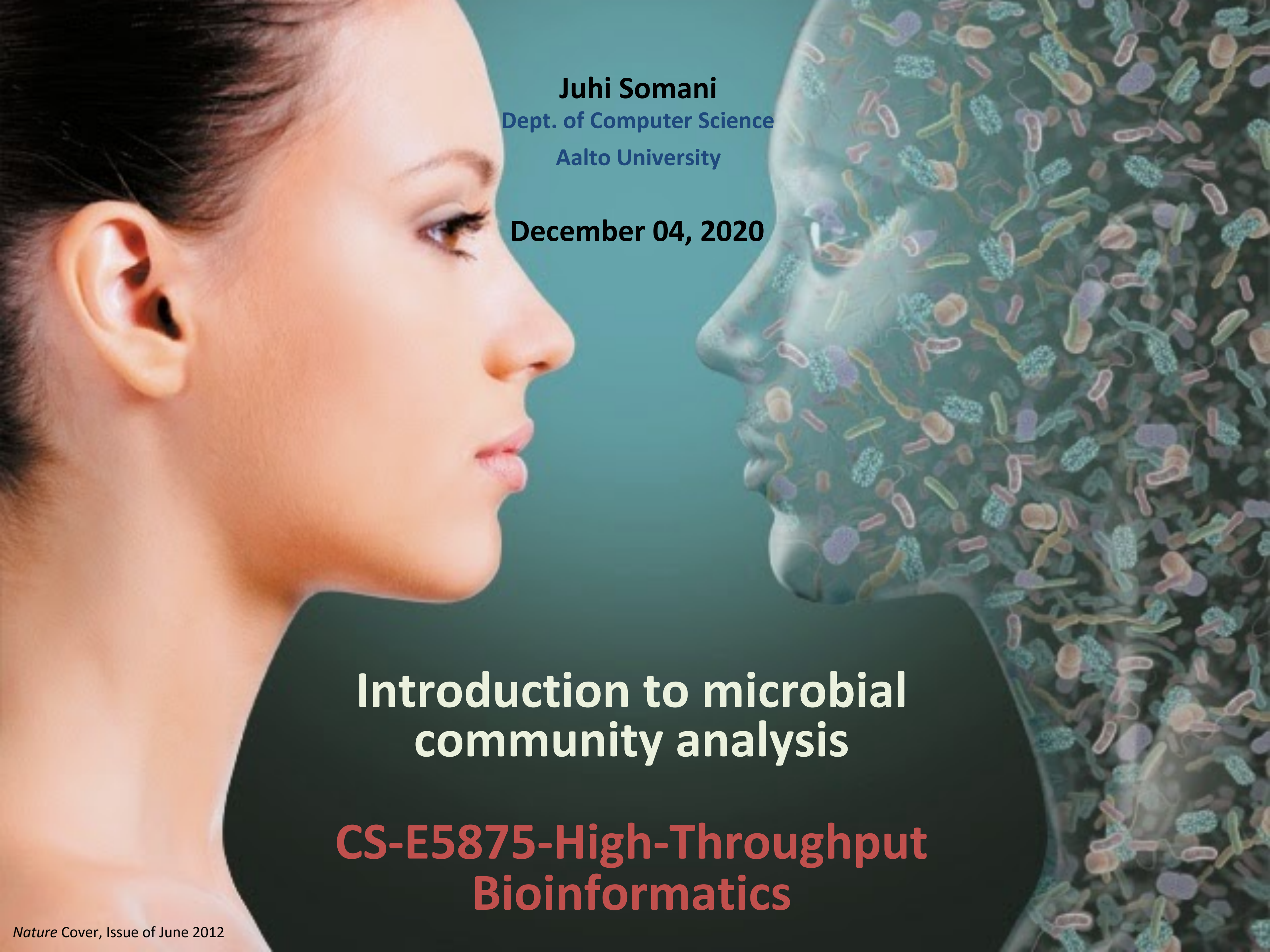


Juhi Somani

Dept. of Computer Science

Aalto University

December 04, 2020

Introduction to microbial community analysis

CS-E5875-High-Throughput Bioinformatics

Nature Cover, Issue of June 2012

Structure of the lecture

- Terminology
- Biological background
- Important microbiome studies
- Sequencing Technologies
- Processing microbial sequencing data & taxonomic profiling
- Functional analysis
- Normalization
- Diversity metrics and ordination
- Association analysis

TERMINOLOGY

Terminology

- **Microorganims or microbes:**

  microscopic organisms that are found all around us, such as bacteria, archaea, fungi, microbial eukaryotes, viruses and phages

- **Diversity**:

  a community's number and distribution of organisms

- **Microbiome & microbiota:**

  (Definitions of these terms are inconsistent in literature and are often used interchangeably. More details in Marchesi *et al.*, *Microbiome* 2015)

  Here, we will define both terms to refer to the collection of microbes as well as their genomes (i.e. genes) in a community

- **Metagenome:**

  The total genomic content of all microbes within a community

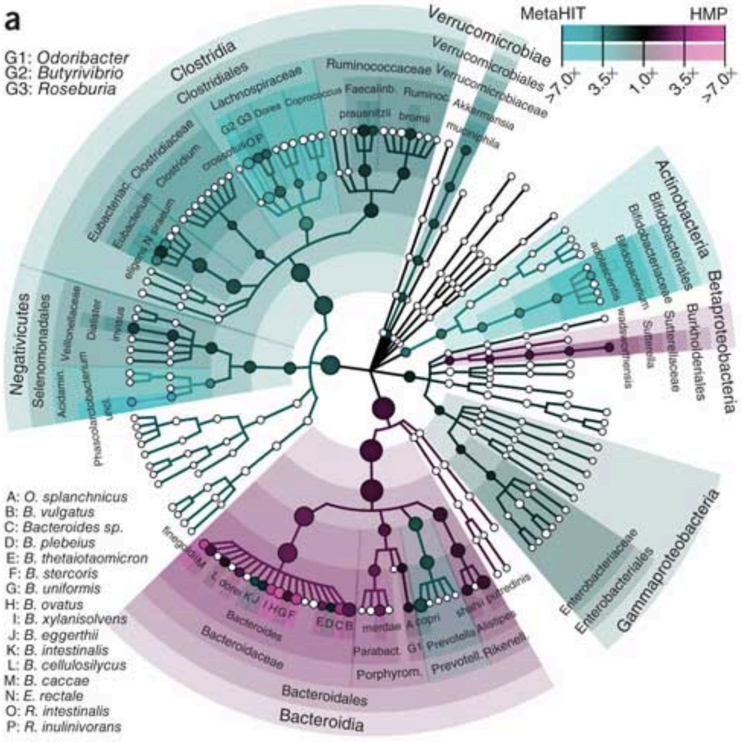
Terminology

- **Taxa/Taxon:**

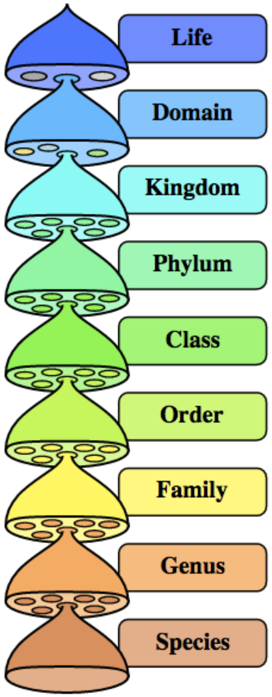
  hierarchy by which all lifeforms on earth can be represented; 8 major taxonomic ranks (bottom right figure)

- **Phylogenetic tree:**

  evolutionary tree that shows relationships between different species; each node is called a taxonomic unit (bottom left figure)



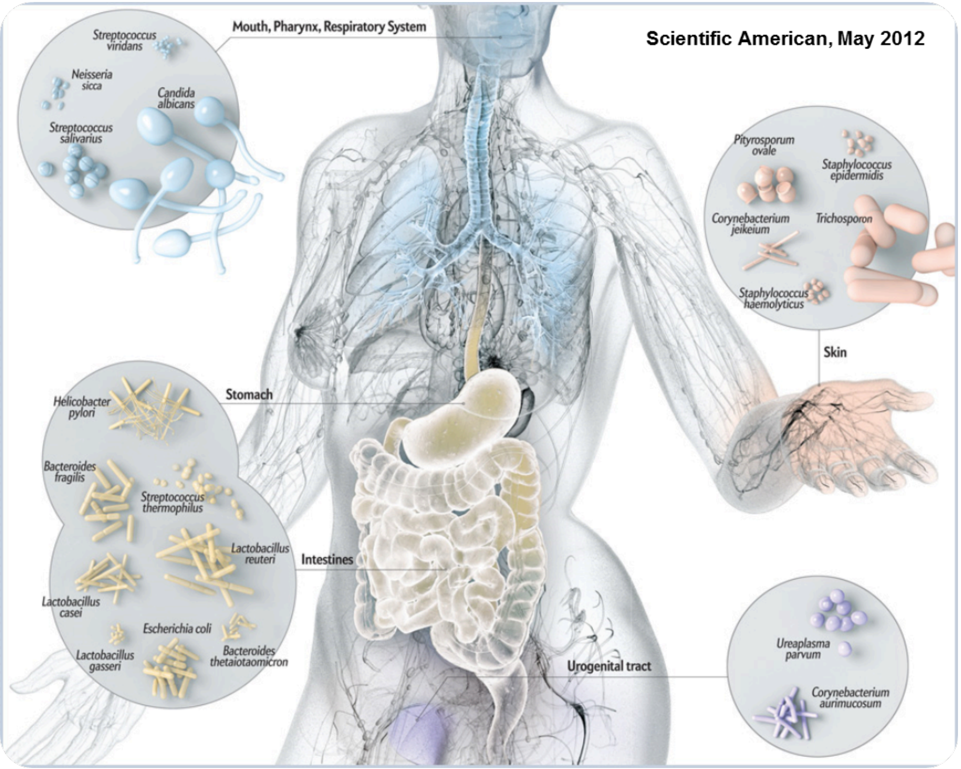
Segata *et al.*, *Nature Methods* 2012



en.wikipedia.org/wiki/Kingdom_(biology)

BIOLOGICAL BACKGROUND

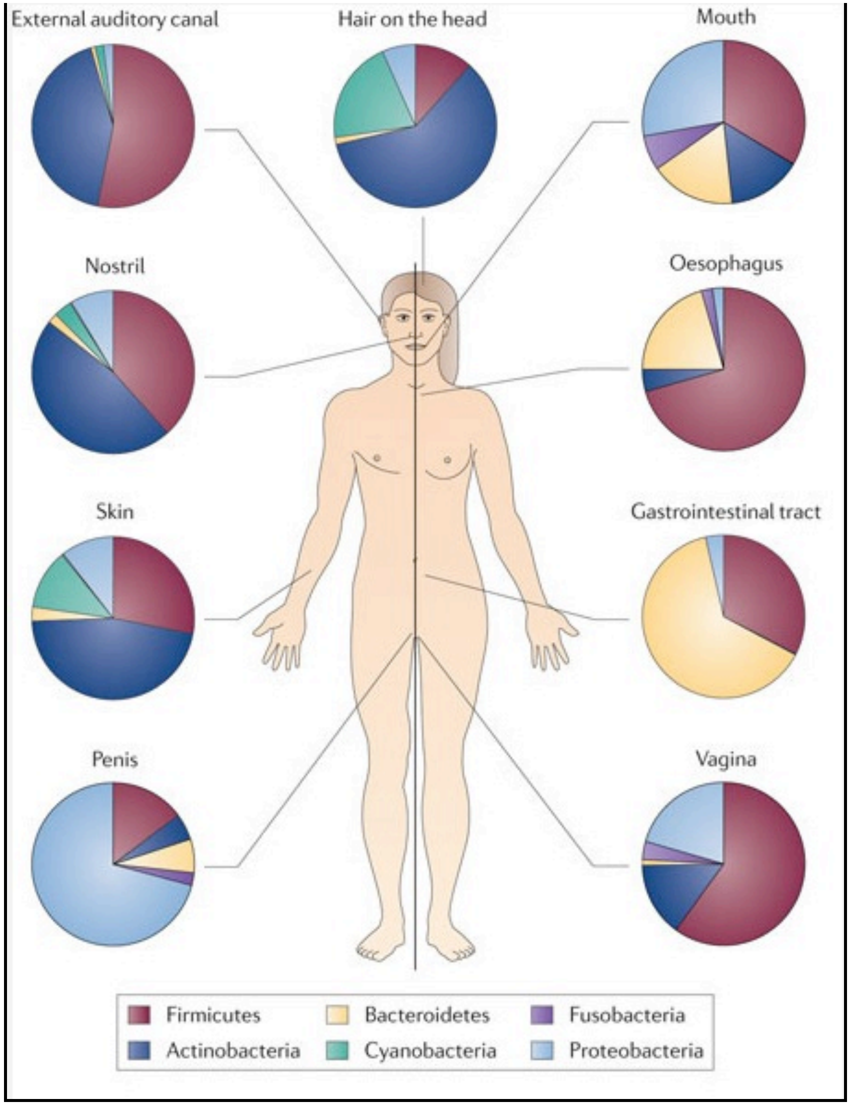
We, the “Super-organisms”



Scientific American, May 2012

- All areas of the human body is colonized by several tens of trillions of microbes
  - collectively known as the **human microbiome**
  - several kilos in body weight
- Approximately as many microbial cells in or on the human body as human cells (1.3:1 ratio of microbes to human cells)
- Outnumber the genes in our genome by about 100:1
- Humans have coevolved with these microbes for millenia, establishing a symbiotic relationship
- In a healthy person, the microbes are **commensal** (“good”) and are responsible in many day-to-day functions
- Inter-individual (human) variability:
  - Human genome: 0.1-0.4%
  - Microbiome: 80-90%

Microbial communities at different body-sites



Spor *et al.*, *Nature Reviews Microbiology*, 2011

- **Each body-site has evolved to harbor specific microbes essential for its physiological activities, for instance:**
  - **Gastrointestinal tract (Gut)**
    - Breakdown of complex polysaccharides
    - Synthesis of vitamins
    - Colonization resistance
    - Maturation of the immune system
  - **Mouth**
    - Breakdown of simple carbohydrates
    - Regulation of pH
  - **Skin**
    - Vitamin D biosynthesis
  - **Vagina**
    - Regulation of pH
- **This results in strikingly different microbial communities between body-sites of an individual**
- **Provide protection from colonization by pathogens**

