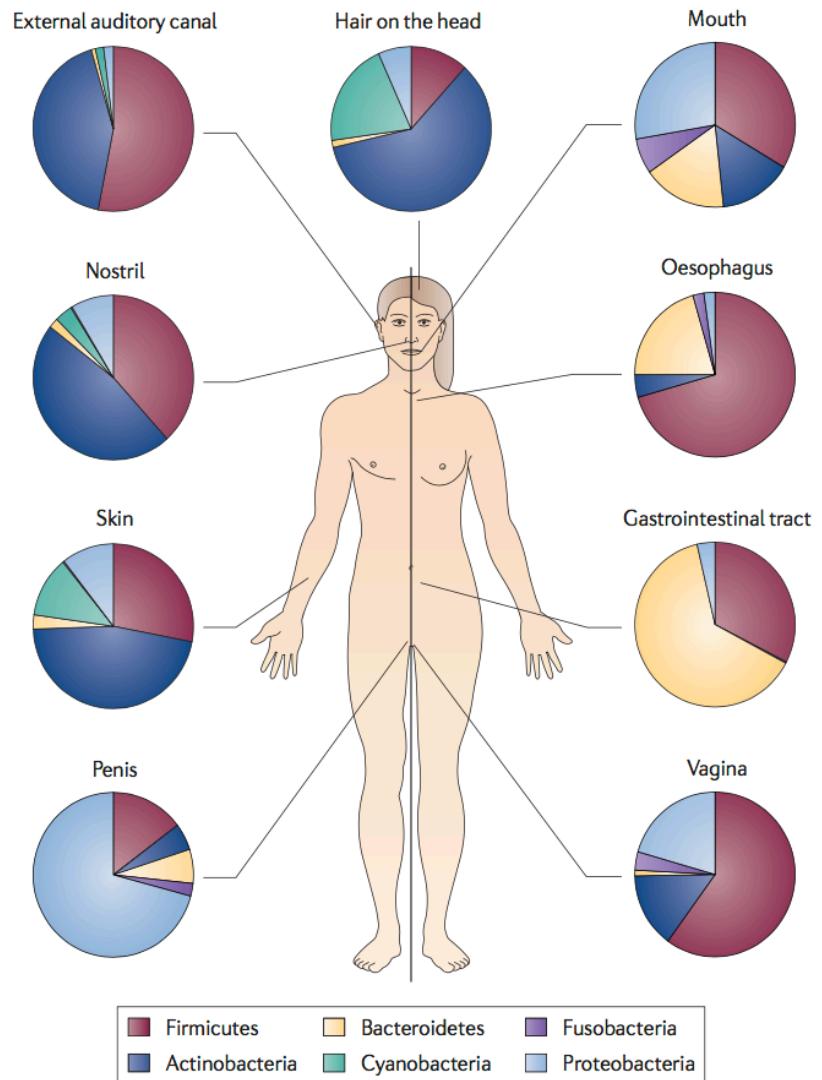

Microbiome Data Analysis

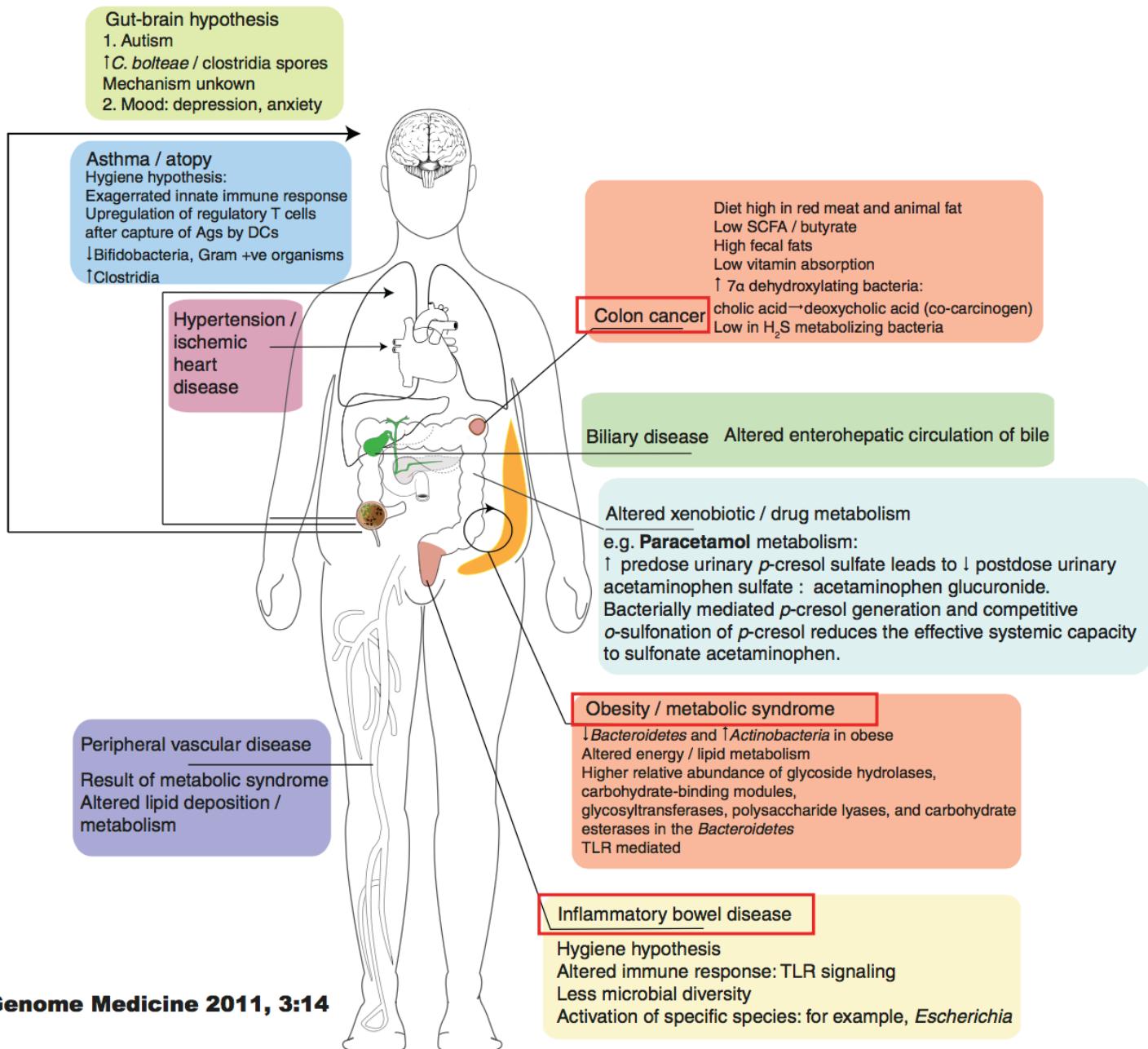
Yijuan Hu
2017-10-12

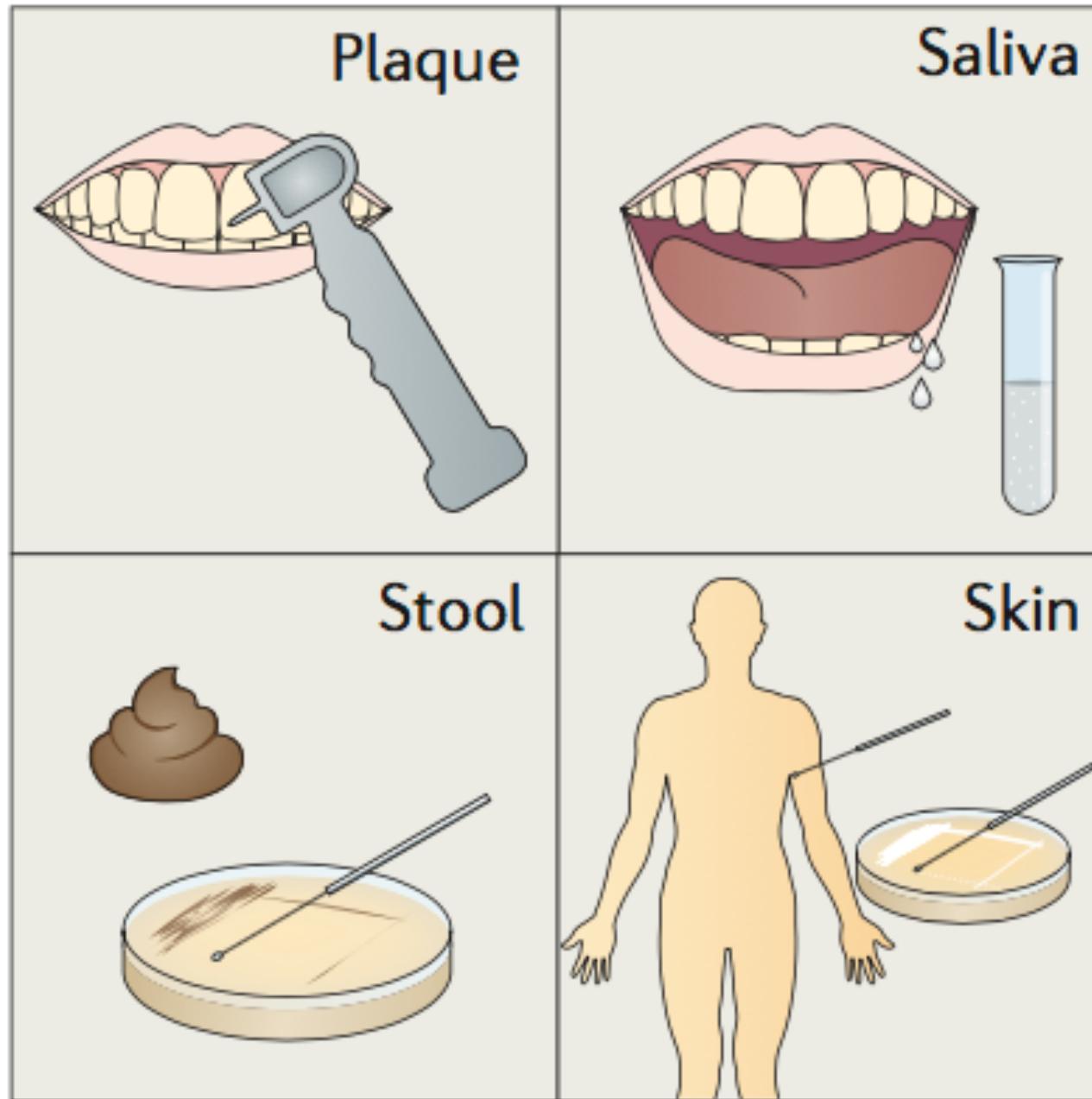
- Microbiome and Human Diseases
- Sequencing + Bioinformatics Pipelines
- Human Microbiome Project (HMP), Metagenomics of the Human Intestinal Tract (MetaHIT)
- Statistical Analyses
- Study Design and Power

- Microbiome as extended human genome
 - 10^{13} human cells vs 10^{14} bacterial cells
