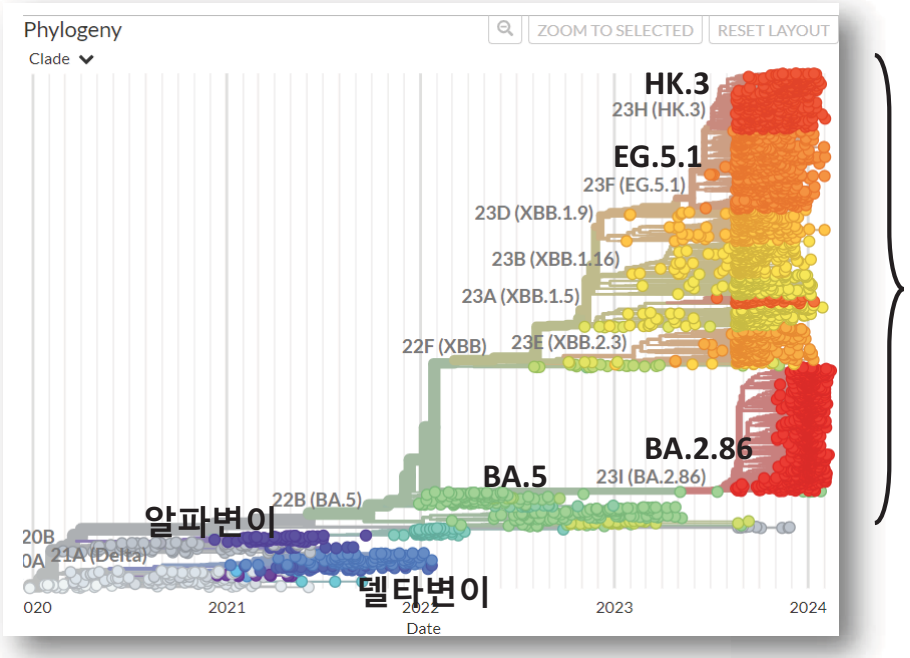


Evolutions of SARS-CoV-2 strains



https://nextstrain.org/sars-cov-2/

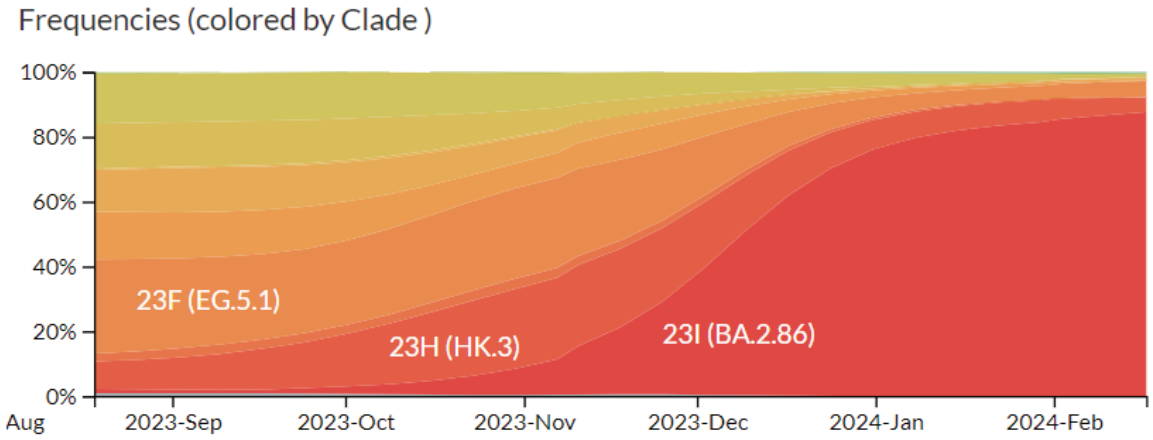
Evolutions of SARS-CoV-2 strains



현재상황

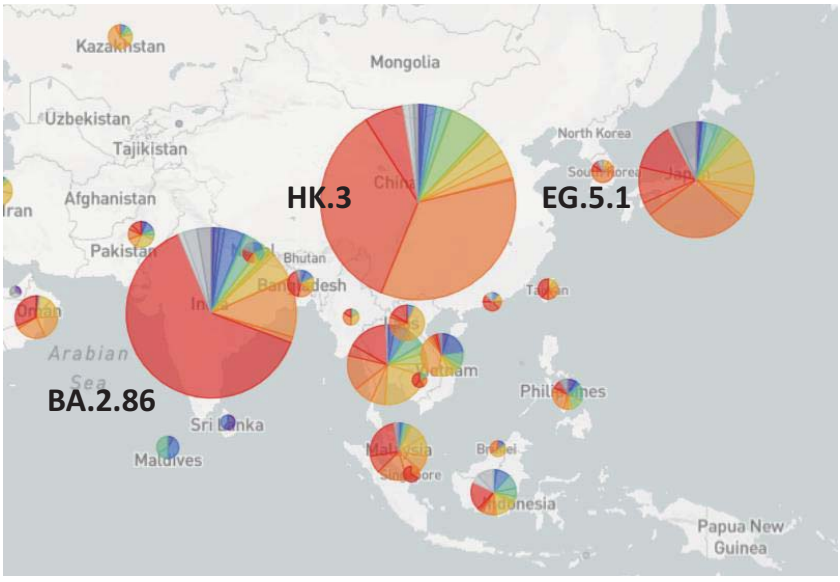
https://nextstrain.org/sars-cov-2/

Evolutions of SARS-CoV-2 strains



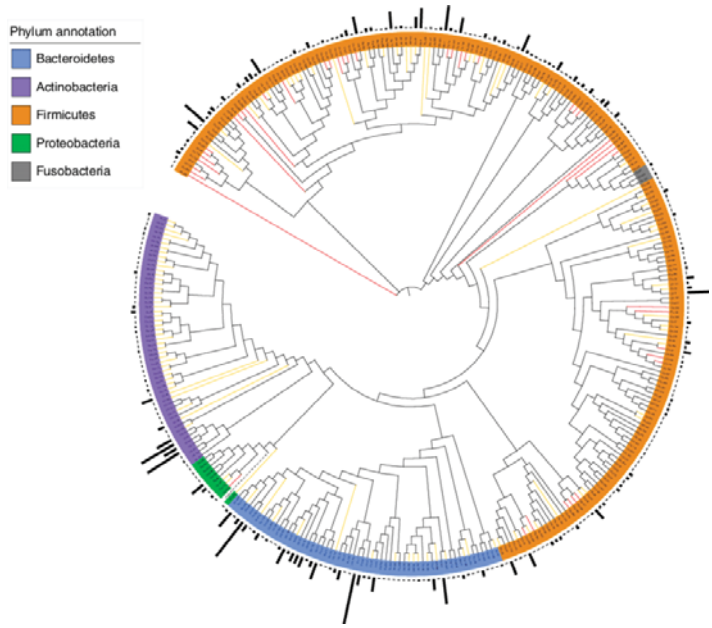
https://nextstrain.org/ncov/gisaid/global/6m

Evolutions of SARS-CoV-2 strains



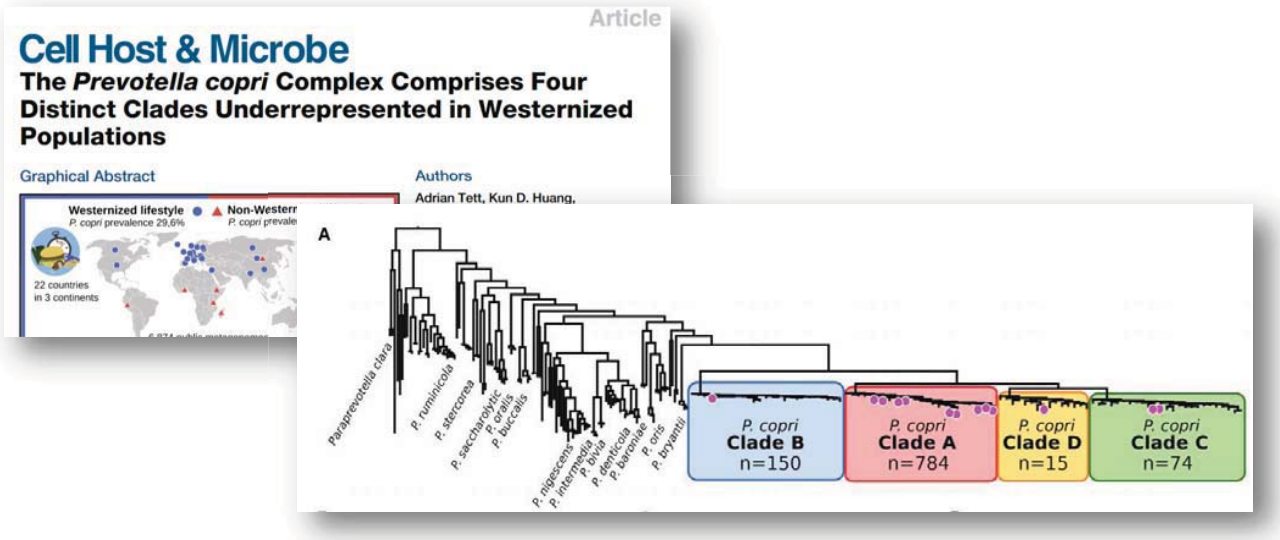
https://nextstrain.org/ncov/gisaid/global/6m

Phylogenetic trees of isolated gut bacteria



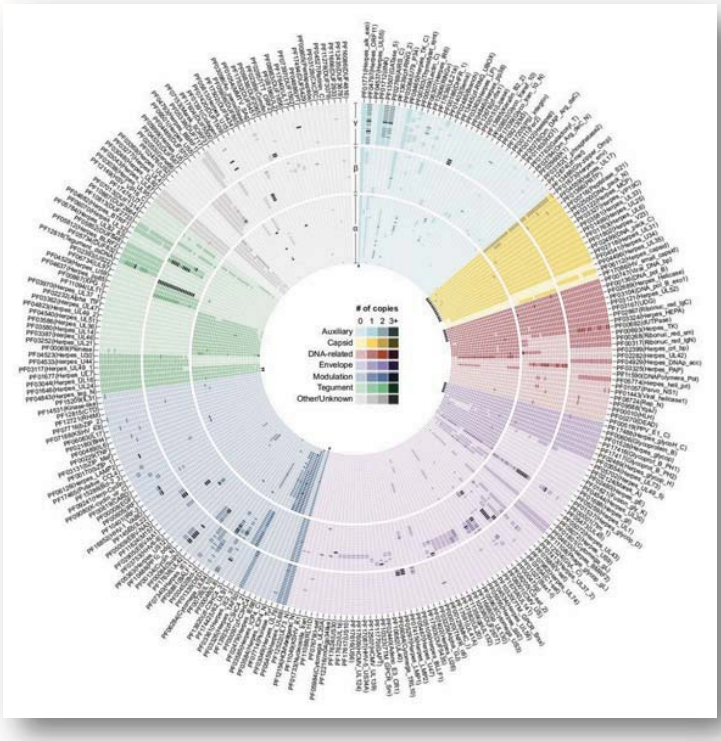
REF | Yuanqiang Zhou et al., Nature Biotechnology, 2019

Phylogenetic trees of “metagenome-assembled genomes”



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

Website - iTOL: interactive Tree of Life

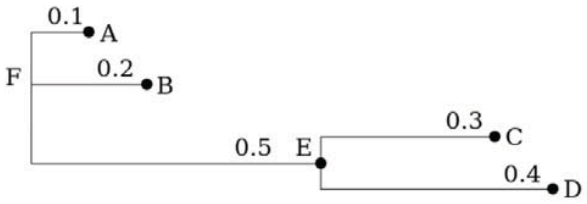


REF | https://itol.embl.de/

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

TAXA + DATA + TREES

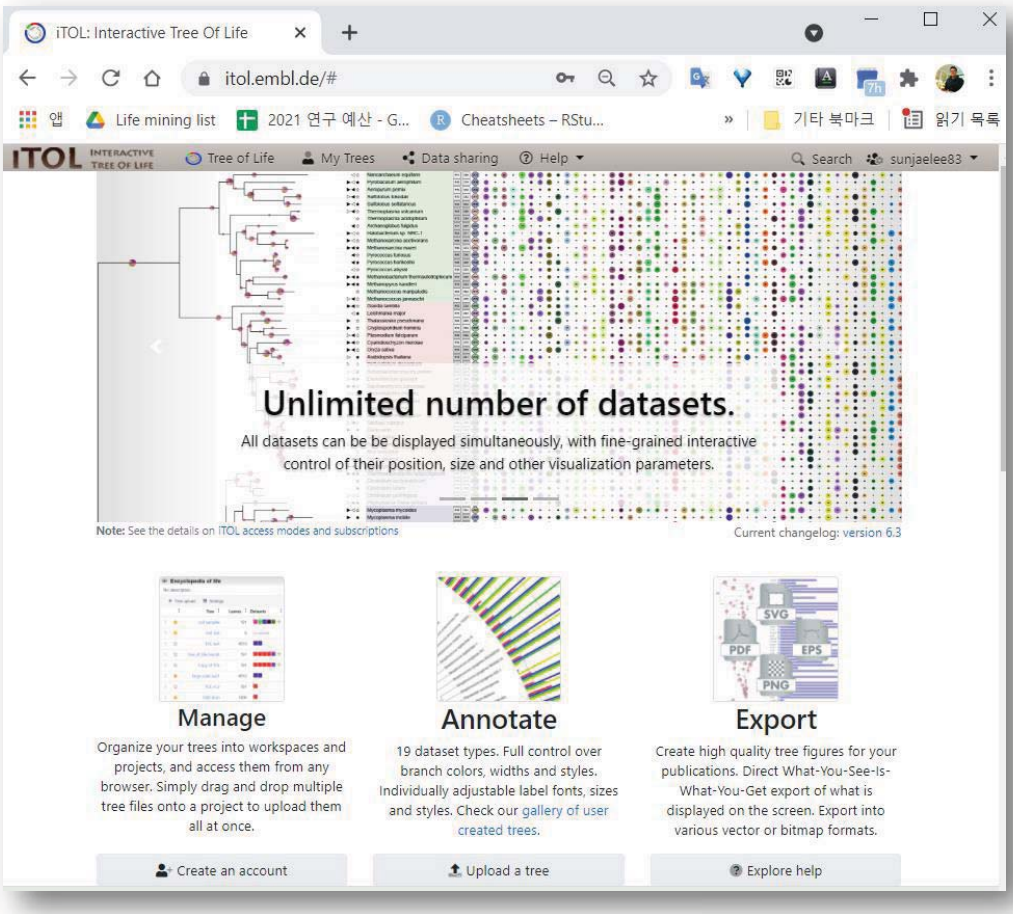
```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf
End;

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;
```

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

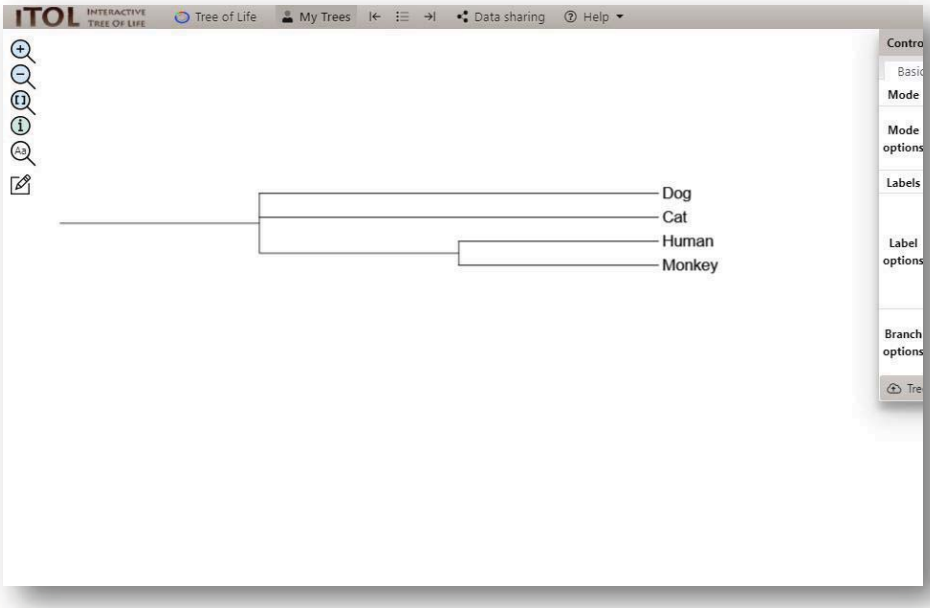
55



56

Example

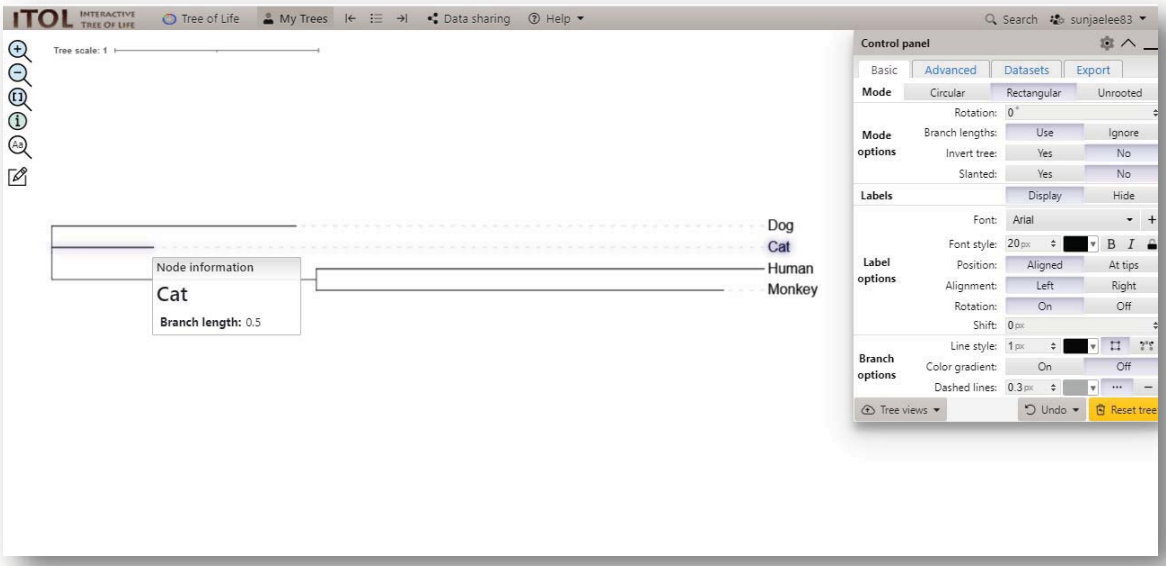
(Cat,Dog,(Monkey,Human));