Gut microbiome

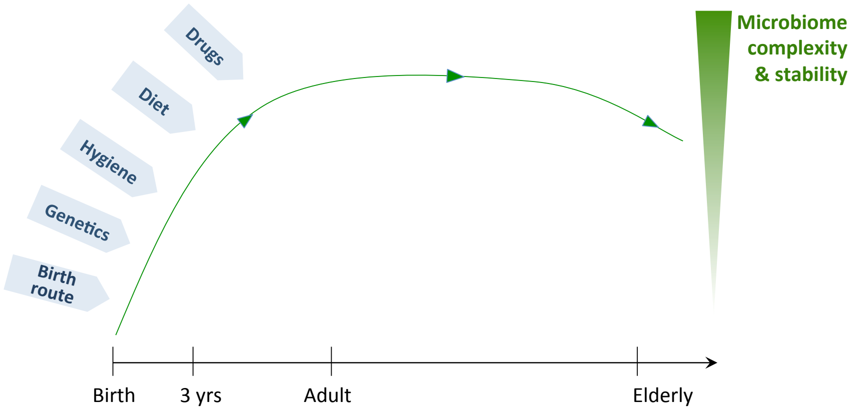
- Gut == gastrointestinal tract
- It is the largest area of the body that is constantly exposed to environmental antigens and microbes
- Houses the largest, most influential and a highly diverse reservoir of microbes and antigens in the human body
  - Tens of trillions of microbial cells that contains millions of unique genes (~150 times more genes than in the human genome)

Importance of gut microbiome

- A rich and diverse gut microbiome plays an essential role in human health and promoting immune homeostasis
- The gut accommodates the largest number of immune cells (up to 70%) of the human body
- From an early age, gut commensals (i.e. commensal microbes of the gut microbiome) establish a cross-talk with the immune system and calibrate nearly all aspects of the immune system, both local and systemic
  1. It trains the immune system to differentiate between commensals and pathogenic microbes
     - ⇒ enables immune system to shape and preserve the microbial ecology of the gut
  2. Gut commensals and the immune system compose the first 2 (of 3) layers of gut barrier
     - ⇒ contributes to the containment of the gut microbial cells, which is crucial for preventing gut microbes from translocating to other parts of the body or into the systemic blood circulation. The commensals that are beneficial for you in the gut can be dangerous in other parts of the body.
  3. Colonization resistance: inhibits pathogens from invading the host and initiating infections as well as clears existing infections

Infant gut microbial colonization

- Initial colonization takes pace *in utero*, but extensive colonization begins immediately after birth and continues until 2-3 years (or approx. 1000 days), after which it stabilizes to resemble that of an adult



Depiction by Tommi Vatanen

Infant gut microbiome – infant immune system interaction

- Infant immune system:
  - Unique in nature => it is also developing and is relatively immature
  - Characterized by blunted inflammatory responses and a regulatory environment => develops tolerance towards new antigens and microbes rather than launching an inflammatory response
  - More durable and permissive to microbial instructions during infancy => providing a ‘window of opportunity’ for proper (or improper) immune development and thus resilience (or susceptibility) towards diseases later in life
- Early microbial colonization of an infant’s gut:
  - is highly complex and dynamic
  - plays an instrumental role in the development (maturation and education) of the immune system
  - has long-term implications on host immune responses and health
- Therefore, a ‘healthy’ colonization by beneficial microbes during this critical window encourages proper immune development and training, which in turn promotes immune homeostasis and long-term health

Factors that influence early gut microbiome

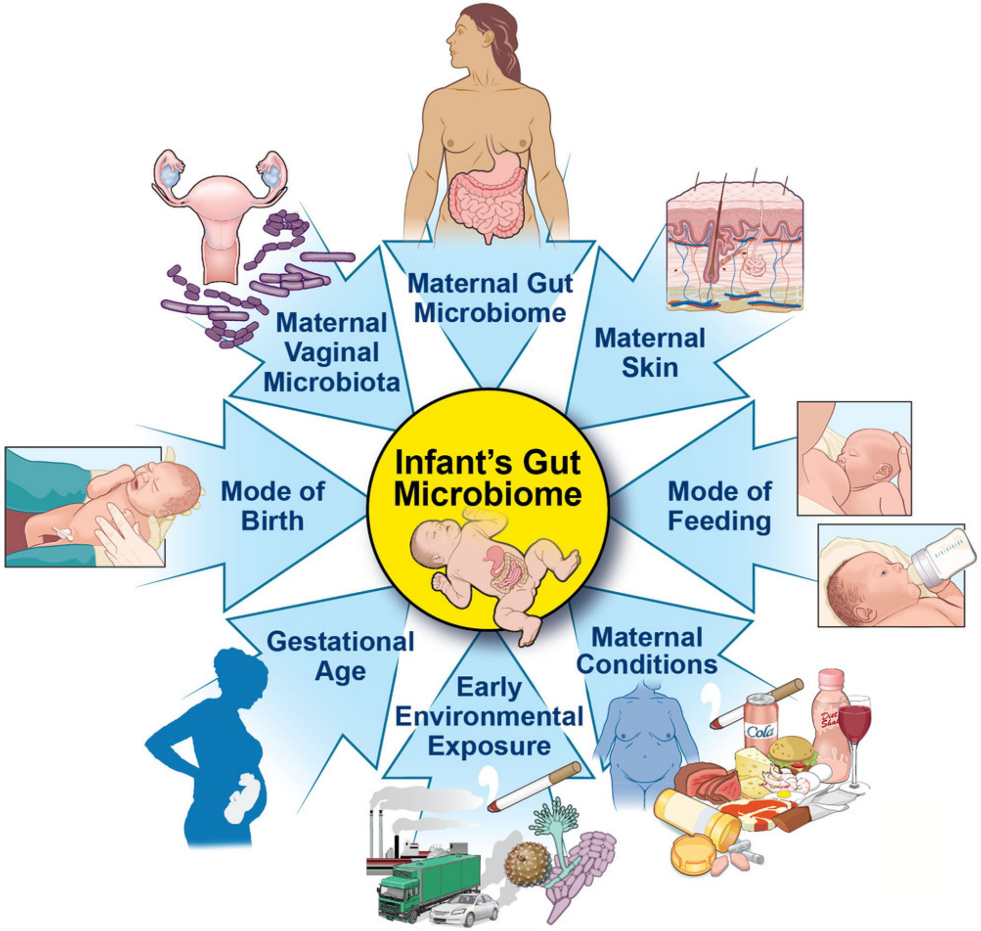


Figure: Kapourchali *et* al., *Nutrition in Clinical Practice* 2020

- Mode of feeding
  - Breastfeeding duration/ pattern
  - Age at weaning
- Environmental exposures
  - Use of antibiotics
  - Infections
- Geographical location
- Exposure to farm environment
- Host genetics
- Gender
- Etc.

Reduced or aberrant colonization

- Certain factors can lead to **reduced** or **aberrant** colonization of the infant gut
  - ⇒ which can result in significant defects or abnormalities in immune development

- For instance, **hygiene hypothesis** states that lack of infections during childhood in urbanized countries/cities due to overuse of antibiotics, changes in diet, socioeconomic status, higher hygiene levels, etc., may result in gut microbiomes that lack maturity and diversity for establishing a stable and homeostatic immune system

- Recent microbial studies have **linked** reduced diversity, aberrant colonization and compositional shifts during **infancy to illnesses that manifest during childhood or later in life**, including T1D, IBD, asthma, etc.
- Mechanisms of the disease pathogenesis remain largely elusive

Dysbiosis

- **After reaching an adult-like composition, certain factors can lead to dysbiosis**
- **Definition:** compositional and functional aberrations in the gut microbiome that is typically driven by pathobionts, loss of gut commensals, and/or loss of overall microbial diversity
- **Consequences:**
  - Increase local and systemic susceptibility to infections
  - compromise the bacteria-mediated immune regulations and induce chronic immune responses that may lead to inflammation and tissue damage
  - Compromise gut barrier that may lead to increased microbial translocation and gut permeability
  - has been linked to the several immune-mediated diseases, such as inflammatory bowel disease (IBD), asthma, type 1 diabetes, multiple sclerosis, antibiotic-resistant infection, etc. => mechanisms not well-known
- **Causes:**
  - Lifestyle: diet, stress, hygiene levels, etc.
  - Early colonization
  - Medicinal practices: vaccinations, antibiotics, drugs, etc.
  - Host genetics
  - Others

IMPORTANT MICROBIOME STUDIES

Some popular studies

- From 2005-2015, more than **USD 1.7 billion** has been spent on human microbiome research

- **Meta**genomes of the **H**uman **I**ntestinal **T**ract (metaHIT) *(2010)*
  - stool samples from 124 European “healthy” adults

- Chinese type 2 diabetes study *(2012)*
  - Stool samples from 145 adults (diabetic and non-diabetic)

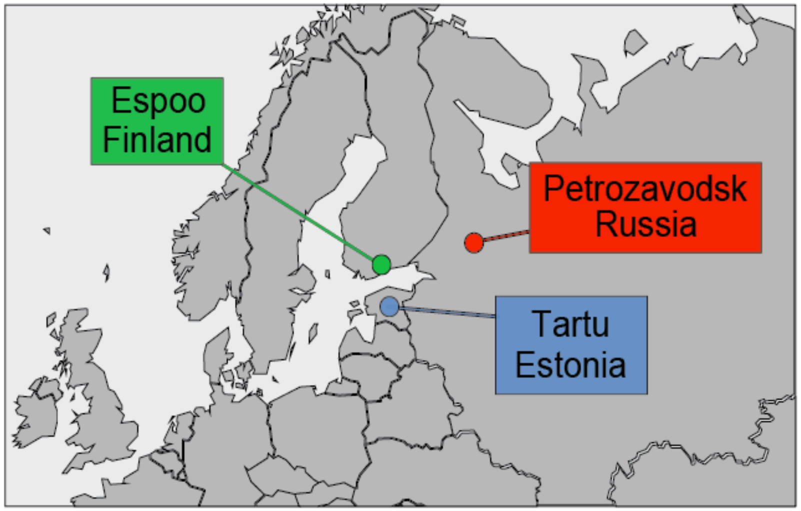
- **H**uman **M**icrobiome **P**roject (HMP) 1, 1-II, & integrative HMP *(2012, 2017, 2019)*

- DIABIMMUNE study *(2015, 2016, 2018, 2019)*

- Several other studies involving both human and other types of microbiome niches

DIABIMMUNE Study

- Was initiated to test **the hygiene hypothesis** in the **development of T1D**
- Follow developing infant gut microbiome in Finland, Estonia and Karelian Republic of Russia
  - ⇒ Considered as "**living laboratory**"
- 200-300 infants, genetically at risk for T1D
- Data collected:
  - Monthly stool samples were collected from birth until 3 years of age
  - Extensive metadata from infants and mothers



https://pubs.broadinstitute.org/diabimmune/

	INFANT INFORMATION	MATERNAL & PREGNANCY INFORMATION
GENERIC VARIABLES	birth weight HLA risk class gender mode of delivery country of residence study cohort	age at delivery gestational age in days gestational diabetes
COMPLEX VARIABLES	antibiotic treatments daycare attendance breastfeeding status (exclusive, non-exclusive or none) urban or rural dwelling of the family at infant's birth elder siblings height and weight disease status	illnesses during pregnancy height weight at the beginning and end of pregnancy antibiotic treatments during pregnancy

Human Microbiome Project

- **Largest body-wide survey of the human microbiome till date**

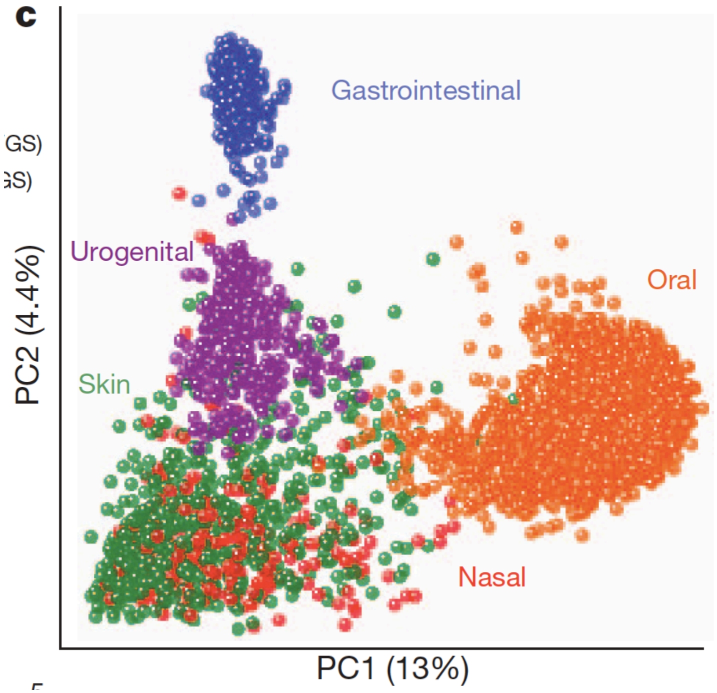
- **Goal:**
  - Create a toolbox of reference data, computational techniques, analytical methods and clinical protocols
  - To identify a **“core” set of microbial taxa** universally present in healthy individuals (lacking obvious disease phenotypes), such that the absence of such microbes would indicate dysbiosis (i.e. hunt for a picture of a “healthy” microbiome)

- Dysbiosis is difficult to define precisely. So, finding features that broadly distinguish healthy from unhealthy microbiomes will aid in the diagnosis of microbiome-related diseases and could provide new means of preventing disease onset or to improve prognosis

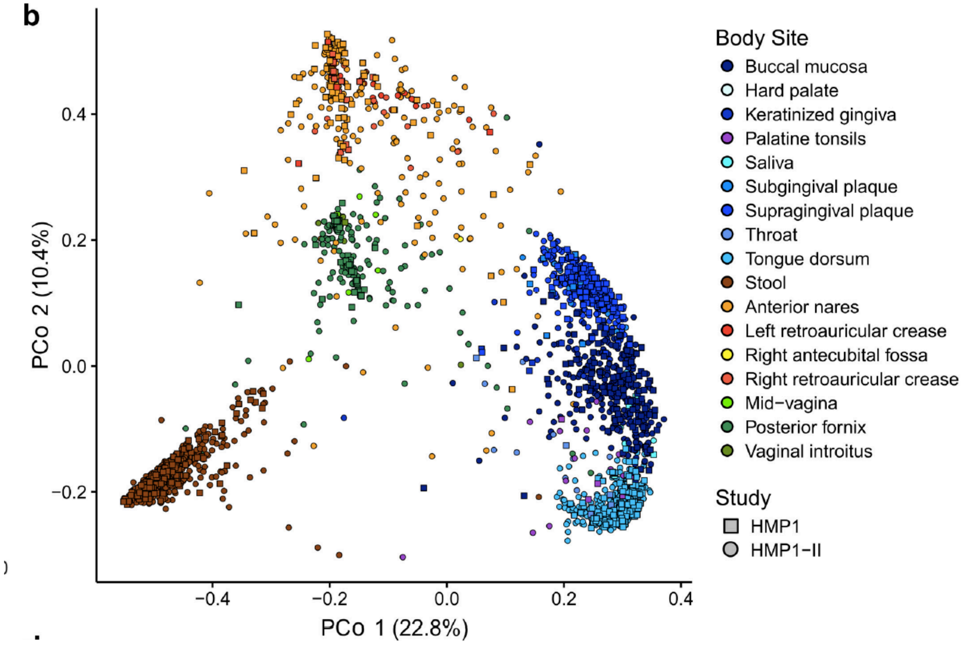
HMP phases

- **HMP 1 (2012):**
  - 242 adults (129 males, 113 females) from 2 distinct geographic locations in USA
  - Sampling at 18 body sites for women, 15 for men
    - 5 major body areas: oral, skin, stool, nares, and vagina
  - Samples from multiple visits
  - Clinical metadata
- **HMP 1-II (2017):**
  - 1631 new samples (for a total of 2355 samples)
  - 265 individuals
  - More longitudinal data
  - Focused on 6 out of 18 body sites
- **Integrative HMP (2019)**
  - Comprised studies of dynamic changes in the microbiome and the host under three conditions:
    - pregnancy and preterm , inflammatory bowel disease, and stressors that affect individuals with prediabetes.