 - Consist of bacteria, fungi, and viruses
 - More than 3×10^6 genes provided by our gut microbiome
 - Distinctive microbiomes at different body sites
 - The human microbiome may explain the missing link between genetic variation and disease
- The human microbiome in health
 - Digestive enzyme activity
 - Synthesis of vitamins
 - Interaction with the immune system
 - Protection from pathogens, etc.

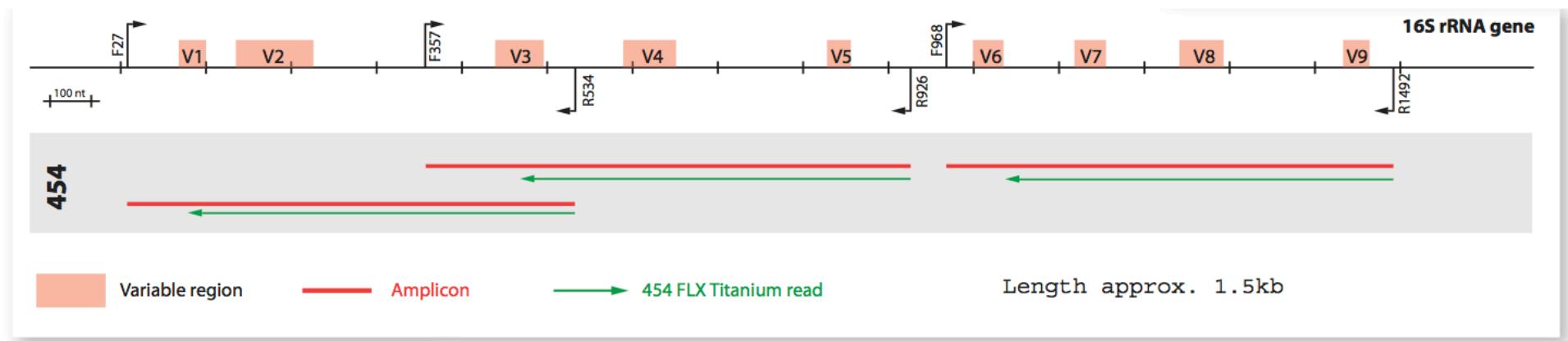


Nat Rev Microbiol. 2011 Apr;9(4):279-90.



