KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머쉰러닝 워크샵(온라인)

Bioinformatics Approaches to Explore Human Microbiomes

노미나

SBi 한국생명정보학회 Korean Society for Bioinformatics
서울대학교 생물정보연구소 SNU BIOINFORMATICS INSTITUTE
Gene·2·us

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

강의개요

Bioinformatics Approaches to Explore Human Microbiomes

인간, 동물 및 특정 환경에 존재하는 미생물 커뮤니티에 대한 연구는 질병과의 관련성, 환경 요인에 따른 다양성의 이해를 위해 지난 10 년동안 많은 연구가 이루어 지고 있다. 특히, 미생물 커뮤니티로부터 생산된 멀티오믹스 데이터를 이용하여 다양한 생물정보학적 분석방법을 적용하며, 임상, 환경 정보를 통합적으로 분석하여 유전형과 표현형 특성과의 연관성에 대한 연구도 활발히 진행되고 있다. 기존의 단일 미생물 유전체 연구를 기반으로 다양한 종으로 이루어진 미생물 커뮤니티의 구성과 기능을 총체적으로 이해하는데 필요한 다양한 연구 결과 및 분석 방법에 대해 소개하고자 한다.

강의는 다음의 내용을 포함한다:

- 메타게놈 연구방법 (metagenomics)에 대한 소개
- 인간 마이크로바이옴에 대한 연구 현황
- 마이크로바이옴 구성 및 기능 분석
- 마이크로바이옴 데이터를 이용한 유전자 분석

Curriculum Vitae

Speaker Name: Mina Rho, Ph.D.

▶Personal Info

Name Mina Rho

Title Associate Professor

Affiliation Hanyang University

▶Contact Information

Address 222, Wangsimni-ro, Seondong-Gu, Seoul, 04763

Email minarho@hanyang.ac.kr

Phone Number 02-2220-2379

Research interest : Biomedical data mining, Machine learning and computational genomics

Educational Experience

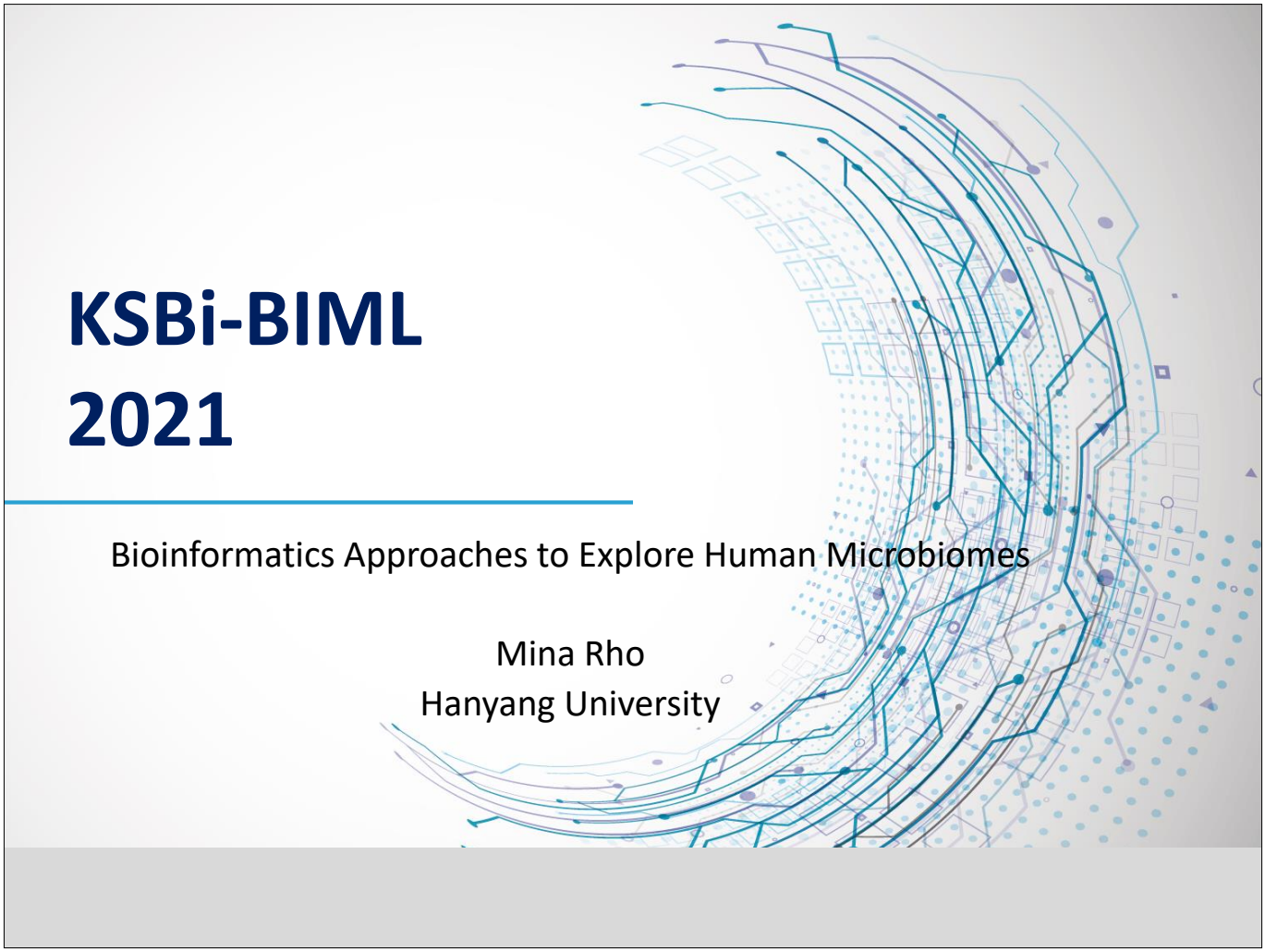
1996	B.S. in Computer Engineering, Ewha Womans University, Korea
2001	M.S. in Computer Engineering, Boston University, USA
2009	Ph.D. in Computer Science, Indiana University Bloomington, USA

Professional Experience

2009-2012.	Postdoctoral Associate, Indiana University Bloomington, USA
2012-2013	Assistant Professor, Roswell Park Cancer Institute, State University of New York at Buffalo, USA
2013-2017	Assistant Professor, Hanyang University, Korea
2017-present.	Associate Professor, Hanyang University, Korea

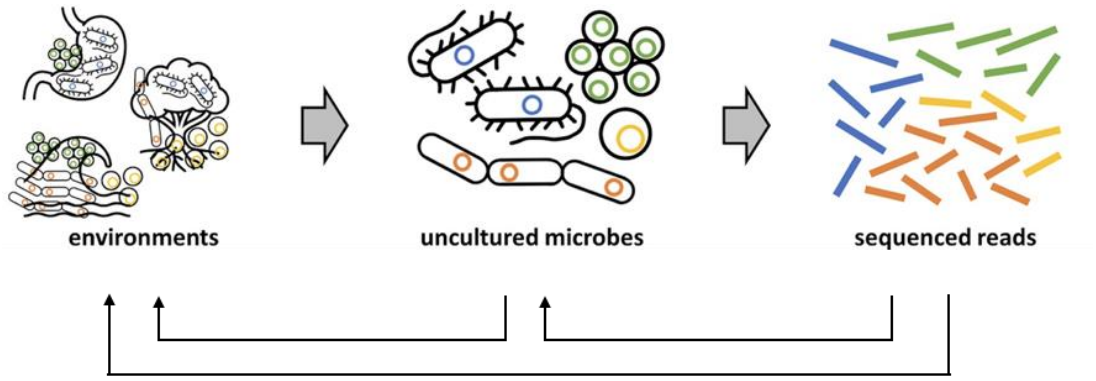
Selected Publications (5 maximum)

1. Lim, S. K., Kim, D., Moon, D. C., Cho, Y. & Rho, M (2020) "Antibiotic resistomes discovered in the gut microbiomes of Korean swine and cattle" GigaScience: 9(5)
2. J. Bae, K. Lee, M. Islam, H. Yim, H. Park, and M. Rho (2019) "iMGEins: detecting novel mobile genetic elements inserted in individual genomes ", BMC Genomics :19(944)
3. J. H. Shin, H. Eom, W. J. song, and M. Rho (2018) "Integrative metagenomic and biochemical studies on rifamycin ADP-ribosyltransferases discovered in the sediment microbiome ", Scientific Reports :8(12143)
4. H. Cho, J. Choi, J. Seo, B.J. Kim, M. Rho*,J. K. Han, and J. G. Kim* (2017) "Helicobacter pylori-derived extracellular vesicles increased in the gastric juices of gastric adenocarcinoma patients and induced inflammation mainly via specific targeting of gastric epithelial cells", Experimental & Molecular Medicine :49, e330



Metagenomics

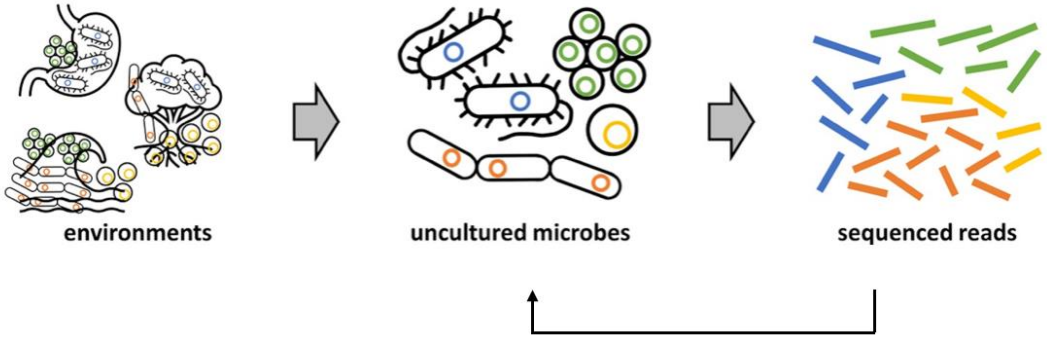
studies on *entire genetic materials*
collected directly from a community of microorganisms



- Taxonomy profiles
- Function profiles
- Genomic variations and exchanges

Metagenomics

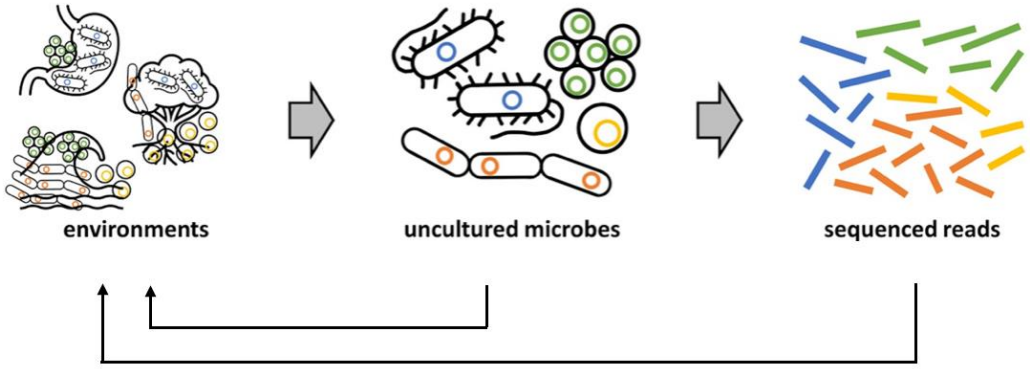
studies on *entire genetic materials*
collected directly from a community of microorganisms