Microbial community composition is more similar within than between habitats



The Human Microbiome Project Consortium, *Nature* 2012

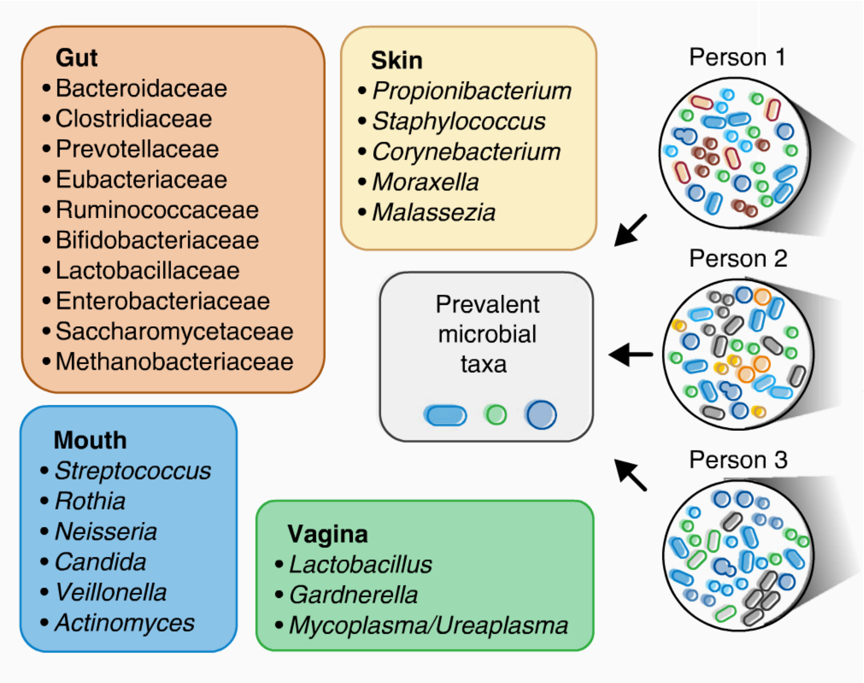


Lloyd-Price *et al.*, *Nature* 2017

Did they succeed in defining the constituents of a 'healthy' microbiome

- They did not succeed to find a **taxonomic composition** of the microbiome that would commonly appear in all healthy individuals
  - Between subject variations were very high
  - No taxa was observed to be universally present in all body habitats and individuals
- Characterizing a "healthy microbiome" as an ideal set of specific microbes is therefore no longer a practical definition

Early definition of 'healthy' microbiome



Lloyd-Price *et al.*, Genome Medicine 2016

- Each body-site habitat possesses strong enrichment of certain taxa over others
  - e.g. healthy gut microbiomes are consistently dominated by bacteria of 2 phyla: Bacteroidetes and Firmicutes
    - **Individuals vary by more than an order of magnitude in their Bacteroidetes/ Firmicutes ratio**
- **Less dominant taxa are highly personalized**, both among individuals and habitats

Healthy “functional core”

- **The abundance of metabolic pathways and other molecular functions is considerably more consistent across people for a given site**
  - allowing the identification of a healthy “**functional core**”, where the functions of a particular habitat are not necessarily provided by the same microbes in different people
- This core includes functions from at least 3 groups:
  1. Housekeeping functions: necessary for all microbial life
  2. Processes specific to human-associated microbomes across body-site habitats
  3. Specialized core functions

❖ **Hallmarks of a “healthy microbiome”:**

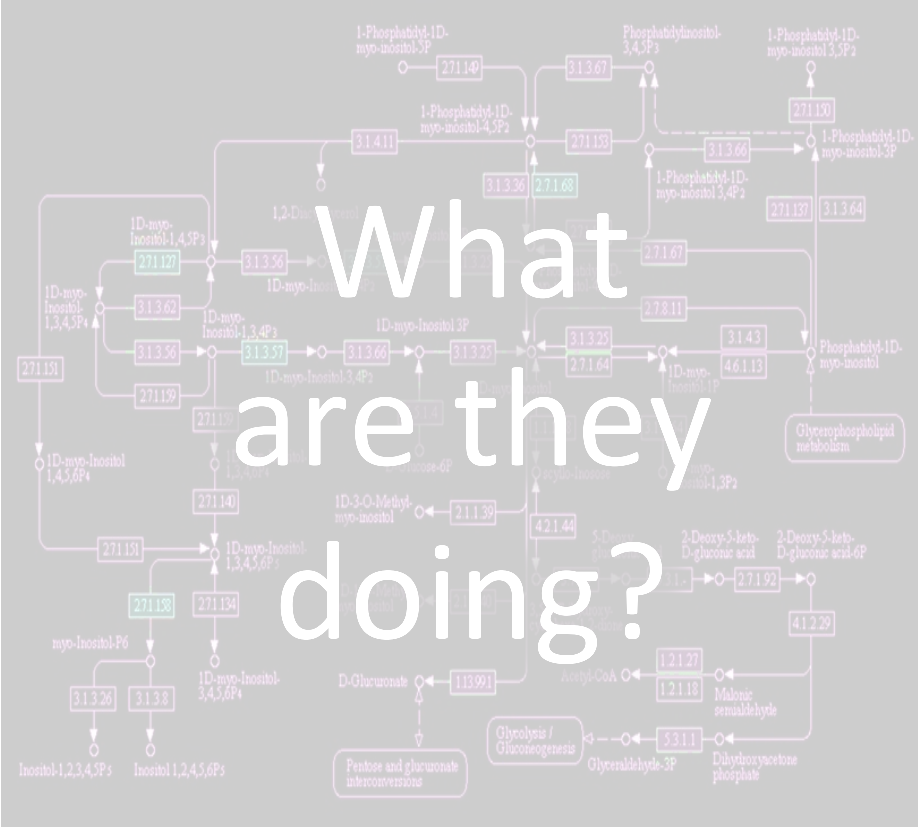
- The microbes at a particular body-site habitat is able to perform the core functions of that site
- It must have a degree of resilience to external (e.g. diet and drugs) or internal (e.g. age) changes

SEQUENCING TECHNOLOGIES

Two big questions of microbial community analysis

Who is there?

400
Fraction of universal proteins in each genome
Escherichia
Salmonella
Sulfolobus
Clostridium
Bacillus
Listeria
Streptococcus
Bacteroides
Prevotella
Chlamydophila
Mycobacterium
Helicobacter
Campylobacter



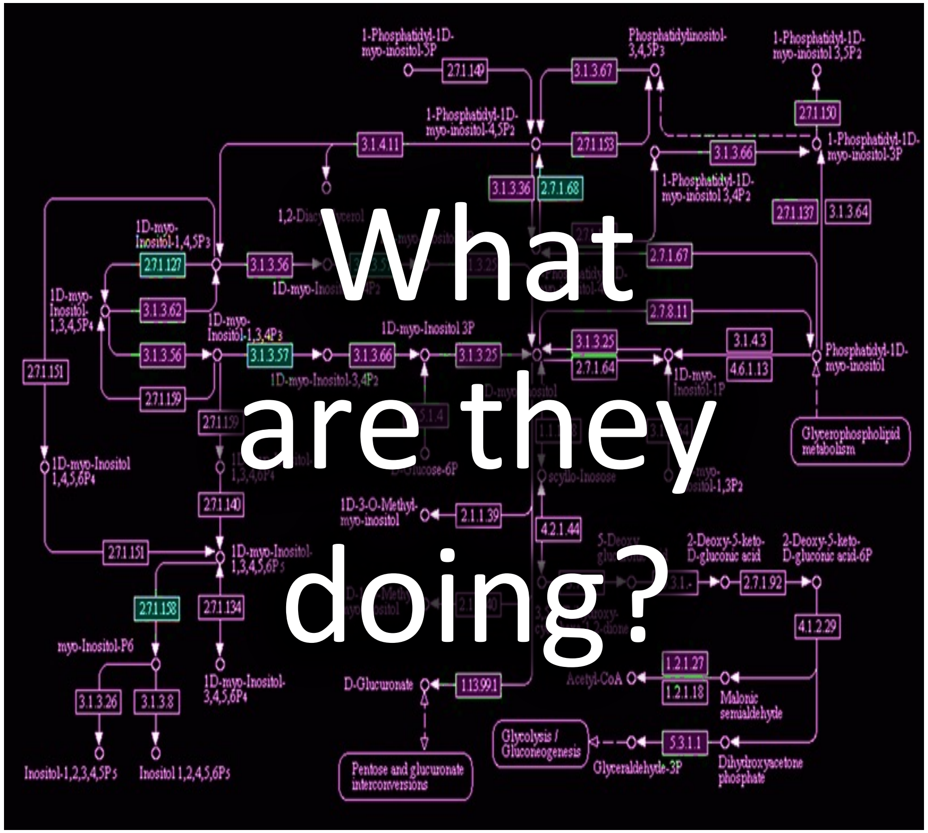
What are they doing?

1-Phosphatidyl-1D-myo-inositol-5P
Phosphatidylinositol-3,4,5P3
1-Phosphatidyl-1D-myo-inositol 3,5P2
1-Phosphatidyl-1D-myo-inositol-4,5P2
1-Phosphatidyl-1D-myo-inositol-3P
1-Phosphatidyl-1D-myo-inositol 3,4P2
Phosphatidyl-1D-myo-inositol
Glycerophospholipid metabolism
D-Glucuronate
Pentose and glucuronate interconversions
Glycolysis / Gluconeogenesis
Glyceraldehyde-3P
Dihydroxyacetone phosphate
Malonic semialdehyde
myo-Inositol-P6
Inositol-1,2,3,4,5P5
Inositol 1,2,4,5,6P5

Two big questions of microbial community analysis

Who is there?

400
Fraction of universal proteins in each genome
Escherichia
Salmonella



What are they doing?

How does one study the microbiome?

- Before popularization and affordability of high-throughput technologies, **culture-based approaches** were mostly used for identifying the microbes in a community
  - >99% of microbes cannot not be easily cultured, which generates a **biased view** of the microbiota

- With the advent of high-throughput technologies, **culture-independent approaches** were developed
  - No culturing; directly analyze the DNA extracted from microbial cells of a sample
  - Revolutionized microbiome studies, bringing about the "**golden age**" of microbial community analysis
  - Allows taxonomic and functional profiling of **entire communities** in an efficient and **unbiased** manner

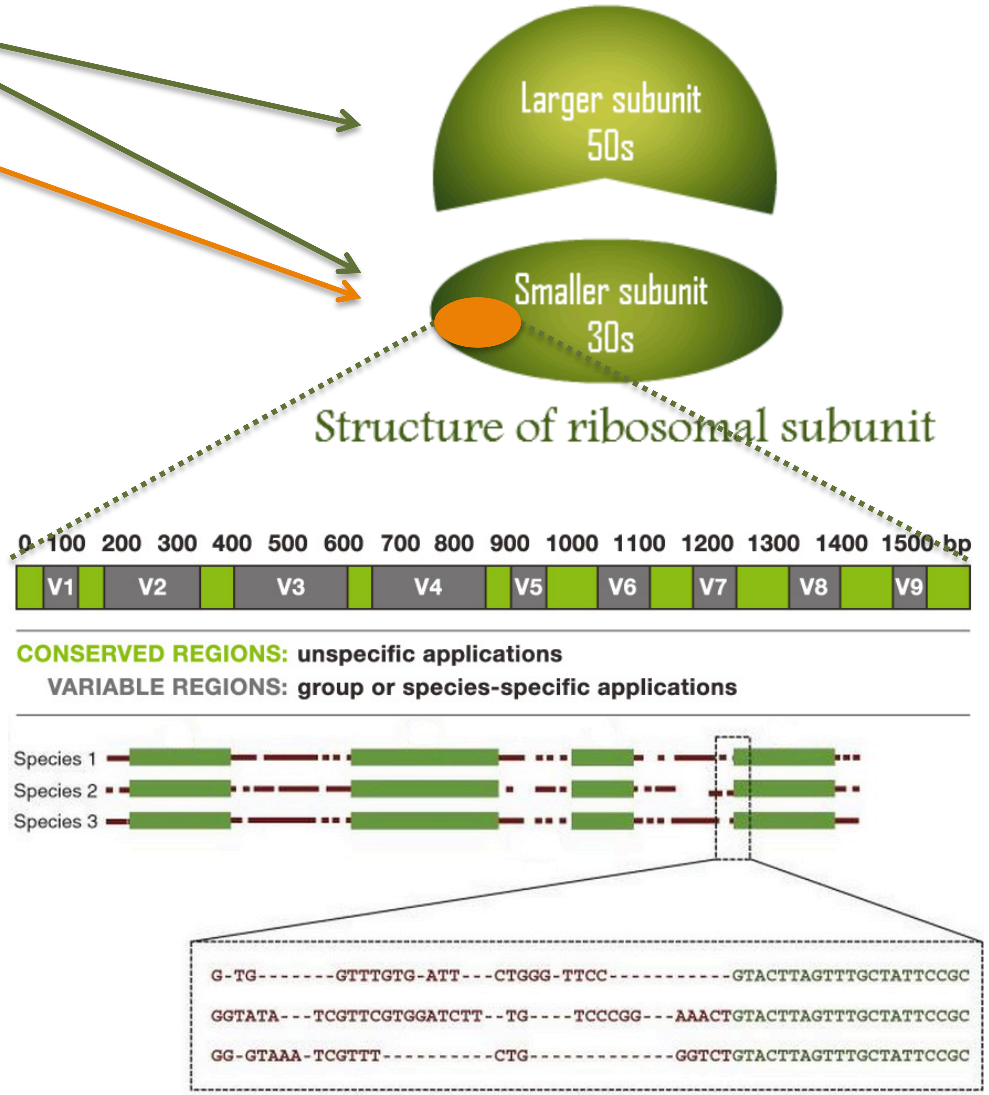
Culture-independent approaches

- 2 main next generation sequencing (NGS)-based methods are used:
  1. **Marker gene sequencing** (also known as amplicon or targeted sequencing):
     - ⇒ specific genes that are able to identify the entire genome are sequenced, i.e. marker genes
     - ⇒ These genes are such that they are:
       - present in almost all bacteria (or other microbe of interest)
       - Highly conserved (changes in sequence serve as an evolutionary clock and distance measure)
     - ⇒ **16S ribosomal RNA (rRNA) gene** is the most commonly used marker gene
     - ⇒ Relatively cheaper and faster, but can assign taxonomy only down to genus-level
  2. **Whole metagenome shotgun (WMS) sequencing** (also known as metagenomic sequencing)**:**
     - ⇒ All genomic DNA in a sample is sequenced
     - ⇒ Can reveal the microbial composition of communities and their genetic content
     - ⇒ More accurate and has better microbial resolution (species- and strain-level)