57

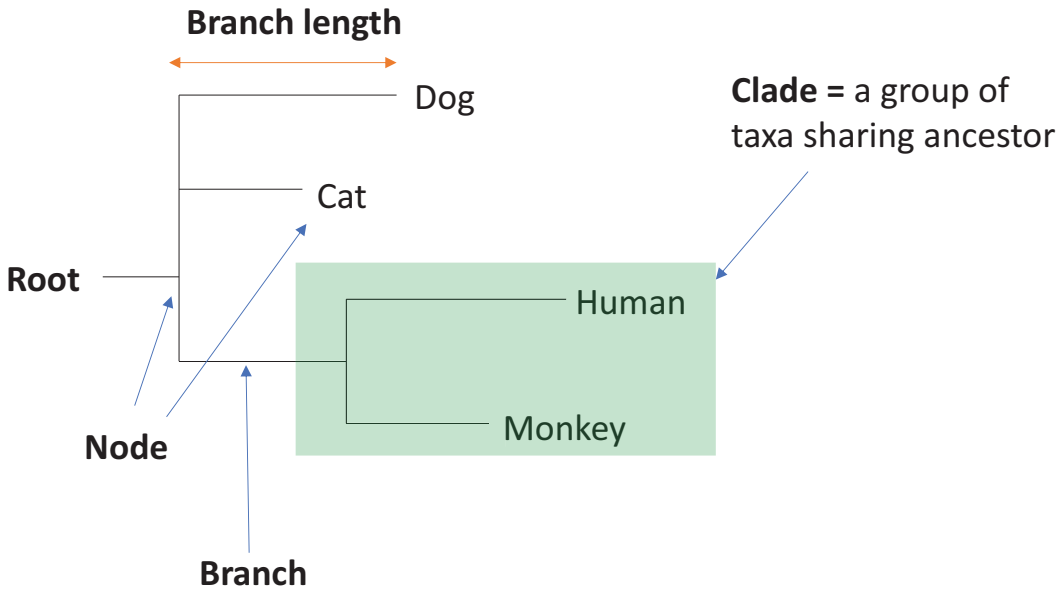
Example

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



58

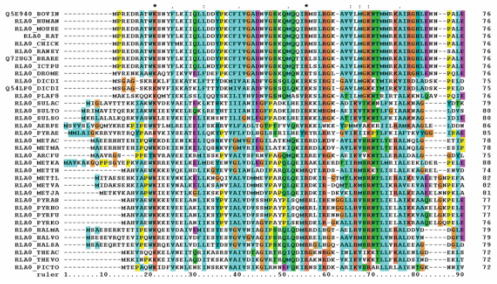
Phylogenetic tree

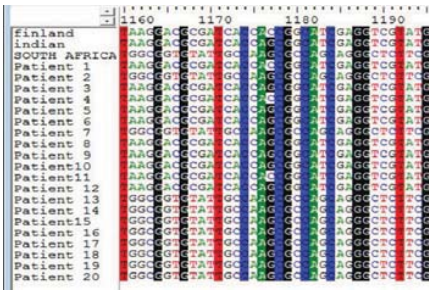


59

Phylogeny – sequence alignment

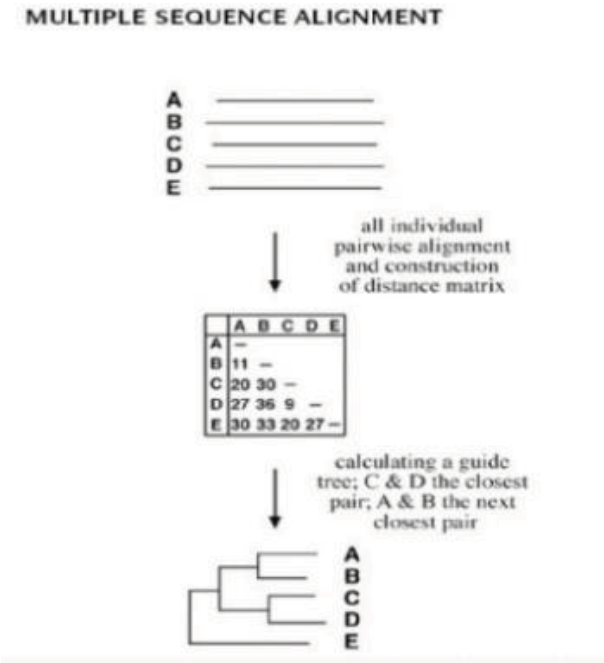
- Inferring regions of homology
  - 16S rRNA, house-keeping genes (multi locus sequence typing) etc
- Tools: Clustal, MUSCLE, T-Coffee, etc





60

Building distance matrix



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

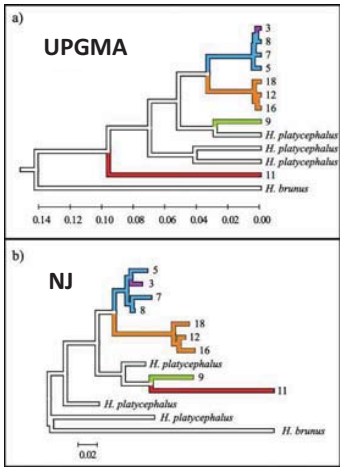
Phylogenetic tree methods

Distance-based methods

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

Criterion-based methods

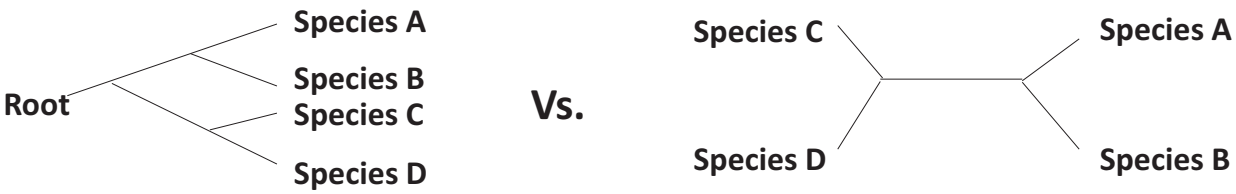
- Seeing sequences as characters
- Maximum-likelihood method
- Maximum parsimony method



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

Phylogenetic tree

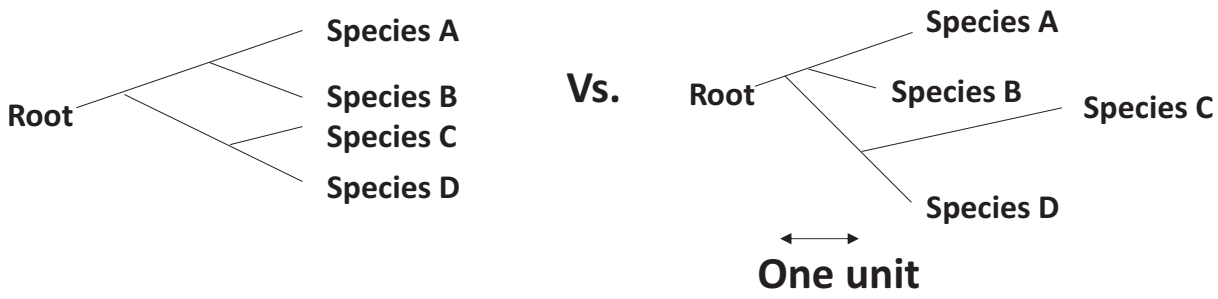
Rooted vs unrooted trees



UPGMA **NJ**

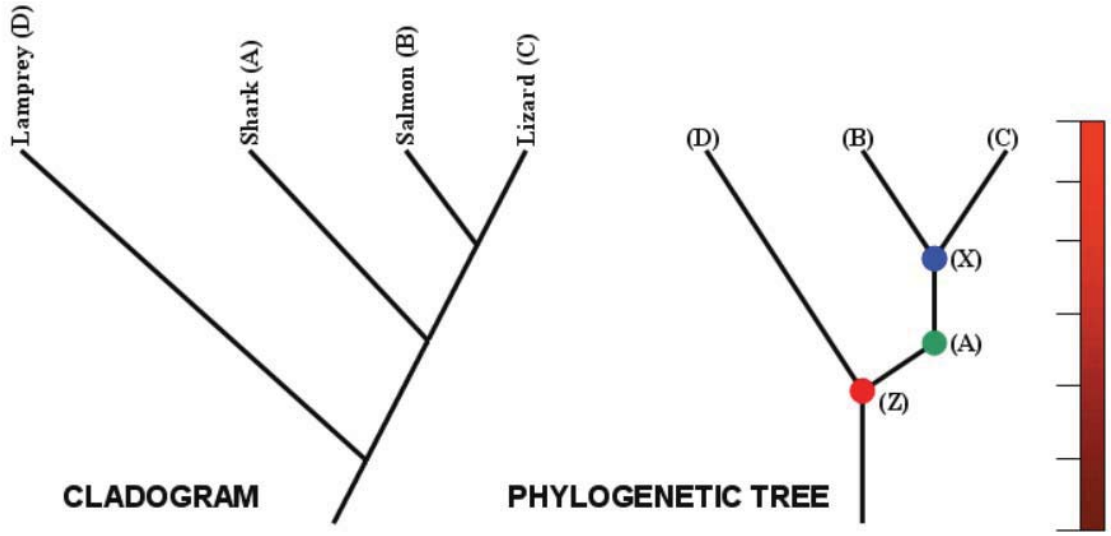
Phylogenetic tree

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



Phylogenetic tree vs cladogram

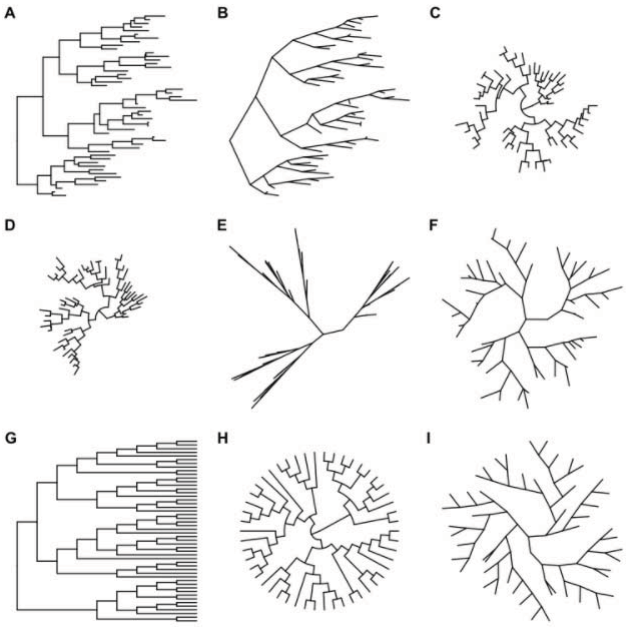
No meaning on branch lengths



65

[Phylogeny]