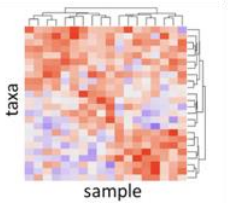


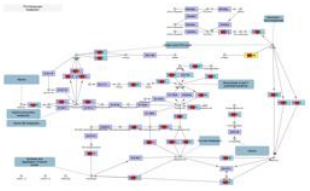
- Supervised approach (i.e. classification)
  - taxonomy assignment
- Unsupervised approach (i.e. clustering)
  - contig binning

Metagenomics

studies on *entire genetic materials*
collected directly from a community of microorganisms







- Supervised approach (i.e. classification)
  - taxonomy assignment
- Unsupervised approach (i.e. clustering)
  - contig binning
- Graph algorithm

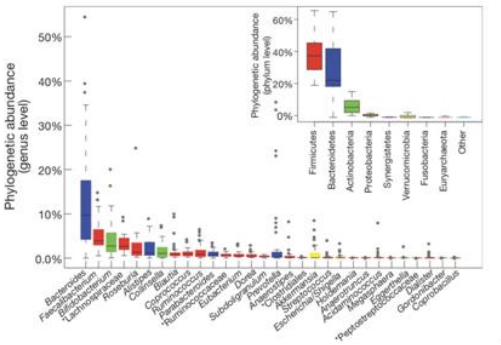
Taxonomy Profiling

Human Microbiome Projects

- 124 healthy from Denmark and Spain
- 576.7 Gb sequences (4.5 Gb per sample)

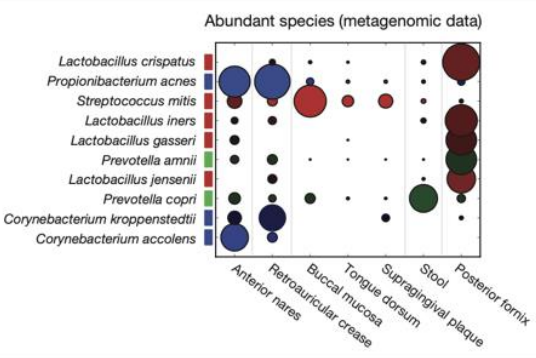
J. Qin et al Nature 2010

- 769 samples from 18 body sites of 242 healthy adults
- 4.6Tbp metagenomic sequences



- 30 most abundant genera by read abundance
- Reference genome based mapping with 85% and 65% sequence similarity cutoff

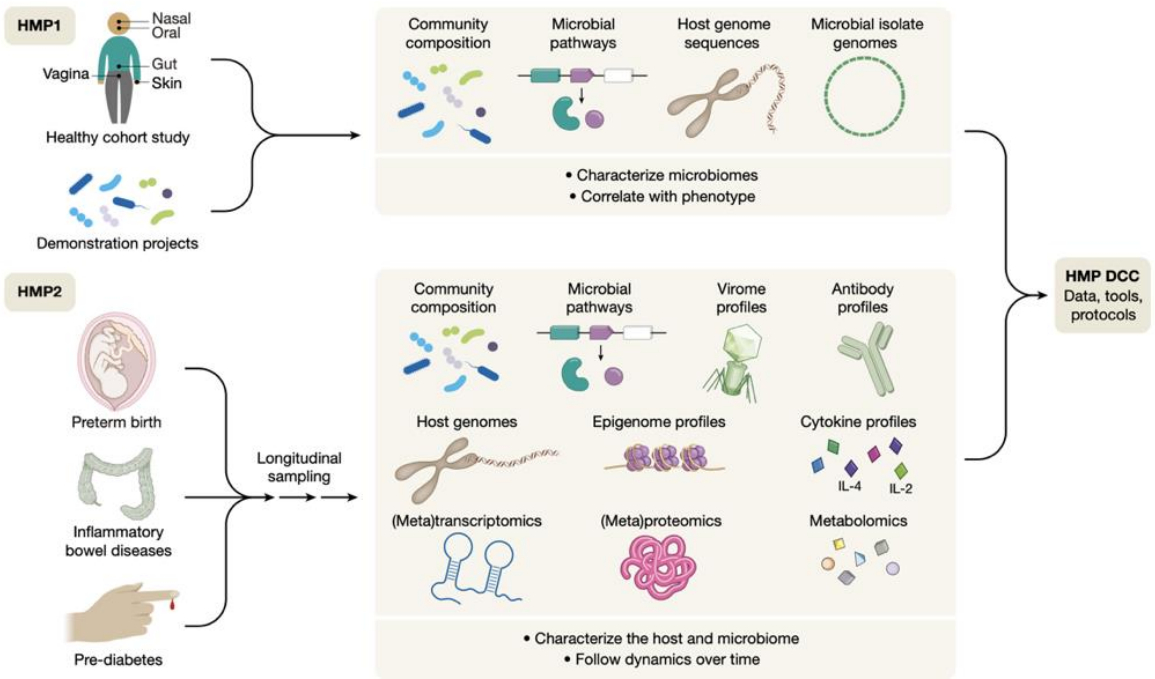
Nature 2012



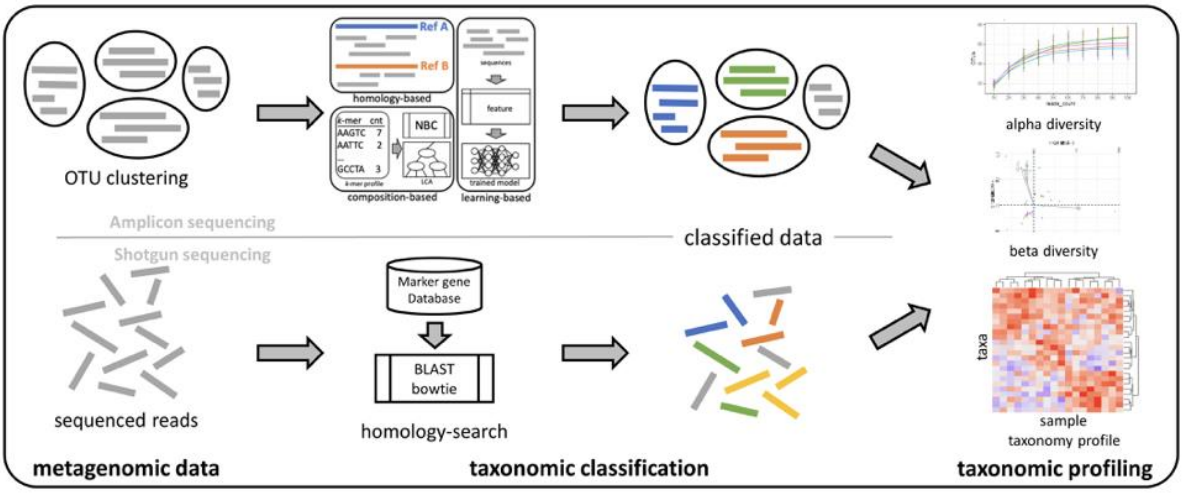
- Relative abundance based on read mapping against clade-specific marker genes (MetaPhlAn)

Nature 2012

iHMP: Integrative Human Microbiome Project

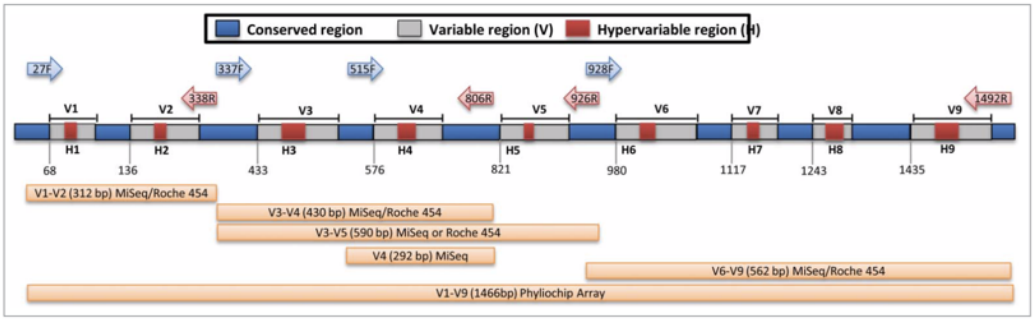


Taxonomy profiling

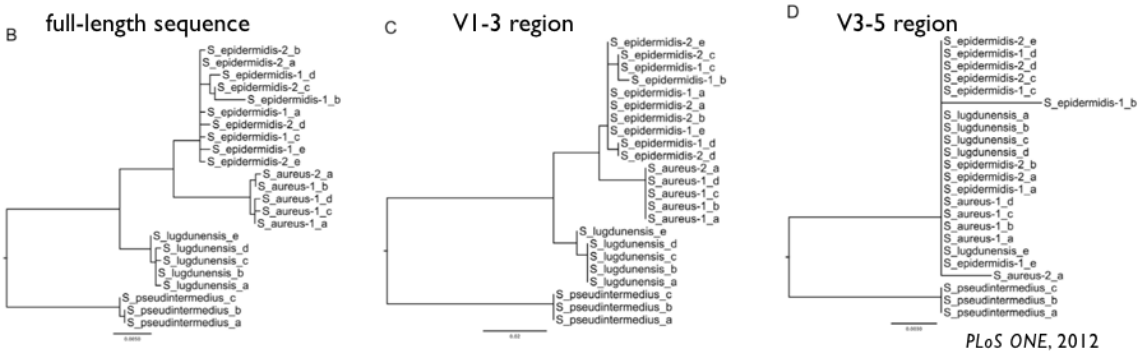


- Amplicon sequencing amplifies what you want
  - e.g. 16S rRNA sequencing (ref. p12)
- Metagenomic sequencing sequences all DNAs in a sample
- Pros and cons for host contamination, primer bias, taxonomic level, functional information