16S ribosomal RNA

- Prokaryotes: 70S ribosomes
- **Small subunit has a 16S ribosomal RNA**
- Segment of gene found in all bacteria
- Has high degree of conservation over time
  - Random sequence changes = accurate measure of evolution
- **9 hypervariable regions**
  - Each exhibits different degrees of sequence diversity and no single region can differentiate among all bacteria
  - V4 is the most popular choice for sequencing
- Long reads => Illumina sequencing is most popular

(454 pyrosequencing – old, 3rd generation sequencing platforms – Pacific Biosciences, Oxford Nanopore MinION and Ion Torrent)


Larger subunit
50s
Smaller subunit
30s
Structure of ribosomal subunit
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 bp
V1
V2
V3
V4
V5
V6
V7
V8
V9
CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications
Species 1
Species 2
Species 3
G-TG-------GTTTGTG-ATT---CTGGG-TTCC-----------GTACTTAGTTTGCTATTCCGC
GGTATA---TCGTTCGTGGATCTT--TG----TCCCGG---AAACTGTACTTAGTTTGCTATTCCGC
GG-GTAAA-TCGTTT----------CTG-------------GGTCTGTACTTAGTTTGCTATTCCGC

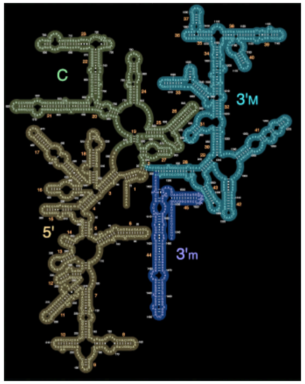
biology.tutorvista.com/animal-and-plant-cells/ribosomes.html
alimetrics.net/en/index.php/dna-sequence-analysis & Pereira *et al.*, *Nucleic Acids Research* 2010

Sequencing as a tool for microbial community analysis (marker gene vs WMS)

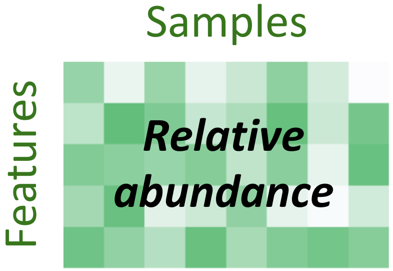
Lyse cells

Extract & fragment DNA

Sequence short DNA reads



- 16S (18S, ITS) rRNA gene
- (PCR amplification)



AGCTAGA CCGATCG

TTAGCAC ACTAGCA

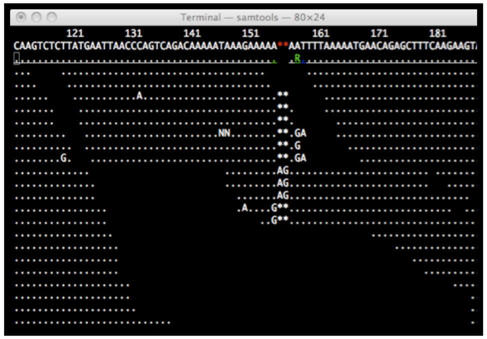
Assemble into contigs

AGCTACAGC

ACAGCACGGCAT

GGCATCATC

AGCTACAGCACGGCATCATC



```
Terminal — samtools — 80×24
           121       131       141       151       161       171       181
CAAGTCTCTTATGAATTAACCCAGTCAGACAAAAATAAAGAAAAA**AATTTTAAAAATGAACAGAGCTTTCAAGAGT
```

Map reads to reference genomes

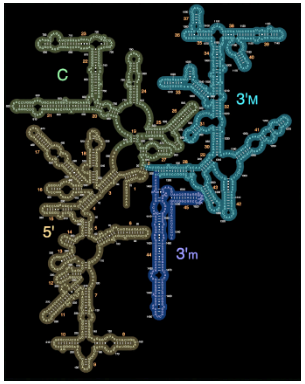
PROCESSING MICROBIAL SEQUENCING DATA & TAXONOMIC PROFILING

Profiling microbial communities by marker gene sequencing

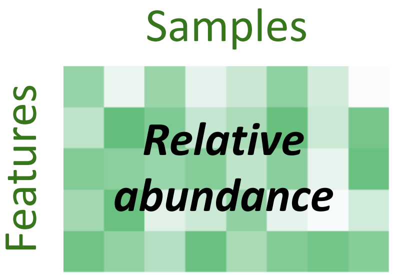
Lyse cells

Extract & fragment DNA

Sequence short DNA reads

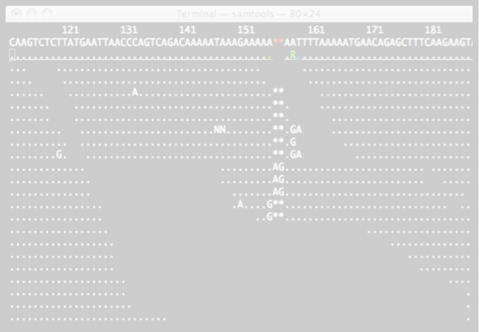


- 16S (18S, ITS) rRNA gene
- (PCR amplification)



AGCTAGA CCGATCG

TTAGCAC ACTAGCA



```
           121       131       141       151         161       171       181
CAAGTCTCTTATGAATTAACCCAGTCAGACAAAAATAAAGAAAAA**AATTTTAAAAATGAACAGAGCTTTCAAGAGT
```

Map reads to reference genomes

Assemble into contigs

```
AGCTACAGC
     ACAGCACGGCAT
              GGCATCATC
AGCTACAGCACGGCATCATC
```

16S rRNA Sequencing – data processing

1. QC analysis
   - Demultiplexing,
   - removal of sequencing artifacts, such as chimeras, low-quality reads, contaminating reads from host-genome, sequencing errors, etc.
   - Tools: FastQC, trimmomatic, cutadapt, **ea-utils** (toolkit)
2. Joining of paired-end reads by overlapping to obtain single reads
   - Tools: fastq-join, PEAR, SeqPrep, etc.
3. Clustering (or binning) of reads into **operational taxonomic units (OTUs)** = lowest level of phylotypes detectable by 16S rRNA sequencing
   - Based on predefined sequence similarity threshold (typically > 97%, which is considered to reflect genus-level classification)
   - Largely 3 categories of OTU clustering:
     - *de novo*, closed reference, and open reference (explained on next slide)
4. Consensus sequence per OTU is determined and taxonomically annotated using reference databases
5. 16S data => usually considered to be insufficient for functional analysis, but some tools do exist, such as PICRUSt and Tax4Fun

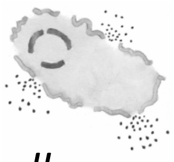
OTU clustering

- **De novo**:
  - Reads are aligned against one another without any reference sequence collection
  - OTUs are annotated using a reference database*
  - Tools: Mothur (agglomerative clustering method) => implemented in QIIME**; UPARSE
- **Closed Reference**:
  - Reads are clustered against a reference sequence collection and reads which do not match any reference sequence are discarded
  - UCLUST => implemented in QIIME
- **Open Reference**:
  - Reads are clustered against a reference sequence collection and any reads with no matches are clustered de novo
  - QIIME

* Reference databases that store annotated 16S rRNA sequences: GreenGenes, Ribosomal Database Project (RDP), SILVA, etc.

** QIIME = open-source bioinformatics software that integrates commonly used tools that are designed for 16S rRNA sequencing analyses

Profiling microbial communities by WMS sequencing



Lyse cells

Extract & fragment DNA

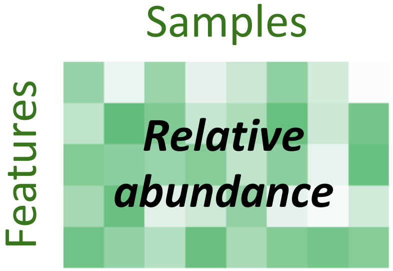
Sequence short DNA reads

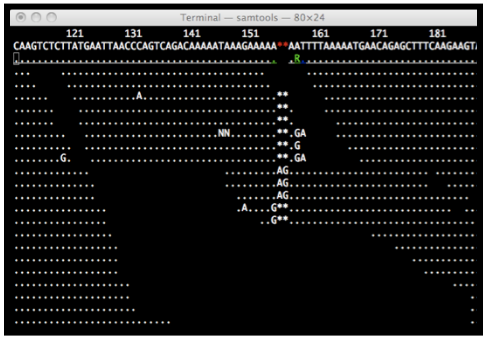
AGCTAGA   CCGATCG

TTAGCAC   ACTAGCA

Assemble into contigs

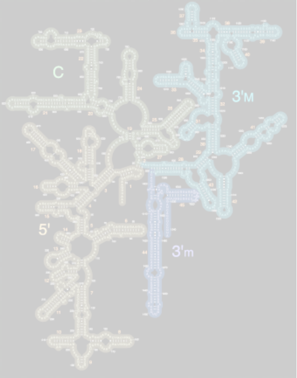
```
AGCTACAGC
    ACAGCACGGCAT
           GGCATCATC
AGCTACAGCACGGCATCATC
```





```
         121       131       141       151         161       171       181
CAAGTCTCTTATGAATTAACCCAGTCAGACAAAAATAAAGAAAAA**AATTTTAAAAATGAACAGAGCTTTCAAGAAGTA
.............................................  .R...............................
...     ....................................       .............................
....     ....................................       ............................
......    ...........A.......................**       ..........................
.......    ..................................**.      ..........................
.......    ..................................**.        ........................
..........  .......................NN........**.GA      ........................
........... .................................**.G         ......................
........G.   ................................**.GA          ....................
.............                        ........AG......................... .......
..............                       ........AG.........................   .....
..............                         ......AG................................
................                       .A...G**.............................  ..
................                          ..G**.................................
.................                                            ...................
..................                                               ...............
..................                                                  ............
..................                                                    ..........
....................                                                        ....
....................                                                           .
.....................                                                          .
...........................                                                    .
```

Map reads to reference genomes

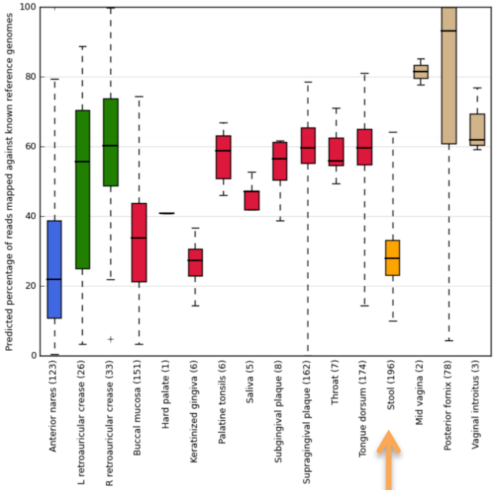


- 16S (18S, ITS) rRNA gene
- Conserved across bacteria
- (Allows PCR amplification)

WGS sequencing – data processing

1. QC analysis
   - Remove low-quality reads and adapters
   - Contaminating sequences from the host- genome is removed (by aligning reads to the host genome)
   - Wrapper tools: KneadData
2. Taxonomic and functional profiling − 2 different types of approaches
   - Assembly-free:
     - aligning short reads to reference genomes and gene catalogues, such as RefSeq, UniRef, etc.
     - Tools for taxonomic profiling : MetaPhlAn, MetaPhlAn2, mOTU, Kraken, MEGAN, etc.
     - Tools for functional profiling: DIAMOND, PALADIN, HUMAnN, HUMAnN2, etc.
   - assembly-based approaches:
     - short reads are assembled into longer sequences, called contigs, before profiling

WGS sequencing – data processing

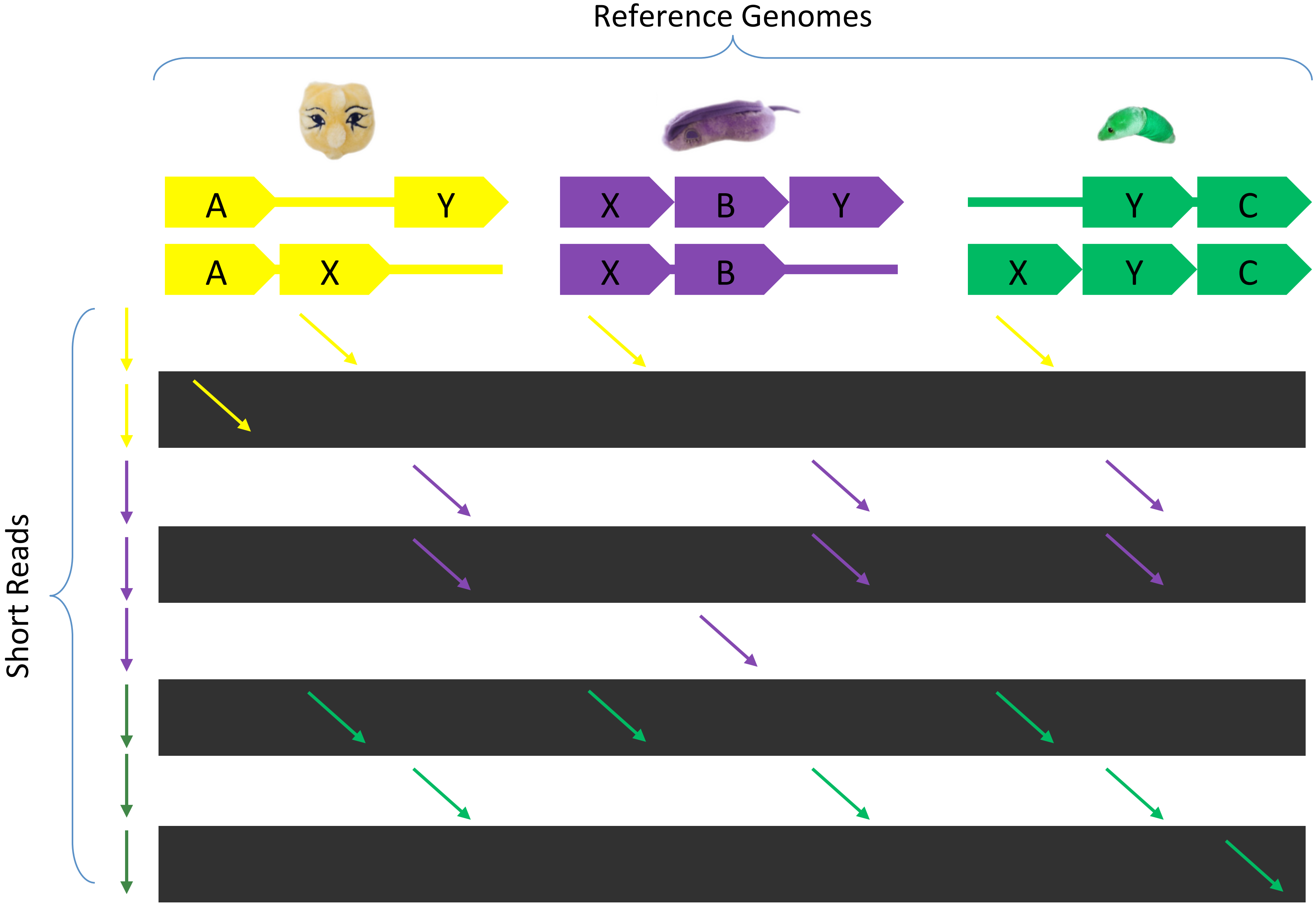


Supplementary Fig. 22. Predicted percentage of reads mapped against known reference genomes for the HMP/HMPII samples.