R package, ggtree

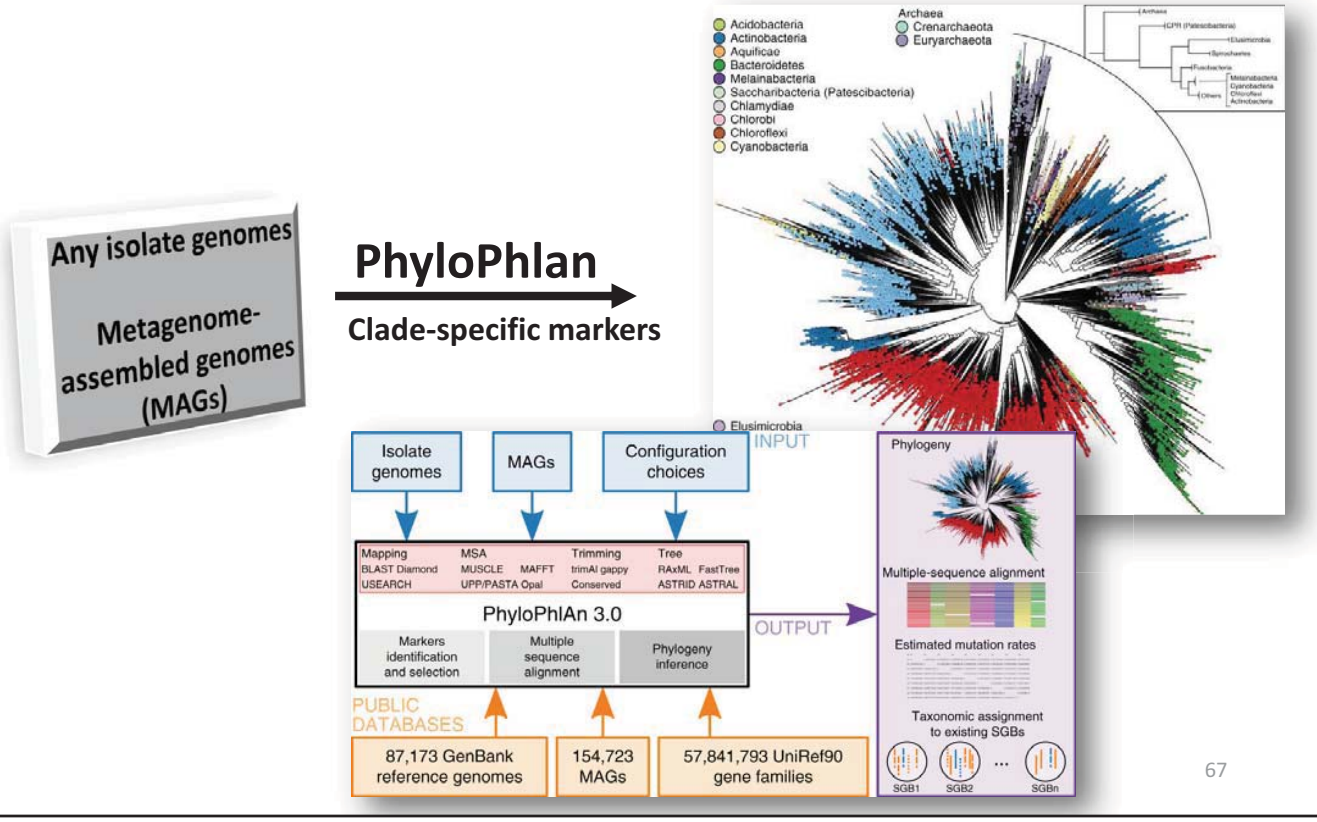



```
library(ggtree)
set.seed(2017-02-16)
tree <- rtree(50)
ggtree(tree)
ggtree(tree, layout="slanted")
ggtree(tree, layout="circular")
ggtree(tree, layout="fan", open.angle=120)
ggtree(tree, layout="equal_angle")
ggtree(tree, layout="daylight")
ggtree(tree, branch.length='none')
ggtree(tree, branch.length='none',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

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

66

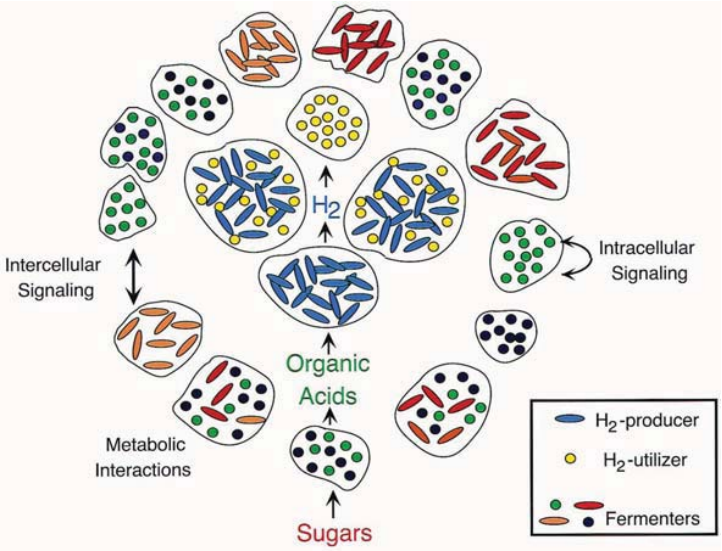
PhyloPhlan (ver3.0)



Diversity

Microbes are living as a community

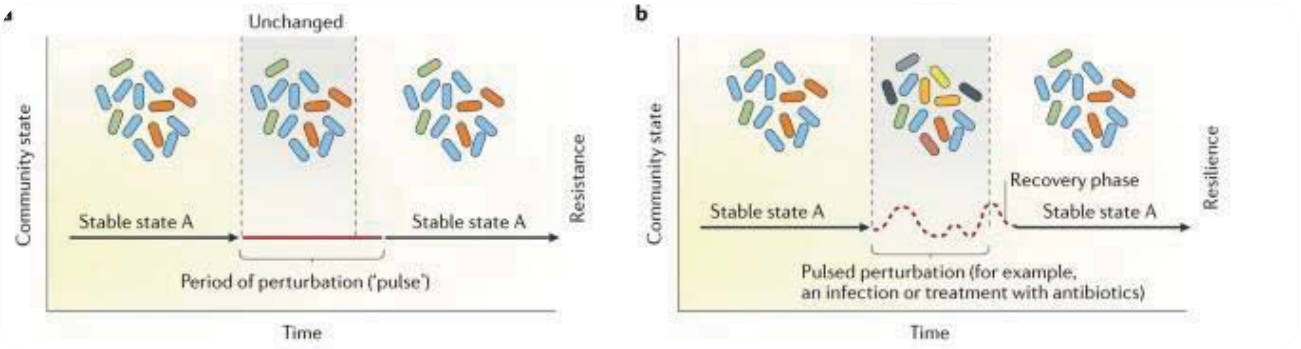
Microbial biofilm



69

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



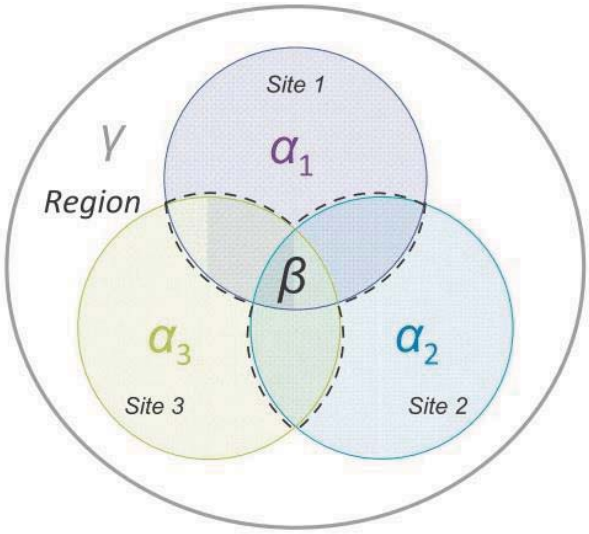
70

Key terminology

- Diversity
- Richness + Evenness
- Coverage

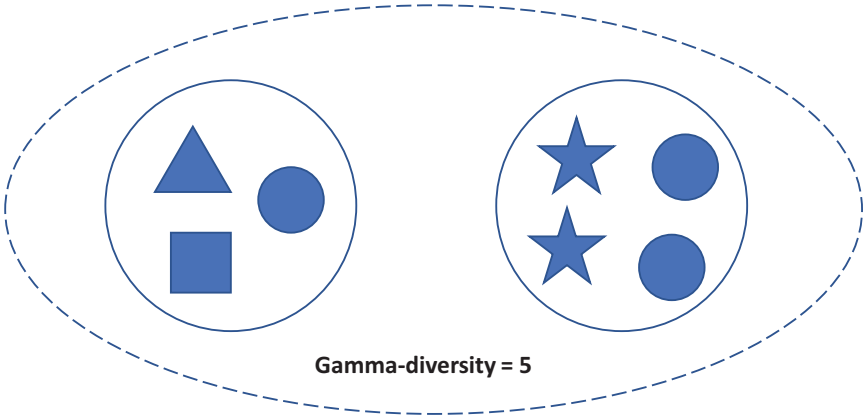
Diversity

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



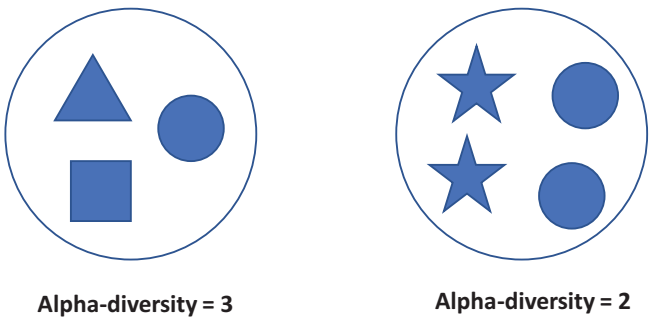
Diversity

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

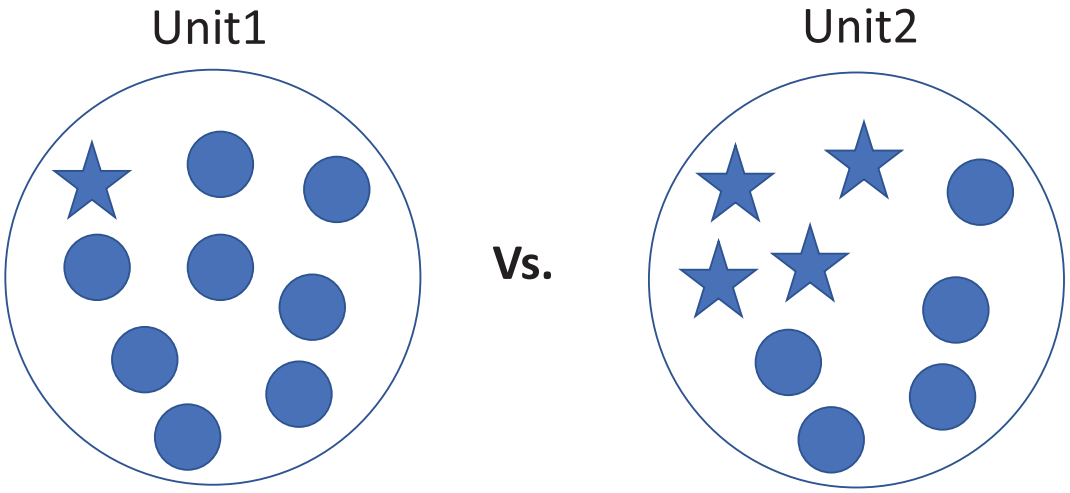


Diversity

- Alpha-diversity = diversity within ecological units or habitats



Diversity



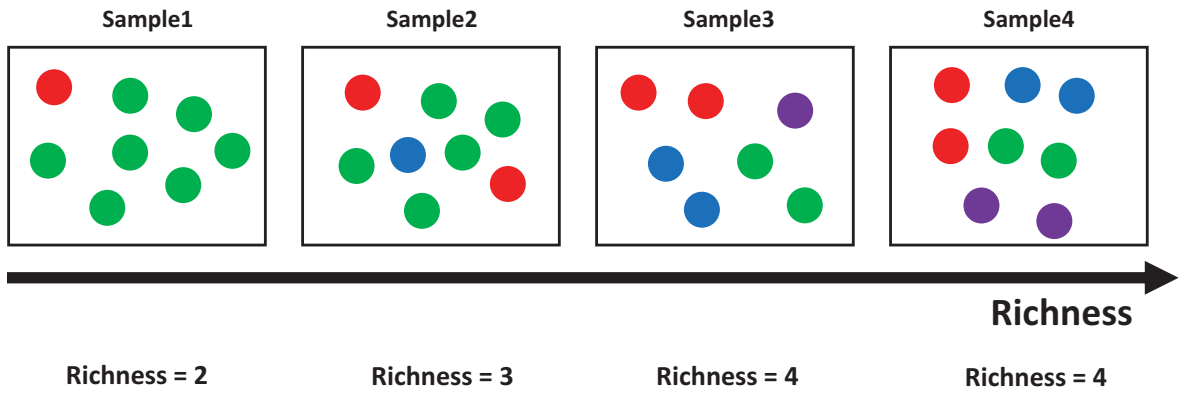
Which unit has a higher diversity?

Diversity

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

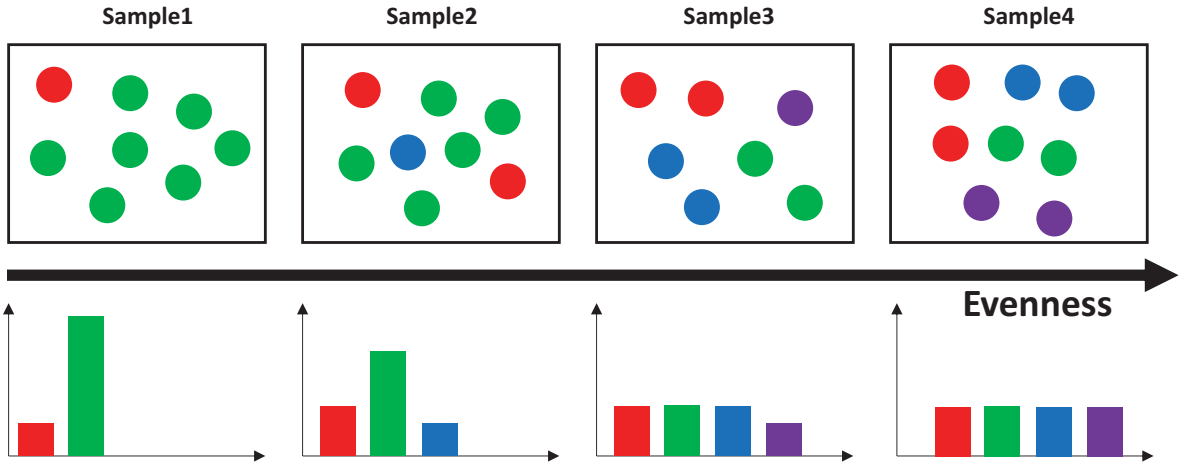
- Richness = number of species present
- Evenness = measuring how different species in community are similar in numbers

Diversity

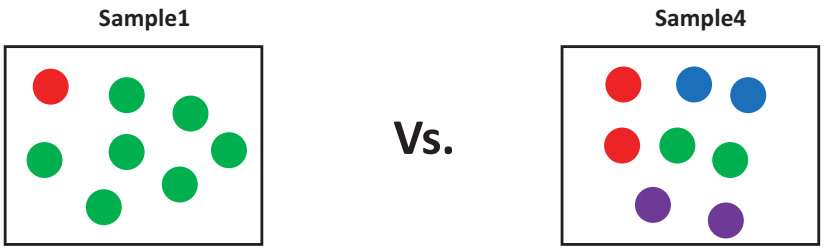


REF | http://www.evolution.unibas.ch/walser/bacteria_community_analysis

Diversity

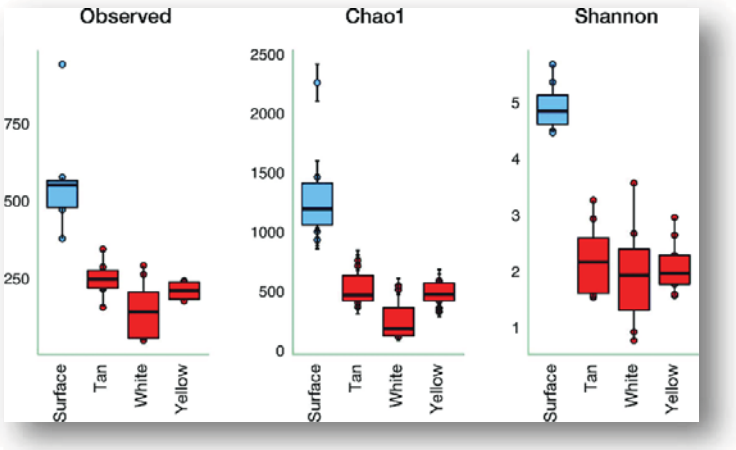


Diversity



Diversity

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



Diversity

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

Shannon

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

