Microbiome Genomics

Timothy Read PhD

What is the microbiome?

"the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space."

Joshua Lederberg

- Bacteria
- Viruses
- Fungi
- Protists

Size of the microbiome

- 10^{30} bacterial cells on Earth
- Maybe 10^{31}-10^{32} bacteriophages
- 10x more bacterial cells than human cells in a person
- Many more bacterial genes in the body than human, especially metabolic genes

Whitman WB, et al(1998) PNAS 95(12) 6578-83
Sears CL (2005). *Anaerobe* **11** (5): 247–51

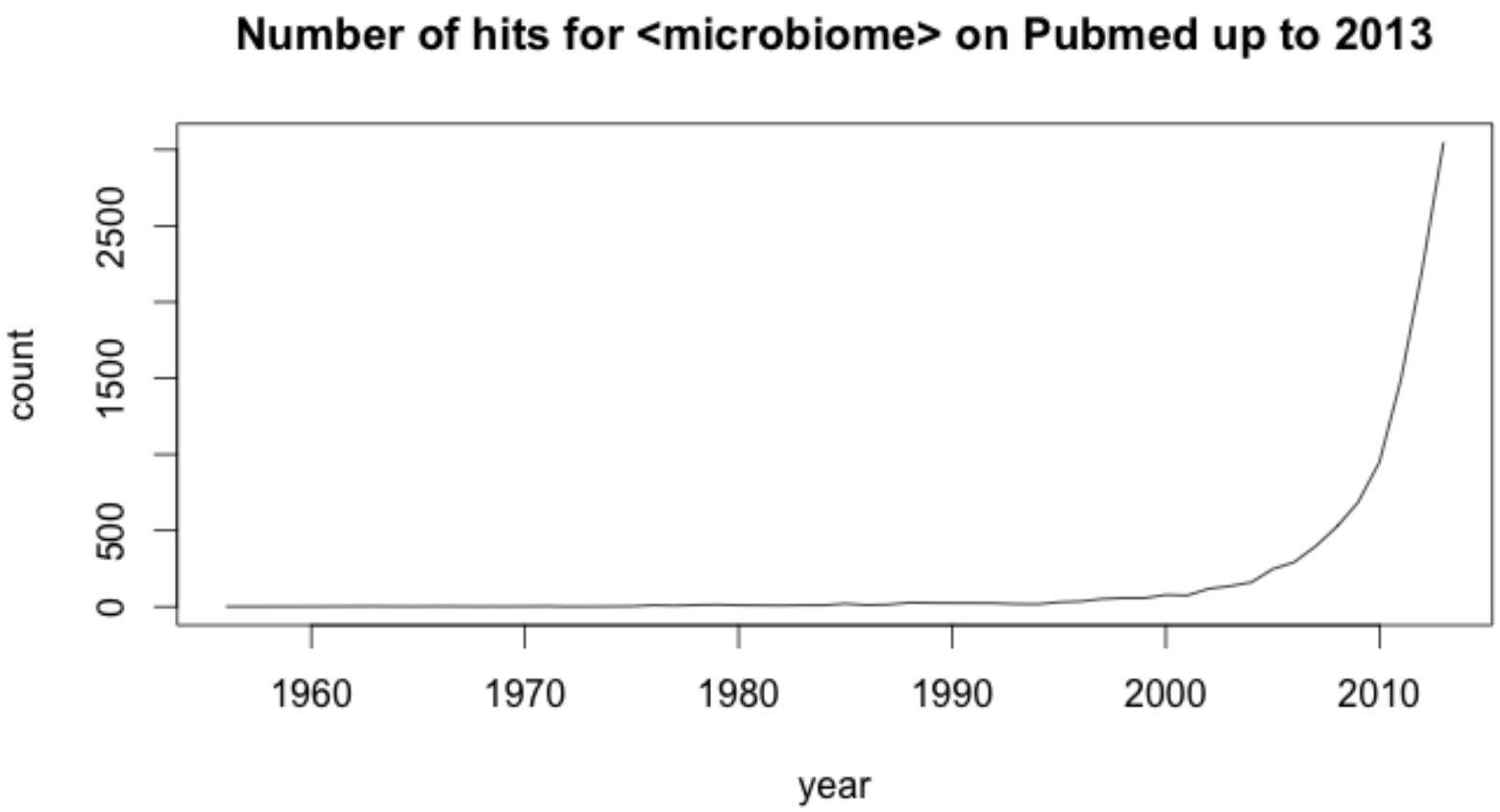
History

- 16S sequencing studies began with Carl Woese in the late 1960s, early 1970s
- Environmental DNA cloned into BAC libraries Jo Handelsman & Ed deLong, - 1990s
- Major explosion after next gen sequencing technologies introduced in 2005 (454 especially)

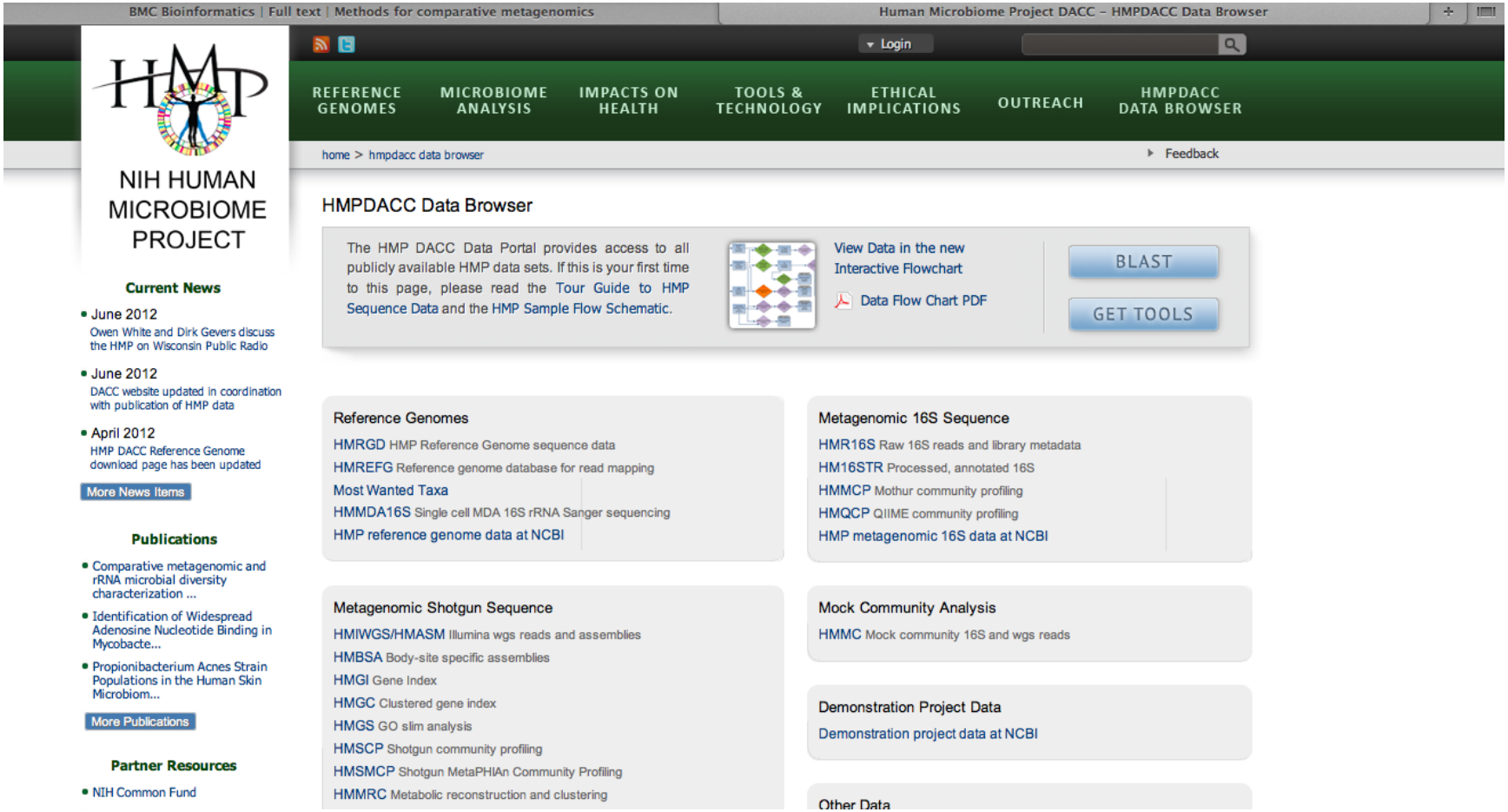
History II

- 2011: Illumina 16S sequencing (Caporaso et al PNAS)
- 2012 Human Microbiome Project Publication