Taxonomy profiling

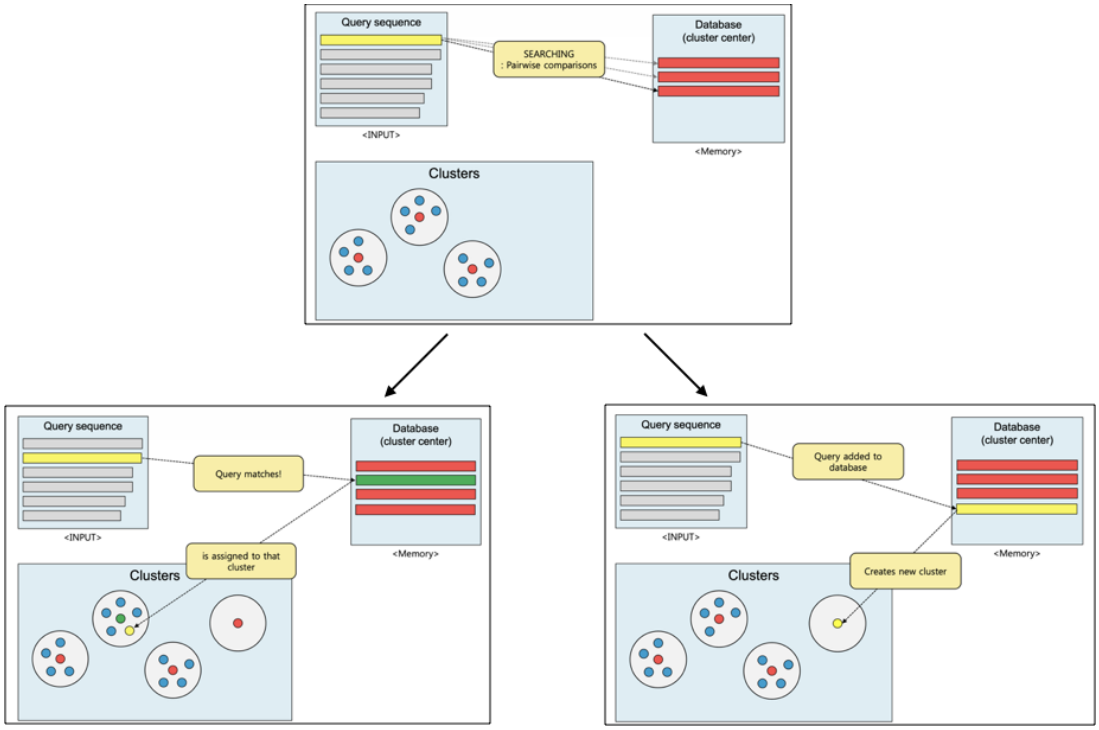


Gut Microbes, 2017

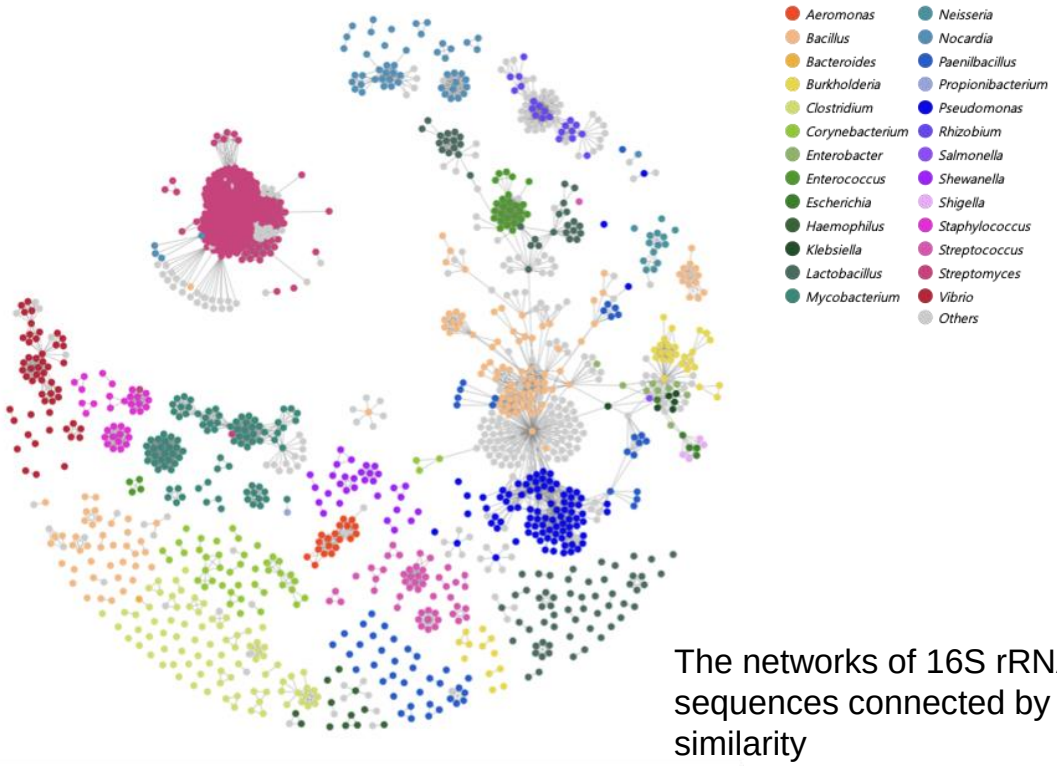


PLoS ONE, 2012

CD-HIT: clustering sequences based on greedy approach

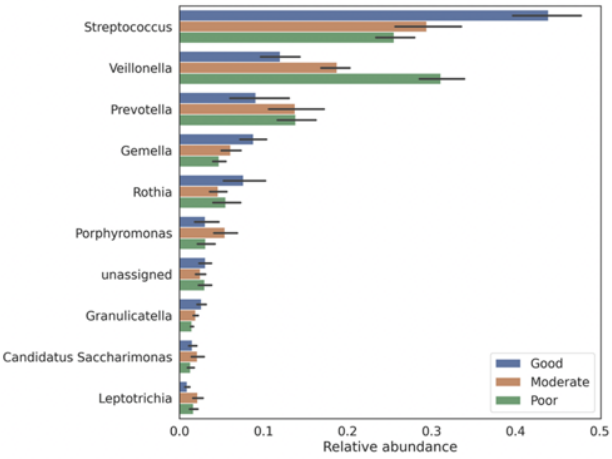


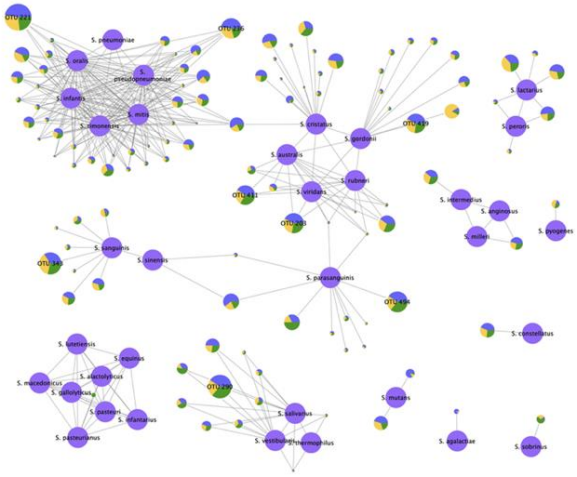
Species network based on sequence similarity



The networks of 16S rRNA sequences connected by the similarity

Profiling based on *homologous species groups*



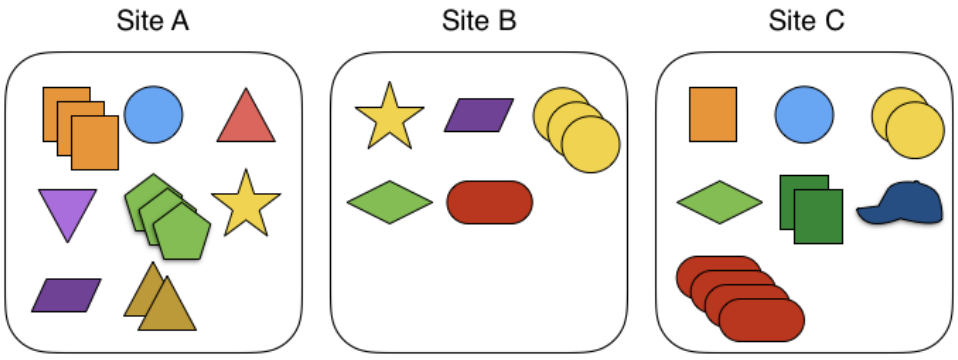


- Taxonomy profiling of the salivary microbiomes obtained from three different hygiene index groups: good, moderate, and poor
- 16S rRNA sequencing

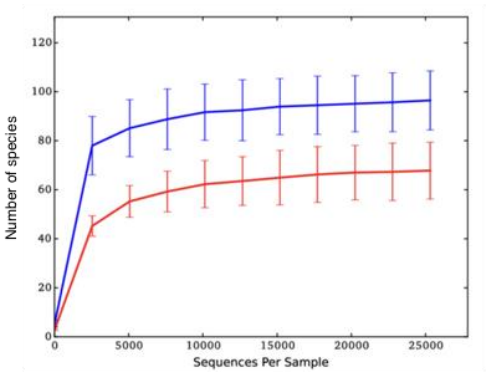
Frontiers in Microbiology, 2020

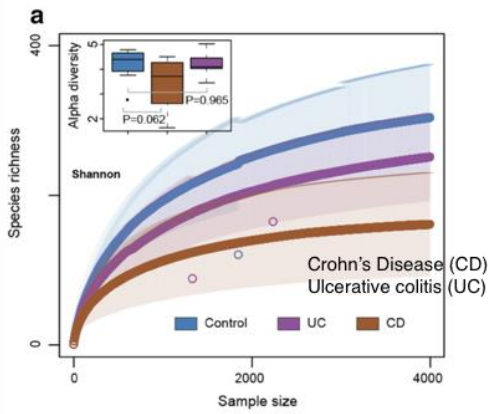
Alpha-diversity

- Variation of microbes in a single sample
- i.e. How many different species are detected in an environment?
- For example, site A=8, site B = 5, site C = 7
- Use metric such as Shannon diversity index, Simpson index, and Chao

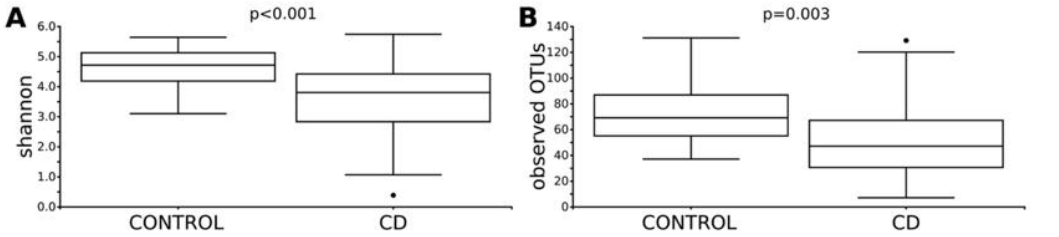


Alpha-diversity





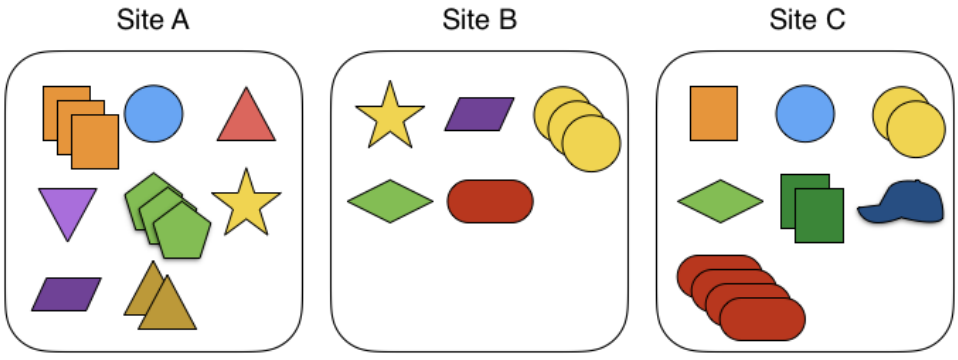
Gut Pathogens, 2020



Scientific Reports, 2019

Beta-diversity

- Variation of microbial communities between samples
- i.e. How different is the microbial composition between two samples?
- For example, sites between A and B=9, sites between B and C = 6, sites between C and A and C = 11
- Use metric such as Bray-Curtis dissimilarity and Jaccard distance



Beta-diversity

- Jaccard distance

$$d_J(A,B) = 1 - J(A,B)$$

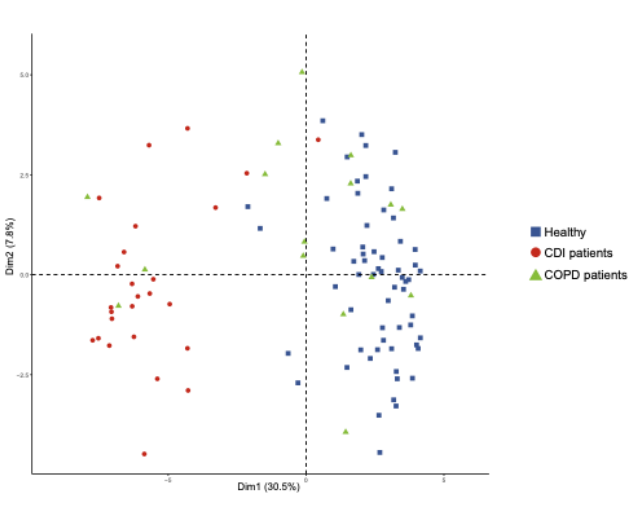
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