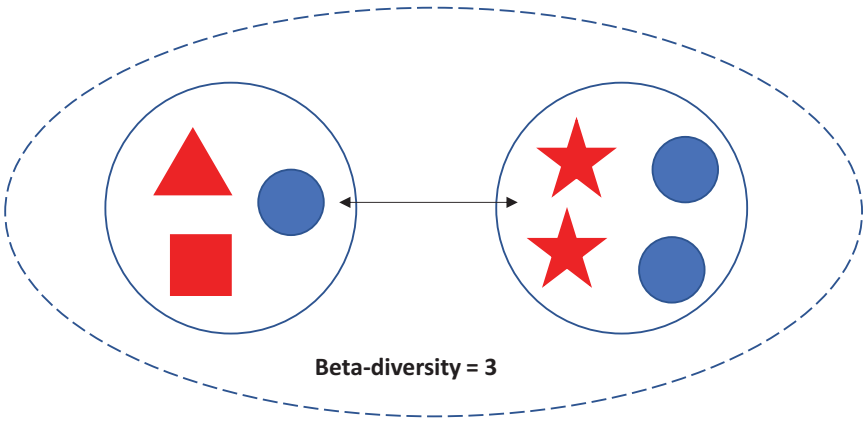
Inverse Simpson

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

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

Diversity

- Beta-diversity = differences in diversity between habitats



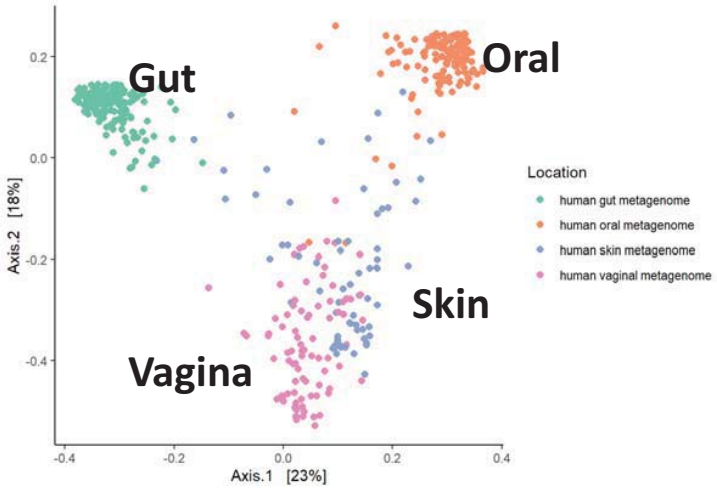
Diversity

Beta-diversity comparison

- Comparing compositional differences

Popular Beta-diversity measure

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

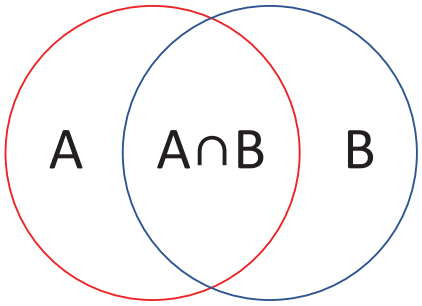


Diversity

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)

= ■ / (■ + ■)

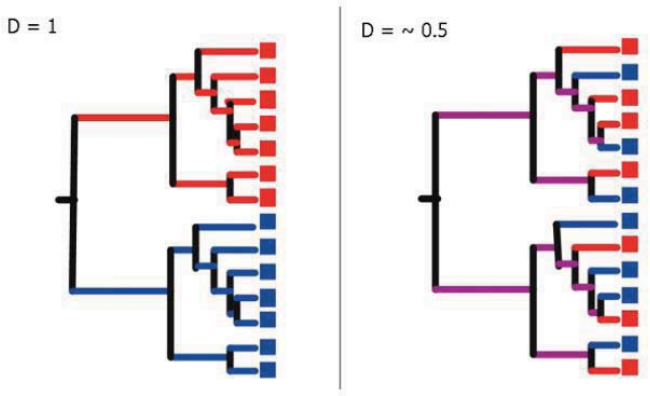
85

Diversity

Unifrac distance

= fraction of unshared branch lengths over tree

= (■ + ■) / (■ + ■ + ■)



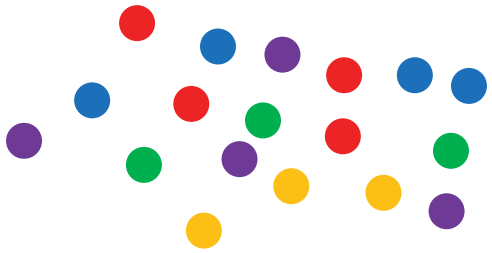
REF | Benjamin Callahan lecture

***weighted unifrac considers abundance changes**

86

Coverage

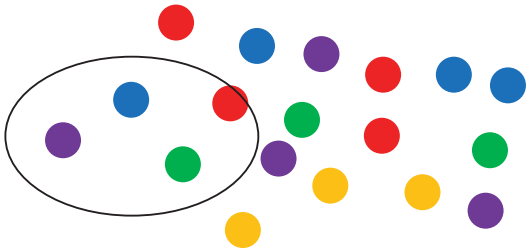
- Sampling is inevitable



Total richness = 5

Coverage

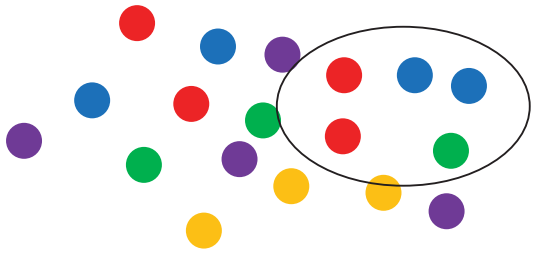
- Sampling is inevitable



Richness = 4

Coverage

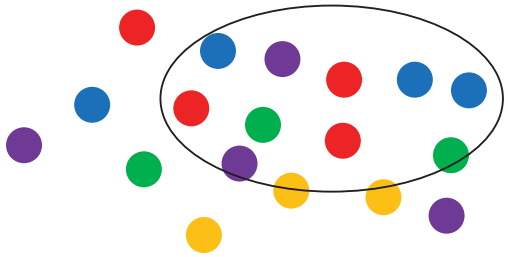
- Sampling is inevitable



Richness = 3

Coverage

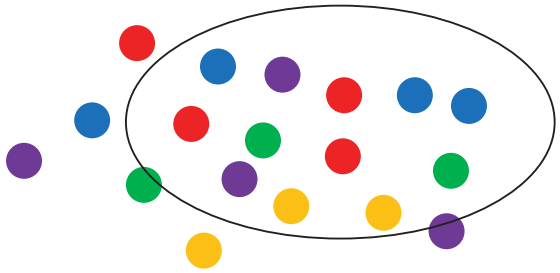
- Sampling is inevitable



Richness = 4

Coverage

- Sampling is inevitable

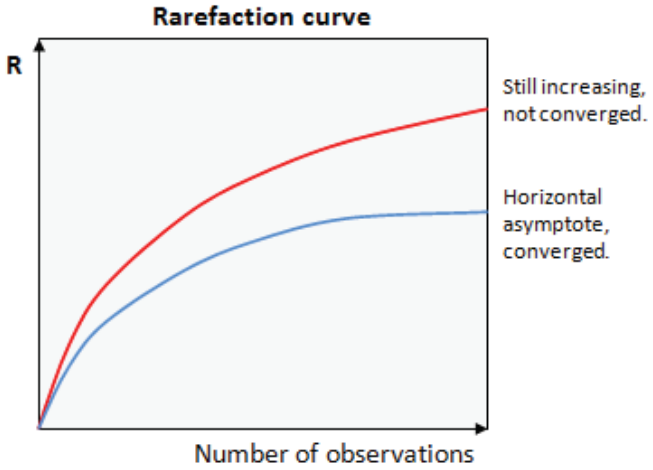


Richness = 5

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

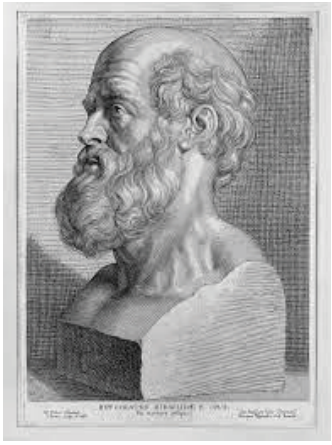
Rarefaction curve



93

Dysbiosis

94



Ancient Greek physician
Hippocrates

"All disease begins in the gut"

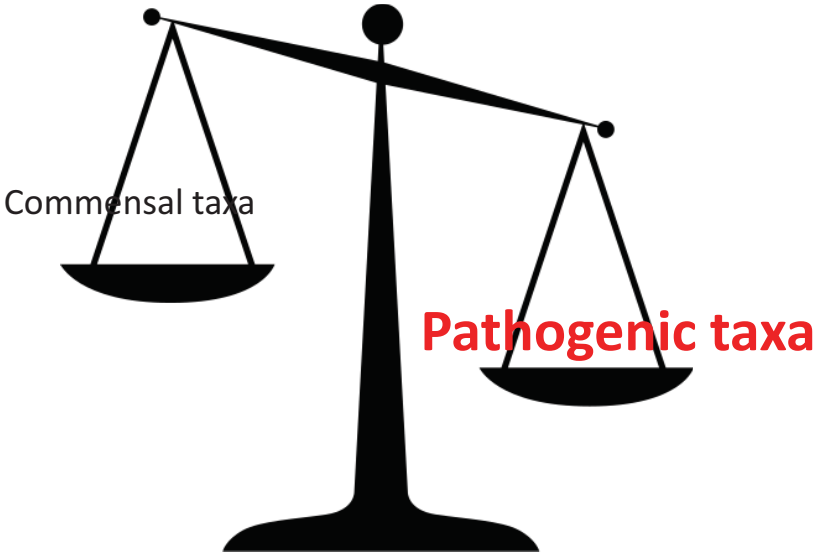
95

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**

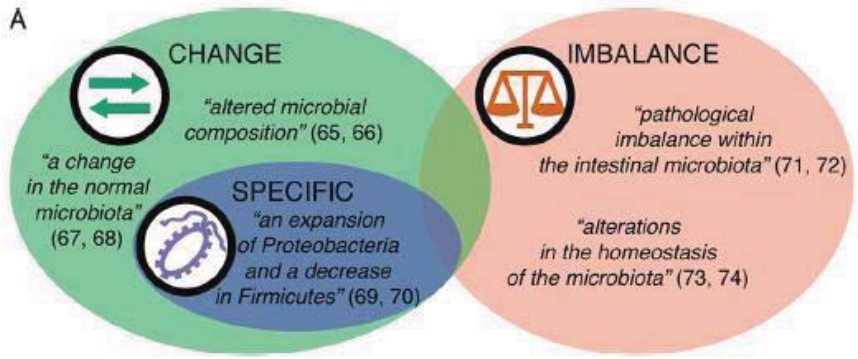
96

Dysbiosis?

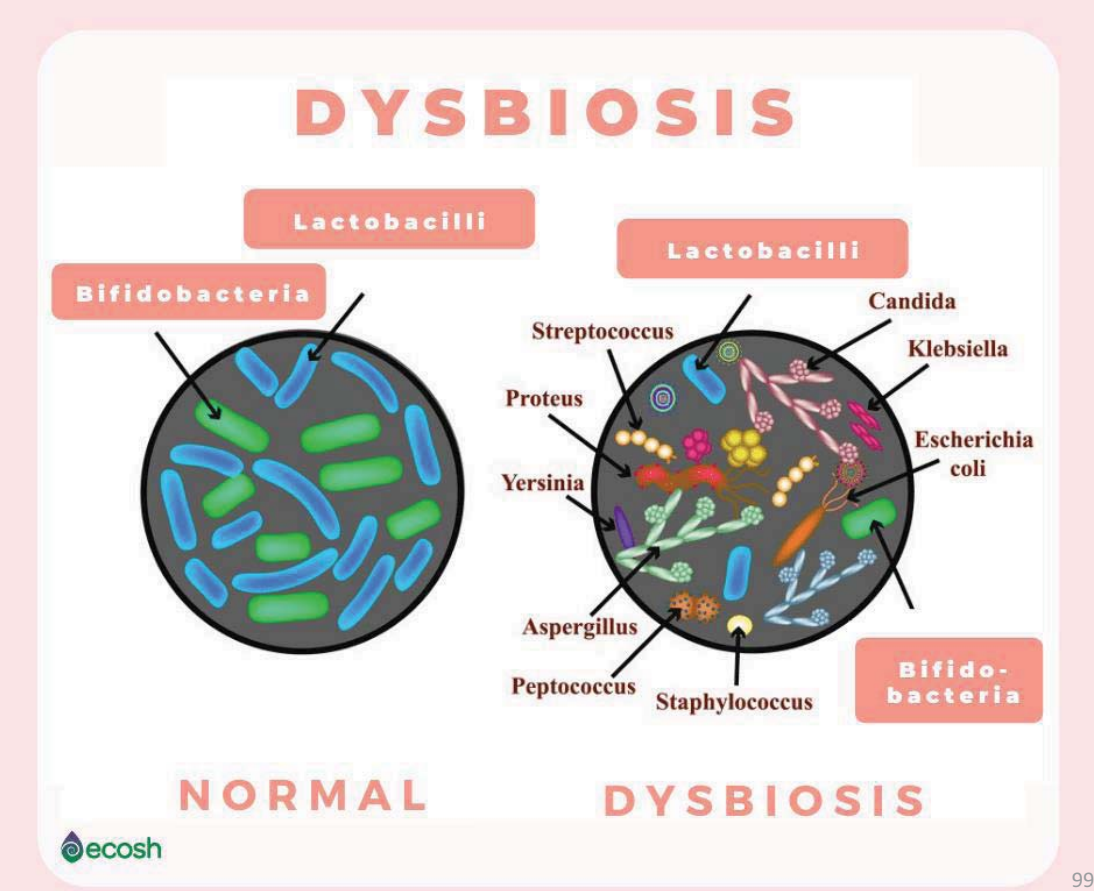


Dysbiosis

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



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



99