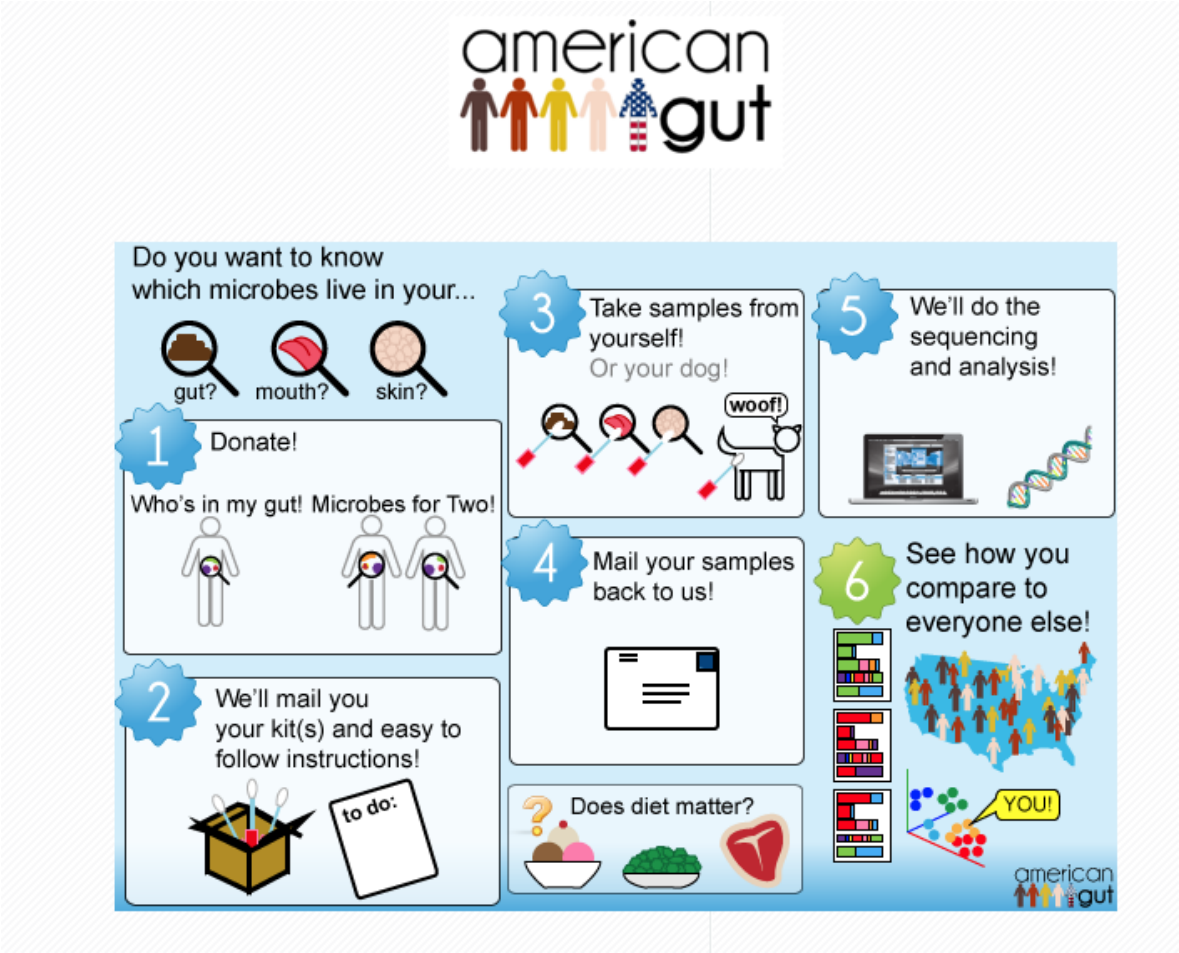
Microbiome Trends



Human Microbiome Project

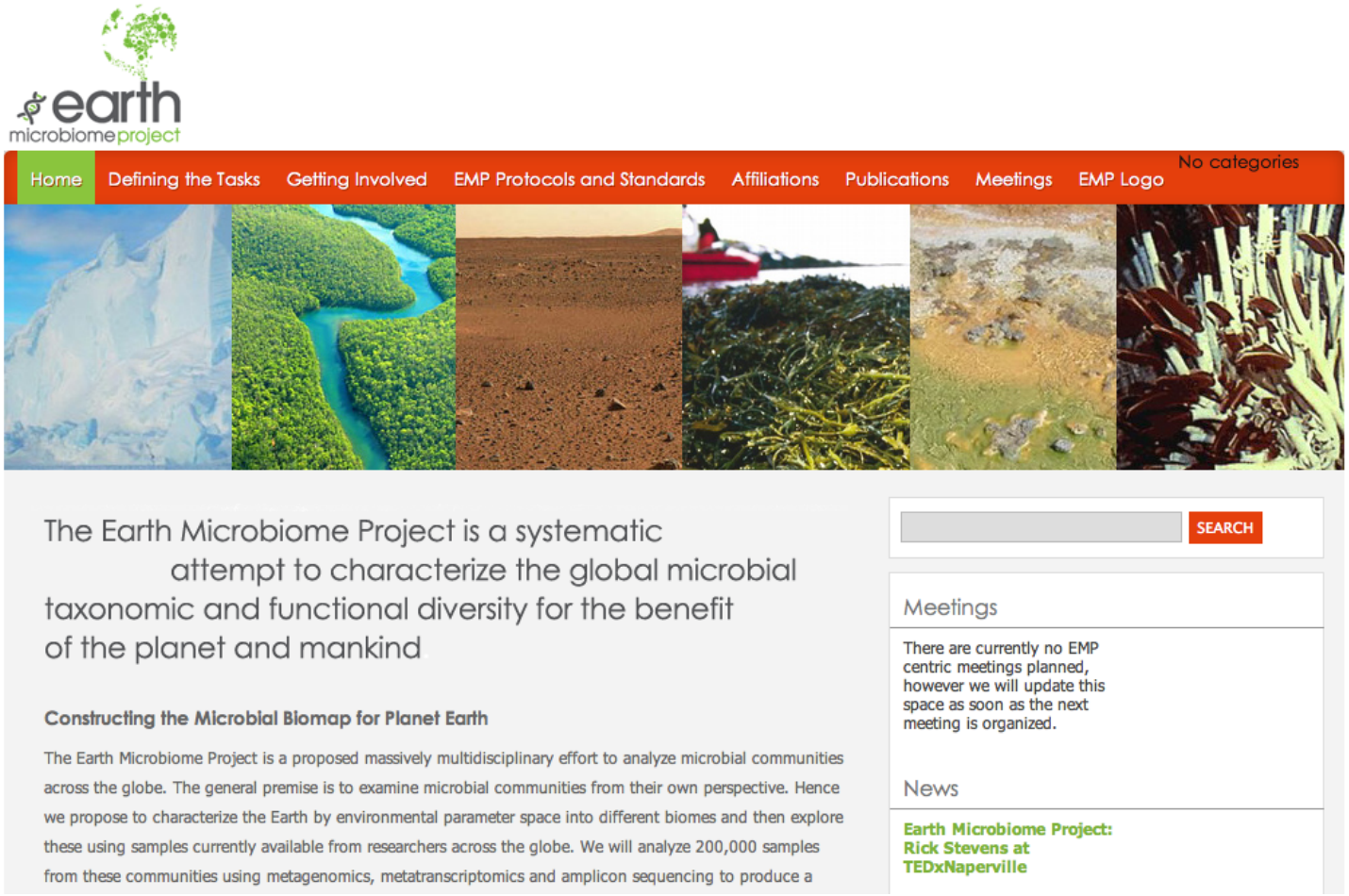


American Gut



http://humanfoodproject.com/americangut/

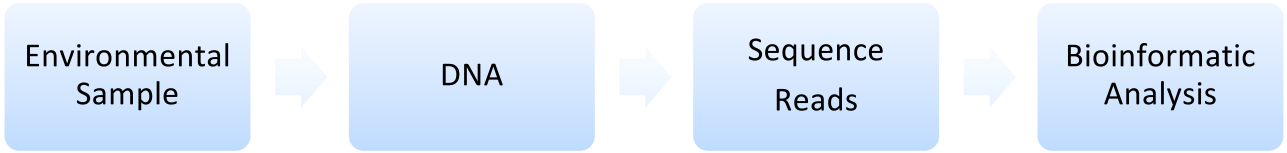
Earth Microbiome Project



What are the outputs of microbiome research?

- Summary statistics – species richness, diversity, species differences between samples
- Genes present and their abundance
- Ecological community drivers – species, peturbations etc