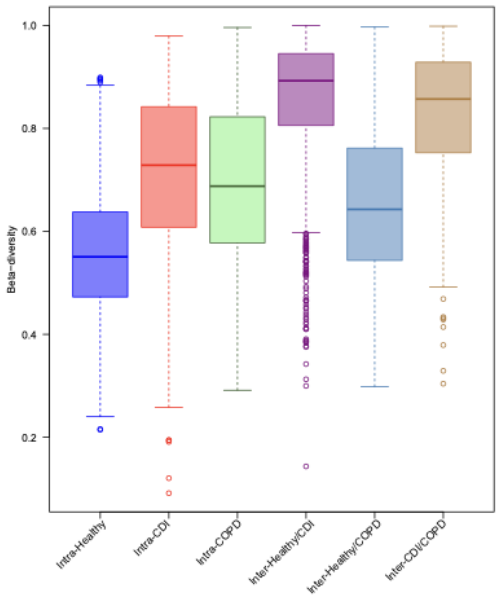
- Bray-Curtis dissimilarity

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

C_{ij} = the sum of only the lesser counts for each species found in both sites
Si = the total number of specimens counted on site i
Sj = the total number of specimens counted on site j

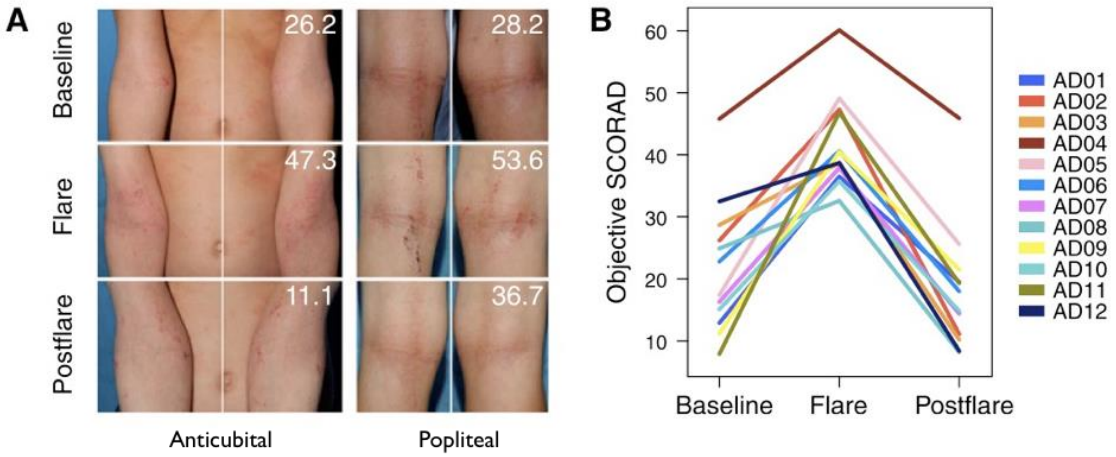
Beta-diversity





Dimension reduction/Embedding method
e.g. PCA, t-SNE

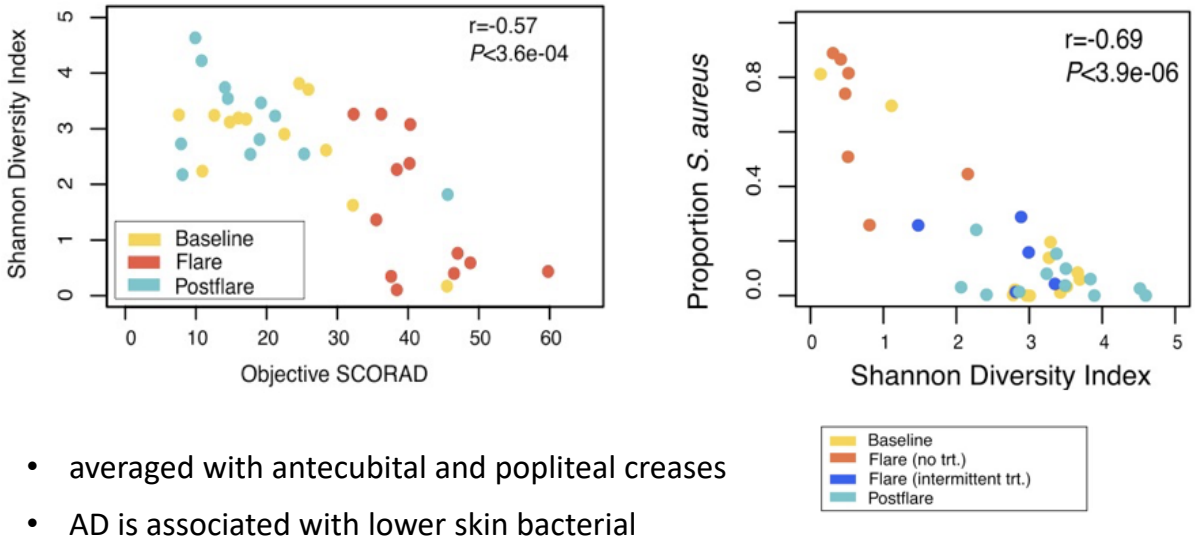
Temporal shifts in the skin microbiome of AD



- 12 2-15 yr patients (moderate-to-severe AD, SCORAD ≥ 15) vs. 11 healthy control
  - > 7d without topical AD treatments and > 4 wk off both antibiotics and corticosteroids
  - Flare: acute exacerbation of disease on any skin region
  - Postflare: 10-14 d after the initiation of intensified AD treatment (dilute bleach baths, topical steroids, bland emollients)
- 8F and 1391R regions of 16S rRNA

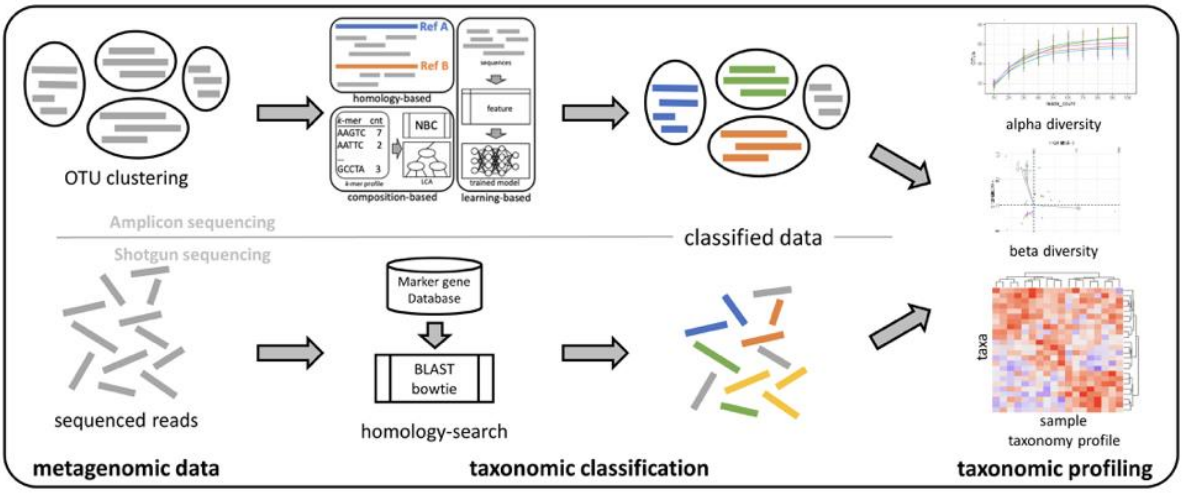
Genome Res. 2012

Temporal shifts in the skin microbiome of AD



- averaged with antecubital and popliteal creases
- AD is associated with lower skin bacterial diversity

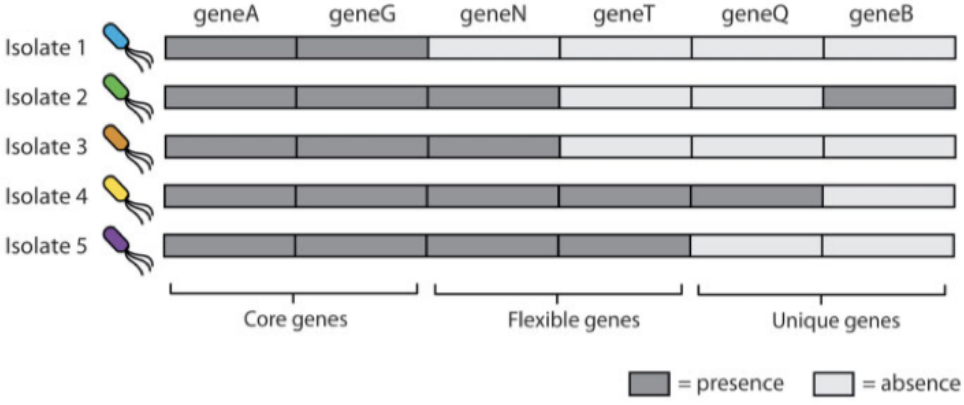
Taxonomy profiling



- Amplicon sequencing amplifies what you want
  - e.g. 16S rRNA sequencing
- Metagenomic sequencing sequences all DNAs in a sample
- Pros and cons for host contamination, primer bias, taxonomic level, functional information

MetaPhlAn: profiling based on marker genes

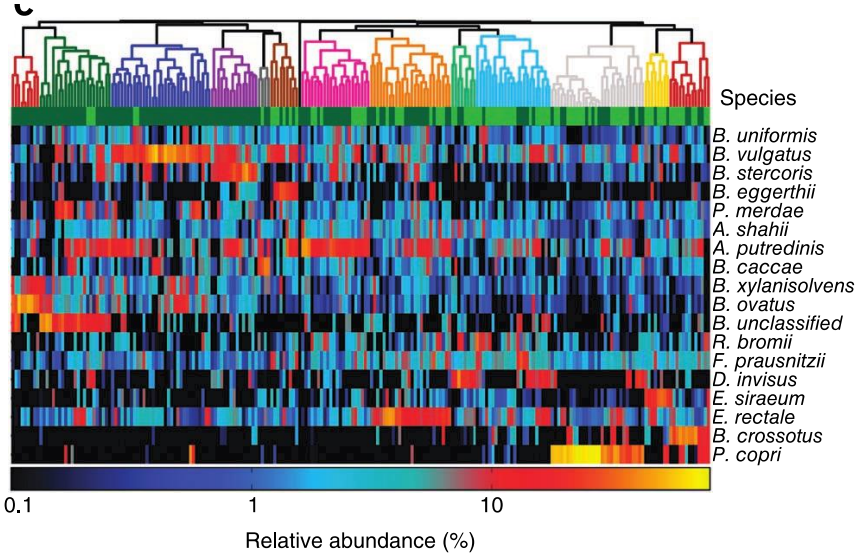
- Using core genes and unique genes, a set of unique clade-specific marker gene was collected
  - Core genes: all of the strains in a clade (species or others) have such genes
  - Unique genes: no other clade contains orthologs close enough to incorrectly map metagenomic reads



Nature Methods, 2012; GBE, 2018

MetaPhlAn: profiling based on marker genes

- MetaPhlAn 2
- Clustering CDS by nucleotide identity threshold of 75% to find seeds of core genes
- Check the seeds for the unique taxonomic marker genes



Nature Methods, 2012; GBE, 2018

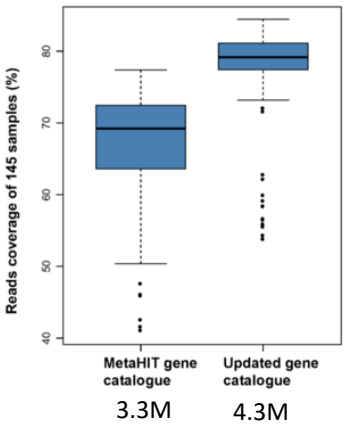
MetaPhlAn: profiling based on marker genes