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

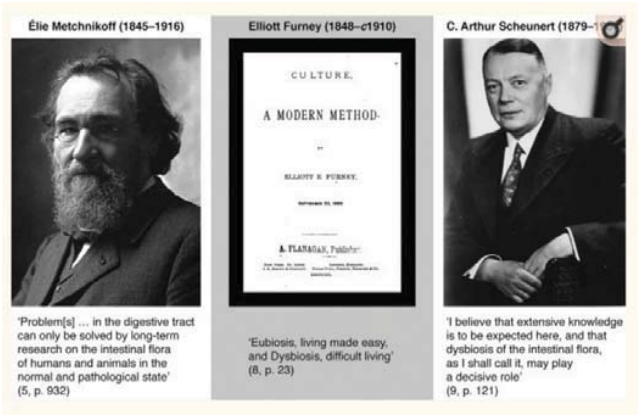
100

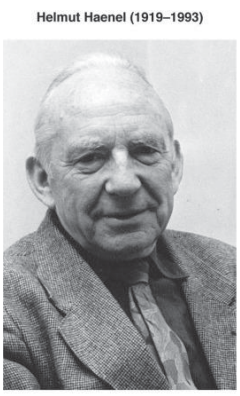
Terminology

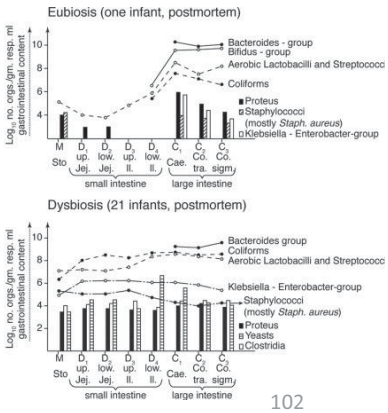
- **Dysbiosis** = condition of having imbalance in microbial community
- **Eubiosis** = microbial balance within the body
- **Symbiosis** = living together

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

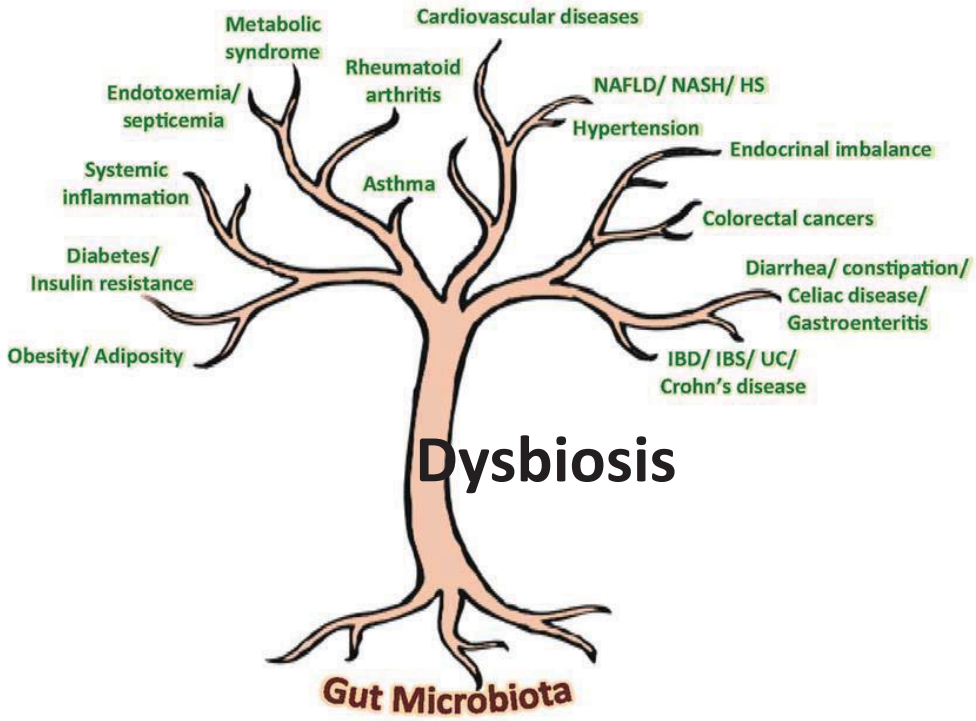






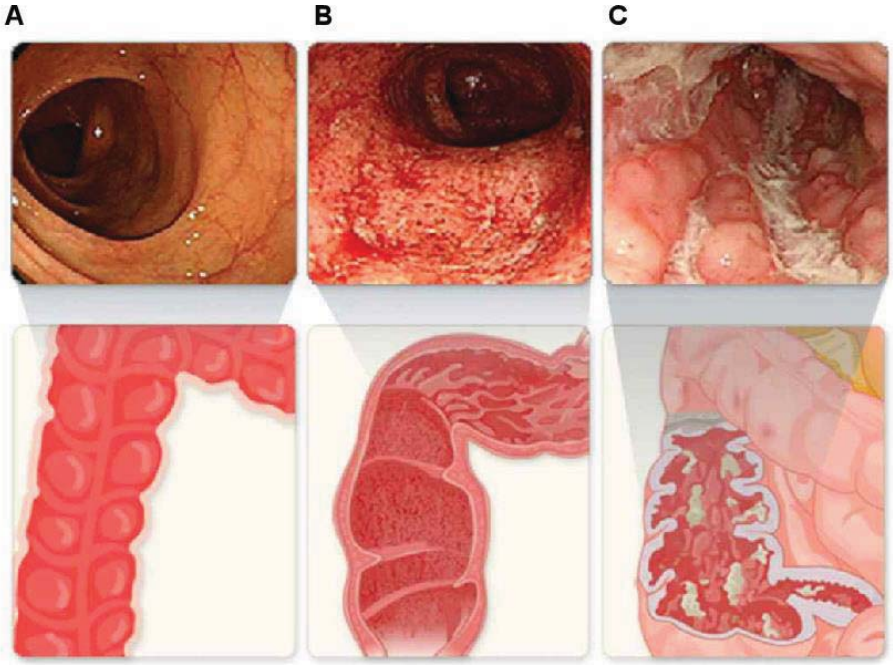
REF | Katarzyna B Hooks & Maureen A O’Malley, mBio, 2017

Dysbiosis → disease



103

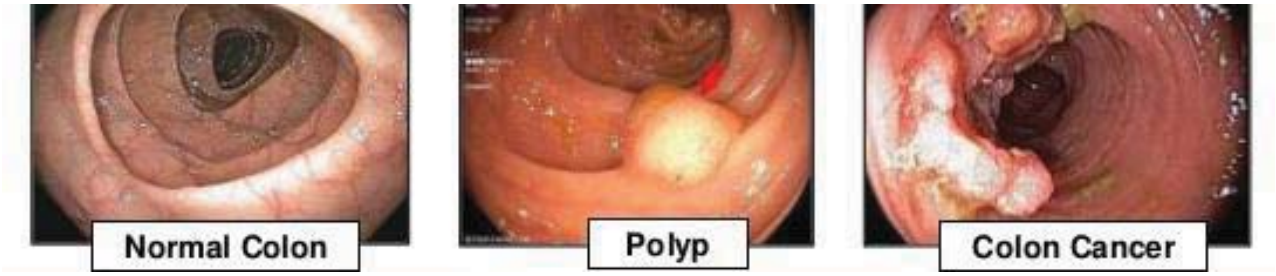
Dysbiosis → disease



Healthy | Ulcerative colitis | Chron's disease

104

Dysbiosis → disease



105

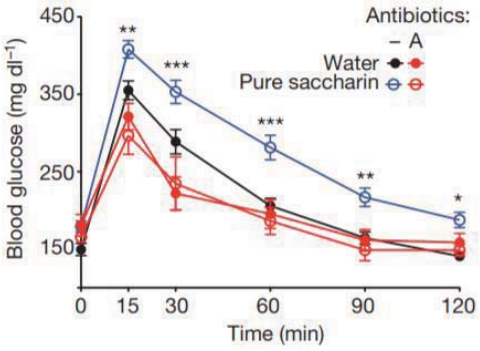
Dysbiosis → disease

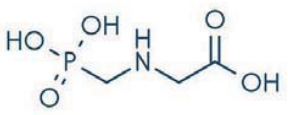
Atopic dermatitis

106

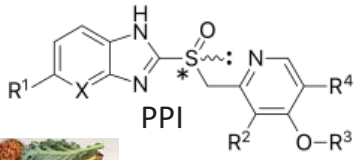
Cause of dysbiosis?

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



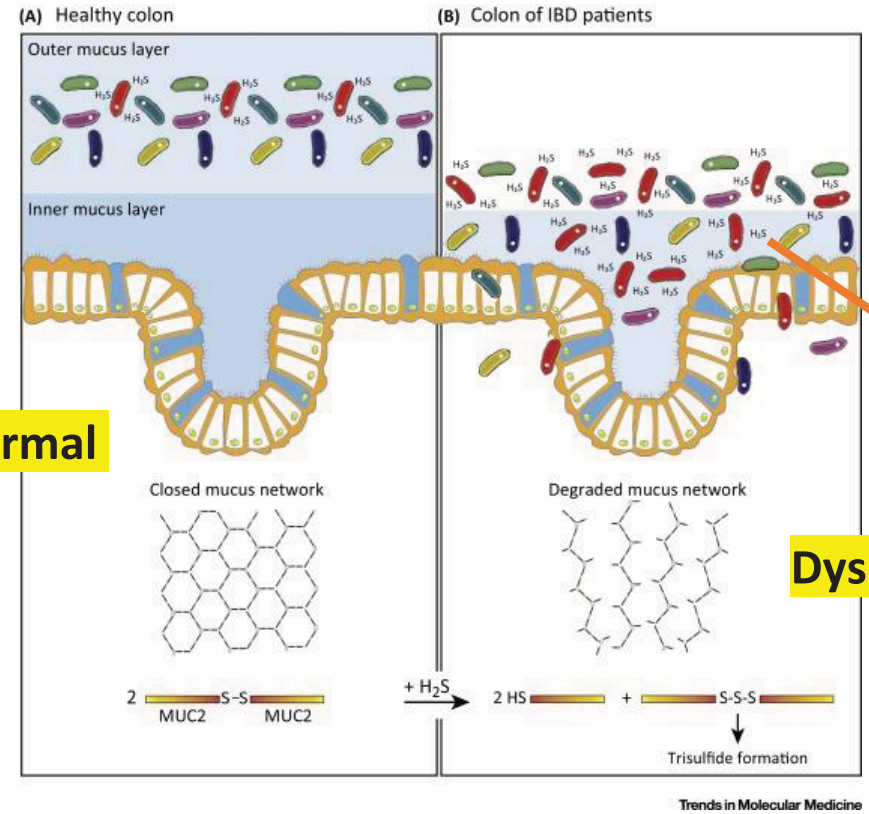


glyphosate



PPI

Low fibre diet → dysbiosis → inflammation

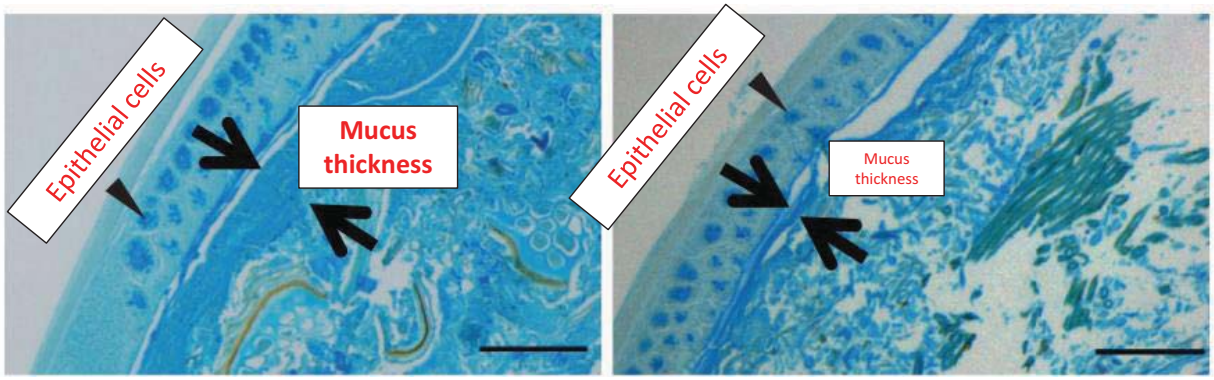


Normal

Degrading mucus by H_2S reducing bacteria

Dysbiosis

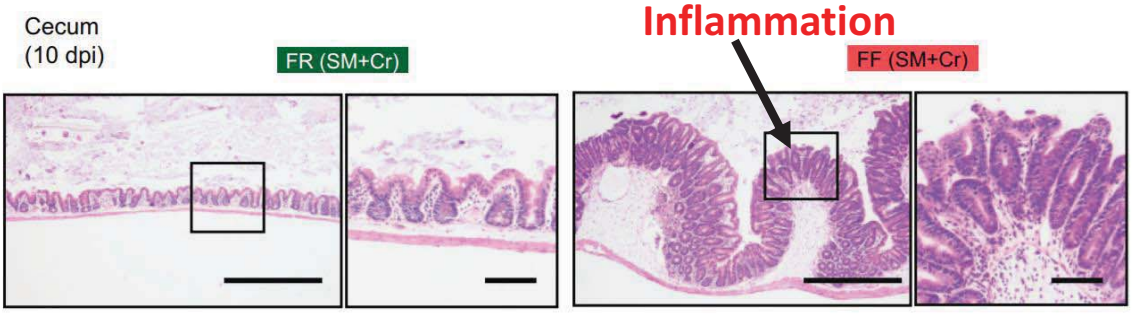
Low fibre diet → dysbiosis → inflammation



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

109

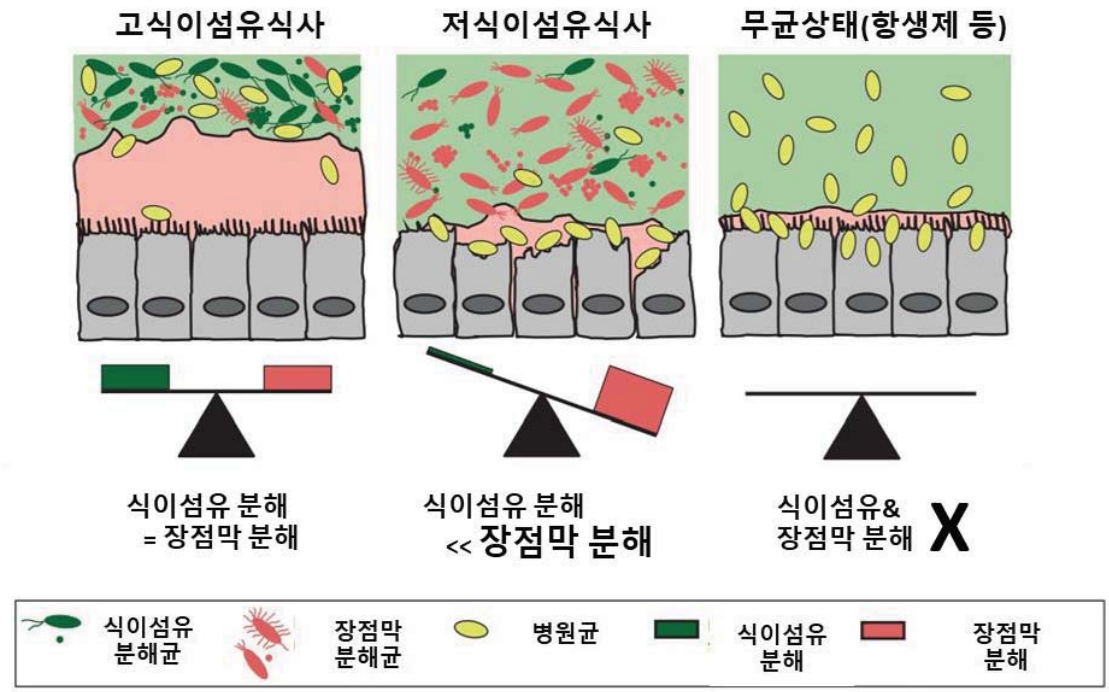
Low fibre diet → dysbiosis → inflammation



High fibre diet + pathogens **Low fibre diet + pathogens**

110

Low fibre diet → dysbiosis → inflammation



111

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

112

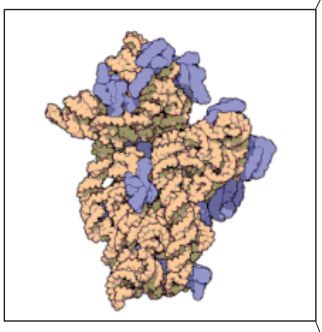
Bioinformatics analysis

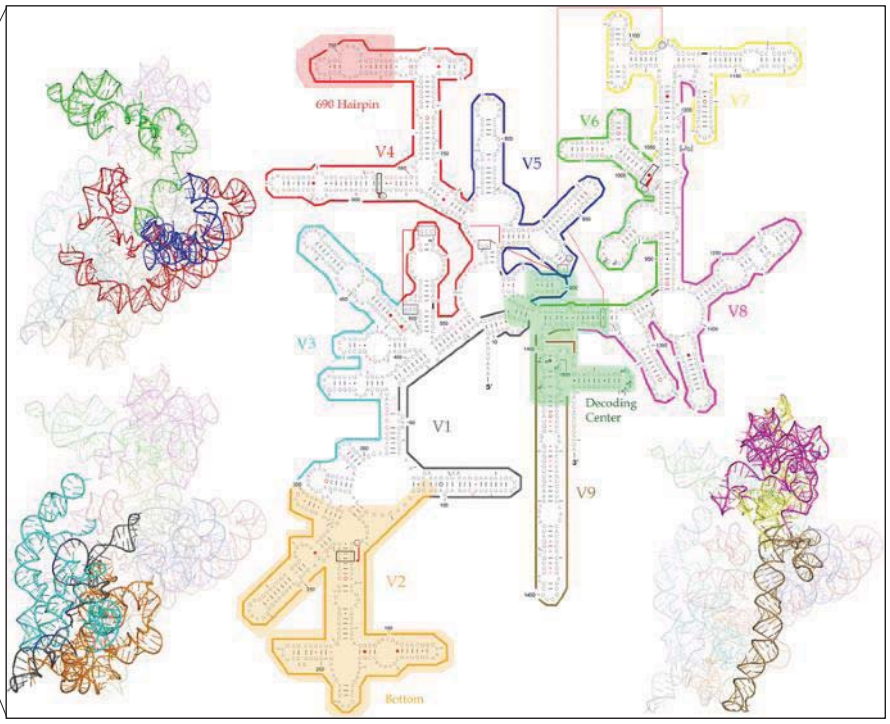
113

16S rRNA amplicon sequencing

114

16S rRNA = universal phylogenetic marker



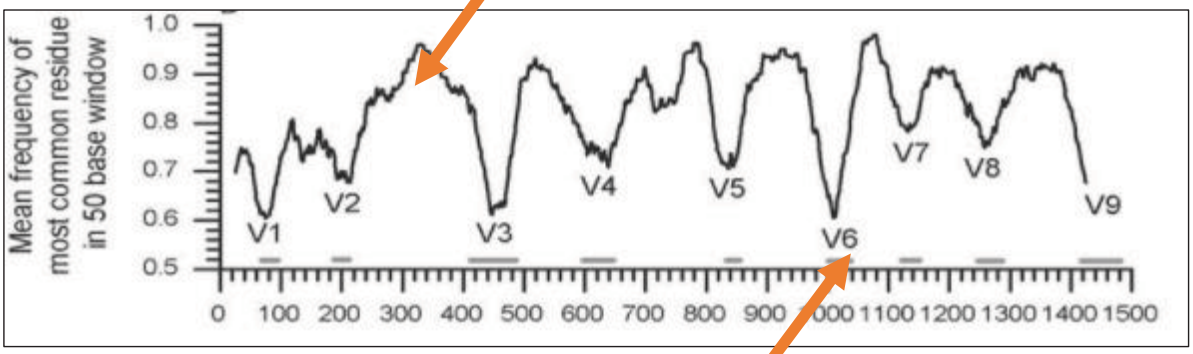


- **Present in all species**
- **Ubiquitous**
- **Extreme sequence conservation**