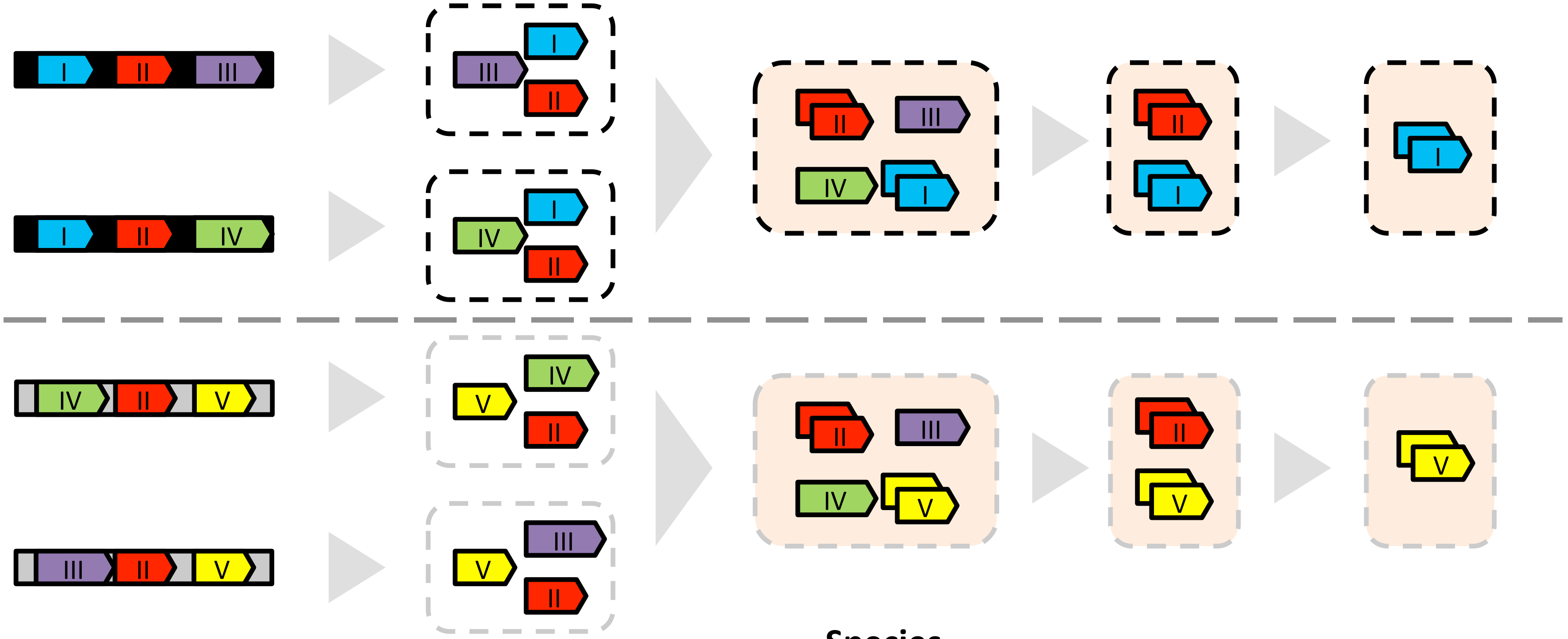
Human gut:
Mapping median
28%

- 50-85% of the species present in human gut microbiota lack reference genomes
  - Single species needs to be isolated and cultured to produce DNA-rich sample to assess the genome, but some species are impossible to cultivate and culture
  - Species of public health interest (*Salmonella enterica, E.Coli, C.Difficile, etc.*) are more represented in the databases, than commensal species
- Even then, reference-based WGS processing is most common
  - Metagenomics data is complex as it contains reads from multiple species
    - Assumptions made when assembling single genomes do not apply when assembling multiple genomes at varying levels of abundance

Assembly-free approach: Mapping Reads to the Genomes



MetaPhlAn: Indexing microbial pangenomes



NCBI isolate genomes

Archaea	300
Bacteria	12,926
Viruses	4,646
Eukaryota	2,177

“Bags” of protein coding genes

49.0 million total genes

Species pangenomes

7,677

containing 18.6 million gene clusters

Core genes

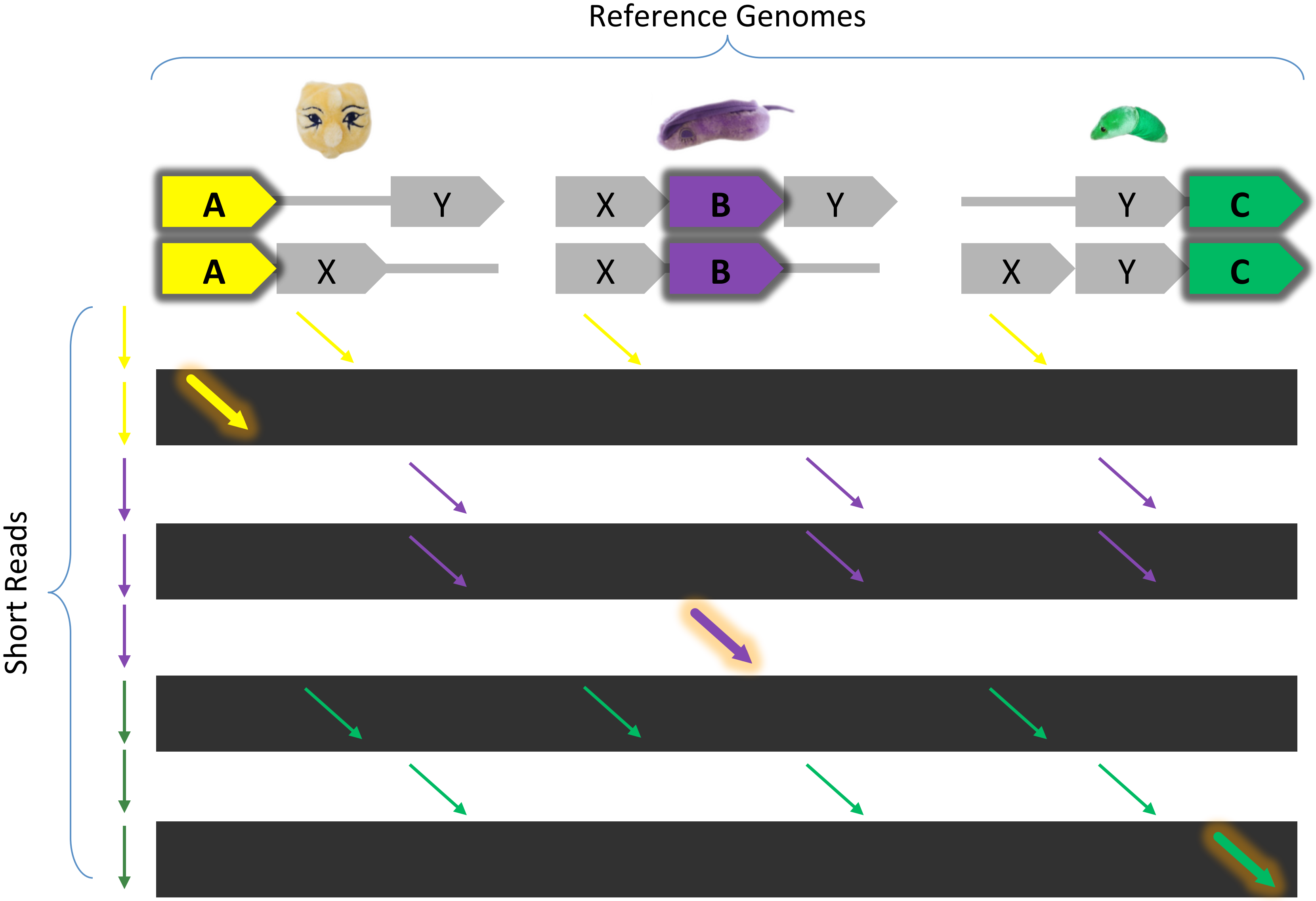
Marker genes

RepoPhlAn

ChocoPhlAn (http://metaref.org)

MetaPhlAn

Metagenomic Phylogenic Analysis



Assembly-based approaches

1. Short reads to contigs
   - In a comparative (using reference sequences) or ***de novo* manner**
   - *De novo* methods, expecially De Bruijn graph strategy, are most widely used
     - Tools: MEGAHIT, SOAP-denovo2, metaSPAdes, etc.
2. Contig binning
   - Each cluster of contigs represent a (partial) genome belonging to a biological taxon
   - Supervised (using reference genomes) or **unsupervised methods**
   - Unsupervised methods are more popular
     - Nucleotide composition-based, abundance-based or **hybrid** methods
     - Hybrid methods: CONCOCT, MaxBin2.0, etc.
3. Gene prediction
   - Open-reading frame predition
   - Tools: Prodigal, Glimmer, etc.
   - Non-redundant gene-catalogue can be built using tools like CD-HIT
4. Mapping short reads back to contig bins or gene catalogue to get abundances

❖ Contig binning

⇒ taxonomic profiling

❖ Gene prediction

⇒ functional profiling

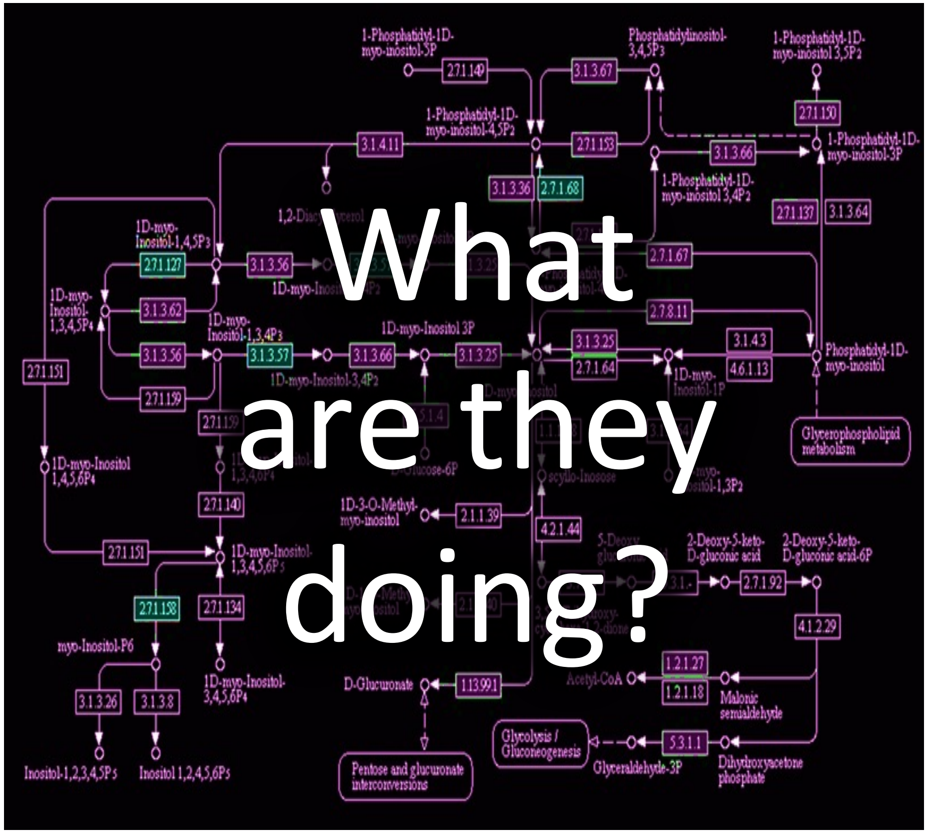
⇒ taxonomic profiling (using tools like MSPminer)

FUNCTIONAL ANALYSIS

Two big questions of microbial community analysis

400
Fraction of universal proteins in each genome

Who is there?


1-Phosphatidyl-1D-myo-inositol-5P
Phosphatidylinositol-3,4,5P3
1-Phosphatidyl-1D-myo-inositol 3,5P2
1-Phosphatidyl-1D-myo-inositol-4,5P2
1-Phosphatidyl-1D-myo-inositol-3P
1-Phosphatidyl-1D-myo-inositol 3,4P2
1D-myo-Inositol-1,4,5P3
1D-myo-Inositol-1,3,4,5P4
1D-myo-Inositol-1,3,4P3
1D-myo-Inositol 3P
1D-myo-Inositol-3,4P2
1D-myo-Inositol 1,4,5,6P4
1D-3-O-Methyl-myo-inositol
1D-myo-Inositol-1,3,4,5,6P5
1D-myo-Inositol-3,4,5,6P4
myo-Inositol-P6
Inositol-1,2,3,4,5P5
Inositol 1,2,4,5,6P5
D-Glucuronate
Pentose and glucuronate interconversions
Glycolysis / Gluconeogenesis
Glyceraldehyde-3P
Dihydroxyacetone phosphate
Malonic semialdehyde
Acetyl-CoA
2-Deoxy-5-keto-D-gluconic acid
2-Deoxy-5-keto-D-gluconic acid-6P
Phosphatidyl-1D-myo-inositol
Glycerophospholipid metabolism
1D-myo-Inositol-1P
scyllo-Inosose
1,2-Diacylglycerol
2.7.1.149
3.1.3.67
2.7.1.150
3.1.4.11
2.7.1.153
3.1.3.66
3.1.3.36
2.7.1.68
2.7.1.137
3.1.3.64
2.7.1.127
3.1.3.56
2.7.1.67
3.1.3.62
2.7.8.11
3.1.3.57
3.1.3.25
2.7.1.64
3.1.4.3
4.6.1.13
2.7.1.151
2.7.1.159
2.7.1.140
2.1.1.39
4.2.1.44
2.7.1.92
4.1.2.29
2.7.1.158
2.7.1.134
1.2.1.27
1.2.1.18
1.1.3.99.1
5.3.1.1
3.1.3.26
3.1.3.8

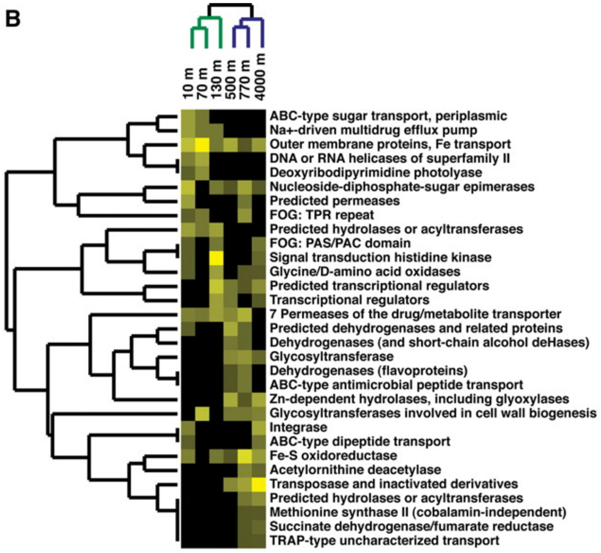
What are they doing?