Microbiome analysis concept



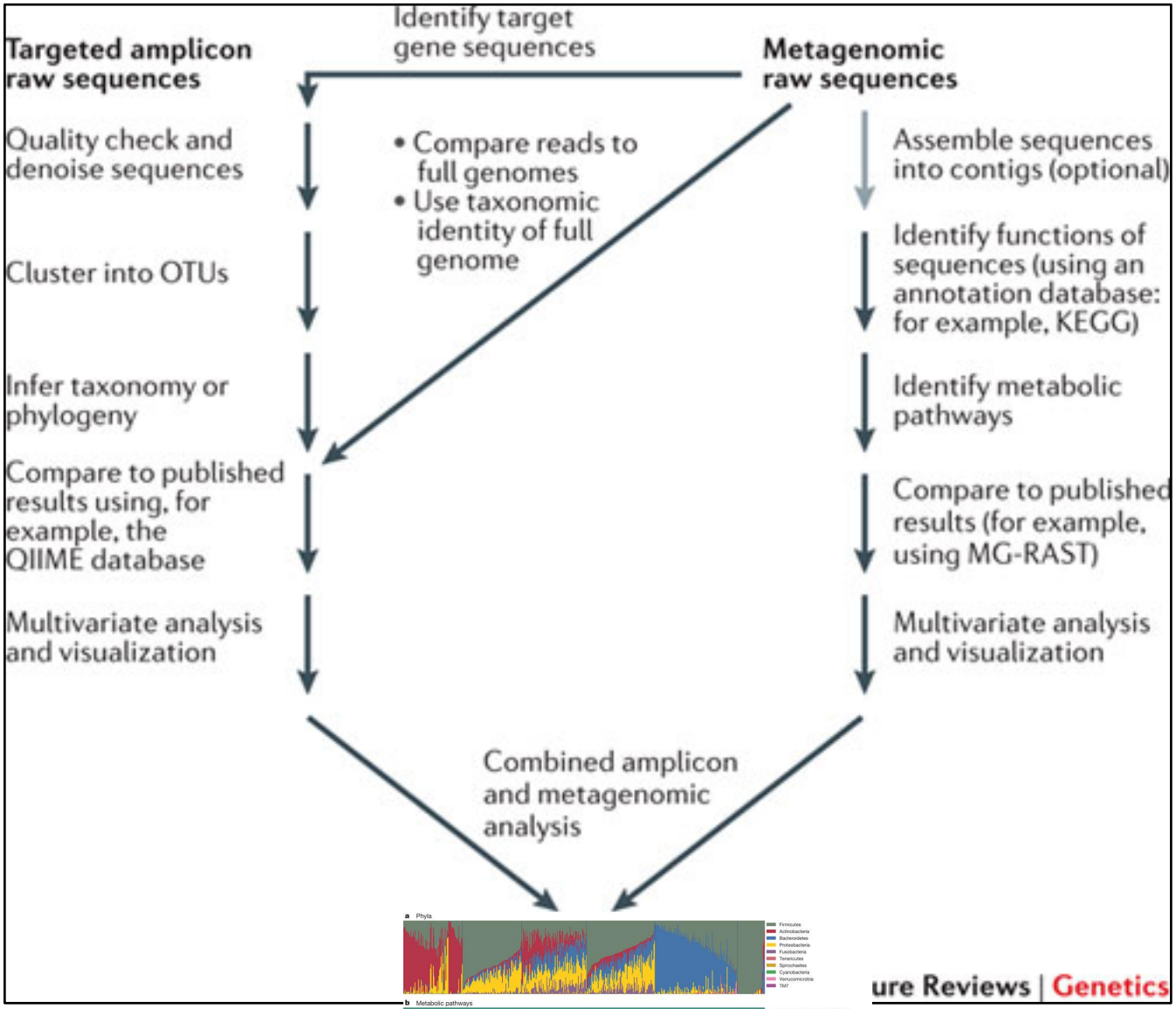
Free microbiome software

- MOTHUR (http://www.mothur.org)
- QIIME (“Chime”; http://qiime.org)
- Phyloseq (R package for post-OUT assignment analysis; http://www.bioconductor.org/packages/2.13/bioc/html/phyloseq.html)

DNA-based microbiome studies

- Goal: infer microbiota composition and diversity
- Targeted amplicon studies
  - Use small # of phylogenetically informative markers
  - Ribosomes are good targets (e.g. ***16S* rDNA**)
    - present in all bacteria
    - ribosomal genes contain both slow and fast-evolving regions
- Species Barcode
  - ITS/18S fungal alternative to
  - Good for **taxonomic profiling**
- Shotgun metagenomic studies
  - Shear up entire genome, sequence, reassemble
  - Good for **functional profiling**
  - Lower coverage = lower resolution for taxonomic profiling

DNA-based microbiome studies



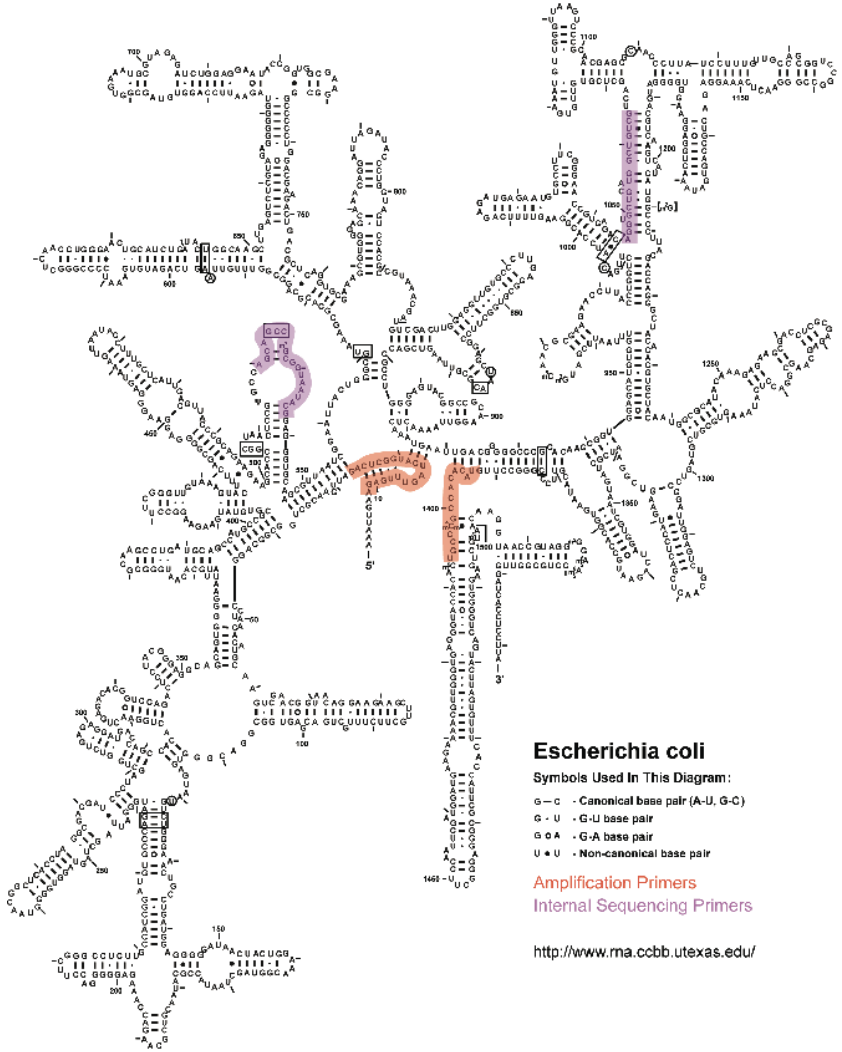
16S

Barcode methodology

- Obtain DNA sample
- Amplification using conserved primers (can add oligo tags to multiplex)
- Compare to sequence database (or can do de novo clustering)
- Obtain map of species composition

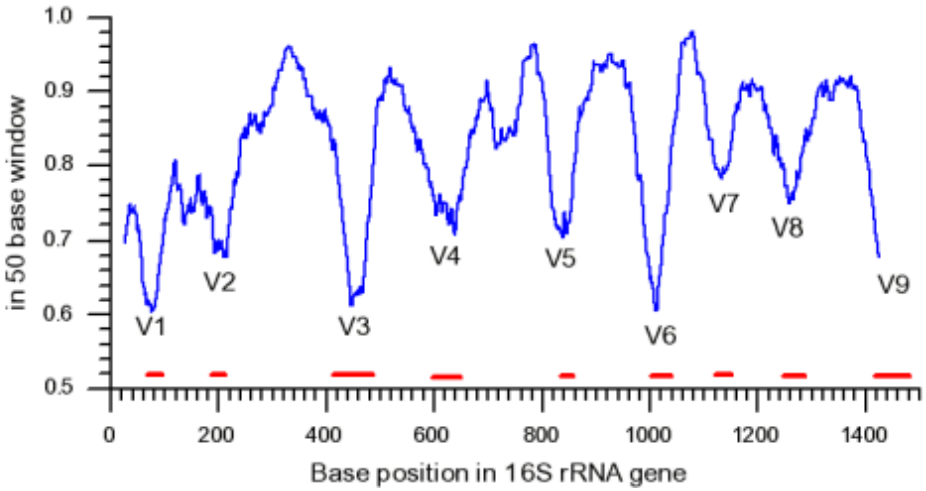
16s rRNA gene analysis:

Genomic approach to identify bacteria



Secondary Structure: 16S small subunit ribosomal RNA

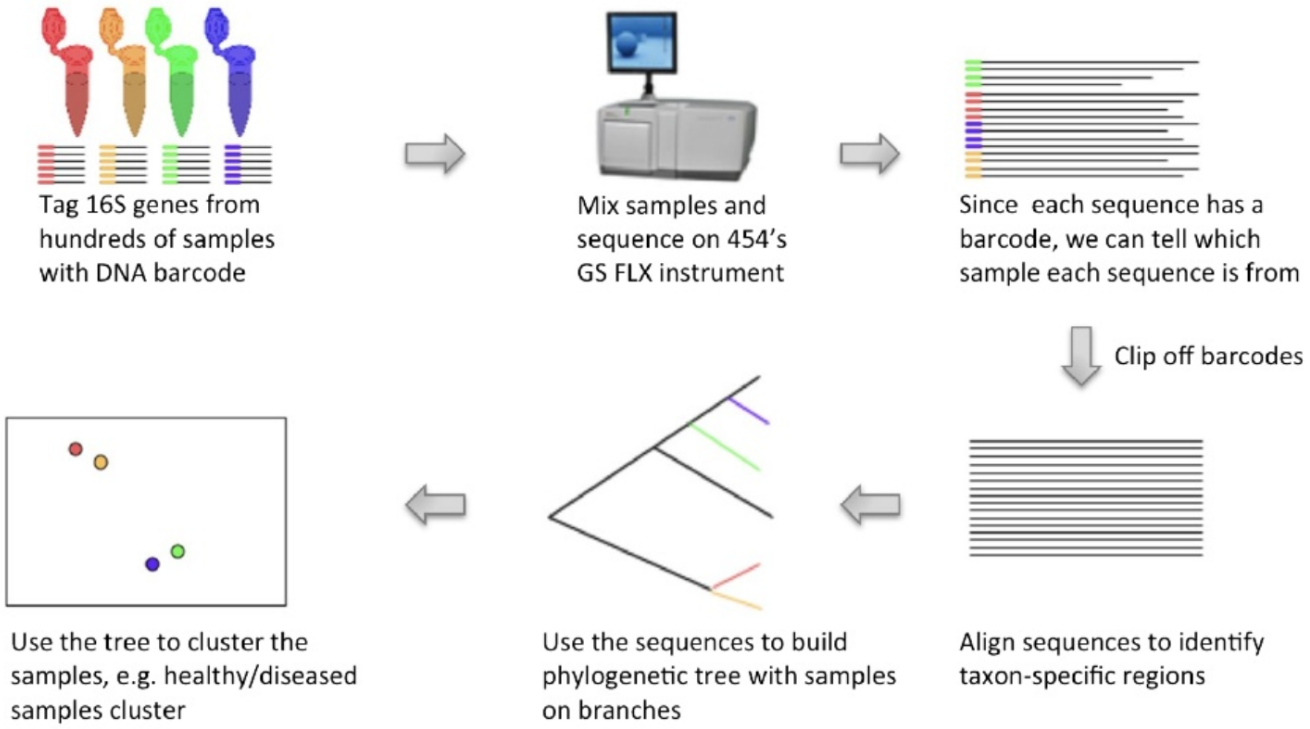
- 16s rRNA gene is universal among prokaryotes
- Phylogeny assessed based on conserved and variable regions