- **Single copy**
- Well-annotated references

115

16S rRNA = universal phylogenetic marker

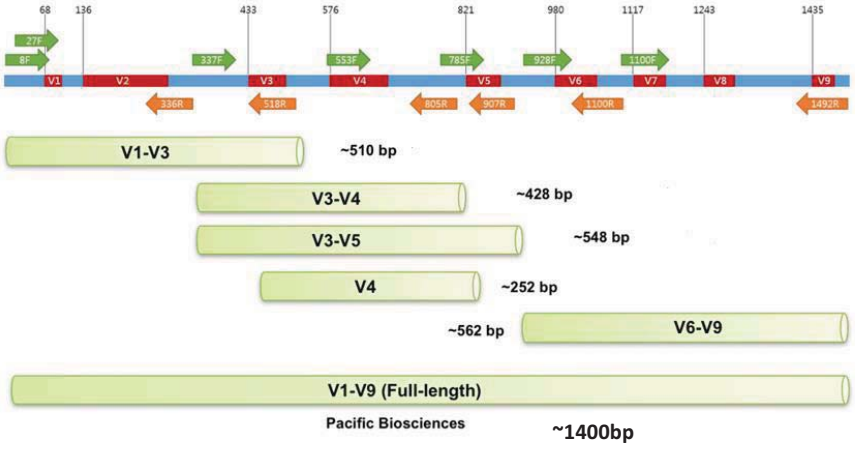
Highly conserved



Hypervariable regions

116

16S rRNA region has 9 hypervariable regions



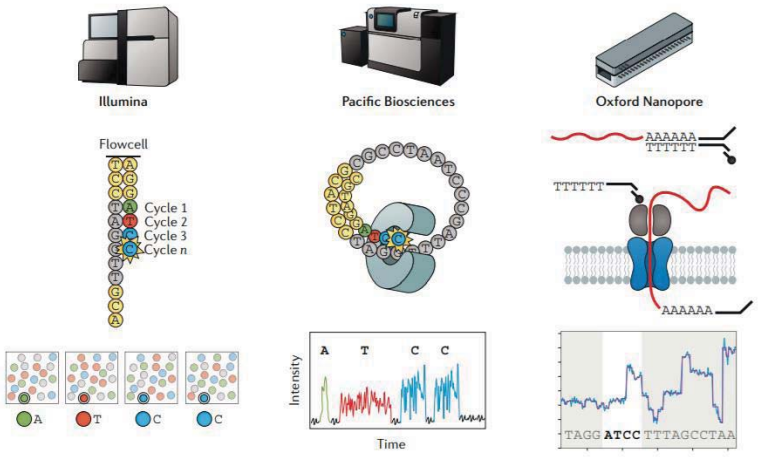
Primers

Primer Name	Sequence (5'-3')	SEQ ID NO:
V1f2	AGAGTTTGATCCTGGCTCAG	SEQ ID NO: 18
V5r4	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

117

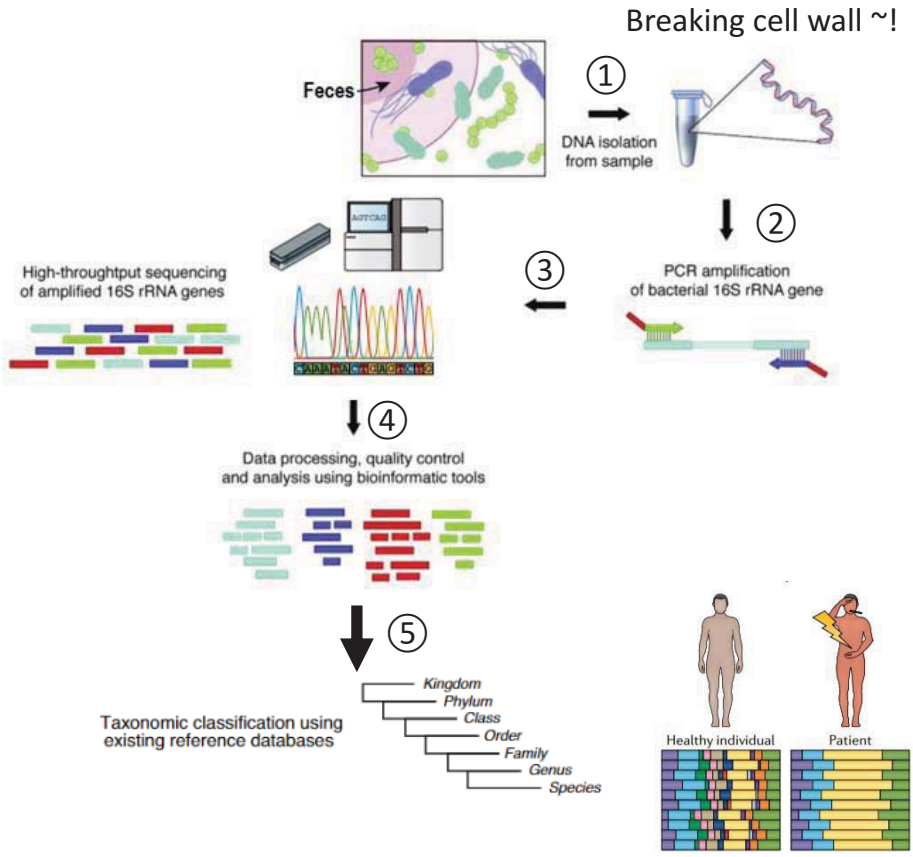
Next-generation sequencing for 16S rRNA amplicon sequencing



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

118

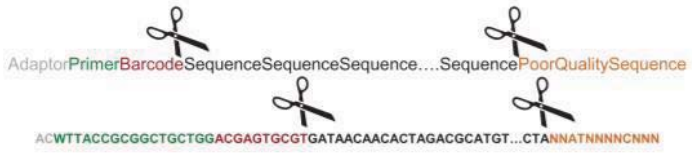
16S rRNA sequencing workflow

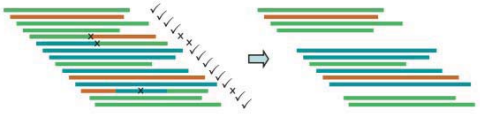


16S rRNA preprocessing

Quality assessment & trimming

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

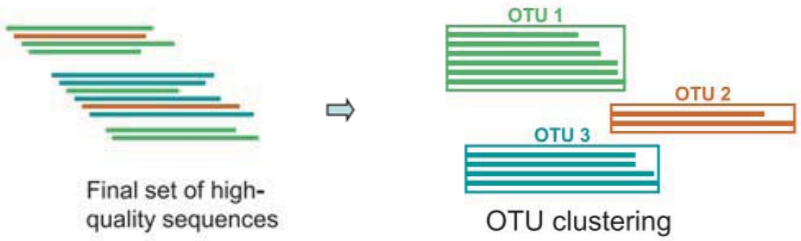




Identification and removal of PCR chimeras

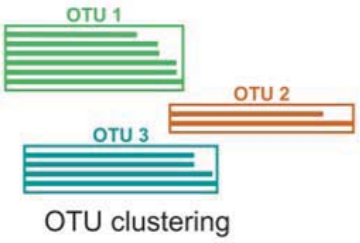
Final set of high-quality sequences

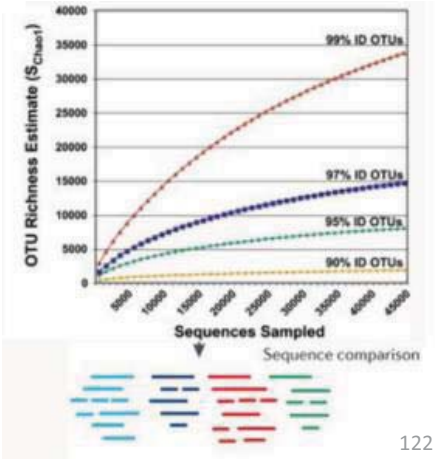
Binning sequences into operational taxonomy unit (OTU)



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*

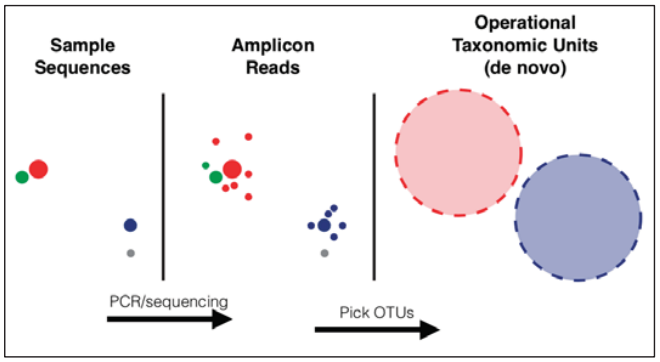




Amplicon sequence variants (ASV)

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

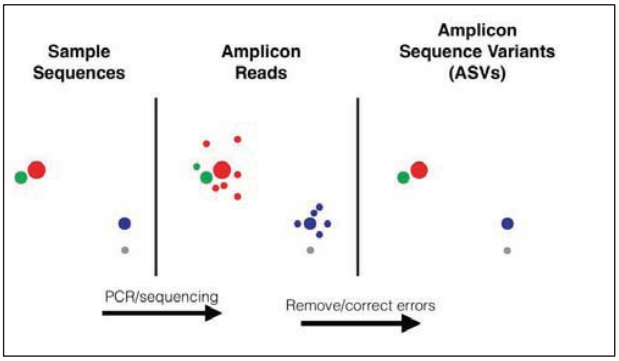


OTU methods

Amplicon sequence variants (ASV)

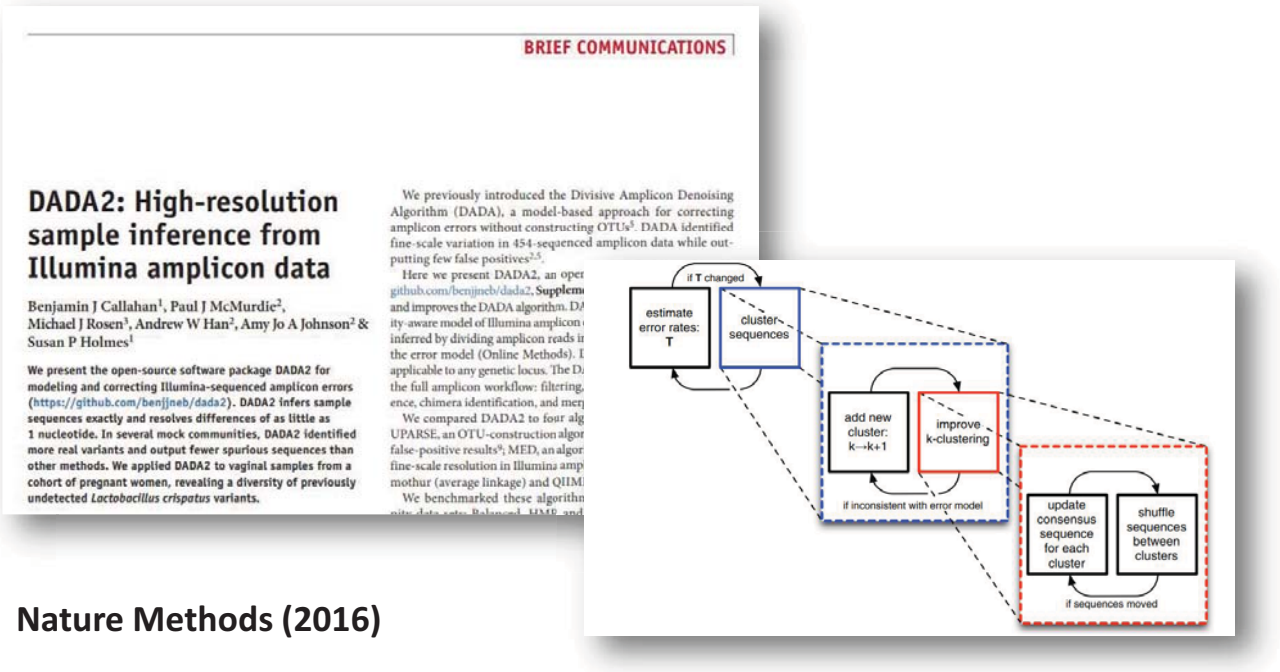
Amplicon sequence variant

- 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[3]. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives[2,5].

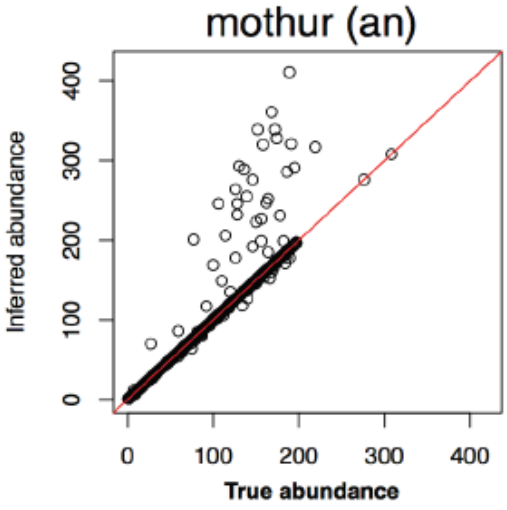
Here we present DADA2, an oper... github.com/benjjneb/dada2, Supplem... and improves the DADA algorithm. DA... ity-aware model of Illumina amplicon... inferred by dividing amplicon reads i... the error model (Online Methods). I... applicable to any genetic locus. The DA... the full amplicon workflow: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