Functional annotations of microbial genes

Orthology:
Grouping genes by conserved sequence features
COG, KO, FIGfam…

Structure:
Grouping genes by similar protein domains
Pfam, TIGRfam, SMART, EC…

Biological roles:
Grouping genes by pathway and process involvement
GO, KEGG, MetaCyc, SEED…



DeLong, 2006

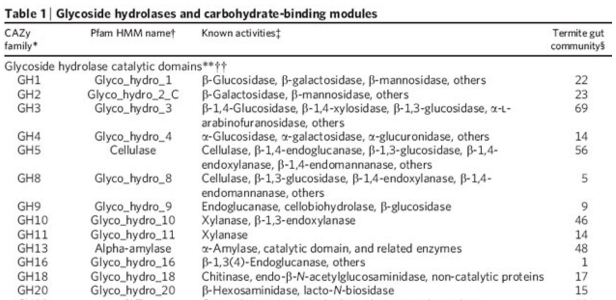
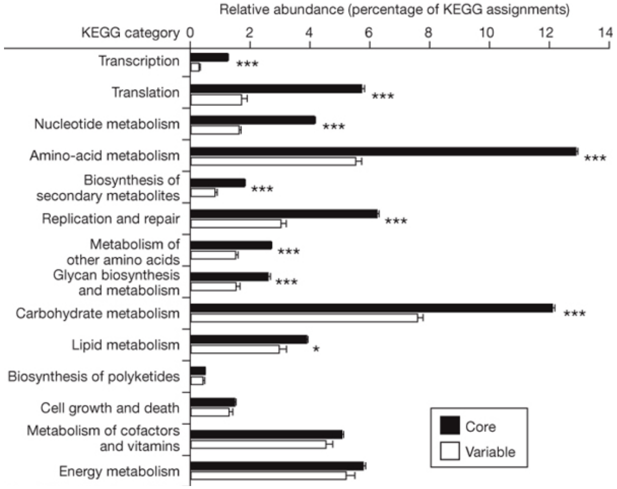


Table 1 | Glycoside hydrolases and carbohydrate-binding modules

CAZy family*	Pfam HMM name†	Known activities‡	Termite gut community§
Glycoside hydrolase catalytic domains**††			
GH1	Glyco_hydro_1	β-Glucosidase, β-galactosidase, β-mannosidase, others	22
GH2	Glyco_hydro_2_C	β-Galactosidase, β-mannosidase, others	23
GH3	Glyco_hydro_3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-L-arabinofuranosidase, others	69
GH4	Glyco_hydro_4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	14
GH5	Cellulase	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	56
GH8	Glyco_hydro_8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	5
GH9	Glyco_hydro_9	Endoglucanase, cellobiohydrolase, β-glucosidase	9
GH10	Glyco_hydro_10	Xylanase, β-1,3-endoxylanase	46
GH11	Glyco_hydro_11	Xylanase	14
GH13	Alpha-amylase	α-Amylase, catalytic domain, and related enzymes	48
GH16	Glyco_hydro_16	β-1,3(4)-Endoglucanase, others	1
GH18	Glyco_hydro_18	Chitinase, endo-β-*N*-acetylglucosaminidase, non-catalytic proteins	17
GH20	Glyco_hydro_20	β-Hexosaminidase, lacto-*N*-biosidase	15

Warnecke, 2007



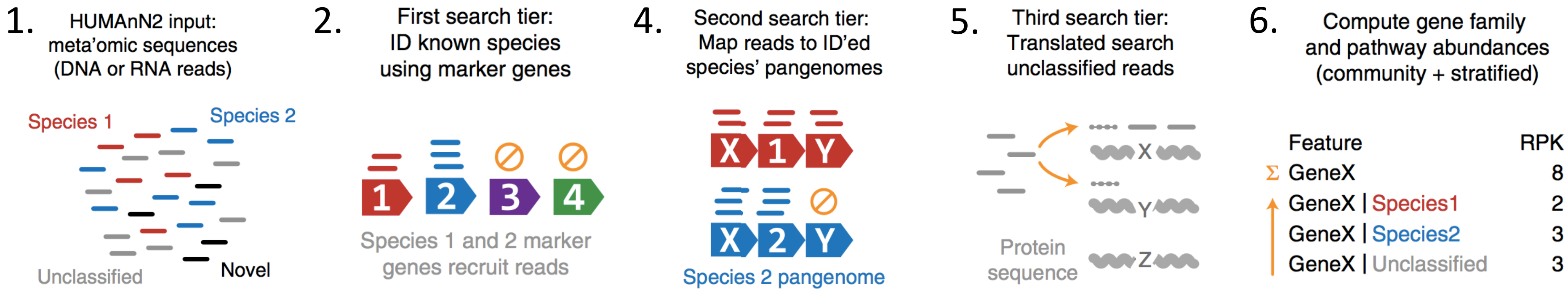
Turnbaugh, 2009

Functional profiling – HUMAnN2

Efficiently and accurately profiling the presence/absence and abundance of microbial pathways in a community from metagenomic or metatranscriptomic sequencing data

3. Constructs a sample-Specific gene-sequence database by merging preconstructed, functionally annotated pangenomes of the identified species

7. Abundance and coverage are then computed for each pathway using MinPath, producing stratified (per species) as well as community-wide values



1. HUMAnN2 input: meta'omic sequences (DNA or RNA reads)

2. First search tier: ID known species using marker genes

4. Second search tier: Map reads to ID'ed species' pangenomes

5. Third search tier: Translated search unclassified reads

6. Compute gene family and pathway abundances (community + stratified)

Feature	RPK
Σ GeneX	8
GeneX | Species1	2
GeneX | Species2	3
GeneX | Unclassified	3

Using MetaPhlAn2

Nucleotide-level mapping of all sample reads against the sample's pangenome Database. Yields: per-species, per-gene alignment statistics

Reads that do not align to identified species' pangenomes are subjected to translated search against a comprehensive protein database

Multiple alignment count is divided across aligned sequences. **Output**: A weighted count normalizated by alignable gene length

NORMALIZATION

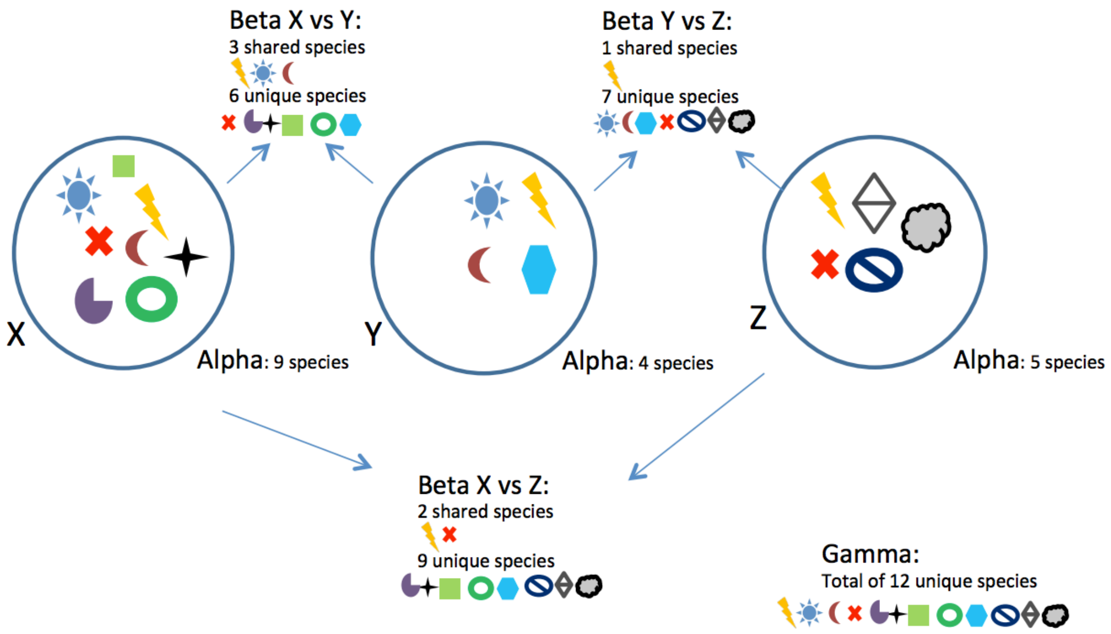
Normalization

- Needs to be performed on all taxonomic- and functional-level raw abundance data to make meaningful comparisons between samples or for other downstream analyses
- Microbiome data is compositional:
  - Human => 1 genome
  - Microbial community => unknown number of genomes
  - Absolute abundance cannot be inferred as biologists believe they cannot capture all species in a community
  - The data is transformed into compositional data, such as relative abundances
- **Total sum scaling (TSS)** − most popular:
  - Individual raw counts are divided by the total number of counts per sample
  - Results in relative abundances that sum to 1 (Simplex space where Euclidean metrics cannot be applied)
- Log-ratio transformation (proposed by Aitchison)
  - Additive, centered, and isometric log-ratio transformation
  - Results in compositional data also
  - In Euclidean space

DIVERSITY METRICS AND ORDINATION

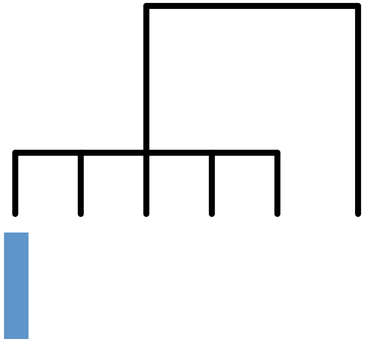
Diversity

- **Diversity**: a community’s number and distribution of organisms
  - Also **community composition** or **structure**
- **Alpha diversity** = refers to the diversity **within** a community or sample
- **Beta diversity** = refers to similarity/dissimilarity **between** two communities or samples
- **Gamma diversity** = refers to the total diversity in a landscape

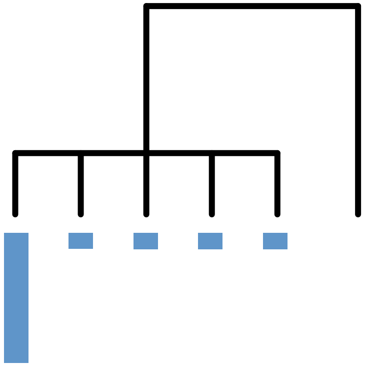


http://users.unimi.it/dmora/materiali/ANU2012/BioInfoTools.pdf

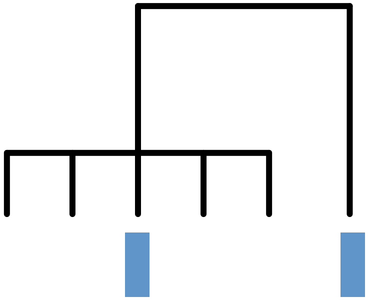
Alpha diversity (within-sample diversity)



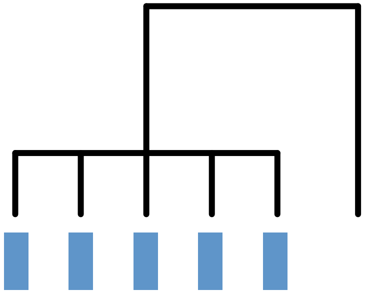
Not diverse



Qualitatively diverse
Taxonomically diverse



Phylogenetically diverse



Quantitatively diverse
Taxonomically diverse

Alpha diversity metrics

- Richness:
  - number of unique taxa = S_{obs}
  - Chao1: $S_{est} = S_{obs} + \frac{f_1^2}{2f_2}$

    f_1 is the number of singleton taxa (observed only once, one read) and f_2 is the number of doubleton taxa

- Evenness:
  - Simpson diversity index $= \sum_{i=1}^{n} p_i^2$
  - Shannon diversity index = $-\sum_{i=1}^{n} p_i \ln p_i$

n = total number of taxa in the sample

p_i is the relative abundance of taxon I

- Many other measures: McIntosh, Berger-Parker, …

```
Vegan::diversity() in R
```

Beta diversity metrics

- **Jaccard index**, proportion of shared taxa

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

- **Bray-Curtis dissimilarity**, shared abundance divided by total abundance

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

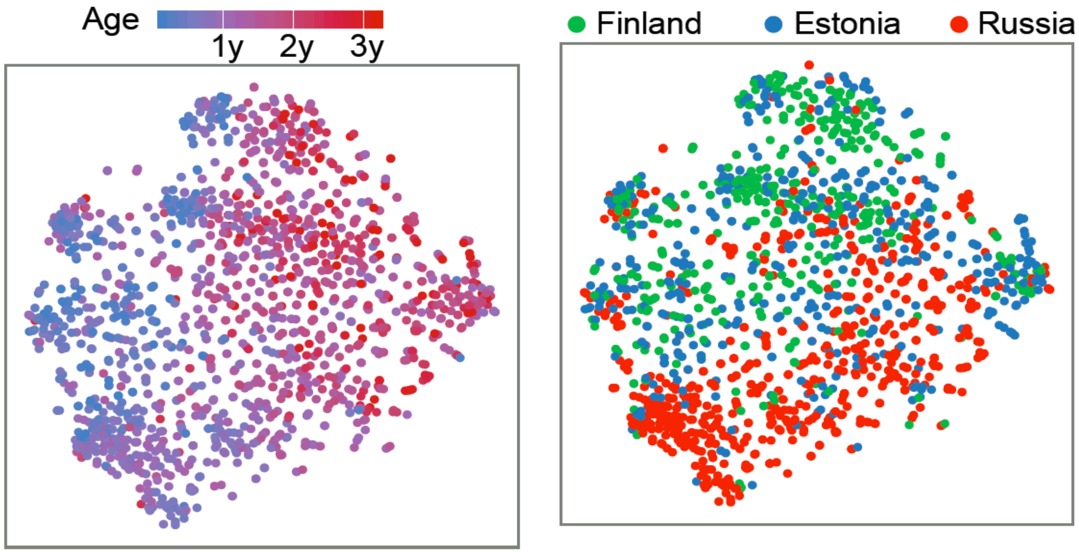
where C is the sum of the lesser values for only those species in common between both samples. S_i and S_j are the total number of species per sample.