Slide courtesy of Julie Segre

Basic bioinformatic analysis pathway

QIIME: Analysis of Hundreds of Samples



Hamady et al. 2008 Nature Methods 5:235; Caporaso et al. 2010 Nature Methods 7:335

Sequence read quality processing

- Most 16S studies to date have been performed using 454 technology. Now and for the near-future, the Illumina MiSeq is the dominant platform.
- 100-600 nt read length depending on the version of the technology used
- Common quality control steps
  - Trim reads for quality
  - Chimera filtering
  - Detect and remove PCR duplicates
  - Remove Singletons post clusters

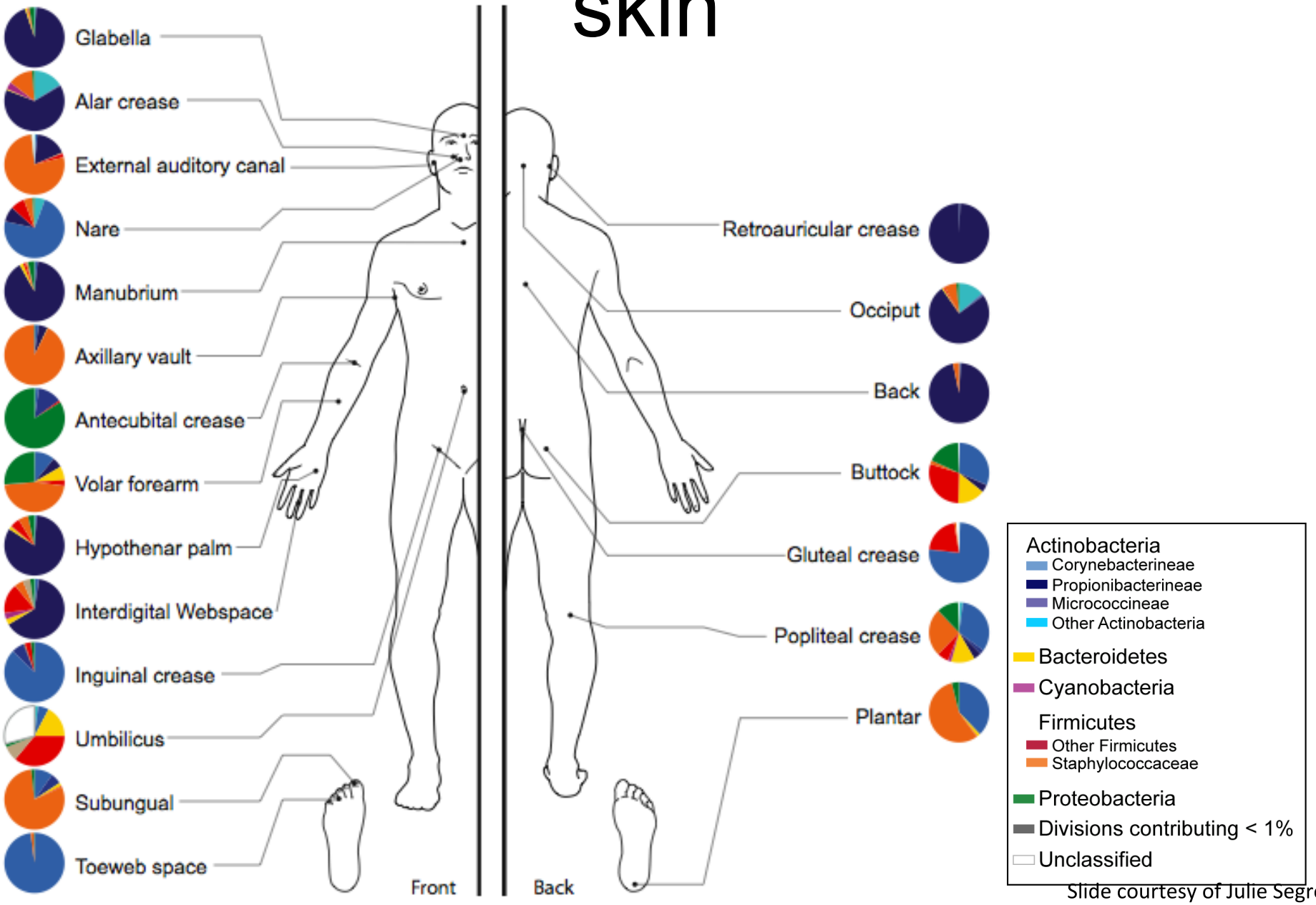
Operational Taxonomic Units (OTUs)

- Essentially a grouping (supposed to approximate species or genus-like groups)
- At 97% similarity, OTUs represent “species”
- Important to remember though that OTUs don’t map exactly to species

Assigning OTUs

- *de novo* vs. reference-based OTU clustering
  - *De novo*: naive cluster analysis using a similarity threshold to form species or genus-like OTUs
  - Reference-based: use of external information to assign sequences to known taxa
    - external info = reference databases of long high-quality DNA sequences (e.g. greengenes, SILVA)

Variation in OTU composition in skin



Slide courtesy of Julie Segre

Comparing microbial communities

- What is there?
- How much of each species? (alpha diversity)
- How do communities differ? (beta - between sample diversity)

Richness and Diversity

- Species richness – Total number of species(OTUs)
- Alpha Diversity – Within Sample Diversity
- *Beta Diversity – Between Sample Diversity
- Gamma Diversity – total diversity of community (all samples)

*There are other definitions – e.g

Beta = Gamma/Alpha

AlphaDiversity

- Many measures of diversity, two common ones
- Shannon Diversity

$$H' = -\sum_{i=1}^{S} (p_i \ln(p_i)),$$

- Simpson Diversity

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

Example

A) Sequence Abundance

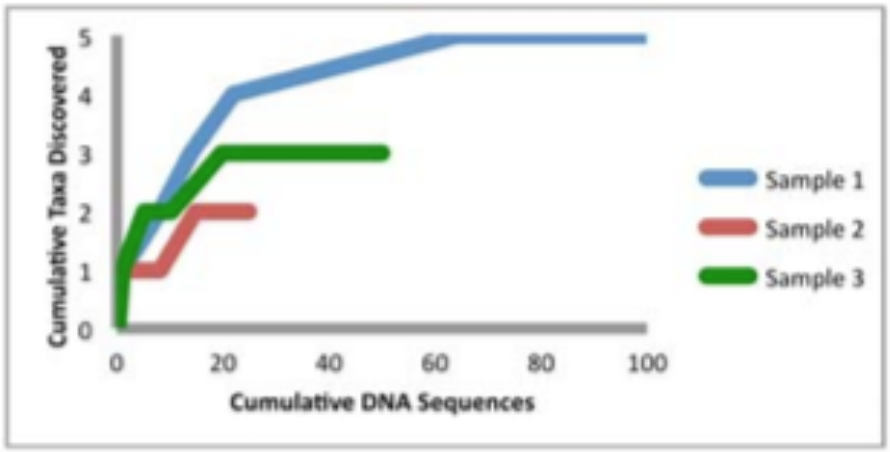
OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

B) Sequence Relative Abundance

OTU	Sample 1	Sample 2	Sample 3
A	0.60	0	0.70
B	0.24	0.20	0.10
C	0.10	0	0
D	0.05	0	0
E	0.01	0	0
F	0	0.80	0.20
Total	1.0	1.0	1.0

Morgan, X. C. & Huttenhower, C. Chapter 12: *PLoS Comp Biol* **8**, e1002808 (2012).