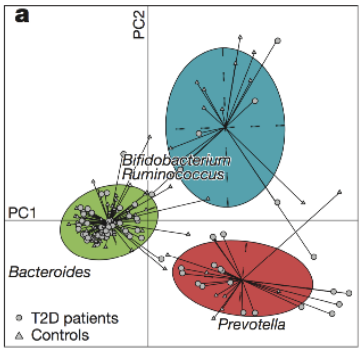
- MetaPhlAn 3
- Based on Uniref90
- 1.1M markers from 1M bacteria, 56.8K archaea, 61.8k virus, and 13.6K eukaryota for 13,475 species
  - Using 16.8K species from 16K bacteria, 739 archaea, and 122 eukaryota (99.2K genomes from 97.9K bacteria, 947 archaea, and 339 eukaryota)
  - 84 ± 47 markers per species from 16.8k species
  - Performance from 123 synthetic metagenome

Tool	Airways	Gastrointestinal tract	Oral	Skin	Urogenital tract	Mouse gut	Non-human
MetaPhlAn v3.0 stat_q 0.2	0.880	0.894	0.869	0.853	0.869	0.722	0.830
MetaPhlAn v3.0 stat_q 0.1	0.896	0.908	0.898	0.883	0.906	0.728	0.855
MetaPhlAn2 v2.7	0.723	0.761	0.672	0.751	0.701	0.544	0.330
mOTUs251_precision	0.768	0.824	0.787	0.761	0.779	0.770	0.632
mOTUs251_recall	0.683	0.768	0.772	0.720	0.727	0.800	0.529
Bracken_208_25_refseq	0.440	0.426	0.525	0.359	0.339	0.091	0.021

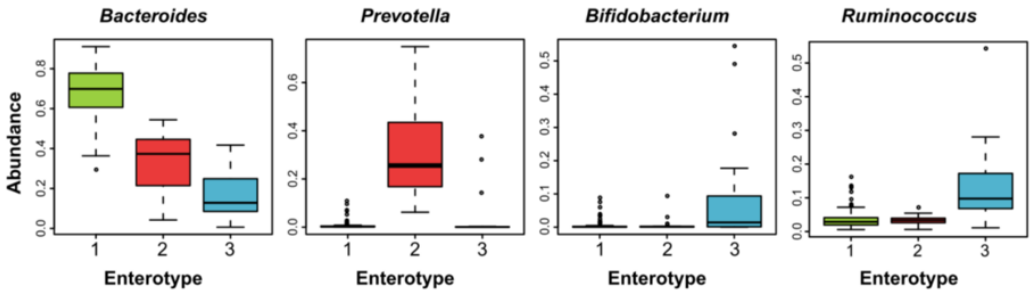
Nature Methods, 2012; bioRxiv 2021

Human Gut Microbiome of healthy group



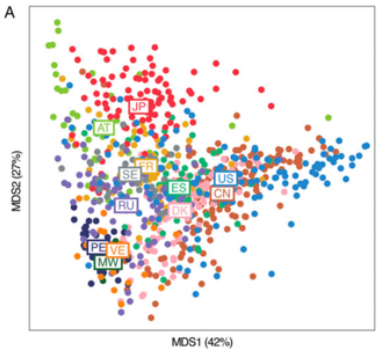


- 145 Chinese individuals (74 control, 71 T2D)
- 200 validation
- 350 bp paired-end reads

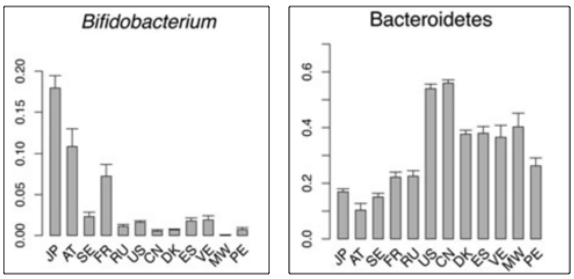


Nature 2013

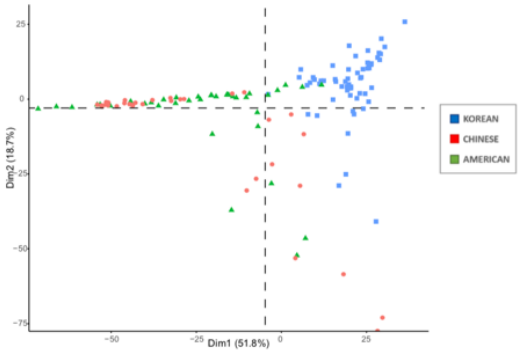
Microbial composition over the countries

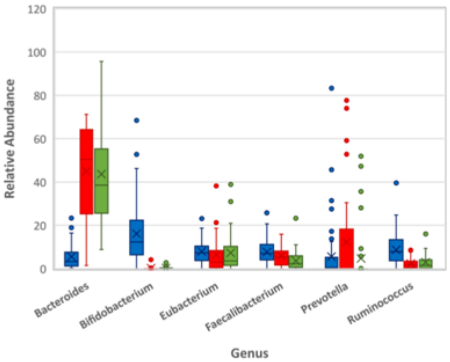


DK: Denmark
ES: Spain
US: USA
CN: China
SE: Sweden
RU: Russia
VE: Venezuela
MW: Malawi
AT: Austria
FR: France
PE: Peru



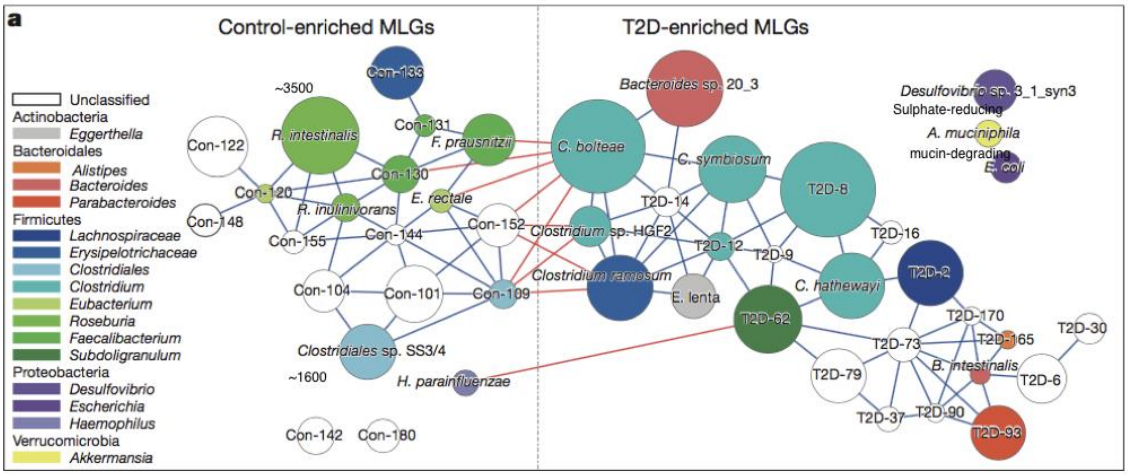
Bacterial composition in the gut microbiomes of different races (DNA Res, 2016)





Sci Rep, 2020

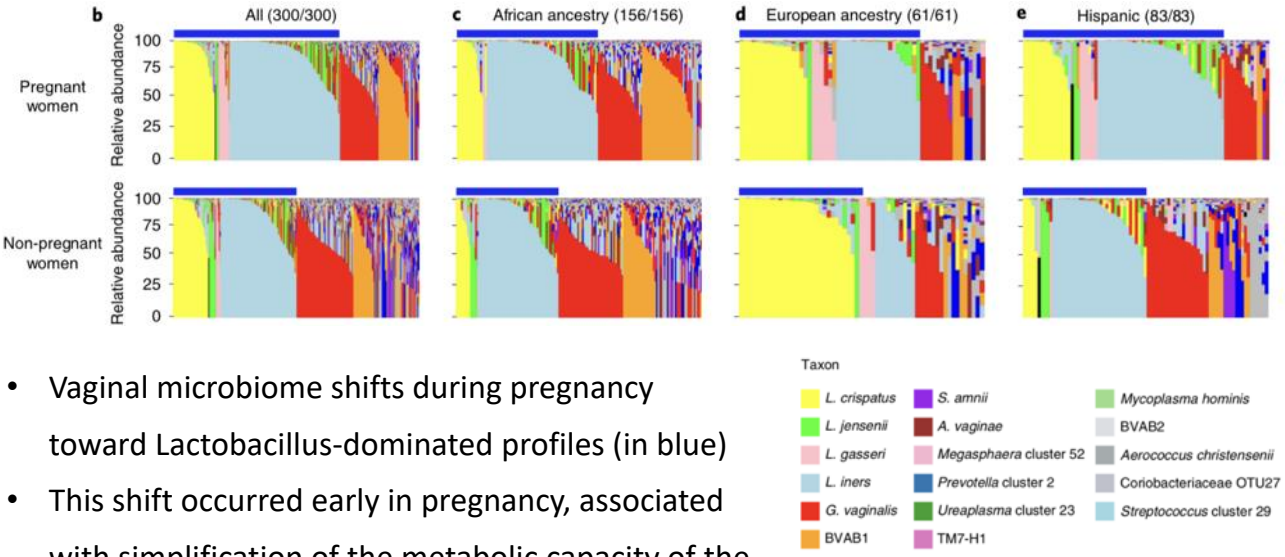
Differential abundance at species or strain level



- 145 Chinese (71 cases and 74 controls)
- MLG (metagenomic linkage group) using abundance level: a group of genes that co-exists among different individual samples and has a consistent abundance level
- Spearman correlation coefficient > 0.4 in blue, <-0.4 in red
- Butyrate-producers are negatively correlated with a group of the T2D-enriched bacteria from Clostridium

Nature, 2013

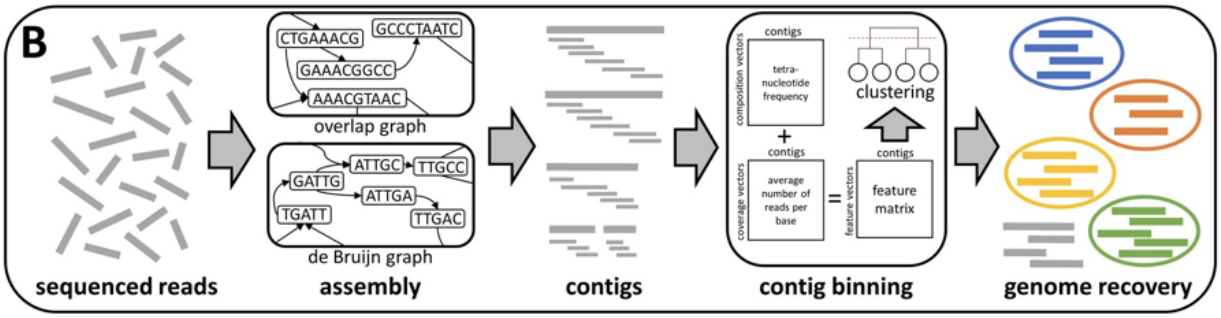
Human Vaginal Microbiome



- Vaginal microbiome shifts during pregnancy toward Lactobacillus-dominated profiles (in blue)
- This shift occurred early in pregnancy, associated with simplification of the metabolic capacity of the microbiome
- This shift is significant in women of African or Hispanic ancestry