Nature Methods (2016)

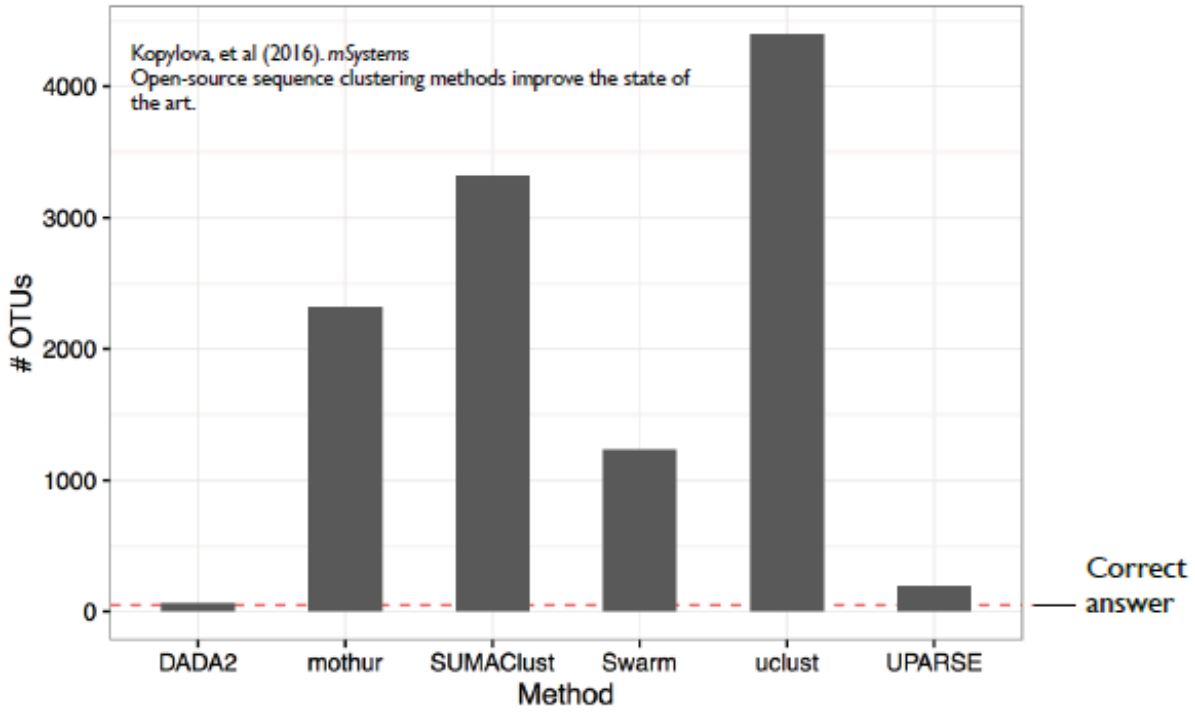
Simulated dataset validation



TP: 978
FP: 272
FN: 77

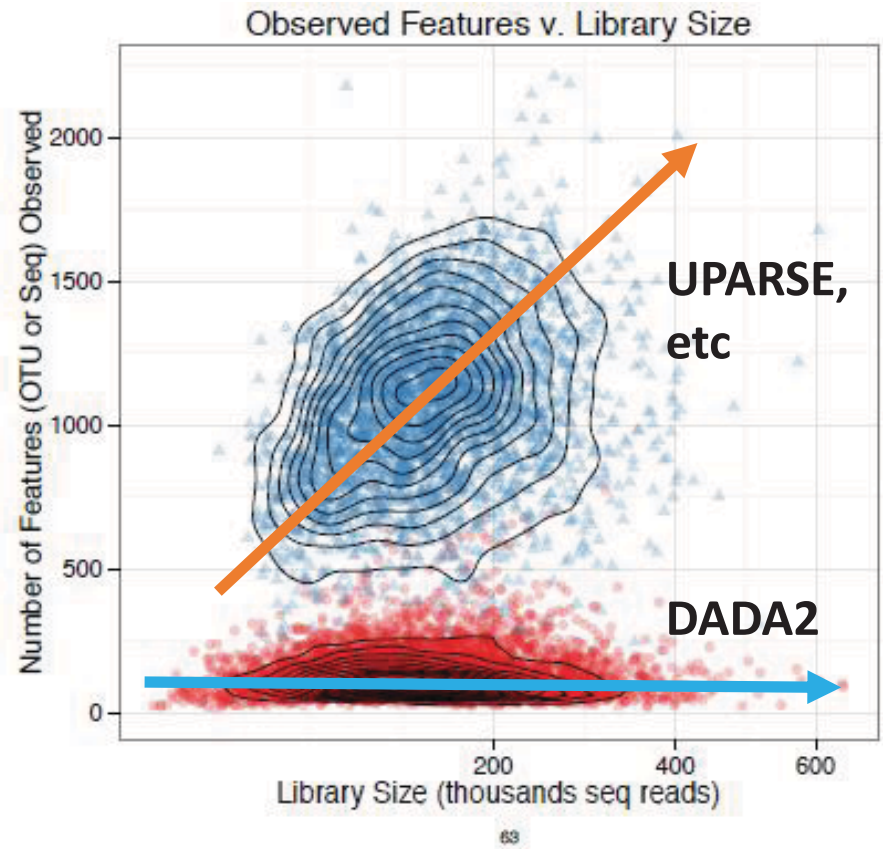
TP: 1042
FP: 0
FN: 13

Mock community validation (Bokulich data)



127

Library sizes less affect ASV detection

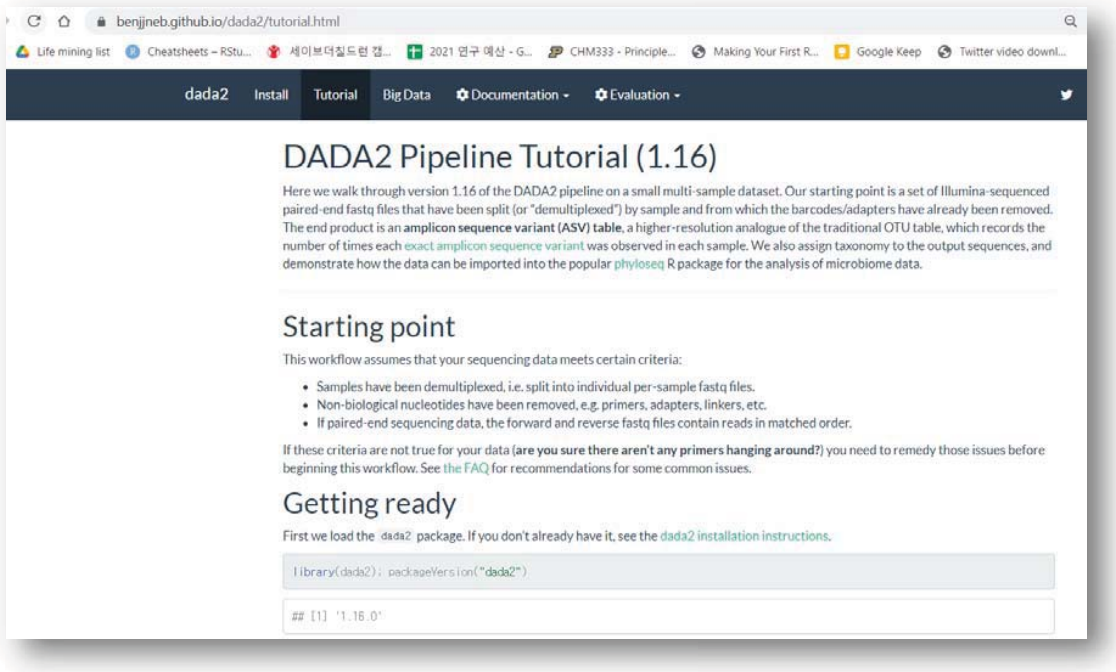


128

ASV vs OTU

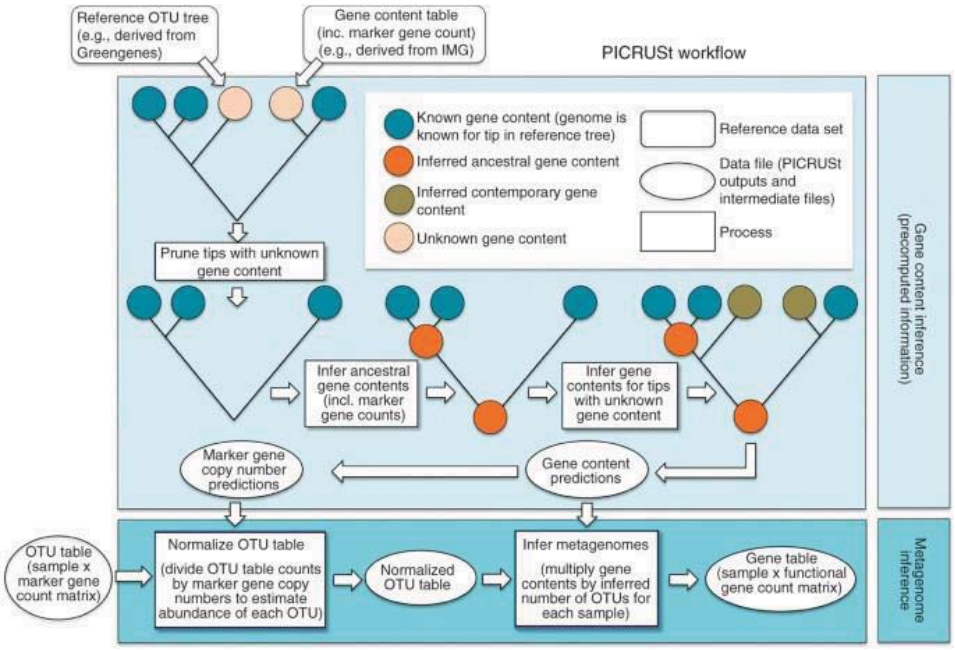
	ASVs	OTUs: De novo	OTUs: Closed-ref
Precise	✓	~	~
Tractable	✓	~	✓
Reproducible	✓	✗	✓
Comprehensive	✓	✓	✗

DADA2 pipeline tutorial



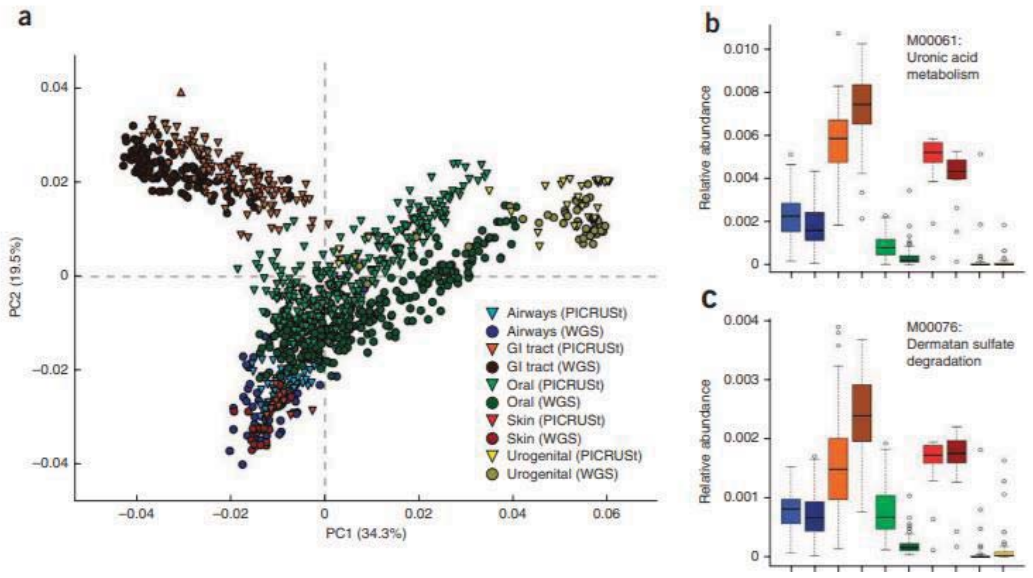
https://benjjneb.github.io/dada2/tutorial.html

Function prediction - PICRUSt



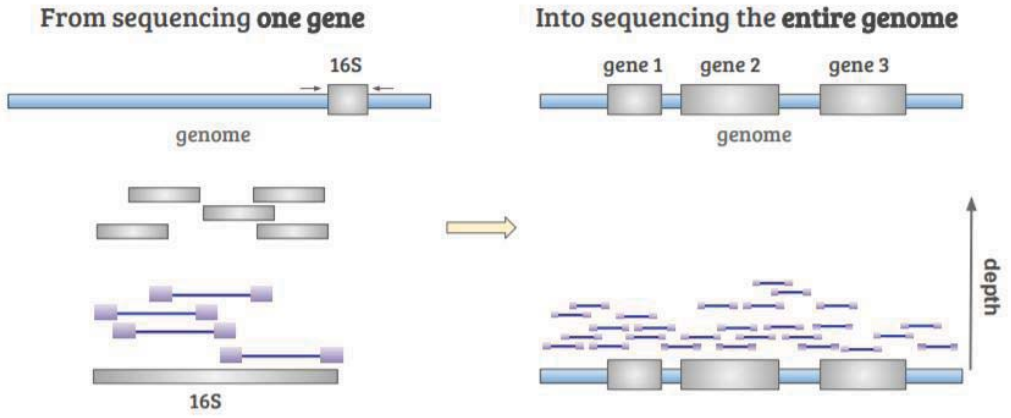
131

Function prediction - PICRUSt



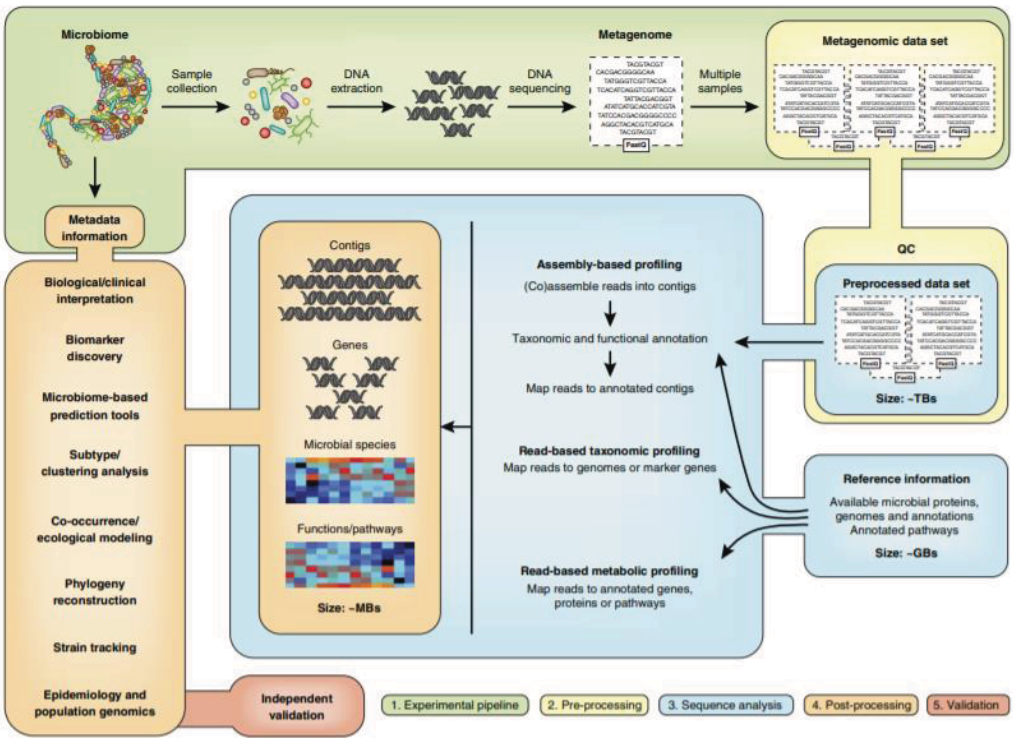
132

Shotgun metagenome analysis



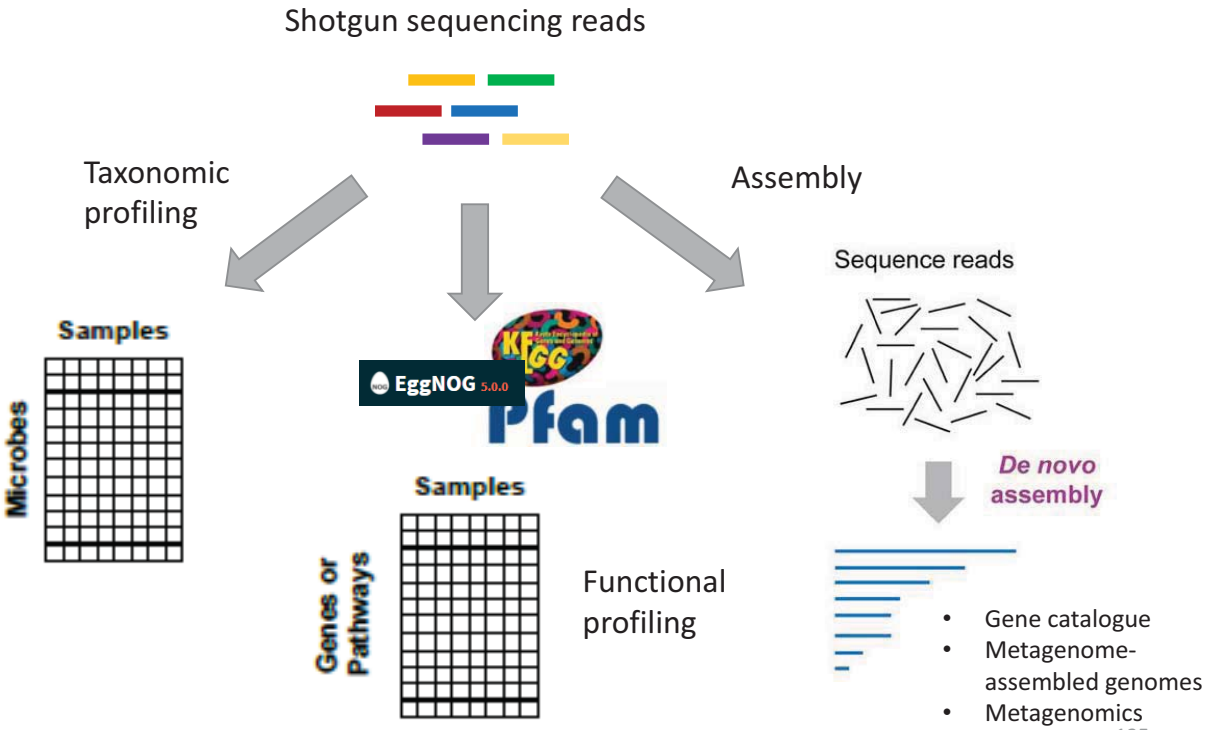
REF | BioBakery

Shotgun metagenomics enables complete overview of a complex microbiome



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

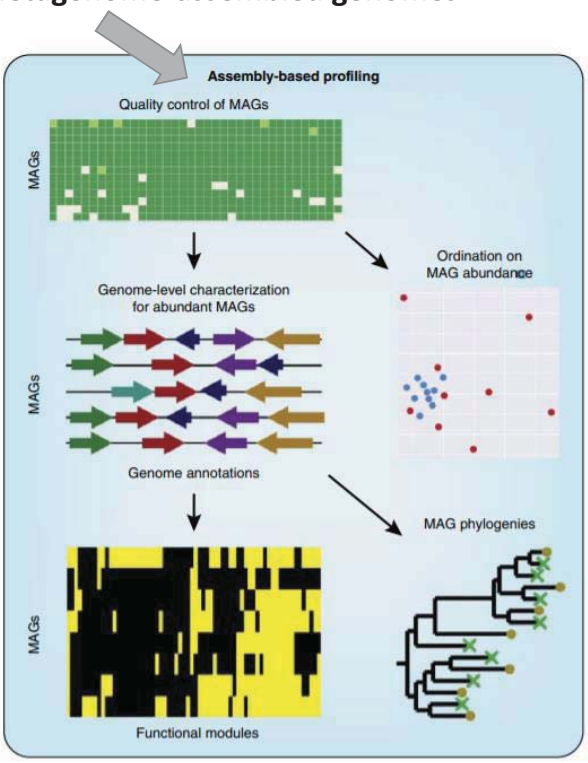
Overview of shotgun metagenome analysis



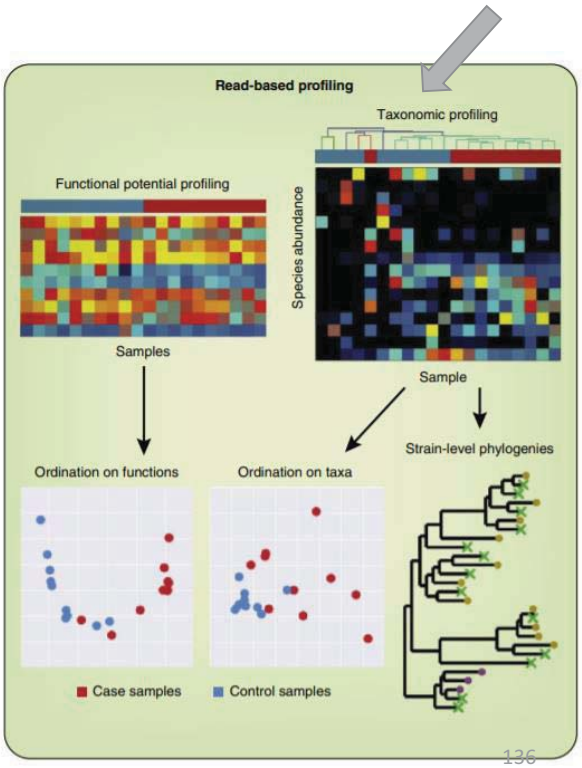
135

Taxonomic profiling (species/strain-level)

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



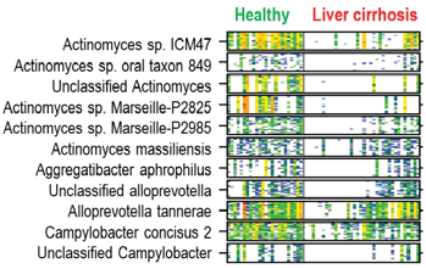
Using reference genomes



136

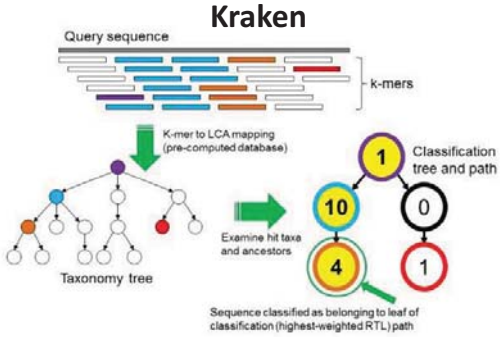
Read-based profiling

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

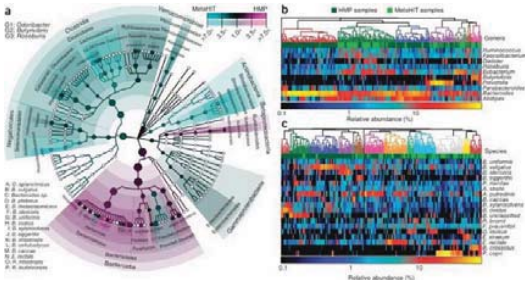


Metagenomic species (MGS)

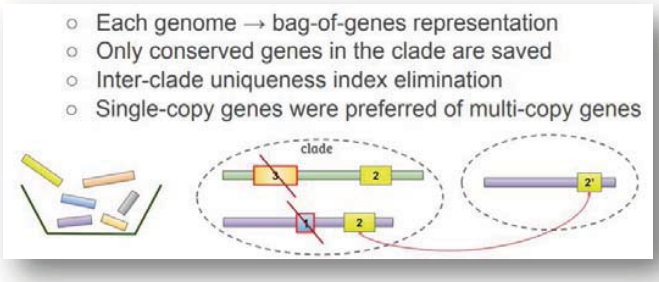
Kraken



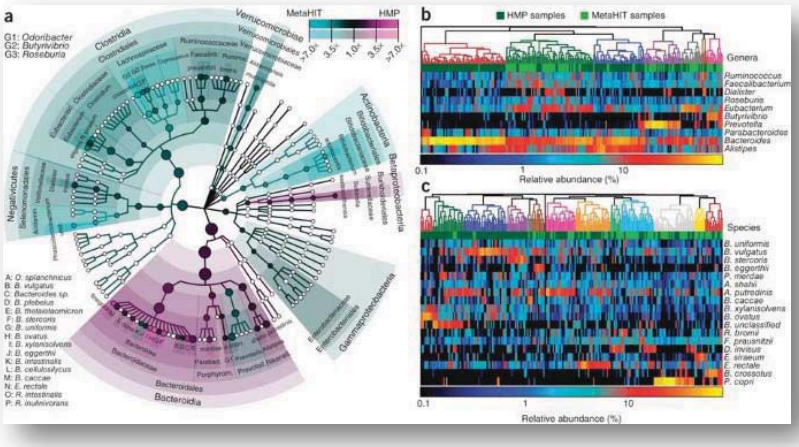
metaPhlan



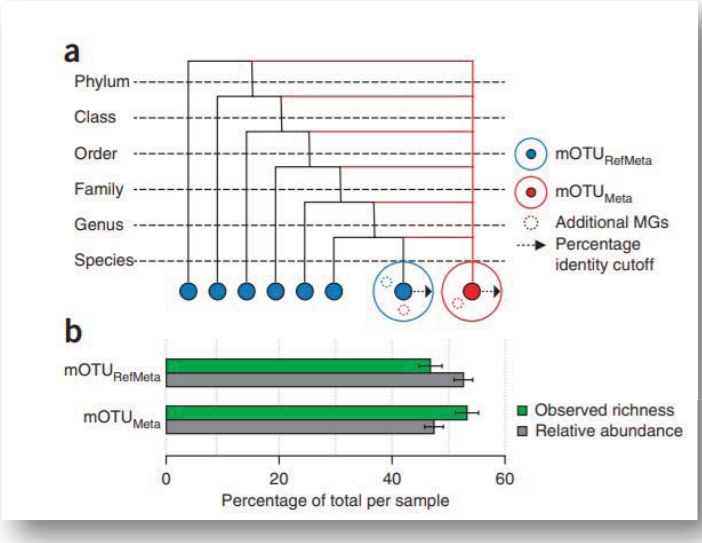
MetaPhlan: clade-specific marker gene based profiling



Clade-specific marker discovery



mOTU: universal phylogenetic marker - based profiling

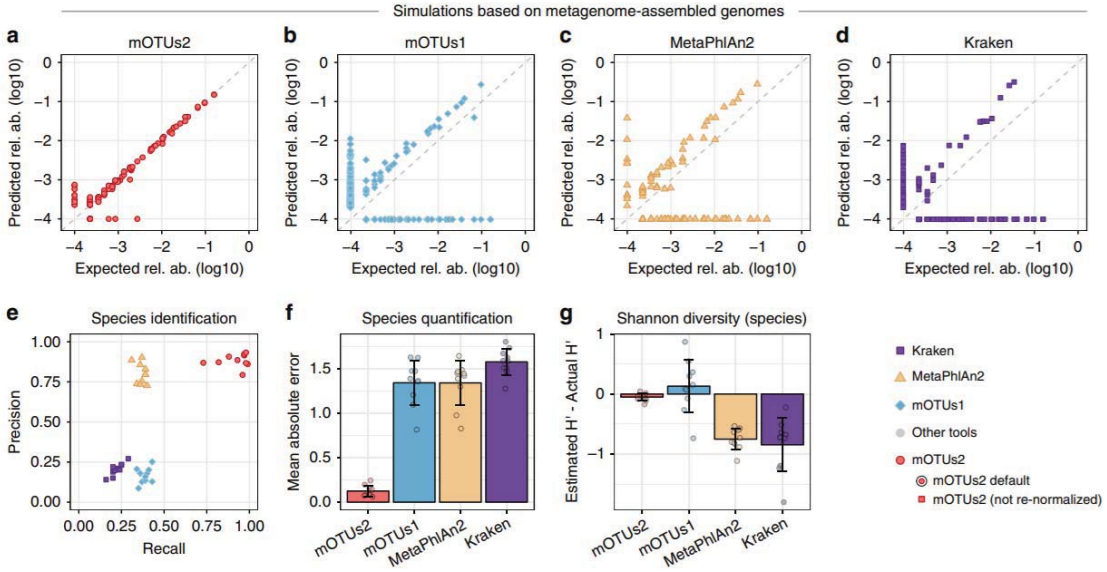