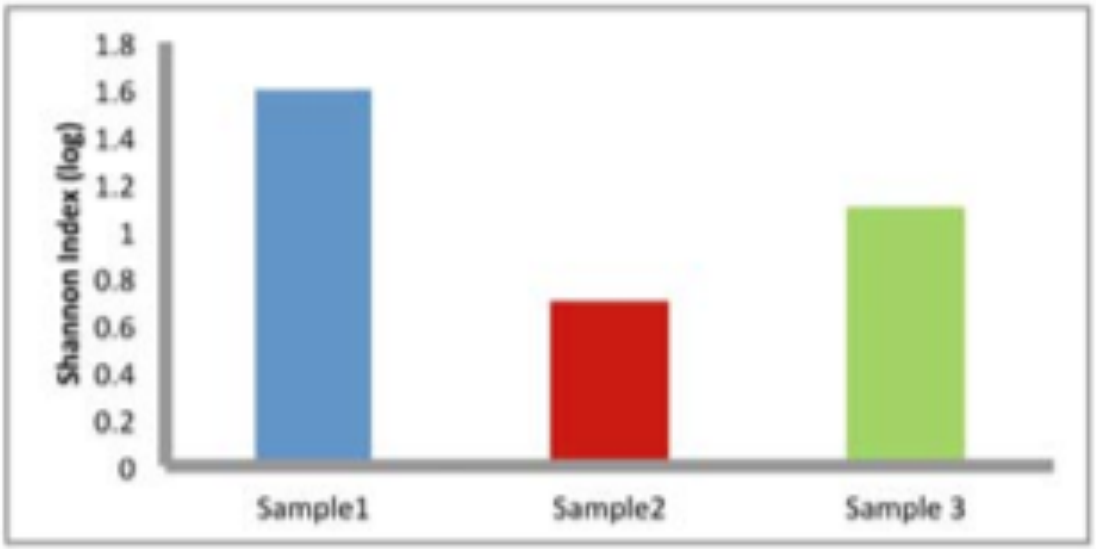
Collector's curve of richness (rarefraction)



A) Sequence Abundance

OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

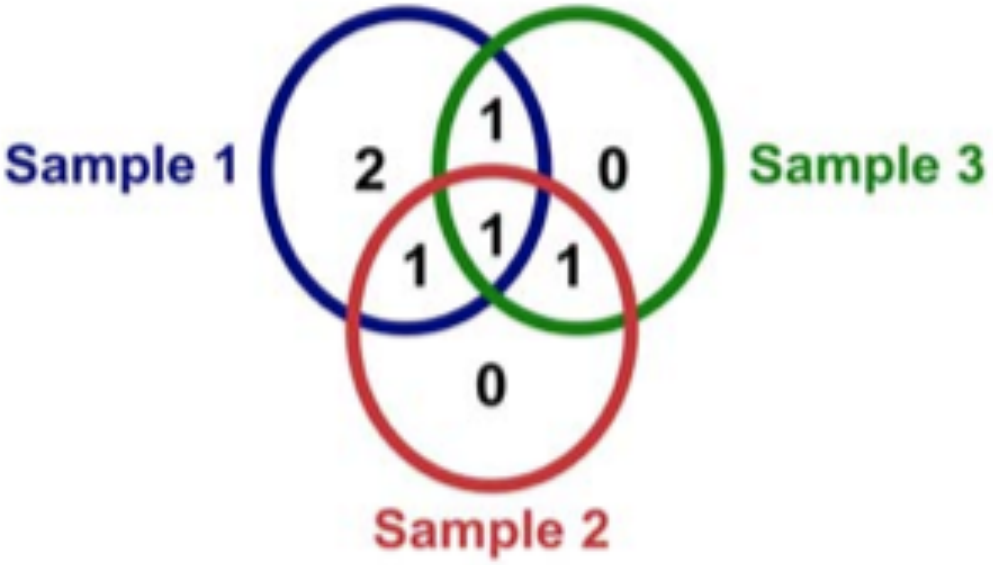
Shannon Diversity Calculations



A) Sequence Abundance

OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

Simple Between Sample Diversity



(Based on conserved taxa)

A) Sequence Abundance

OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

Another Between Sample Diversity measures

- Bray-Curtis

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

Where S_i and S_j are the number of species in populations i and j, and C_{ij} is the total number of species at the location with the fewest species

A) Sequence Abundance

OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

Other measures: Unifrac

- UniFrac (Lozupone, C. & Knight, R. *Appl Environ Microbiol* **71**, 8228–8235 (2005).
  - Uses phylogenic tree to measure distance between two communities

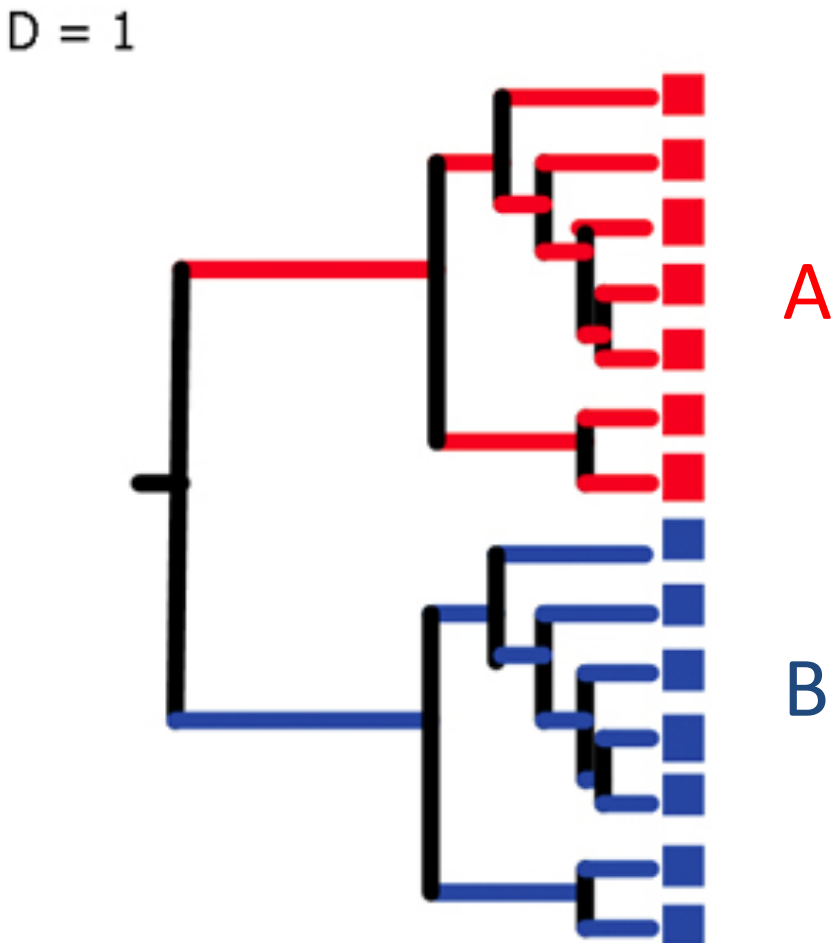
Simple Unifrac Metric

- n is the total number of branches in the tree, b_i is the length of branch i, A_i and B_i are the number of descendants of branch i from communities A and B respectively, and A_T and B_T are the total number of sequences from communities A and B respectively. In order to control for unequal sampling effort, A_i and B_i are divided by A_T and B_T.

$$u = \sum_{i}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

Perfectly separate example

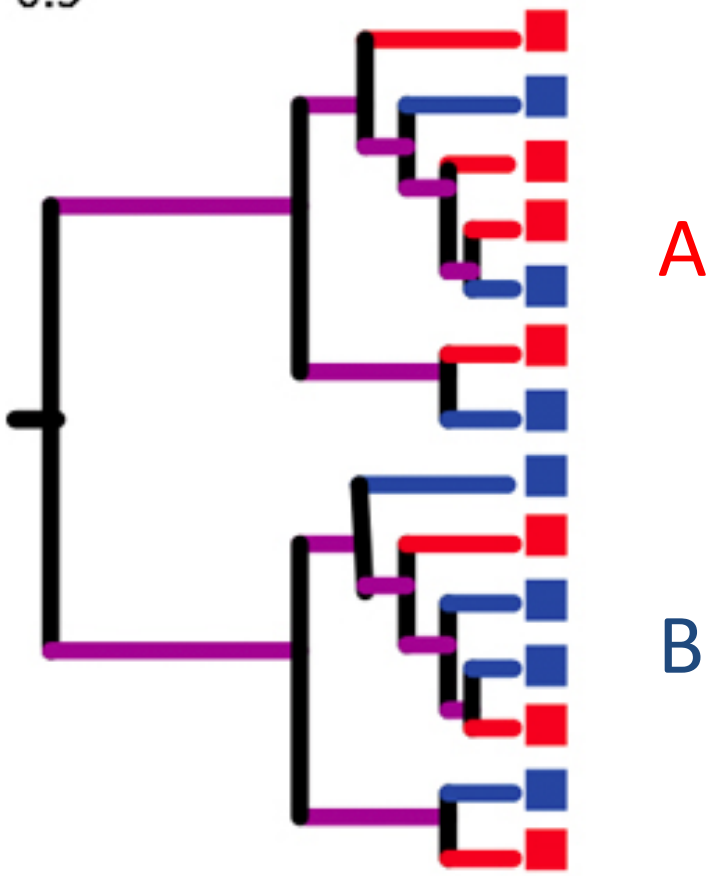
$$u = \sum_{i}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$