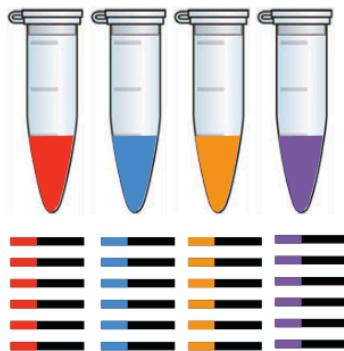


- 16S rRNA gene targeted sequencing
 - Specific to bacteria, not in fungi and viruses
 - Omnipresent in bacteria
 - Some regions are constant, allowing amplification
 - Some regions are variable, allowing identification of a particular genus and species
 - Reveals “who is there” in terms of relative abundances of bacterial taxa

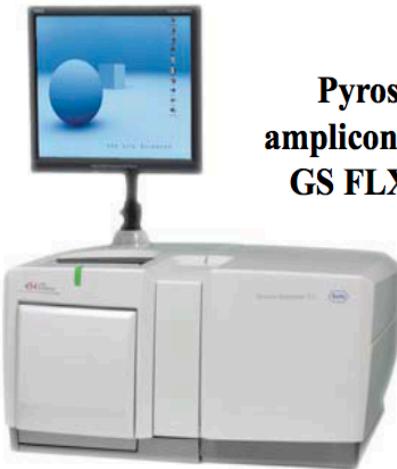


- Metagenomic (whole-genome) shotgun sequencing
 - The total extracted DNA is fragmented and sequenced
 - Reveals “what can they do” in terms of the encoded functions of the sequenced microbial DNA
 - 20–30 times more expensive than 16S rRNA gene sequencing, as well as requiring additional computational costs and high-level expertise for performing metagenomic analyses
 - We focus on 16S rRNA gene sequencing data here

Extract DNA and
PCR amplify with
barcoded primer



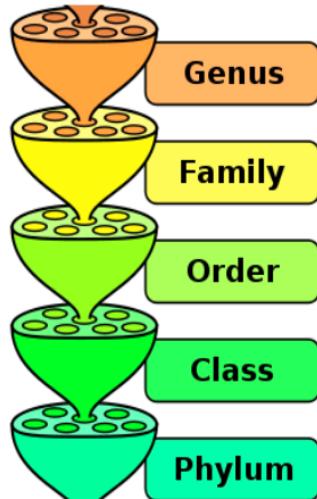
Pool amplicons



Pyrosequence
amplicons using 454's
GS FLX instrument

Quality control, assign
sequences to samples
using barcode and denoise

Species level OTUs can be
summarized into higher levels



Assign lineages by
RDP classifier

Thousands of OTUs,
100's of samples

OTU1
OTU2
OTU3

OTUq

>AGTGAGAGAAGCAGGGTCGTAATGTT ...
>AGTGCATGCGTAGGGTCGTAATGCG ...

>AGTGCATGCGTAGGGTCGTAATGTA ...
>AGTGGATGCTCTAGGGTCGTAATGCA ...

>AGTGTACGGTGAGGGTCGTAATGGG ...
>AGTGGATGCTCTAGGGTCGTAATGTT ...

>AGTGTACGGTGAGGGTCGTAATGCC ...
>AGTGAGAGAAGCAGGGTCGTAATCAC ...

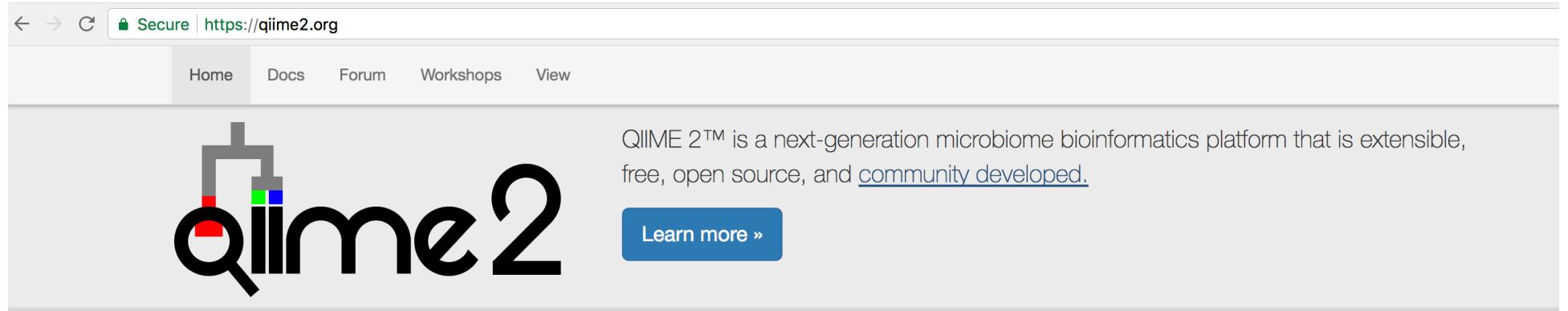
OTU1

OTU2

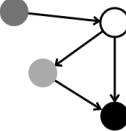
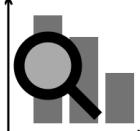
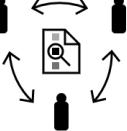
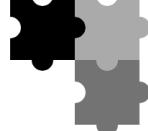
OTU3

Cluster sequences into OTUs (97% level),
align to reference alignment (e.g., using
PYNAST), infer phylogeny (FastTree)

OTU: Operational Taxonomic Unit



The screenshot shows the QIIME 2 website at <https://qiime2.org>. The header includes navigation links for Home, Docs, Forum, Workshops, and View. The main content features the QIIME 2 logo and a brief description: "QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed." A "Learn more »" button is present. Below the main section are four cards with icons and descriptions:

-  Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!
-  Interactively explore your data with beautiful visualizations that provide new perspectives.
-  Easily share results with your team, even those members without QIIME 2 installed.
-  Plugin-based system — your favorite microbiome methods all in one place.