`vegan::vegdist in R`

Ordination

- **Ordination** is a constrained projection of high-dimensional data into a lower dimensions
- Principal component analysis (PCA) guarantees the new dimensions to maximize normal variation => Euclidean metrics
- Principal coordinates analysis (PCoA), i.e. classical MDS, denotes to any ordination method based on (dis)similarity matrix => works with non-Euclidean metrics also
- t-SNE: Modern, distance / similarity matrix based technique for visualizing (high-dimensional) data

PCoA plots

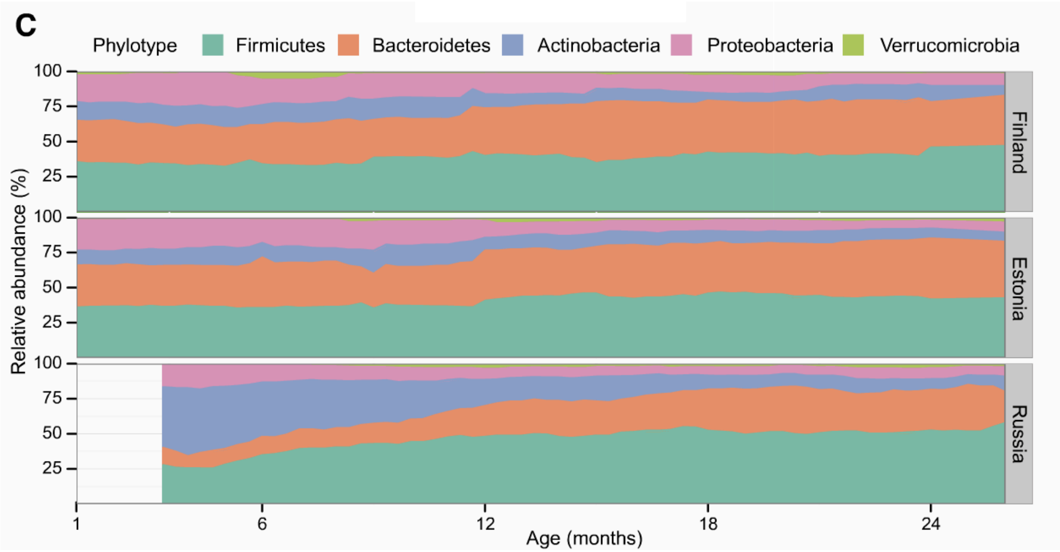


Vatanen *et al.*, *Cell* 2016

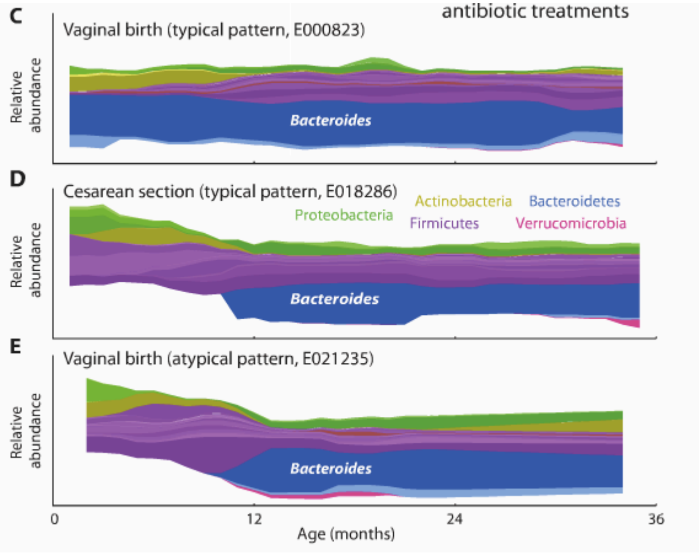
ASSOCIATION ANALYSES

Post-hoc testing of external factors

- Microbial composition is influenced by a variety of confounding factors /external factors
  - Scientists cannot control all possible influences
- We can try to explain the variations in the microbiome with the available metadata
- Some externals factors have been shown to influence the gut microbial compositions:
  - E.g. country and mode of delivery (DIABIMMUNE study)



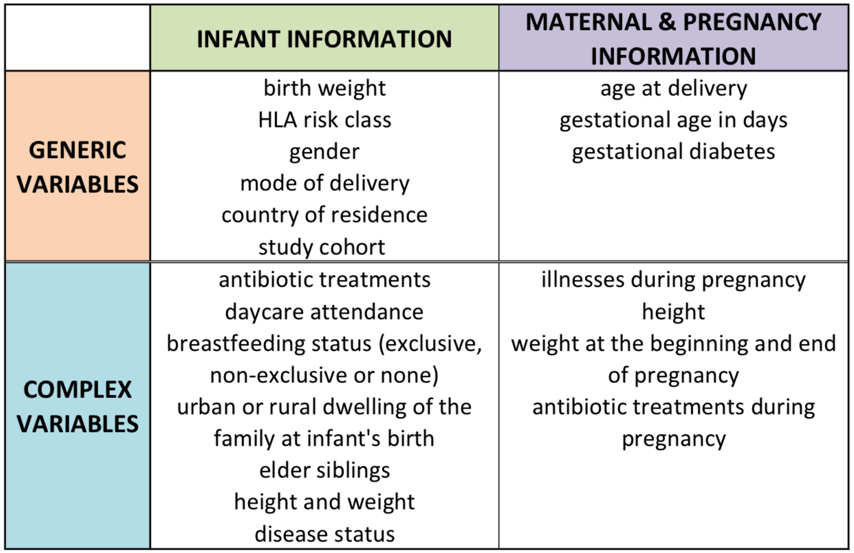
Vatanen *et al.*, *Cell* 2016



Yassour *et al.*, *Science Translational Medicine* 2016

Association Analyses

- Most big study cohorts collect clinical data along with microbiome samples
  - E.g. DIABIMMUNE study

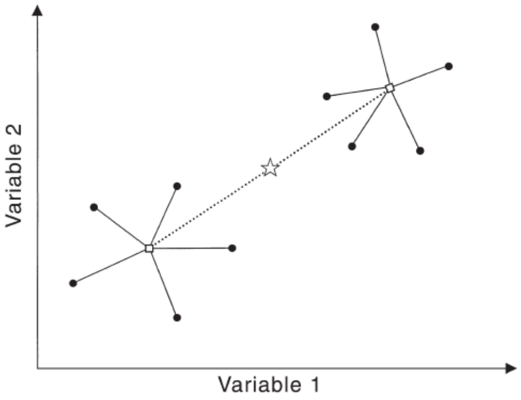


	INFANT INFORMATION	MATERNAL & PREGNANCY INFORMATION
GENERIC VARIABLES	birth weight HLA risk class gender mode of delivery country of residence study cohort	age at delivery gestational age in days gestational diabetes
COMPLEX VARIABLES	antibiotic treatments daycare attendance breastfeeding status (exclusive, non-exclusive or none) urban or rural dwelling of the family at infant's birth elder siblings height and weight disease status	illnesses during pregnancy height weight at the beginning and end of pregnancy antibiotic treatments during pregnancy

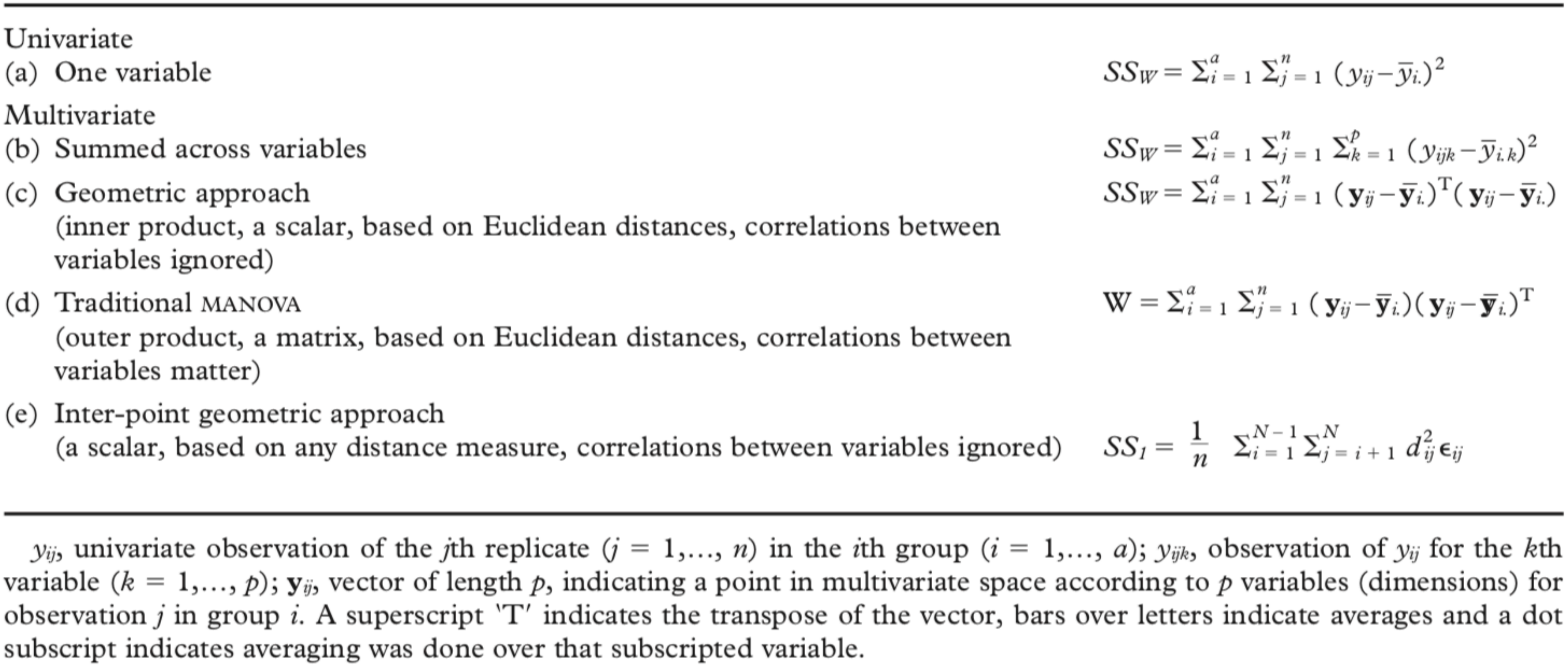
- Association analyses can uncover correlations correlations to diseases and other clinical metadata as a first step to innovation
- Usually a cocktail of external factors influence the microbiome
- 2 ways of association analyses:
  - Whole microbial composition – wise (i.e. multivariate association analysis)
  - each bacteria on a species/genus level – wise (i.e. univariate association analysis)

Composition Association Analysis

- **Aim: Identifying the differences in the microbial compositions of samples from different groups or treatments**
- Powerful multivariate statistical methods, such as MANOVA, use statistics that assume the data to be normally distributed
  - Not generally met by ecological data



Anderson, Austral Ecology (2001)



Univariate	
(a) One variable	$SS_W = \Sigma_{i=1}^{a} \Sigma_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$
Multivariate	
(b) Summed across variables	$SS_W = \Sigma_{i=1}^{a} \Sigma_{j=1}^{n} \Sigma_{k=1}^{p} (y_{ijk} - \bar{y}_{i.k})^2$
(c) Geometric approach (inner product, a scalar, based on Euclidean distances, correlations between variables ignored)	$SS_W = \Sigma_{i=1}^{a} \Sigma_{j=1}^{n} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})^{\mathrm{T}} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})$
(d) Traditional MANOVA (outer product, a matrix, based on Euclidean distances, correlations between variables matter)	$\mathbf{W} = \Sigma_{i=1}^{a} \Sigma_{j=1}^{n} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})^{\mathrm{T}}$
(e) Inter-point geometric approach (a scalar, based on any distance measure, correlations between variables ignored)	$SS_I = \frac{1}{n} \Sigma_{i=1}^{N-1} \Sigma_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij}$

y_{ij}, univariate observation of the jth replicate ($j = 1,\ldots, n$) in the ith group ($i = 1,\ldots, a$); y_{ijk}, observation of y_{ij} for the kth variable ($k = 1,\ldots, p$); $\mathbf{y}_{ij}$, vector of length p, indicating a point in multivariate space according to p variables (dimensions) for observation j in group i. A superscript 'T' indicates the transpose of the vector, bars over letters indicate averages and a dot subscript indicates averaging was done over that subscripted variable.

Non-parametric multivariate analysis of variance (PERMANOVA)

- This method circumvents the calculations of any distance measures and instead
  - **obtains an additive partitioning of sums of squares for any distance measure, without calculating the central locations of groups.**
  - **Can be used for Bray-Curtis similarity measure (non-Euclidean)**
- Calculates a permutation based p-value
- Able to cope with more complex multifactorial designs

`Vegan::adonis() in R`

A new method for non-parametric multivariate analysis of variance

MARTI J. ANDERSON
Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories A11, University of Sydney, New South Wales 2006, Australia

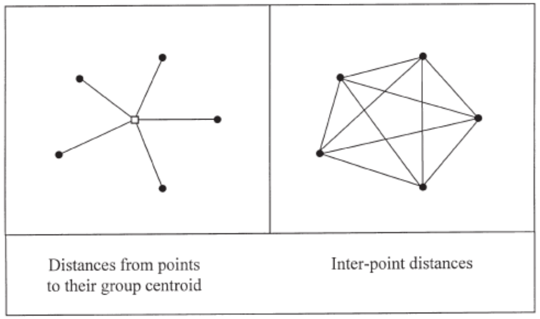


Fig. 2. The sum of squared distances from individual points to their centroid is equal to the sum of squared inter-point distances divided by the number of points.