By COG	
COG ID	COG name
COG0012	Predicted GTPase, probable translation factor
COG0016	Phenylalanyl-tRNA synthetase alpha subunit
COG0018	Arginyl-tRNA synthetase
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein 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

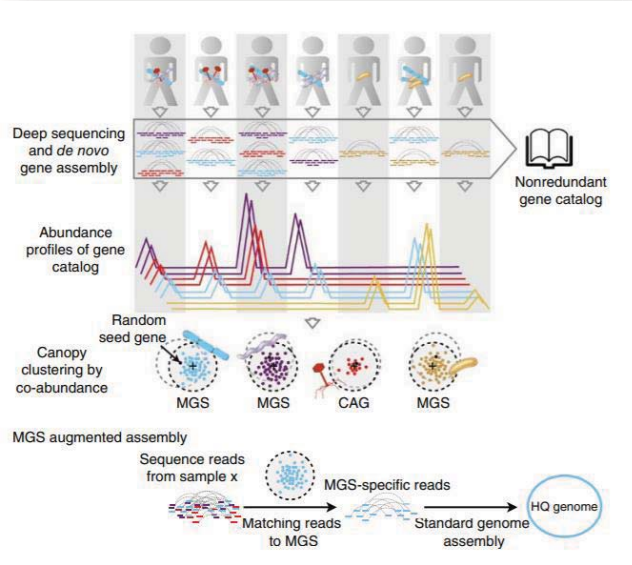
https://motu-tool.org/

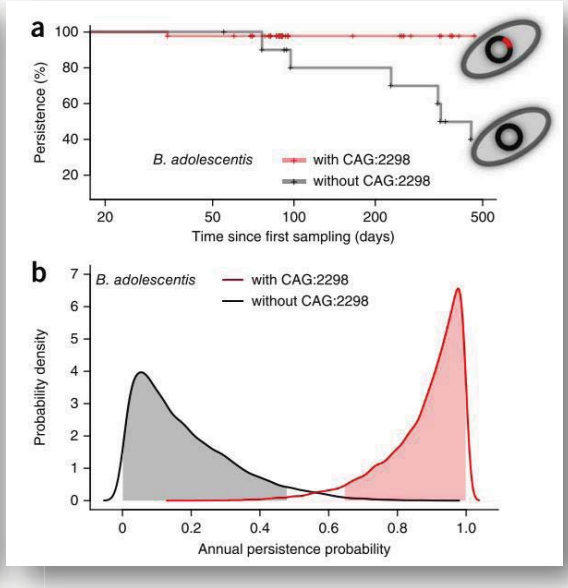
mOTU: universal phylogenetic marker - based profiling



https://motu-tool.org/

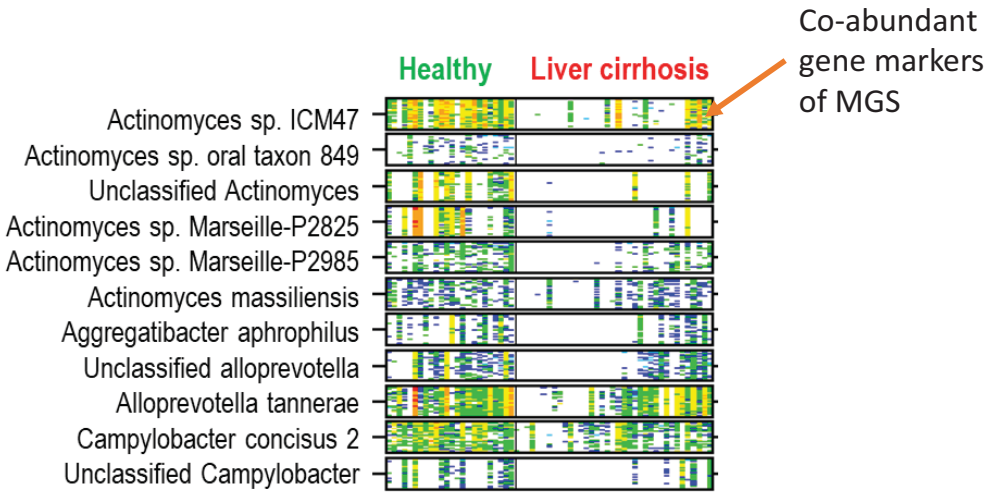
Metagenomic species-based profiling





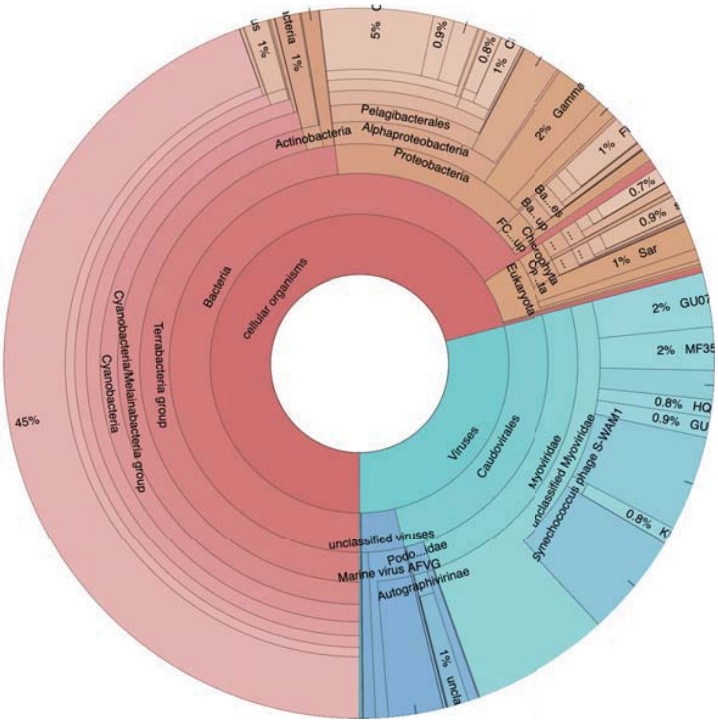
141

Metagenomic species-based profiling



142

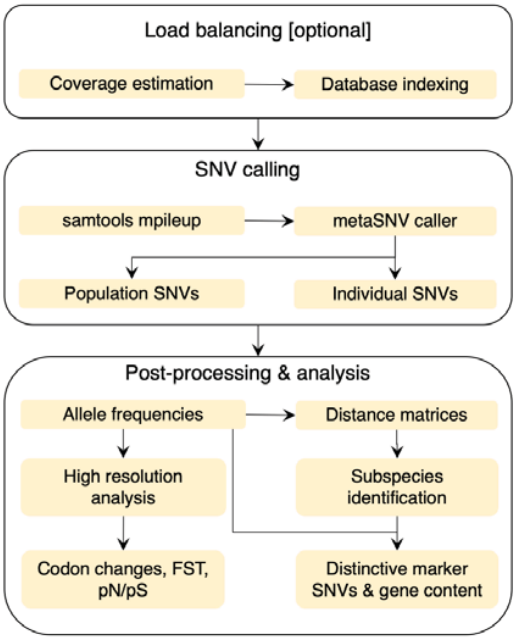
Plotting microbiome composition - Krona plot

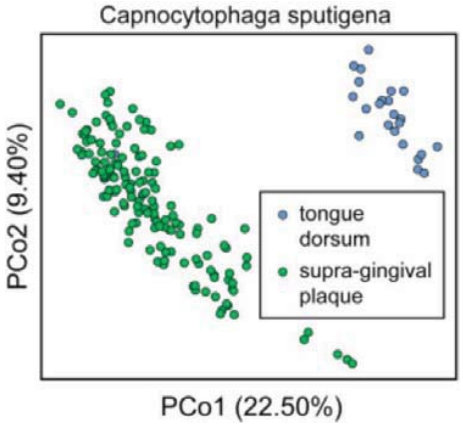


https://github.com/marbl/Krona/wiki/KronaTools

143

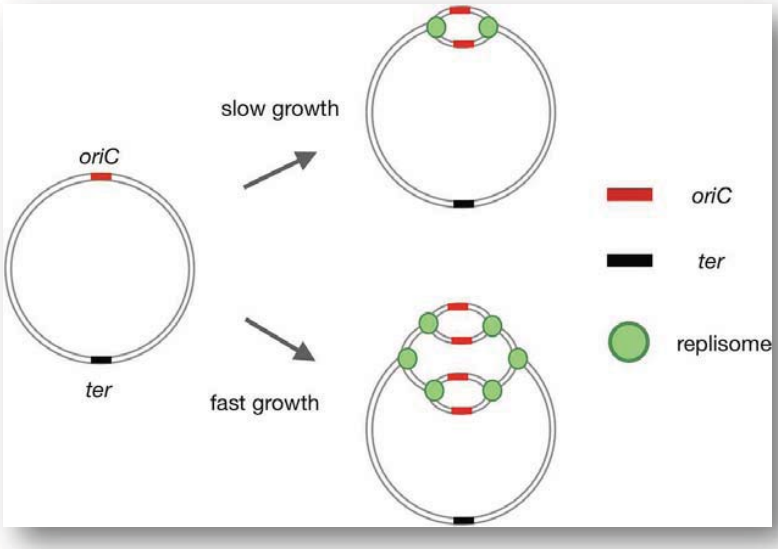
Subspecies analysis - meta-SNV





144

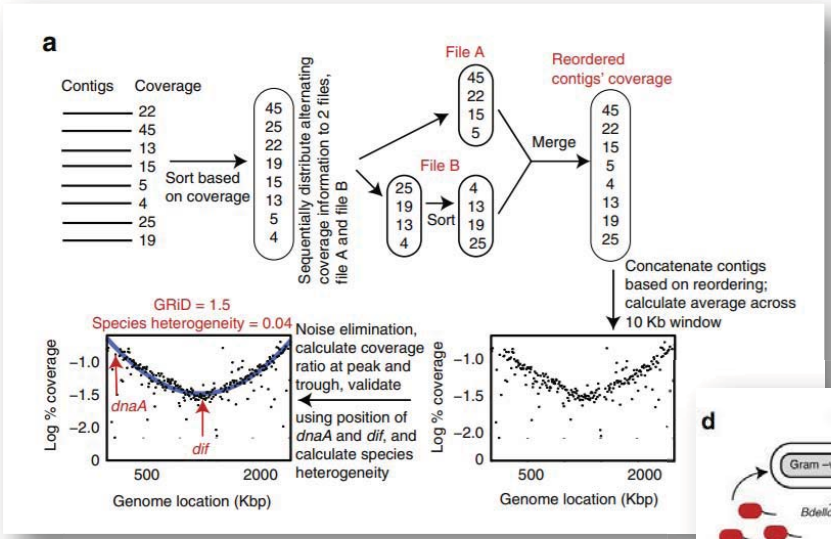
Growth-rate estimation: GRiD

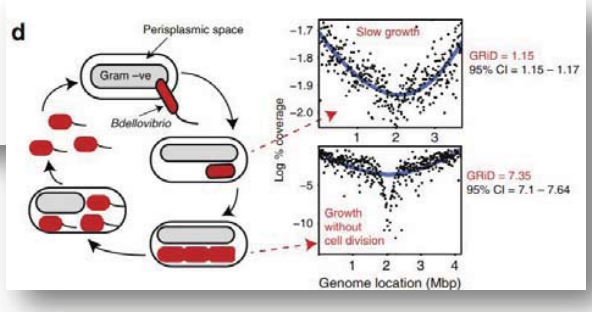


Different sequencing depths from replication of origins

145

Growth-rate estimation: GRiD

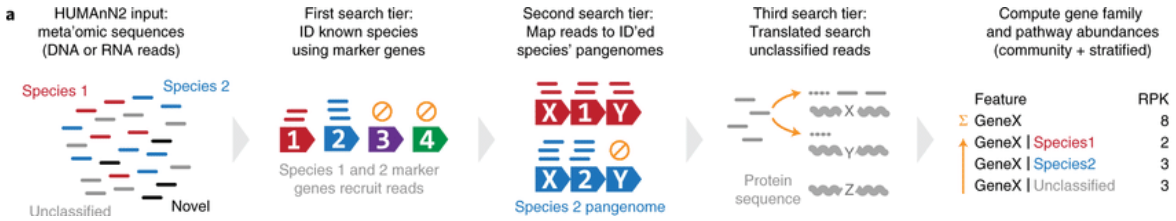




https://github.com/ohlab/GRiD

146

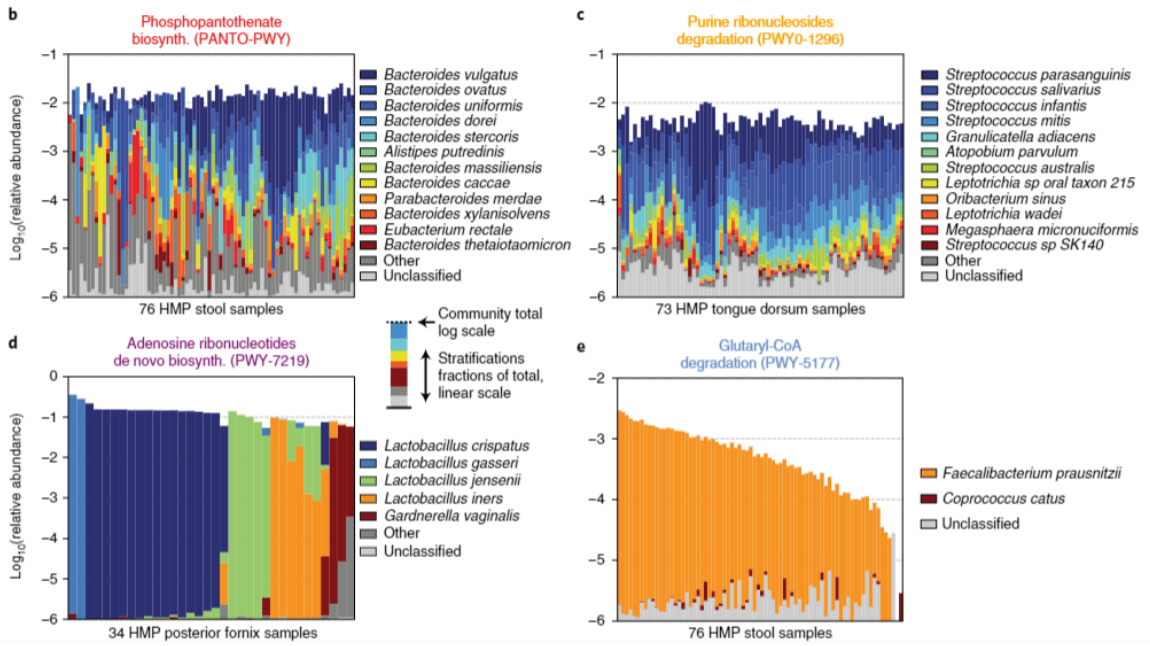
HUMAnN2 - Functional profiling



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

147

HUMAnN2 - Functional profiling



148

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
…

149

Appendix

150

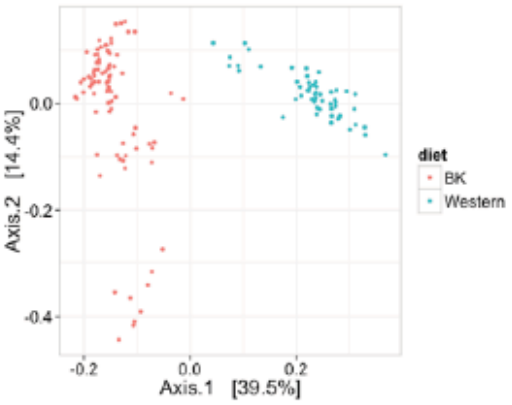
Ordination Methods

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,…
.
.

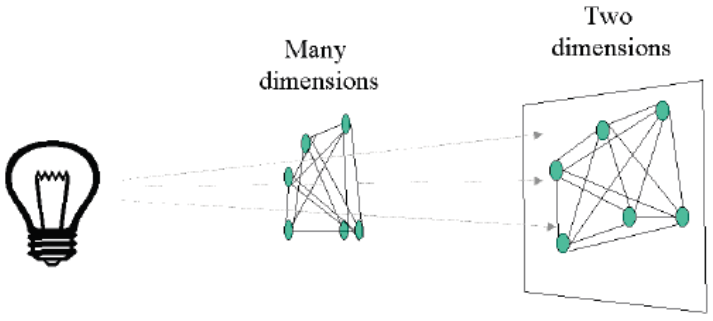


P-dimensions

2-dimensions

Multi-dimensional Scaling

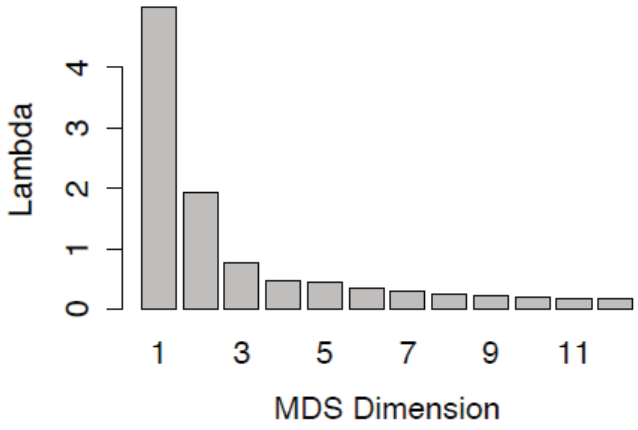
Why MDS? It works with any distance!



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

MDS Scree Plot

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



153