OTU Table

— 9/25 —

#	OTU ID	A1	A2	B1	B2	C1	C2	D1	D2	ConsensusLineage
	denovo0	1	0	0	0	0	0	0	0	k_Bacteria
	denovo1	0	1	0	0	0	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Ruminococcaceae; g_Oscillospira; s_
	denovo2	1	0	1	0	0	1	0	0	k_Bacteria; p_Bacteroidetes; c_Bacteroidia; o_Bacteroidales; f_Bacteroidaceae; g_Bacteroides
	denovo3	0	0	0	0	0	2	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Veillonellaceae; g_Dialister; s_
	denovo4	0	1	0	0	0	0	0	0	k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae; g_Streptococcus
	denovo5	2	0	0	0	0	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Ruminococcaceae; g_Oscillospira; s_
	denovo6	0	0	0	0	1	1	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Ruminococcaceae
	denovo7	0	0	0	0	3	1	10	11	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Lachnospiraceae; g_ ; s_
	denovo8	1	7	0	0	0	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Lachnospiraceae; g_Blautia; s_
	denovo9	0	0	0	1	0	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Ruminococcaceae
	denovo10	1	0	0	2	0	1	1	0	k_Bacteria; p_Proteobacteria; c_Deltaproteobacteria; o_Desulfovibrionales; f_Desulfovibrionaceae; g_ ; s_
	denovo11	0	0	0	0	0	0	0	3	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_[Tissierellaceae]; g_Finegoldia; s_
	denovo12	0	0	0	0	0	0	0	1	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales
	denovo13	0	0	0	0	0	1	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Lachnospiraceae
	denovo14	12	13	6	13	121	58	1	12	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Veillonellaceae; g_Dialister; s_
	denovo15	30	16	0	0	0	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Lachnospiraceae
	denovo16	0	0	0	1	0	0	0	0	k_Bacteria; p_Firmicutes; c_Bacilli
	denovo17	8	4	0	3	1	0	1	2	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales
	denovo18	0	0	1	0	0	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales
	denovo19	0	0	0	0	1	0	0	0	k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales

>denovo0 A1_21775

TACGTAGGTGGCAAGCGTTGTCGGAATTACTGGGTGAAAGGGAGCGCAGGCAGGAGATCAAGTCGGCTGTGACAACACTACAGGCTTAACCTGTAGACTGCGGTGAAACTGGTTCTTGAGTGAAGTATAGG

The screenshot shows the official website for the NIH Human Microbiome Project. The header includes a 'Secure' lock icon and the URL <https://hmpdacc.org>. The main title 'NIH Human Microbiome Project' is displayed prominently. Below the title are two main sections: 'HMP1' and 'iHMP'. Each section features a circular logo (the Human Microbiome Project logo and a DNA helix respectively) and a brief description of the project's scope. At the bottom of each section is a call-to-action button.

HMP1

Characterization of the microbiomes of healthy human subjects at five major body sites, using 16S and metagenomic shotgun sequencing.

iHMP

Characterization of microbiome and human host from three cohorts of microbiome-associated conditions, using multiple 'omics technologies.

National Institutes of Health (NIH) Common Fund supported

- Phase I (HMP1): established in 2008
- Phase II (iHMP): ongoing

Metagenomics of the Human Intestinal Tract (MetaHIT) — 11/25 —

www.metahit.eu



Metagenomics of the Human Intestinal Tract

Home | Paris 2012 | Live News | Project | WPs | Our Team | Publications | Conf 2010 | Media | Links | Intranet

▶ Home

Search Legal mention Site map

Menu

Home
Paris 2012
Live News

Project

Objectives
Catalog of genes
Genes in individuals
Microbial profiling

Welcome to MetaHIT website

MetaHIT is a project financed by the European Commission under the 7th FP program. The consortium gathers 13 partners from academia and industry, a total of 8 countries. Its total cost has been evaluated at more than 21,2 million € and the funding requested from the European Commission has been set with an upper limit of 11,4 million €. The project will last from January 1, 2008 until June 30, 2012.

Grant agreement ref.: HEALTH-F4-2007-201052

Starting date: January 1st, 2008

follow us on

News

May 15, 2012
we just uploaded a **series of interviews** given during the International Human Microbiome Congress in Paris

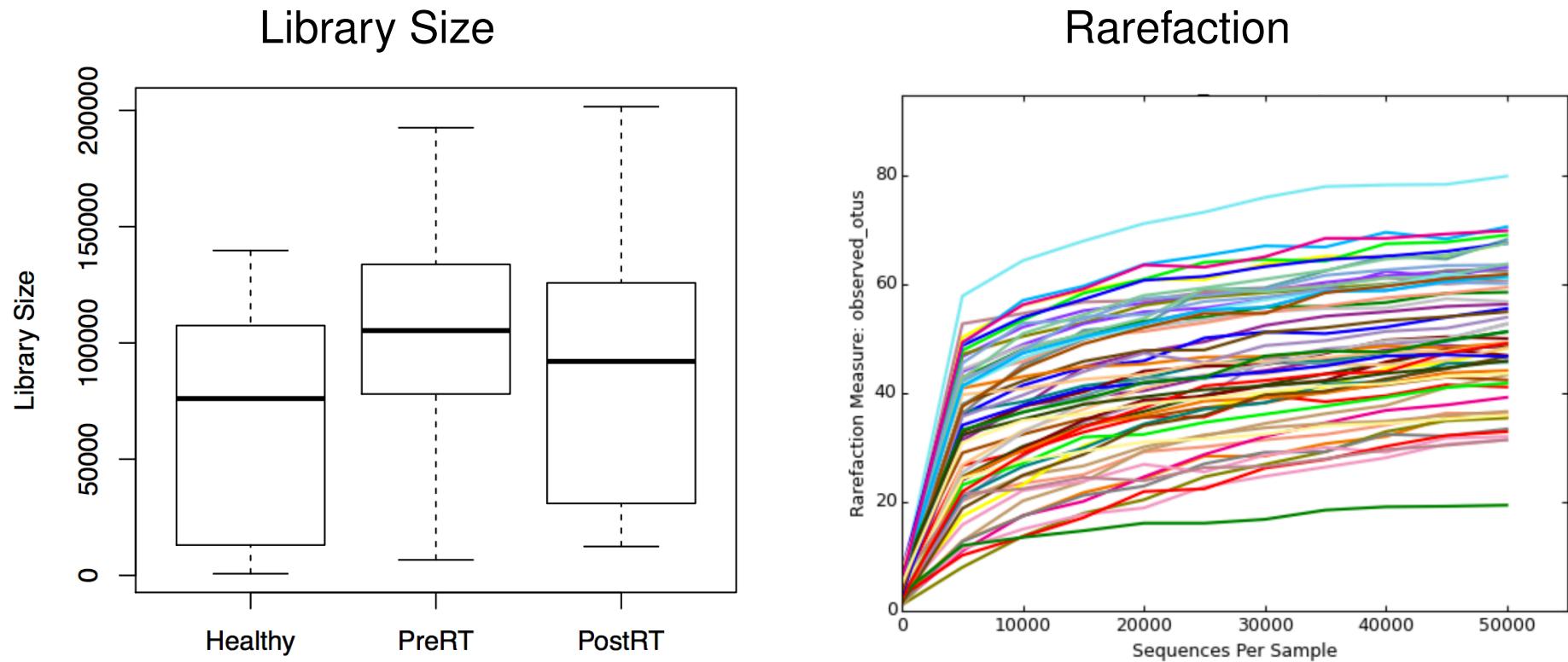
- to establish associations between the genes of the human intestinal microbiota and our health and disease
- focus on two disorders of increasing importance in Europe, Inflammatory Bowel Disease (IBD) and obesity

- Quality control
 - Filtering of OTUs and samples
 - Library size and Rarefaction
- Exploratory analysis
 - Relative abundance (e.g., heatmap, painter plot)
 - Alpha diversity (e.g., boxplot)
 - Beta diversity (e.g., PCoA)
- Global testing
 - Compare the overall microbiome composition across different clinical groups
- OTU-based testing
 - Detect differentially abundant OTUs across different clinical groups

There is no gold standard yet!