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \qquad (1)$$

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij} \qquad (2)$$

$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)} \quad (SS_A = SS_T - SS_W) \qquad (3)$$

$$P = \frac{(\text{No. of } F^{\pi} \geq F)}{(\text{Total no. of } F^{\pi})} \qquad (4)$$

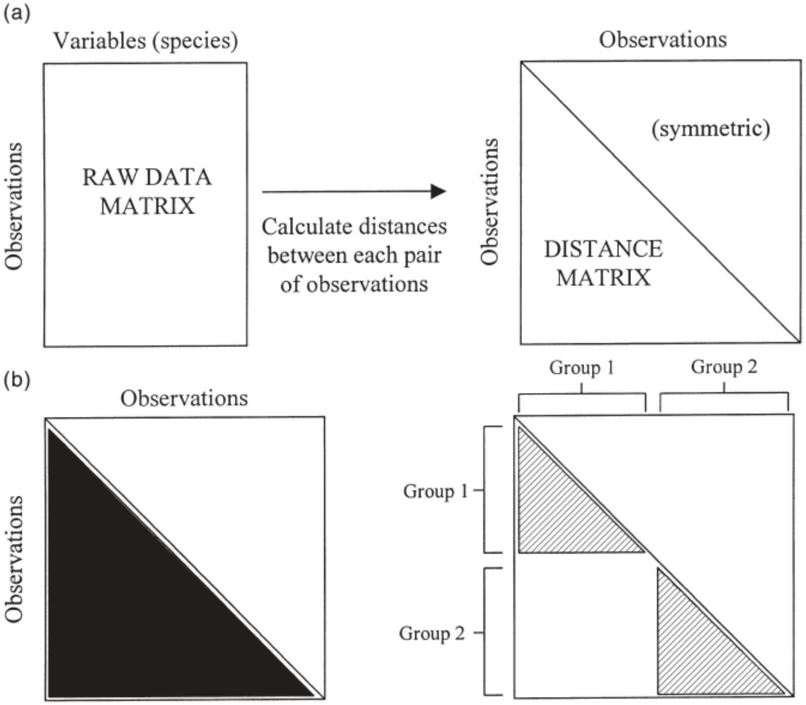


Fig. 3. Schematic diagram for the calculation of (a) a distance matrix from a raw data matrix and (b) a non-parametric MANOVA statistic for a one-way design (two groups) directly from the distance matrix. SS_T, sum of squared distances in the half matrix (■) divided by N (total number of observations); SS_W, sum of squared distances within groups (▨) divided by n (number of observations per group). $SS_A = SS_T - SS_W$ and $F = [SS_A/(a-1)]/[SS_W/(N-a)]$, where a = the number of groups.

$$SS_{W(A)} = \frac{1}{bn} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij}^{(A)} \qquad (5)$$

$$SS_{W(B)} = \frac{1}{an} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij}^{(B)} \qquad (6)$$

$$SS_R = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij}^{(AB)} \qquad (7)$$

$$SS_{AB} = SS_T - SS_A - SS_B - SS_R$$

(Univariate) Bacterial Association Analysis

- **Aim: Identify specific bacterial species or genus that are associated with particular covariates**

- **Linear model**

$$y_i = \beta_0 + \sum_p \beta_p X_{ip} + \varepsilon_i, \qquad \varepsilon_i \sim N(0,\sigma^2)$$

  y_i = observed quantities; relative abundances of one microbial taxon (e.g. species)

  i = 1, … ,n (samples)

  p = fixed effects/predictor (continuous or categorical)

- **Assumptions:**
  - Linearity
  - **Absence of collinearity** = fixed effects are not correlated
  - **Homoskedasticity** = variability of the data should be approximately equal across the range of predicted values
  - Normality of residuals
  - Absence of influential data points
  - **Independence** = each data point is from a different person!!!!

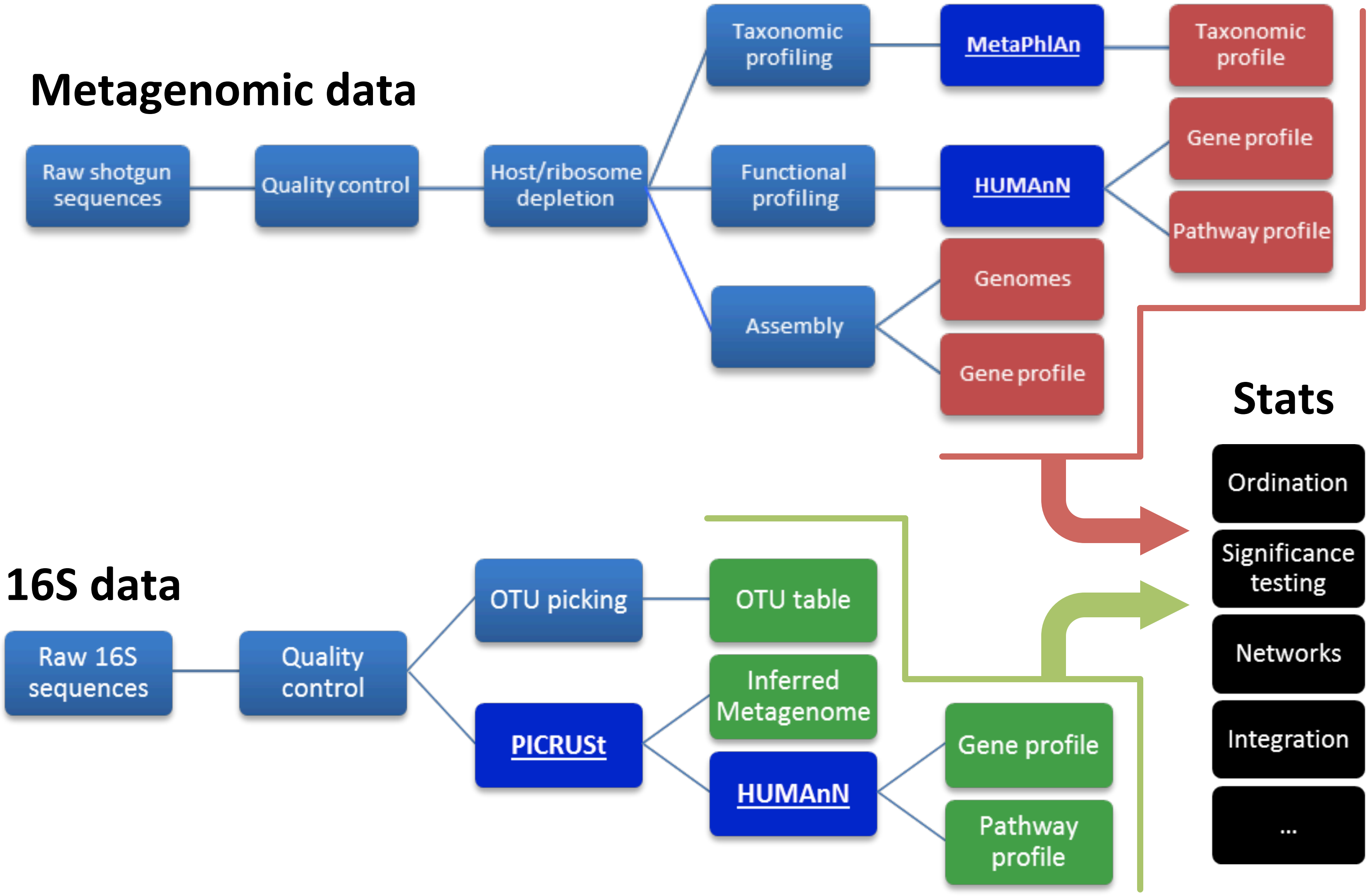
Linear Mixed Effect Model

- Independence assumption would not be met in studies where multiple samples are collected from the same subject as technical replicates or time-series data.
- The non-independence issues are resolved by adding another type of effect: **random effect**
- **Random effect =** covariate with non-systematic, idiosyncratic, unpredictable or “random” influence on the data
- **Fixed effect =** covariate with systematic and predictable influence on the data
- For e.g. if “subject” is the random effect, you model the data such that each subject has a different intercept (or baseline)

$$Y_i = \gamma_i Z_i + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2), \quad \gamma_i \sim N(0,\delta^2)$$

- LMM model can be applied on log-transformed relative abundances or using methods like MaAsLin, where arcsin of the square root of the relative abundances is taken.

Typical microbiome community analysis tasks



Emerging study areas

- **Metatranscriptomics**

  study of the **total transcribed RNA pool** of all organisms within a community

- **Metaproteomics**

  study of the **total proteome** of all organisms within a community

- **Meta-metabolomics** (or community metabolomics)

  study of the **total metabolite pool** of all organisms within a community

References

- Background Papers
  - Lloyd-Price, J. *et al.* 2016, ‘The healthy human microbiome’, *Genome Medicine*, 8:51.
  - Lynch, S. V. & Pedersen, O. 2016, ‘The Human Intestinal Microbiome in Health and Disease’, *The New England Journal of Medicine*, 375;24.
  - Morgan, X. C. & Huttenhower, C. 2012, ‘Chapter 12: Human Microbiome Analysis’, *PLOS Computational Biology*, vol. 8, issue 12.
  - Hamady, M. & Knight, R. 2009, ‘Microbial community profiling for human microbiome projects: Tools, techniques, and challenges’, *Genome Research*, vol. 19, pp.1141-1152.
  - Spor A. *et al.* 2011, ‘Unravelling the effects of the environment and host genotype on the gut microbiome’, *Nature Reviews Microbiology*, vol. 9, pp. 279-290.
  - Kapourchali *et al.* 2020, ‘Early-Life Gut Microbiome—The Importance of Maternal and Infant Factors in Its Establishment’, *Nutrition in Clinical Practice*, vol. 35, pp. 386-405.
  - Belkaid, Y. and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, **157**(1), 121–141
  - Liang, D., Leung, R. K.-K., Guan, W., and Au, W. W. (2018). Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut pathogens*, **10**(1), 3
  - Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., *et al.* (2012). Structure, function and diversity of the healthy human microbiome. *nature*, **486**(7402), 207.
  - Tibbs, T. N., Lopez, L. R., and Arthur, J. C. (2019). The influence of the microbiota on immune development, chronic inflammation, and cancer in the context of aging. *Microbial Cell*, **6**(8), 324.
  - Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research: a proposal.
  - Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, **164**(3), 337–340.
  - Levy, M., Kolodziejczyk, A. A., Thaiss, C. A., and Elinav, E. (2017). Dysbiosis and the immune system. *Nature Reviews Immunology*, **17**(4), 219.
  - And more…
- HMP Papers
  - The Human Microbiome Project Consortium, 2012, ‘Structure, function and diversity of the healthy human microbiome’, *Nature*, vol. 486, pp.207-214.
  - Lloyd-Price, J. *et al.* 2017, ‘Strains, functions and dynamics in the expanded Human Microbiome Project’, *Nature*, vol. 550, pp.61-66.
  - Gevers, D. *et al.* 2012, ‘The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome’, *PLOS Biology*, vol. 10, issue 8.
  - The Integrative HMP (iHMP) Research Network Consortium 2019,” The Integrative Human Microbiome Project , ” *Nature*, vol. 569, pp.641-648.
- DIABIMMUNE Papers
  - Vatanen, T.*et al.* 2016, ‘Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans’, *Cell*, vol. 165, pp.842-853.
  - Kostic, A. D. *et al.* 2015, ‘The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes’, *Cell Host & Microbe,* vol. 17(2), pp. 260-273.
  - Yassour, M. *et al.* 2016, ‘Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability’, *Science Translational Medicine*, vol. 8, issue 343.
  - Tommi Vatanen, Damian R. Plichta, Juhi Somani, *et al. 2019, ’*Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life’, *Nature Microbiology*, vol. 4, pp. 470-479.

References

- Tool papers
  - Correspondance 2010, ‘QIIME allows analysis of high-throughput community sequencing data’, *Nature Methods*, vol. 7, no. 5, pp.335-336.
  - Schloss, P. D. *et al.* 2009, ‘Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities’, *Applied and Environmental Microbiology*, pp. 7537-7541.
  - Segata, N *et al. 2012*, ‘Metagenomic microbial community profiling using unique clade-specific marker genes’, *Nature Methods*, vol. 9, no. 8, pp. 811-814.
  - Anderson, M. J. 2001, ‘A new method for non-parametric multivariate analysis of variance’, *Austral Ecology*, vol. 26, pp.32-46.
  - Dixon, P. 2003, ‘VEGAN, a package of R functions for community ecology’, *Journal of Vegetation Science*, vol. 14, issue 6, pp.927-930.
  - Andrews, S. (2010). Fastqc: A quality control for high throughput sequence data https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed: 31 july 2020).
  - Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**(15), 2114–2120.
  - Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**(1), 10–12.
  - Aronesty, E. (2013). Comparison of sequencing utility programs. *The open bioinformatics journal*, **7**(1).
  - Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics*, **30**(5), 614–620.
  - John, J. S. (2011). Seqprep: Tool for stripping adaptors and/or merging paired reads with overlap into single reads. *URL: https://githubcom/jstjohn/ SeqPrep*.
  - DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, **72**(7), 5069–5072.
  - Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, **41**(D1), D590–D596.
  - Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Thurber, R. L. V., Knight, R., *et al.* (2013). Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, **31**(9), 814–821.
  - Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4fun: predicting functional profiles from metagenomic 16s rrna data. *Bioinformatics*, **31**(17), 2882–2884.
  - Huttenhower (2020). Kneaddata: Tool designed to perform quality control on metagenomic sequencing data https: //huttenhower.sph.harvard.edu/ kneaddata/ (accessed: 20 september 2020).
  - O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith- White, B., Ako-Adjei, D., *et al.* (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**(D1), D733–D745.
  - Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6), 926– 932.
  - Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, **12**(10), 902– 903.
  - Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, **15**(3), 1–12.
  - Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, **12**(6), e1004957.

References

- Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, **12**(1), 59–60.
- Westbrook, A., Ramsdell, J., Schuelke, T., Normington, L., Bergeron, R. D., Thomas, W. K., and MacManes, M. D. (2017). Paladin: protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics*, **33**(10), 1473–1478.
- Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., *et al.* (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*, **8**(6), e1002358.
- Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., *et al.* (2018). Species-level functional profiling of metagenomes and metatran- scriptomes. *Nature methods*, **15**(11), 962–968.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, **31**(10), 1674–1676.
- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.* (2012). Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**(1), 2047–217X.
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaspades: a new versatile metagenomic assembler. *Genome research*, **27**(5), 824–834.
- Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods*, **11**(11), 1144–1146.
- Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**(4), 605–607.
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, **11**(1), 119.
- Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic acids research*, **40**(1), e9–e9.
  Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C., Gauthier, F., Magoulès, F., Ehrlich, S. D., and Pichaud, M. (2019). Mspminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, **35**(9), 1544–15

- Extra References
  - Imhann, F. *et al. 2016*, 'Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease', *Gut*, pp. 1-12.
  - Hall, A. B. *et al.* 2017, 'Human genetic variation and the gut microbiome in disease', *Nature Reviews*, vol. 18, pp. 690-699.
  - Noecker, C. *et al.* 2017, 'High-Resolution Characterization of the Human Microbiome', *Translational Research*, vol. 179, pp. 7-23.
  - Cho, I. & Blaser, M. J. 2012, 'The Human Microbiome: at the interface of health and disease', *Nature Review Genetics,* vol. 13(4), pp. 260-270.
  - Gilbert, J. A. *et al.* 2016, 'Microbiome-wide association studies link dynamic microbial consortia to disease', *Nature Review,* vol. 535, pp. 94-103.
  - Weiss, S. *et al.* 2017, 'Normalization and microbial differential abundance strategies depend upon data characteristics', *Microbiome*, 5:27.
  - Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, **8**, 2224.
- **DIABIMMUNE webpage:** https://diabimmune-17.ltdk.helsinki.fi/About%20DIABIMMUNE.html
- **Microbiome blog for interesting publications:** https://microbiomedigest.com/microbiome-papers-collection/microbiome-blogs-tweeps-and-books/
- Some pieces of text are based on the un-submitted doctoral dissertation of Juhi Somani (i.e. the presenter of this lecture)