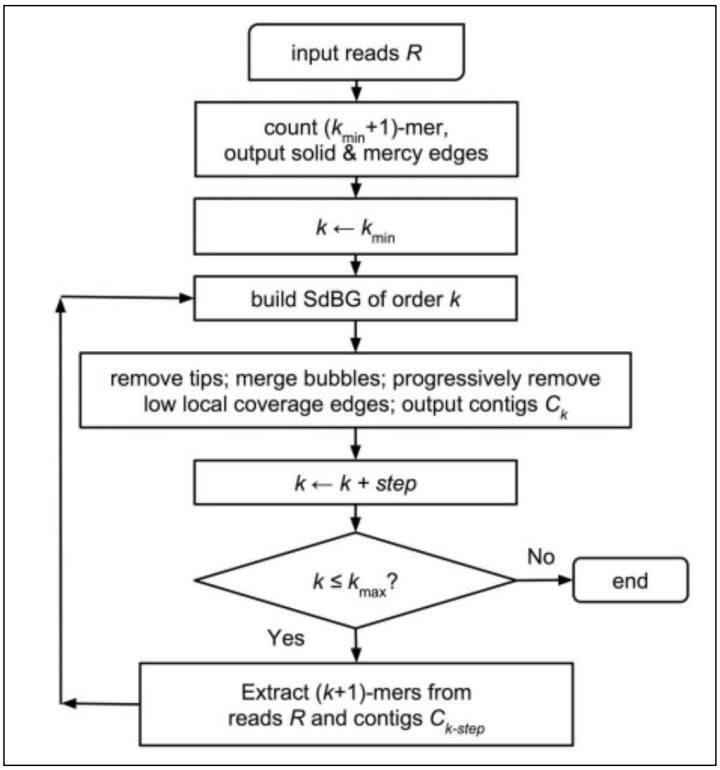
Nature Medicine, 2019

Assembly and genome recovery



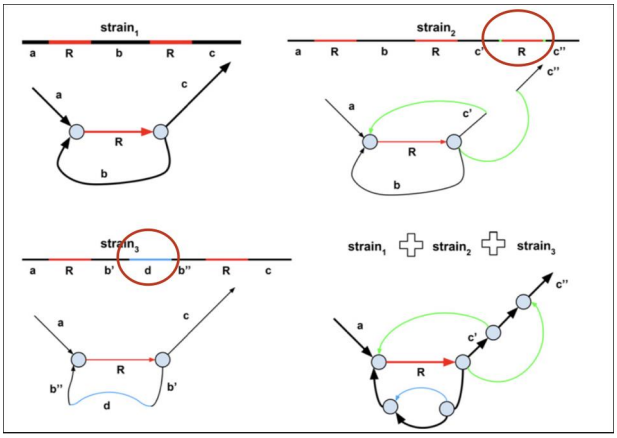
- Assembly
  - overlap graph-based assembly
  - mostly, de Bruijn graph-based assembly;
  - e.g. metaSPAdes, MEGAHIT
- Contig clustering

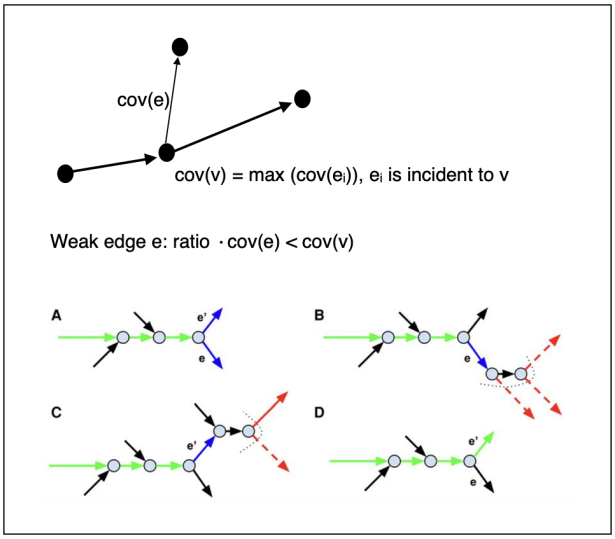
MEGAHIT



Bioinformatics, 2015

metaSPAdes

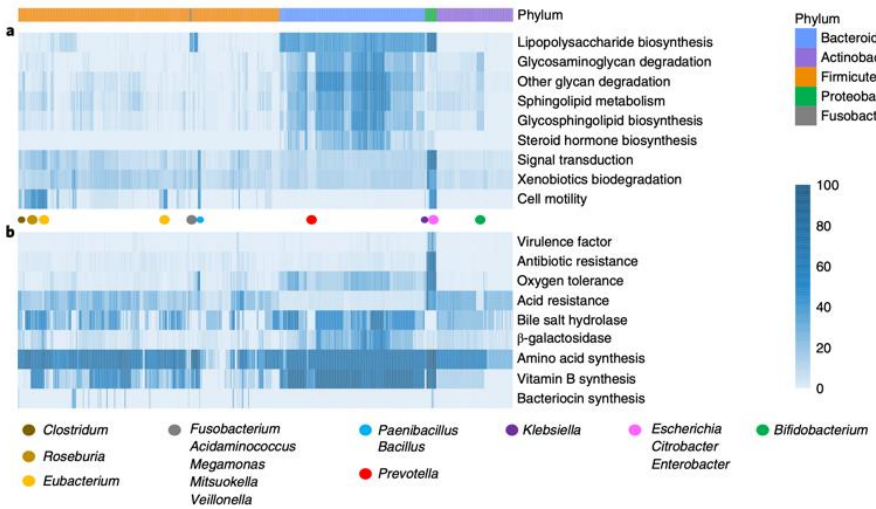




- De Bruijn graphs of multiple strains with different abundance
  - Collapses bulges and removes tips
  - Disconnect weak edges
  - Resolve repeats

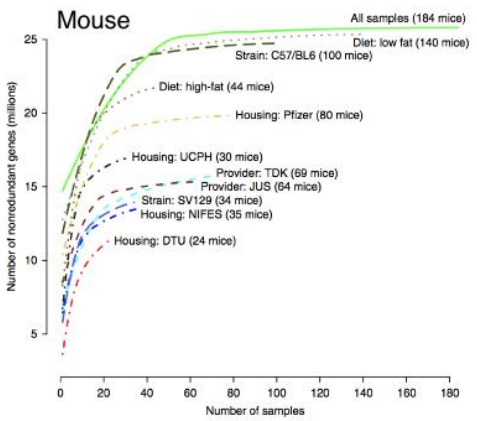
Reference catalogue

- The first catalog of 178 reference bacterial genomes by HMP (2010)
- 1,520 reference genomes of bacteria cultivated from gut (2019)
- 50% -> 70% coverage of gut microbiome
- 338 species-level clusters (ANI >= 95%)
- 134 clusters (264 genomes) were newly annotated

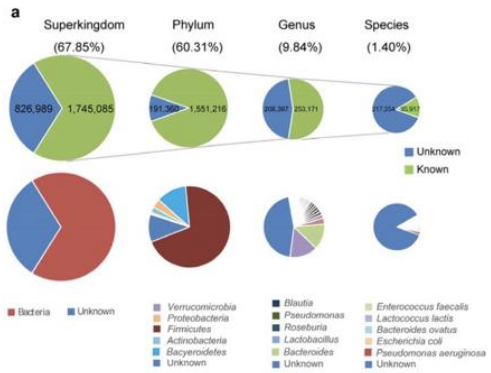


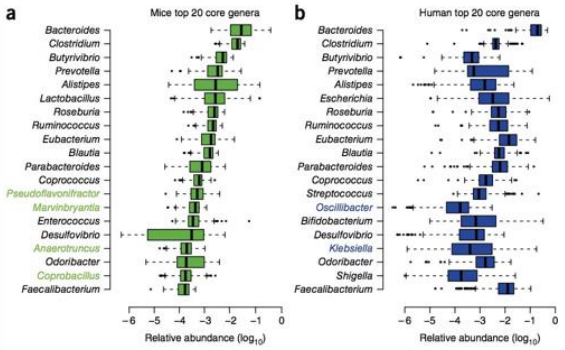
Nature Biotech, 2019

Mouse Gut Microbiome



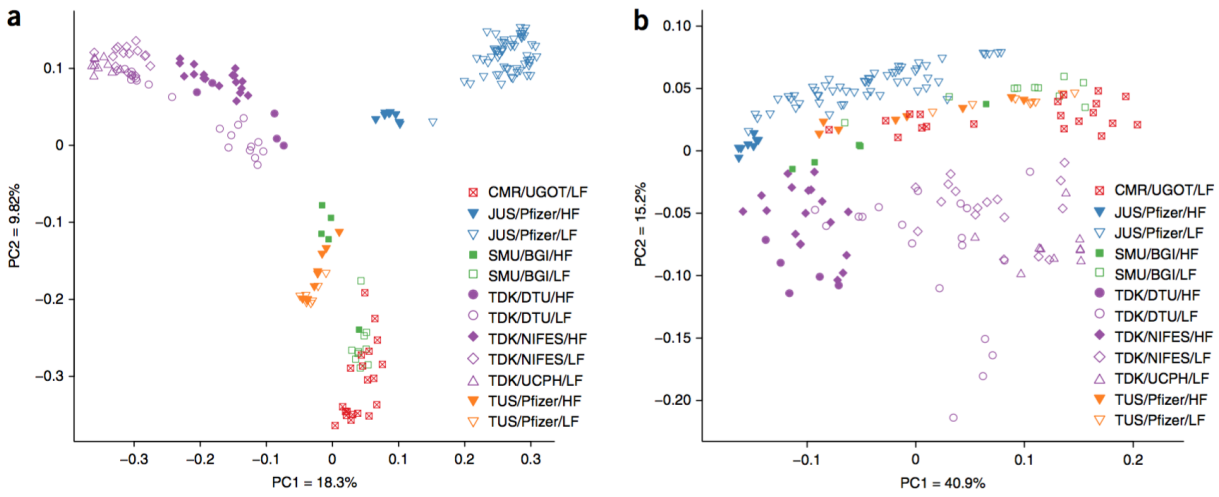
- Estimation of bacterial composition by origin of functional genes
- Gene catalog by 95% identity and 90% coverage
- Genus assignment with 85%, 80% of the alignment coverage
- Phylum assignment, 65% identity