For example:

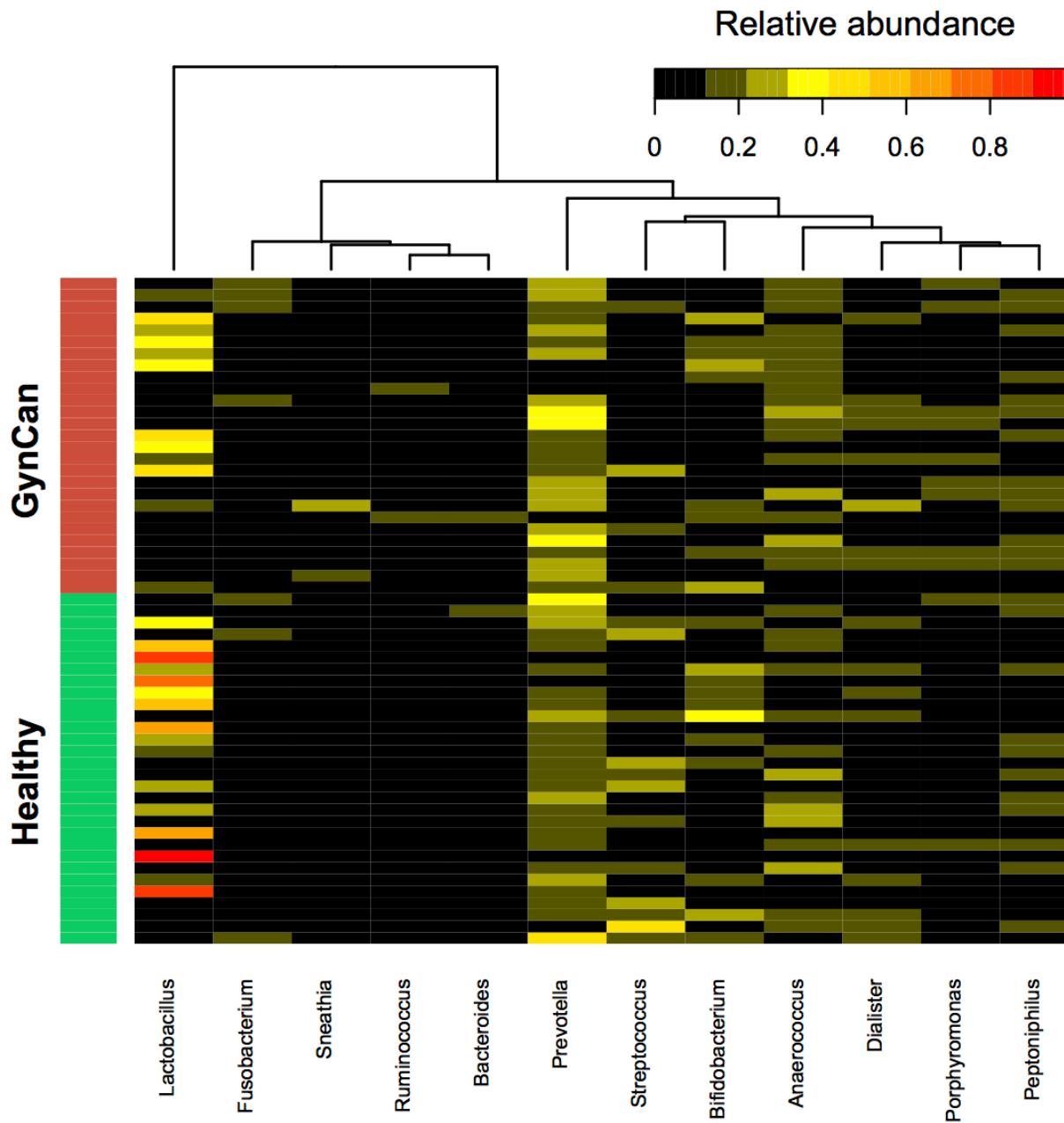
- OTU filter:
 - remove singletons (i.e., exist only in one sample)
 - remove OTUs that are present in < 10% of samples
 - remove OTUs with relative abundance < 0.5%
- Sample filter
 - remove samples with < 500 sequencing reads



- Sequencing experiments lead to an arbitrary total number of sequence reads per sample (**library size**); strong batch effect on library size
- Uneven library size is a strong confounder for microbiome analysis
- Rarefaction curves are used to determine the library size that all samples are rarefied to.