Perfectly mixed example

$$u = \sum_{i}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

D = ~ 0.5

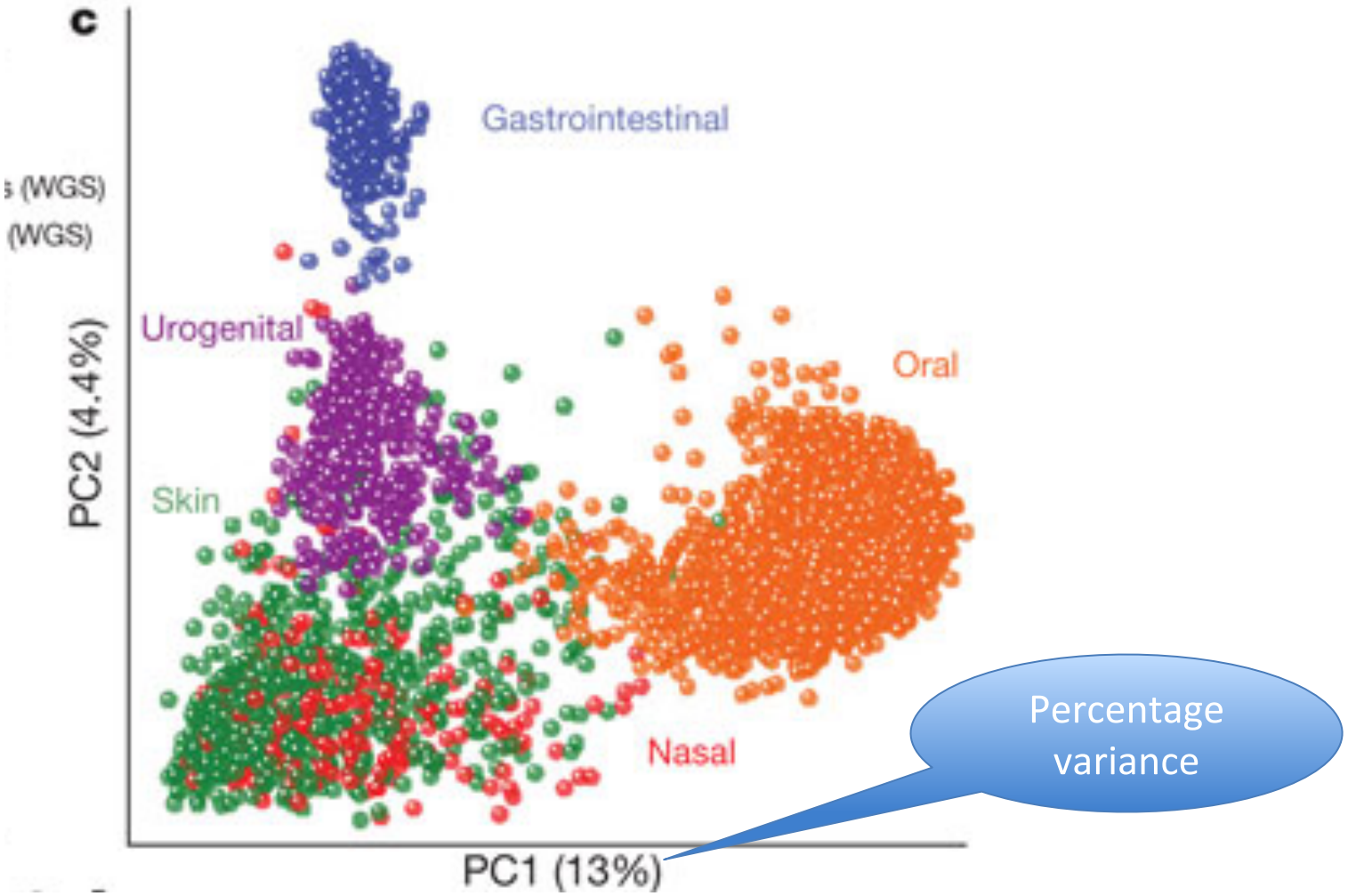


Principal Component Analysis (PCA or PCoA)

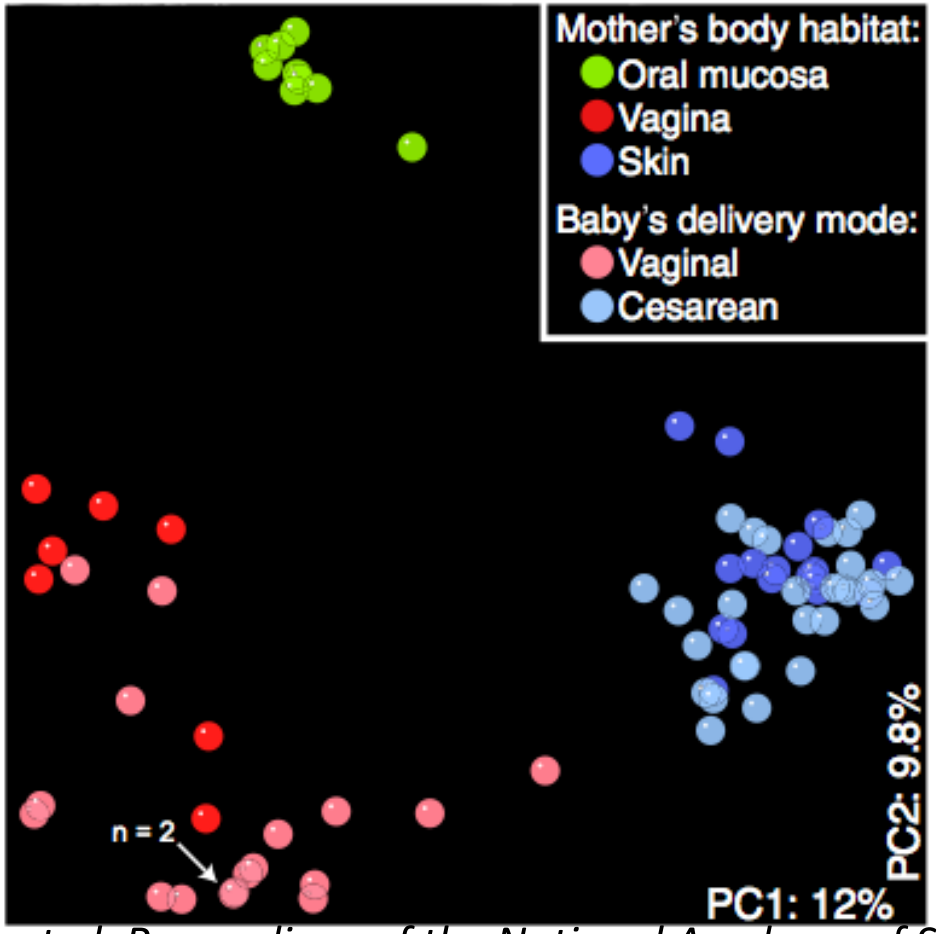
- Method to explore data in distance matrixes
- Output largest dimensions of variance
- Commonly, the two-three largest dimensions are plotted

- See: https://www.youtube.com/watch?v=BfTMmoDFXyE

Beta diversity based PCA

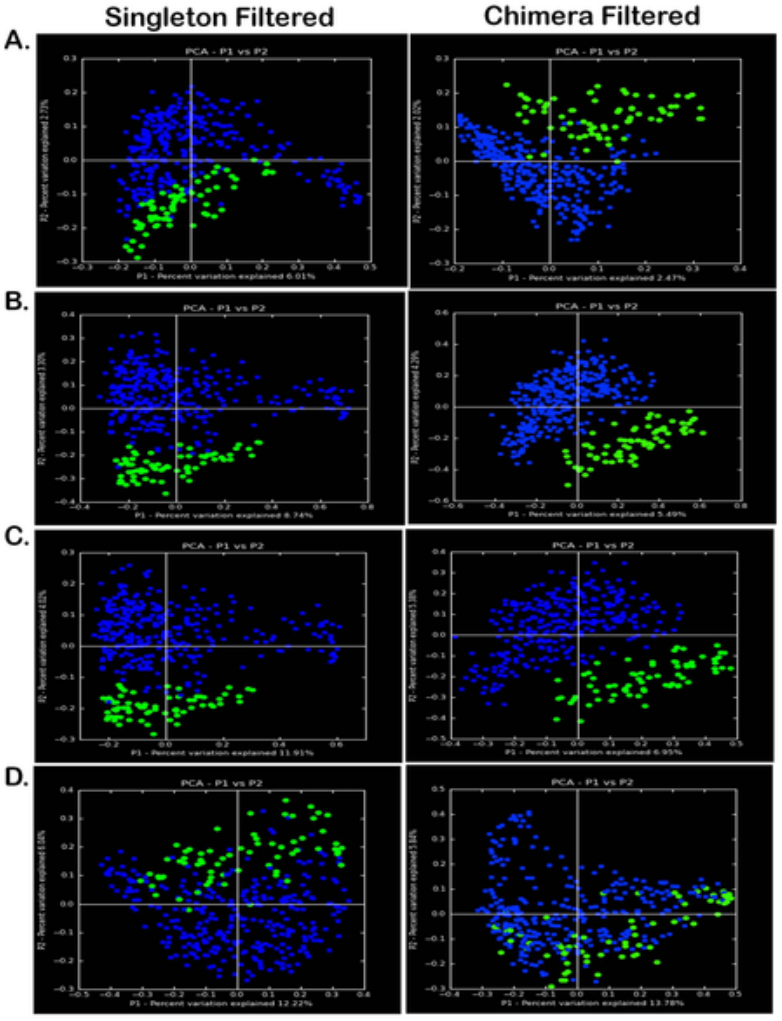


Comparing communities: Microbiome of vaginal versus caesarian birth



Dominguez-Bello, M. G. *et al. Proceedings of the National Academy of Sciences* **107**, 11971–11975 (2010). 2010 PNAS

Figure 1. Beta diversity metrics of bacterial 16S rRNA genes reveal distinctly clustered vaginal microbiome communities structured by pregnancy.