Nature Biotech, 2015

Mouse Gut Microbiome

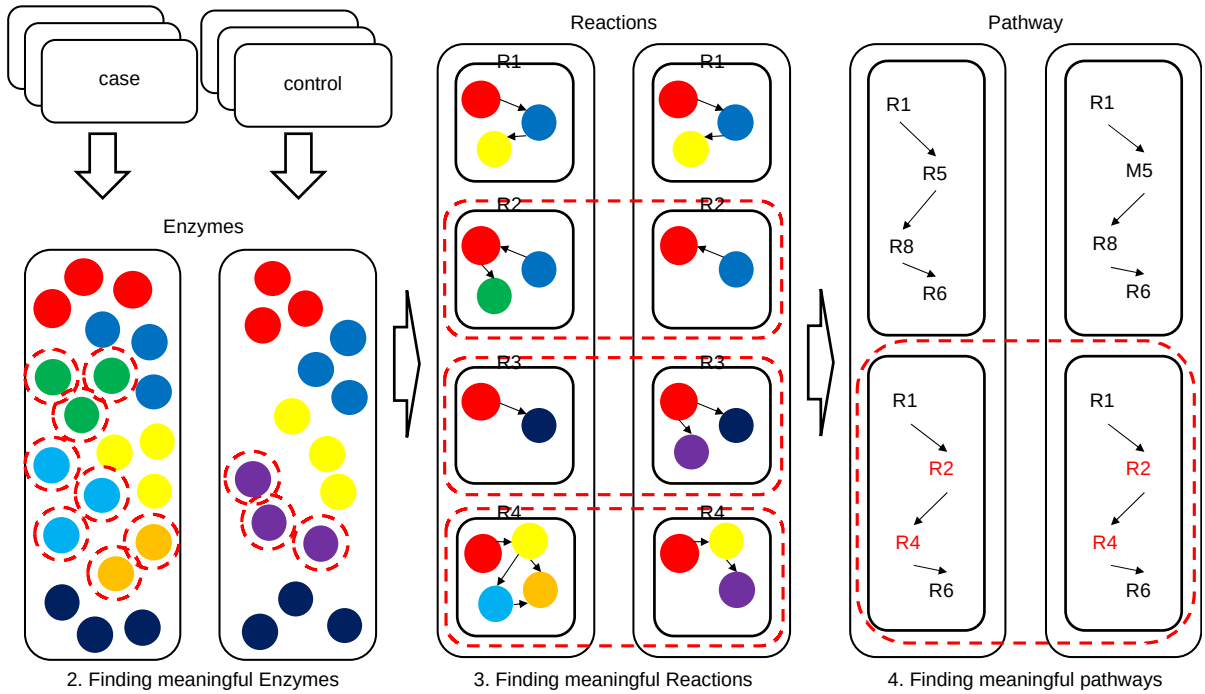


- PCoA based on gene profile (a) and KO profile (b)
- Mice from different providers are clearly separated along PC2
- Compositions are quite different by the provider and housing laboratory

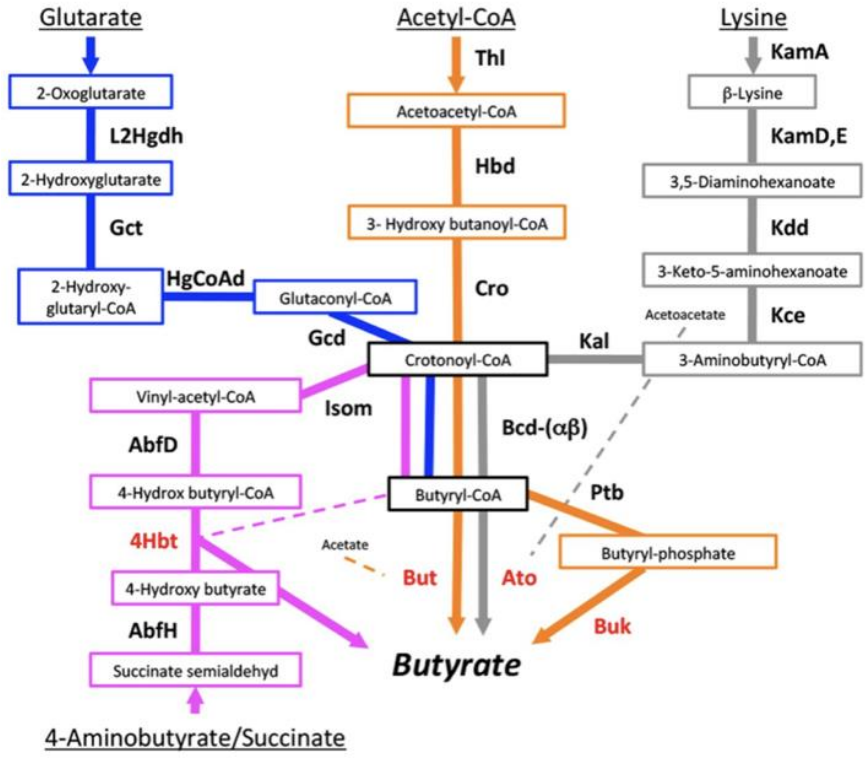
Nature Biotech. 2015

Function Profiling

Functional analysis



Four different pathways for pan-genomes

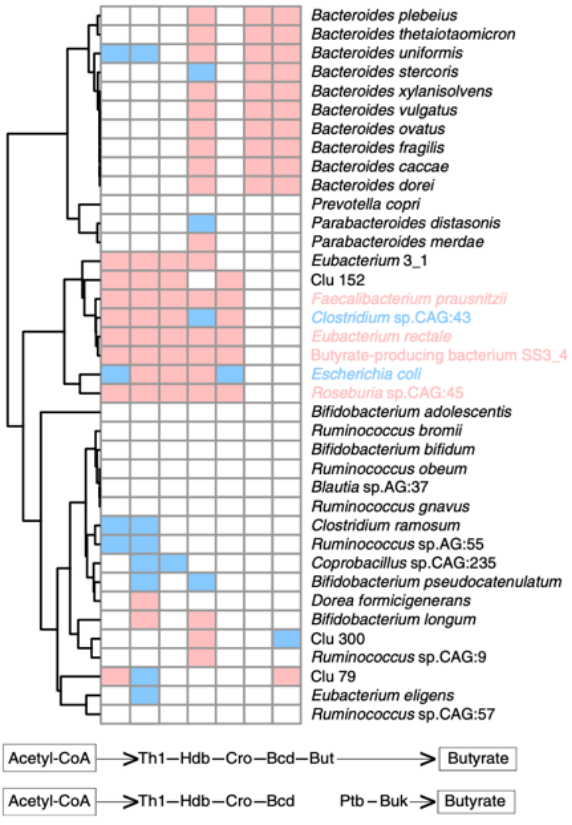


mBio,2014

Four different pathways for butyrate synthesis

- Distribution of genes involved in butyrate biosynthesis pathway
- 36 speices with more than 10 genomes and two other species of *Fecalibacterium prausnitzii* and butyrate-producing bacterium SS3_4 (enriched in healthy people)
- Genes in the core genomes (pink) and dispensable genomes (cyan)

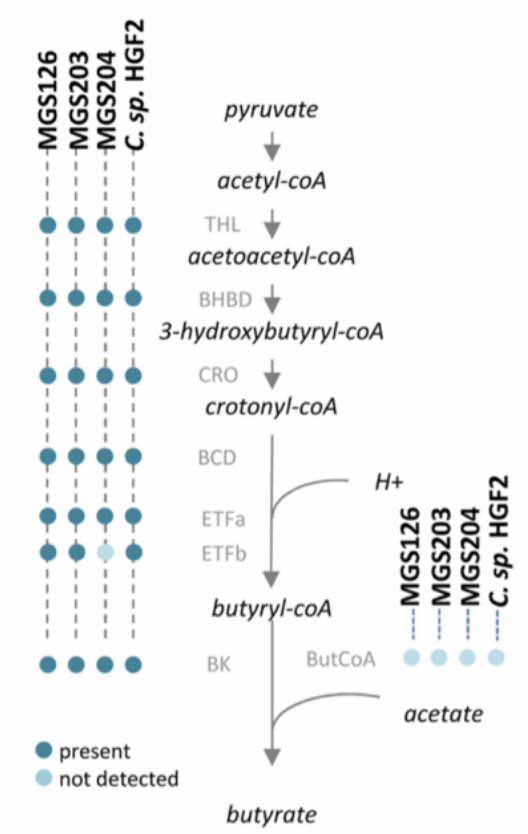
-Thl, thiolase;
-Hdb, β-hydroxybutyryl- CoA dehydrogenase;
-Cro, crotonase;
-Bcd, butyryl-CoA dehydrogenase (including electron transfer protein α and β subunits);
-But, butyryl-CoA:acetate CoA transferase;
-Ptb, phosphate butyryltransferase;
-Buk, butyrate kinase



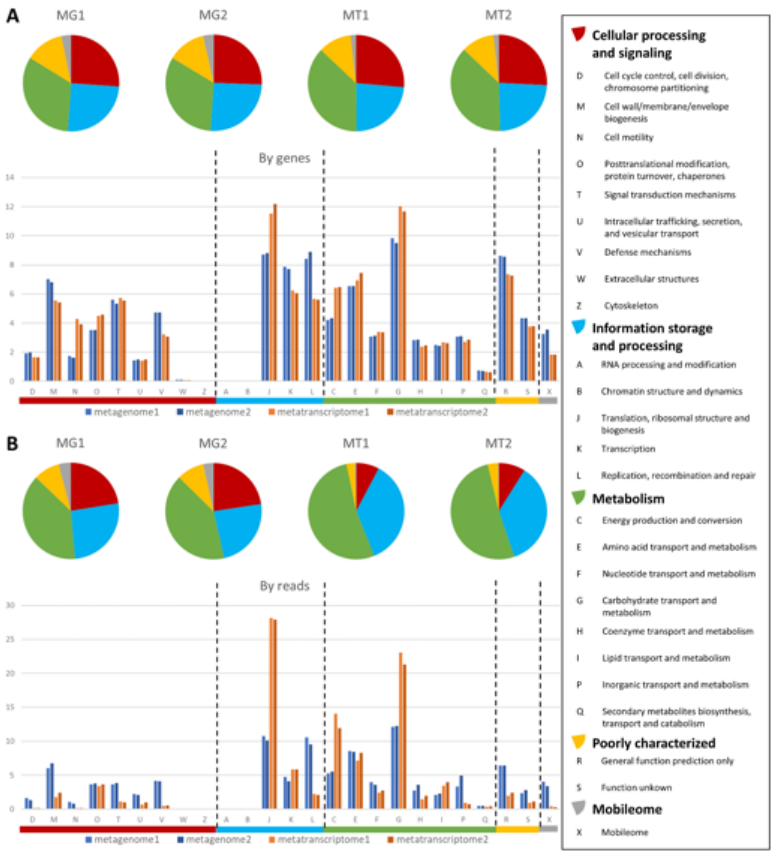
Nature Biotech, 2019

Functional analysis with MetaGenome species (MGS)

- Modulated by a fermented milk product (FMP) in a 4-week intervention study in subjects suffering from Irritable Bowel Syndrome (IBS) with constipation (n=28).