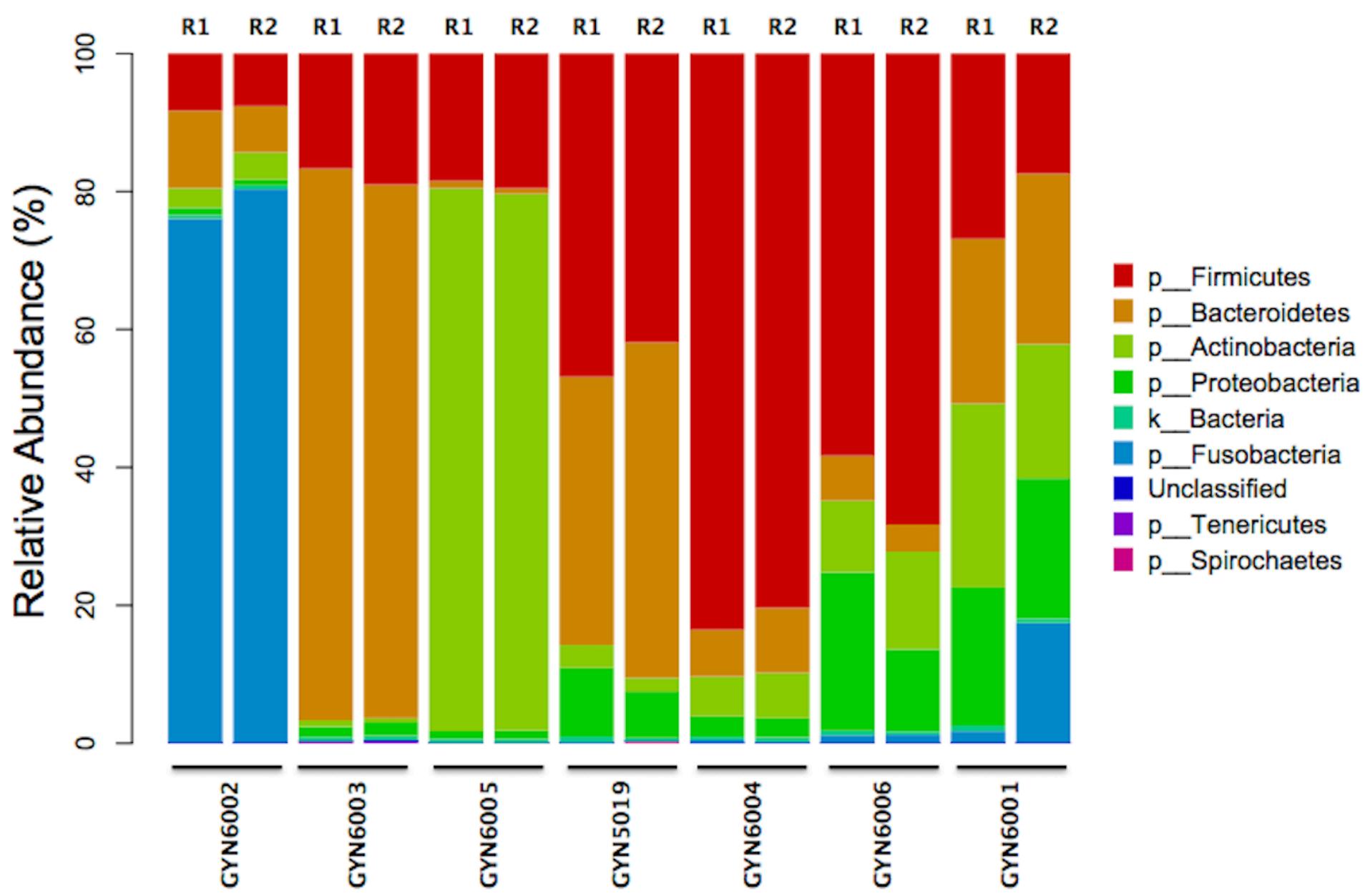
Exploratory Analysis – rel. abundance (heatmap)

— 15/25 —



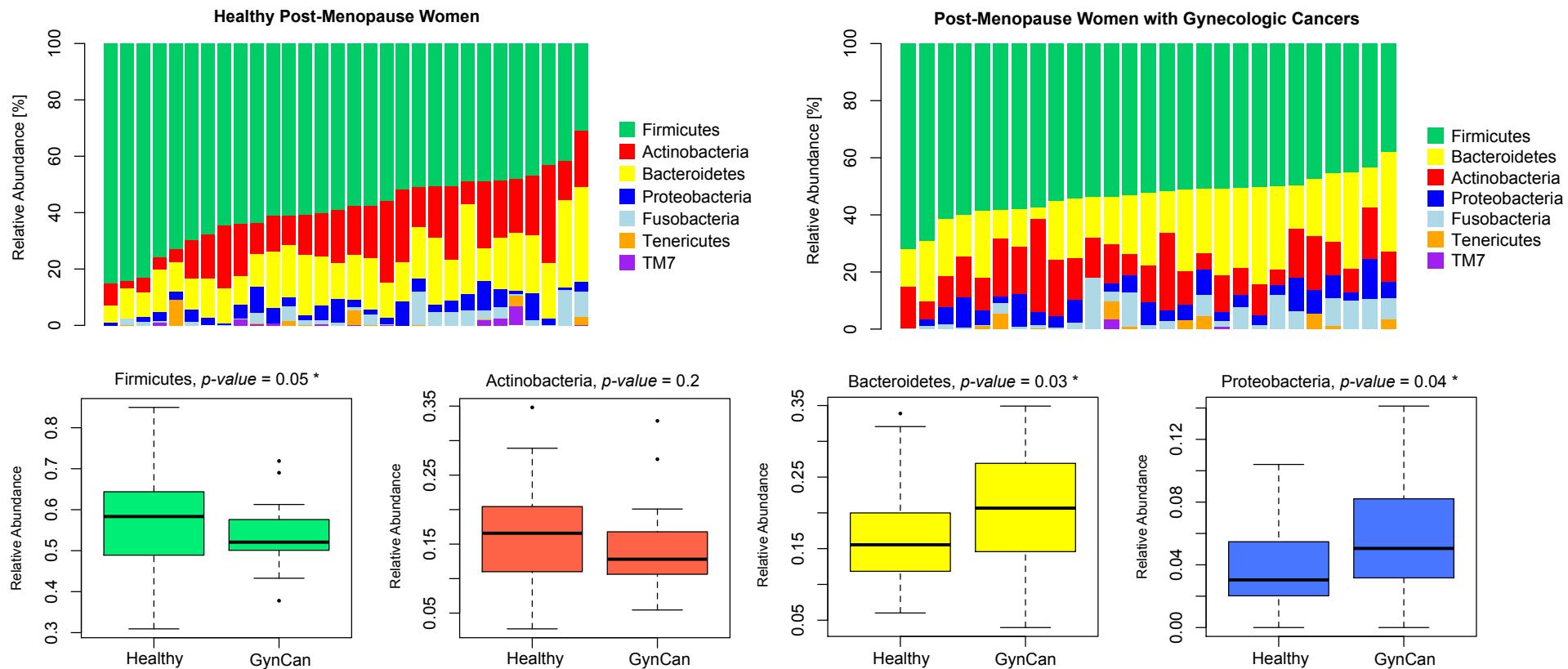
Exploratory Analysis – rel. abundance (painter plot)