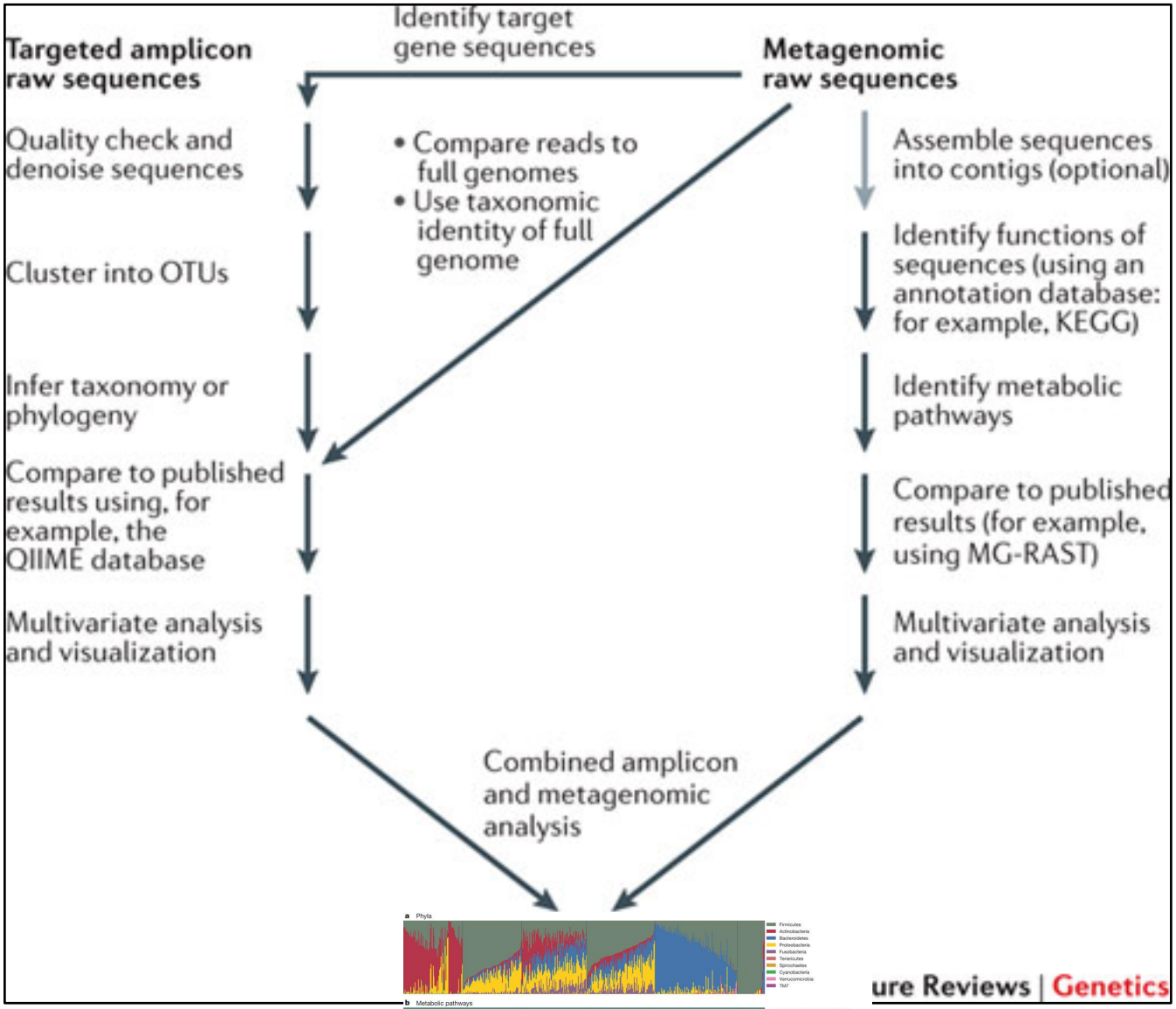
Aagaard K, Riehle K, Ma J, Segata N, et al. (2012) A Metagenomic Approach to Characterization of the Vaginal Microbiome Signature in Pregnancy. PLoS ONE 7(6): e36466. doi:10.1371/journal.pone.0036466
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0036466

Development of the infant microbiome

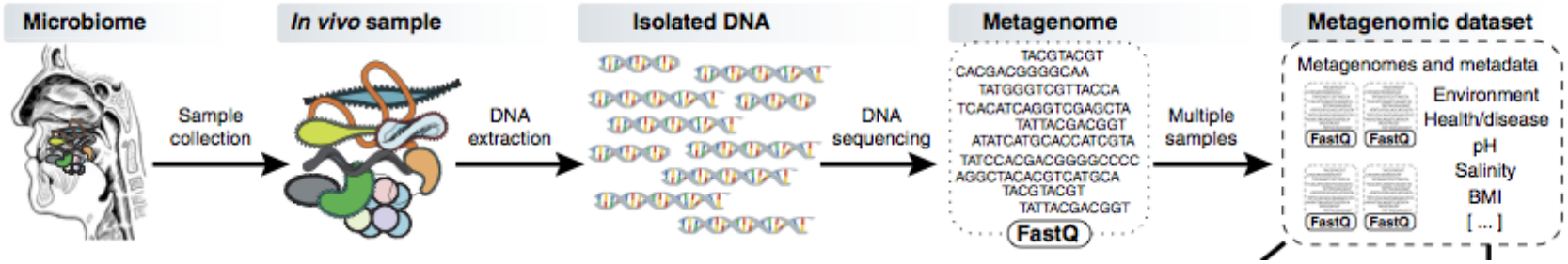
http://www.youtube.com/watch?feature=player_embedded&v=Pb272zsixSQ

Metagenome studies

DNA-based microbiome studies



Metagenome *de novo* assembly



Metagenome assembly not the same as standard *de novo* assembly

- Multiple species present at different levels of coverage
- Very large range of genome sizes, from virus to eukaryote
- Multiple members of the same species could be present > SNPs, indels need to be tolerated
- Also “contaminating” DNA

Metagenomics - Human DNA removal

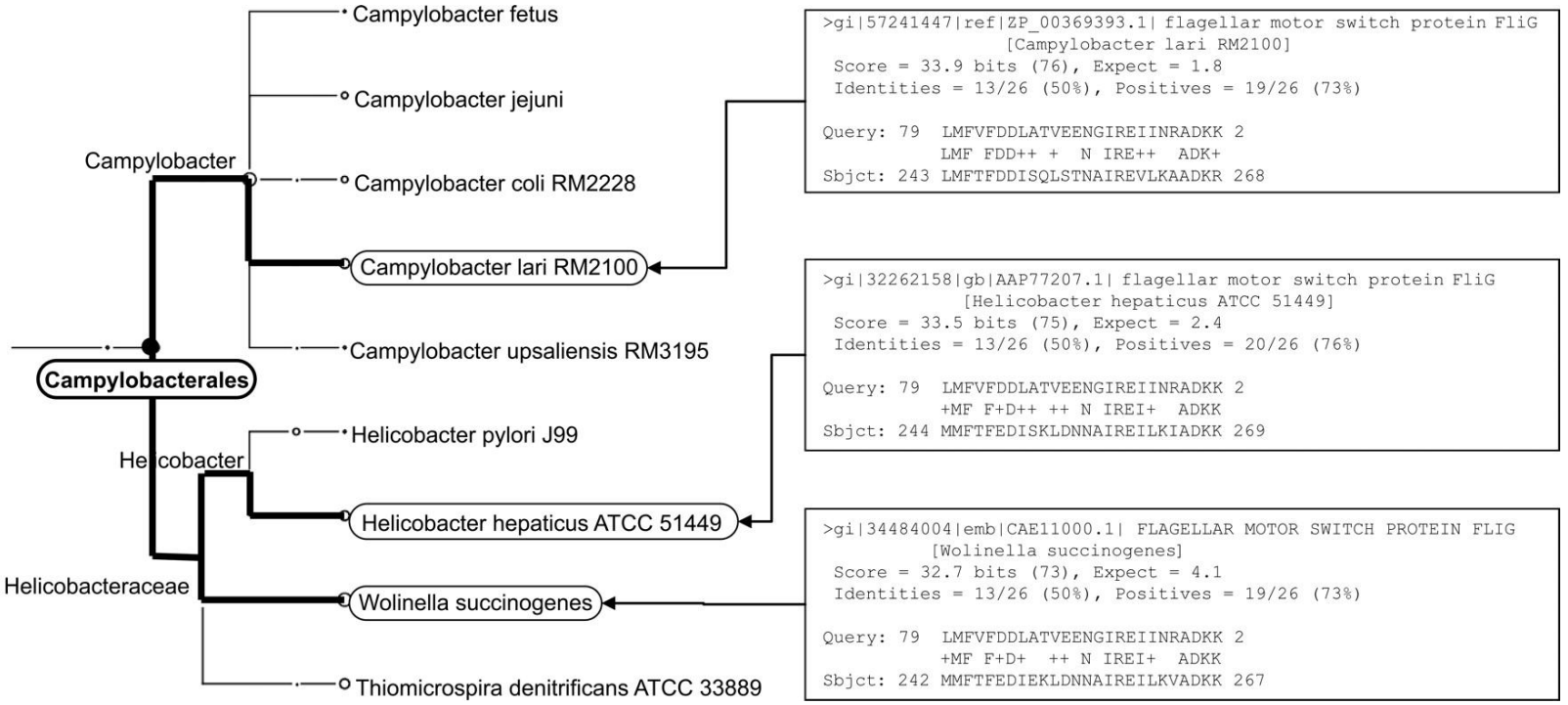
- Different tissues contain different proportions of human DNA
  - Stool < 1%
  - Throat 75%
  - Saliva 80%
  - Anterior Nares 82%
  - Mid-Vagina 96%

The HMP Consortium,. *Nature* **486**, 215–221 (2012).