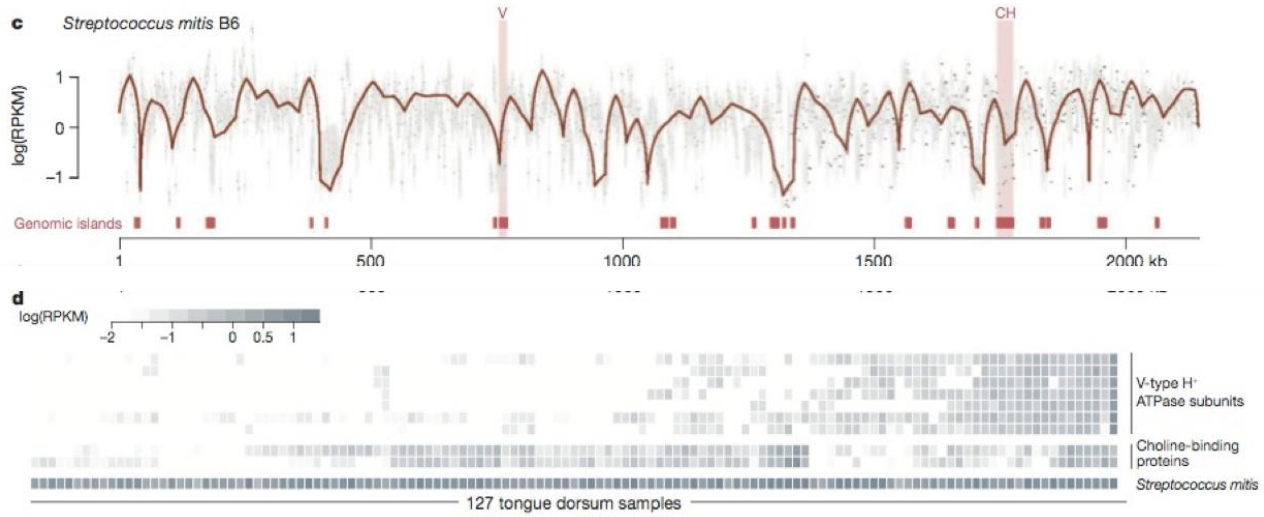
Metagenomics vs. Metatranscriptomics



PlosOne, 2019

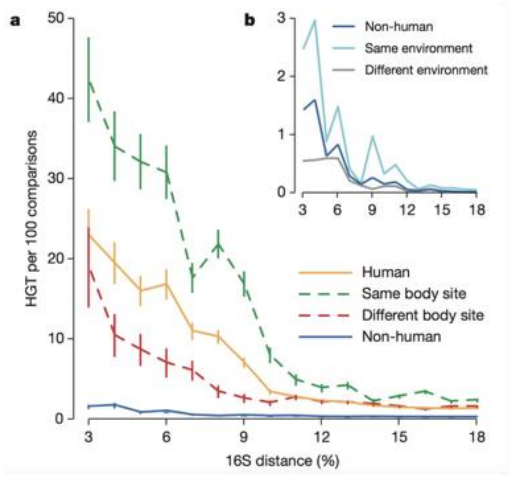
Genomic variation and exchange

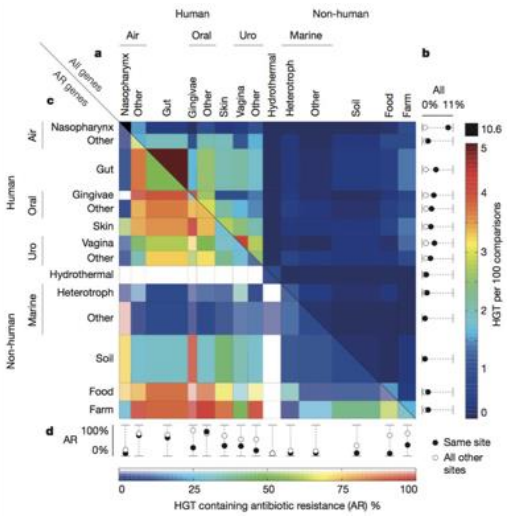
Microbial carriage varies between subjects



Nature, 2012

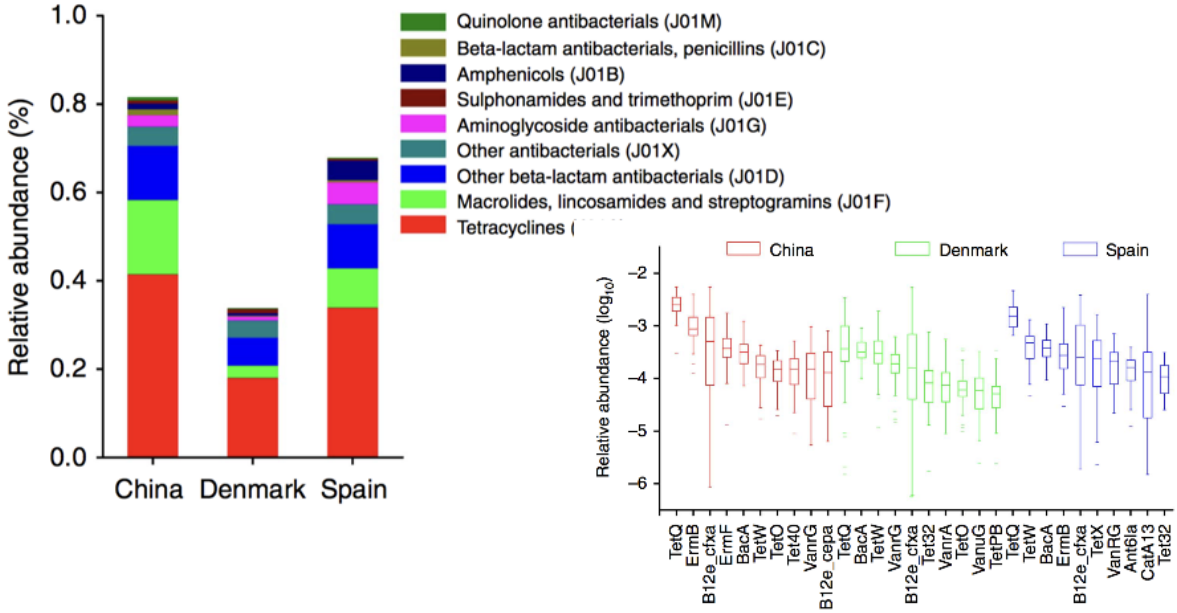
Global network of gene exchange





- Sequence blocks (>500 nucleotide, > 99%) in genomes (< 97% 16S rRNA)
- 10,770 unique genes in 2,235 bacterial genomes have been transferred
- 25 times more HGT between human associated than non-human isolates
- 27% of observed transfer event include at least one MGE
- 1,183 human-associated bacteria and 1,052 non-human-associated isolates

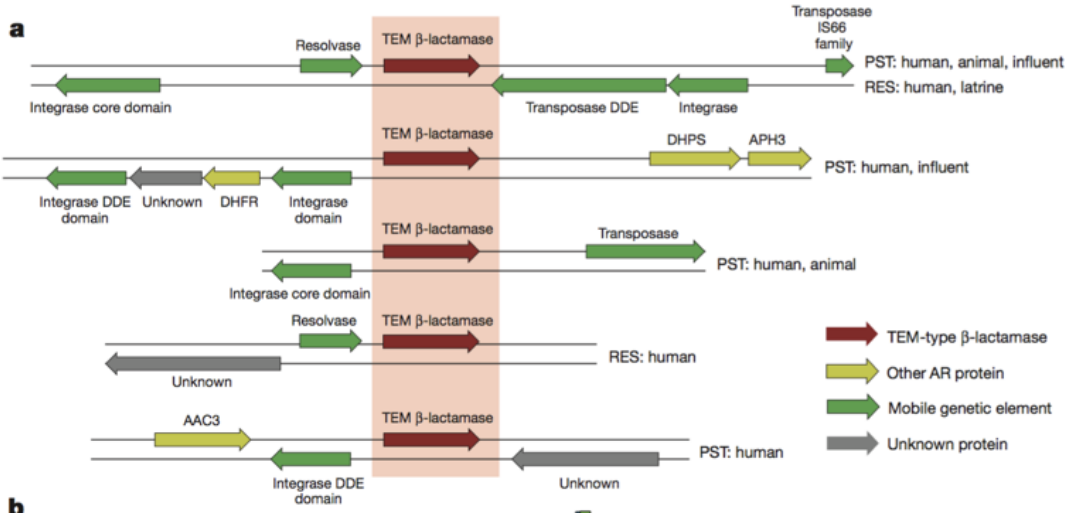
Antibiotic resistance genes in the gut microbiome



- 1093 antibiotics resistance genes found from 162 individuals (85 Danish, 39 Spanish, 38 Chinese)
- 7825 antibiotic resistance proteins from ARDB
- e-value < e-10, coverage >70%, sequence similarity > 80%

Nature Comm., 2013

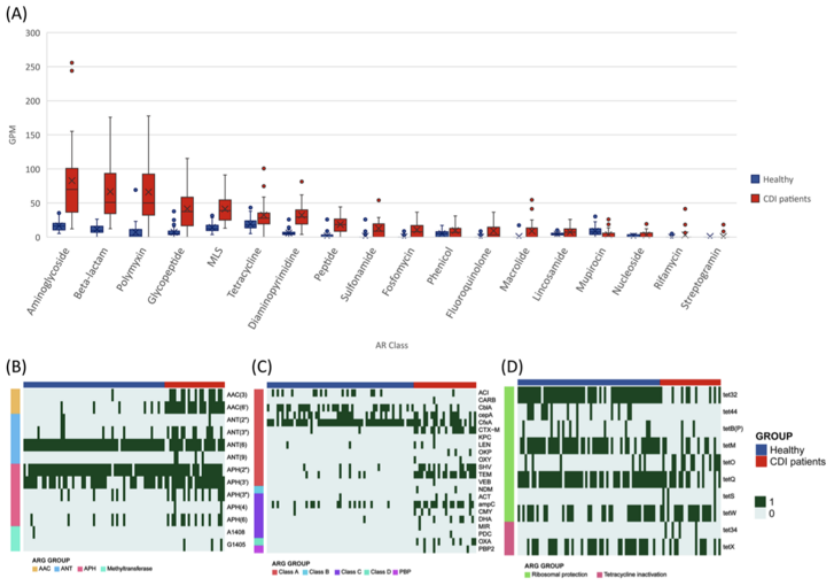
Gene exchange for resistome



Gene(s)	Clinical isolate(s)
$bla_{\text{TEM-1c}}$	*E. coli* BM2729
$bla_{\text{TEM-8}}$	*K. pneumoniae* HM12
$bla_{\text{TEM-10b}}$ and $bla_{\text{TEM-23}}$	*E. coli* F2
$bla_{\text{TEM-11}}$	*E. coli* 1724A
$bla_{\text{TEM-12c}}$	*E. coli* F1, *E. coli* MG32
$bla_{\text{TEM-13}}$	*M. morganii* BM2717
$bla_{\text{TEM-15a}}$	*K. pneumoniae* BM2730
$bla_{\text{TEM-15b}}$	*K. pneumoniae* BM2731 and BM2732
$bla_{\text{TEM-24b}}$	*E. aerogenes* SLE48
$bla_{\text{TEM-53}}$	*K. pneumoniae* BM2733
$bla_{\text{TEM-33b}}$	*E. coli* BM2725
$bla_{\text{TEM-33c}}$	*E. coli* BM2724
$bla_{\text{TEM-54}}$	*E. coli* BM2728

Nature, 2016

Antibiotic resistance genes in the gut microbiome



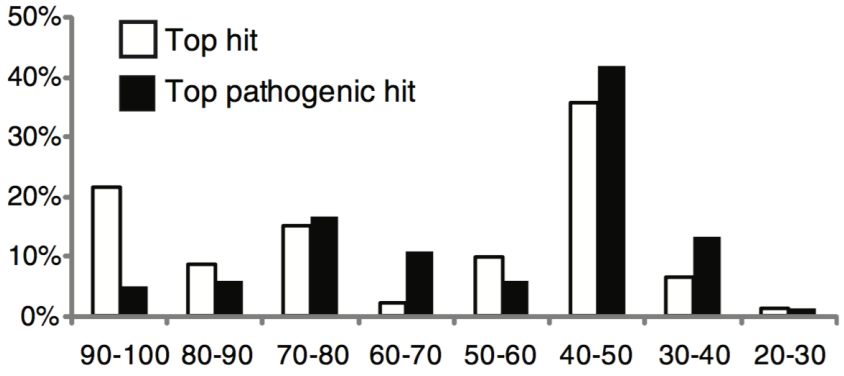
•Differential distribution of antibiotic resistance genes in patients with *Clostridioides difficile* infection and healthy people
•Distribution of AR class in healthy individuals and patients with CDI (*p* value < 0.0 in t-test)

Scientific Reports, 2020

Screening AR genes from the metagenomic DNA

- Functional metagenomics
  - Cloning of DNA fragments into a vector and the subsequent expression into heterologous hosts
  - Screening of resistance genes expression by growing them on antibiotic containing media
  - Advantage: The ability to isolate completely novel antibiotic resistance genes

•9.3 Gb library
•93 resistance genes identified from saliva and fecal microbiome of 2 subjects



Science, 2009

Metagenomics