— 16/25 —



Exploratory Analysis – rel. abundance

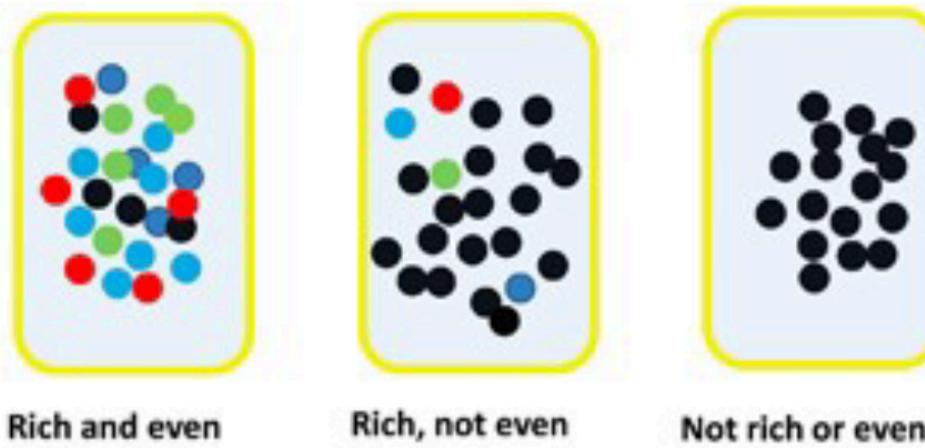
— 17/25 —



- Wilcoxon rank-sum test (nonparametric) for comparing two groups
- Krustal-Wallis test (nonparametric) for comparing more than two groups

Alpha diversity: measure the diversity within a sample

- Richness: a measure of number of species present in a sample
- Evenness: distribution of different microbes

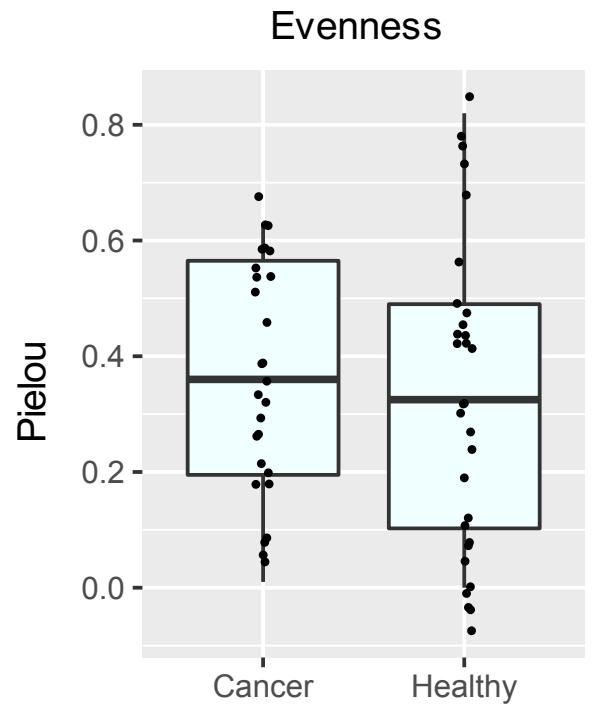
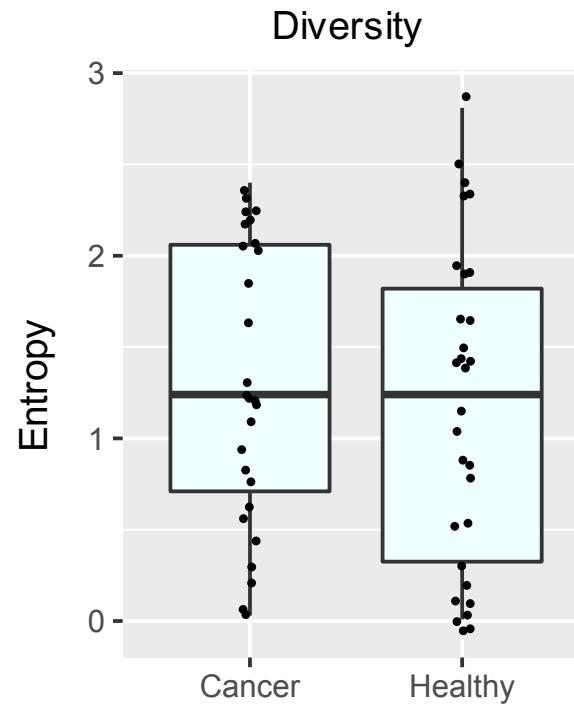
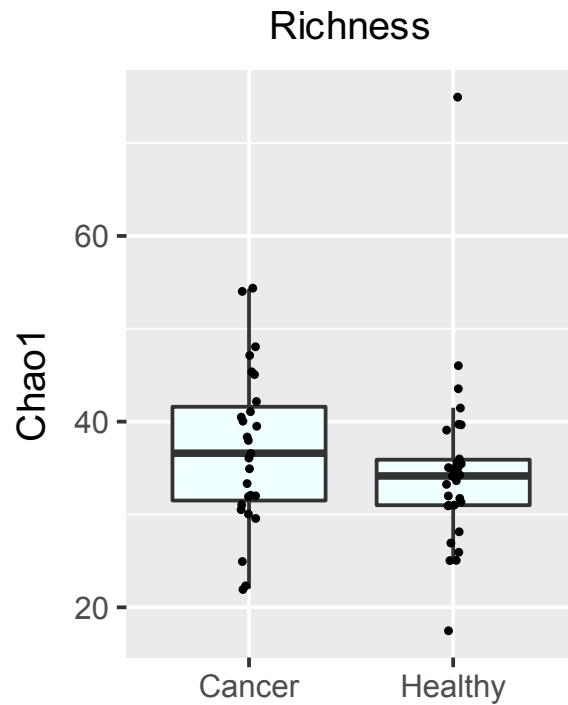


Common used alpha diversity metrics

- **Observed species:** measure richness only, S_{obs}
- **Chao1:** measure richness only, $S_{\text{obs}} + (1/2)S_{\text{singleton}}^2/S_{\text{doubleton}}$
- **Shannon:** measures richness and evenness, $H' = - \sum_{j=1}^J p_j \ln p_j$
- **Pielou:** measure evenness only, $H'/H'_{\max} = H'/\ln S_{\text{obs}}$

Exploratory Analysis – Alpha Diversity

— 19/25 —



- Wilcoxon rank-sum test (nonparametric) for comparing two groups
- Krustal-Wallis test (nonparametric) for comparing more than two groups

Beta diversity: measure the distance or dissimilarity between each sample pair
⇒ distance/dissimilarity matrix

Common used beta diversity metrics

- Non-phylogeny based

- **Bray-Curtis**: based on abundance

- * based on absolute abundance n_{ij} , $b_{ii'} = \frac{\sum_{j=1}^J |n_{ij} - n_{i'j}|}{n_{i+} + n_{i'+}}$

- * based on relative abundance p_{ij} , $b_{ii'} = \sum_{j=1}^J |p_{ij} - p_{i'j}|$

- **Jaccard**: based on presence-absence

- * : $J_{ii'} = \frac{|S_i \cap S_{i'}|}{|S_i \cup S_{i'}|}$, S_i is the set of present OTUs in sample i

- Phylogeny based (b_j the length of branch j)

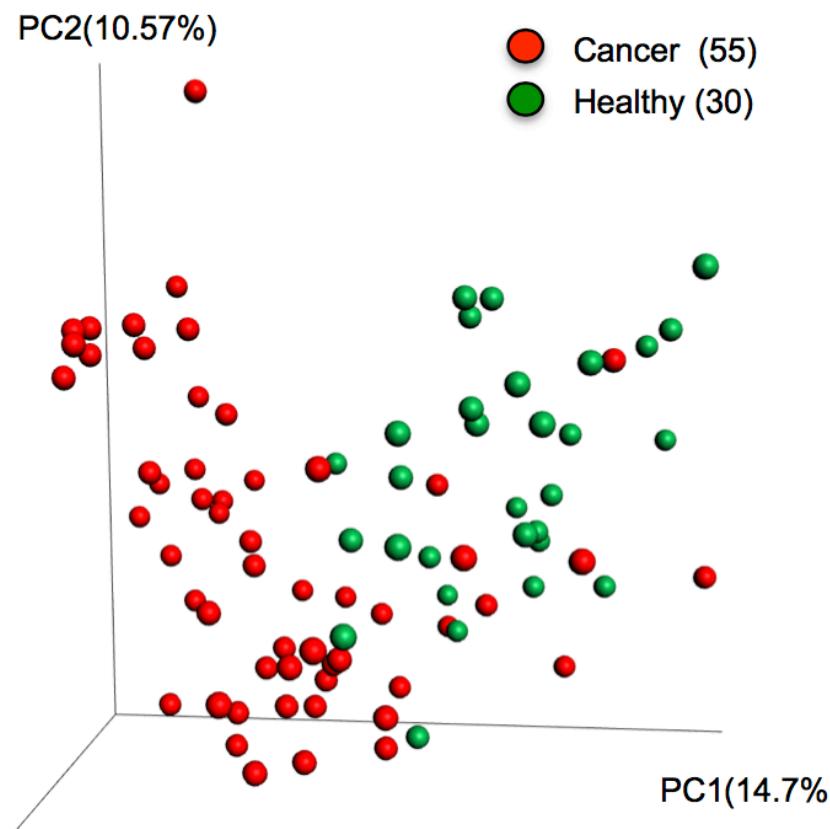
- **Weighted UniFrac**: based on abundance, $d_{W,ii'} = \frac{\sum_{j=1}^J b_j |p_{ij} - p_{i'j}|}{\sum_{j=1}^J b_j (p_{ij} + p_{i'j})}$

- **Unweighted UniFrac**: based on presence-absence,

$$d_{U,ii'} = \frac{\sum_{j=1}^J b_j |I(p_{ij}>0) - I(p_{i'j}>0)|}{\sum_{j=1}^J b_j}$$

Principal Coordinates Analysis (PCoA) can be used for visualization of the data present in the beta diversity distance matrix in the form of 2-dimensional or 3-dimentional plots known as PCoA plots.

Perform eigen-decomposition of a pre-specified distance matrix and obtain eigenvectors (PC1, PC2, ...)



Statistical hypothesis: the microbiome compositions are different in the healthy and in the diseased group

PERMANOVA (Permutation-based ANOVA): based on a pre-specified distance matrix ($d_{ii'}$)

- Square of distance matrix: $A = (a_{ii'})$, where $a_{ii'} = -\frac{1}{2}d_{ii'}^2$
- Gower standardization: $G = \left(I - \frac{11'}{n}\right)A\left(I - \frac{11'}{n}\right)$
- Hat matrix of the design matrix X : $H = X(X^T X)^{-1}X^T$
- The pseudo-F statistic (m covariates and n samples):

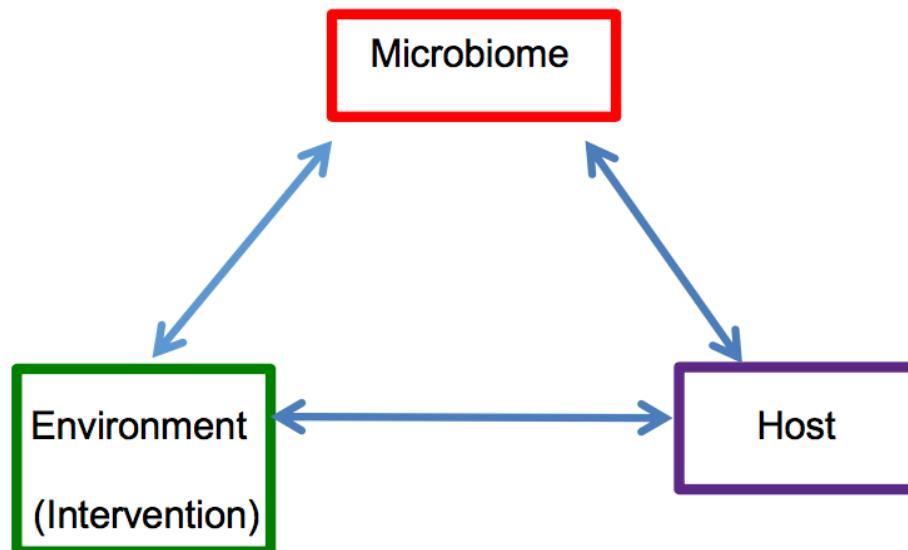
$$F = \frac{\text{tr}(HGH)/(m - 1)}{\text{tr}[(I - H)G(I - H)]/(n - m)}$$

- The significance of the pseudo-F statistics is assessed based on permutations

There is no gold standard!

- **DESeq2** (Love et al., 2014)
 - Developed for detecting differentially expressed genes using RNA-seq data
 - Normalization for gene expression data (non-sparse data)
 - Assume Negative-Binomial model
- **MetagenomeSeq** (Paulson et al., 2013)
 - Developed for detecting differentially abundant OTUs using 16S sequencing data
 - Normalization accounts for sparse data
 - Assume a zero-inflated Gaussian (ZIG) distribution mixture model
- **ANCOM** (ANalysis of Composition Of Microbiomes, Mandal et al., 2015)
 - Developed for detecting differentially abundant OTUs using 16S sequencing data
 - make no distributional assumptions; use log-ratios

- Analysis of paired, clustered, or longitudinal data
- Adjustment of confounders (e.g., gender, ancestry)
- Adjustment of batch effects (e.g., library size)
- Causal inference
 - Randomized clinical trials
 - Mediation analysis



- Network analyses identify co-varied OTUs

- To control batch effects
 - Randomization
 - Use control samples with known composition in each batch
 - Replicate some samples across sequencing batches
- Paired sample designs will increase power
- Longitudinal design help reveal dynamics or even causality