Assigning OTUs to metagenome sequences

1. Can just use 16S sequences (however, only a small portion of the data)
2. Use all the sequence data
   - Taxonomy tree traversal approaches: BLAST, then map all hits to common ancestor
   - Use reference genomes

Assigning OTU by Finding Last Common Ancestor (e.g. MEGAN , KRAKEN)



Huson D H et al. Genome Res. 2007;17:377-386

Reference Genomes

BMC Bioinformatics | Full text | Methods for comparative metagenomics

Human Microbiome Project DACC – Home

NIH HUMAN MICROBIOME PROJECT

Login

REFERENCE GENOMES | MICROBIOME ANALYSIS | IMPACTS ON HEALTH | TOOLS & TECHNOLOGY | ETHICAL IMPLICATIONS | OUTREACH | HMPDACC DATA BROWSER

home > reference genomes

Feedback

Current News

- June 2012
  Owen White and Dirk Gevers discuss the HMP on Wisconsin Public Radio

Microbial Reference Genomes

The HMP plans to sequence, or collect from publicly available sources, a total of 3000 reference genomes isolated from human body sites. The information gained from the reference genomes will aid in taxonomic assignment and functional annotation of 16S rRNA and metagenomic wgs sequence, respectively, from microbiome samples. More information can be found below and on the NIH Common Fund Site.

GET DATA

GET TOOLS

About Reference Genomes

The HMP plans to sequence 3000 genomes from both cultured and uncultured bacteria, plus several viral and small eukaryotic microbes isolated from human body sites. This, in conjunction with reference genomes sequenced by HMP Demonstration Projects and other members of the International Human Microbiome Consortium (IHMC), will supplement the available selection of non-HMP funded human-associated reference genomes to provide a comprehensive pool of genome sequences to aid in the analysis of human metagenomic data.

Domain

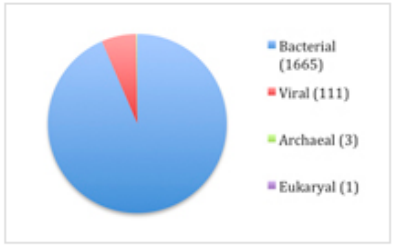


Figure 1. Taxonomic domain of currently ongoing or completed HMP reference genomes. Sequencing efforts are largely bacterial-focused, however we anticipate sequencing of ~1000 viruses, and small numbers of representative archaeal and eukaryotic references.

Project

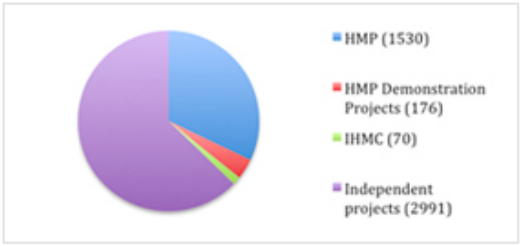
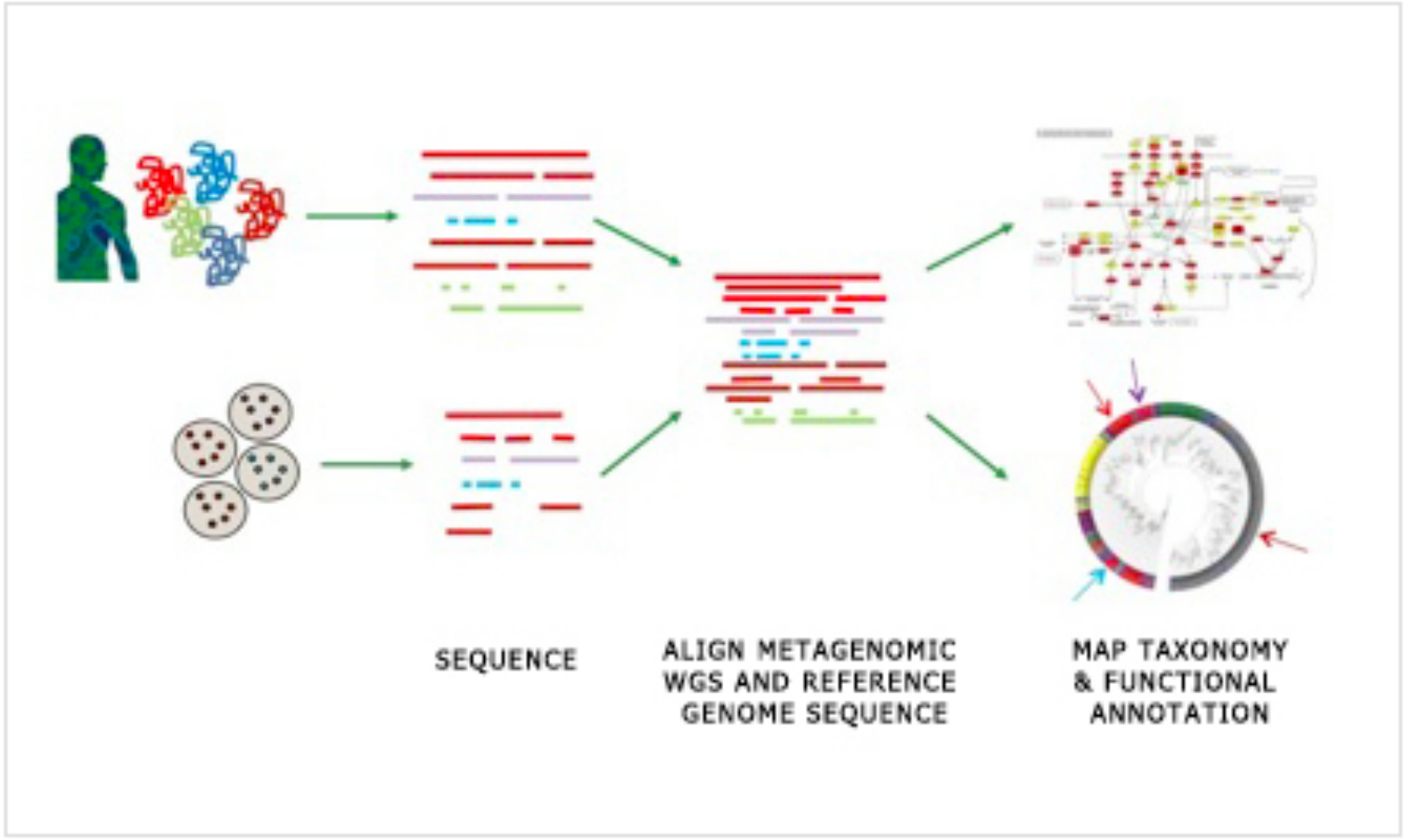


Figure 2. Project affiliation of currently ongoing or completed human-associated reference genomes. Human Microbiome efforts, including the HMP & HMP Demonstration projects, and the IHMC, are contributing considerable numbers of novel reference genome sequence to the pool of human associated genome data available to aid in analysis of metagenomic samples.

Mapping to reference genomes



Kraken – kmer based taxonomic assignment

Wood and Salzberg *Genome Biology* 2014, **15**:R46
http://genomebiology.com/2014/15/3/R46

METHOD Open Access

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood[1,2*] and Steven L Salzberg[2,3]

Kraken method

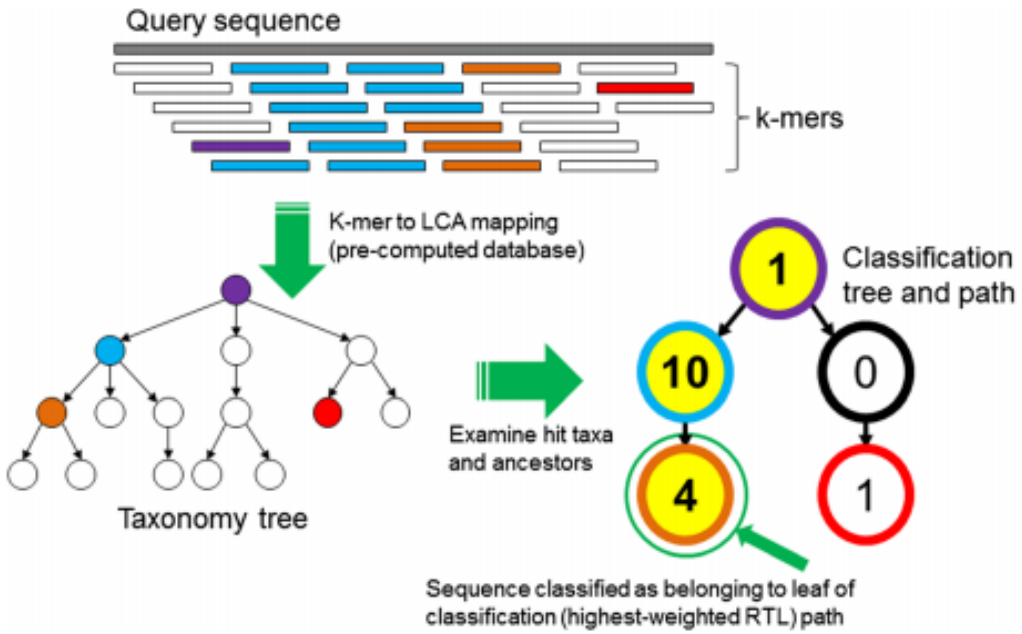
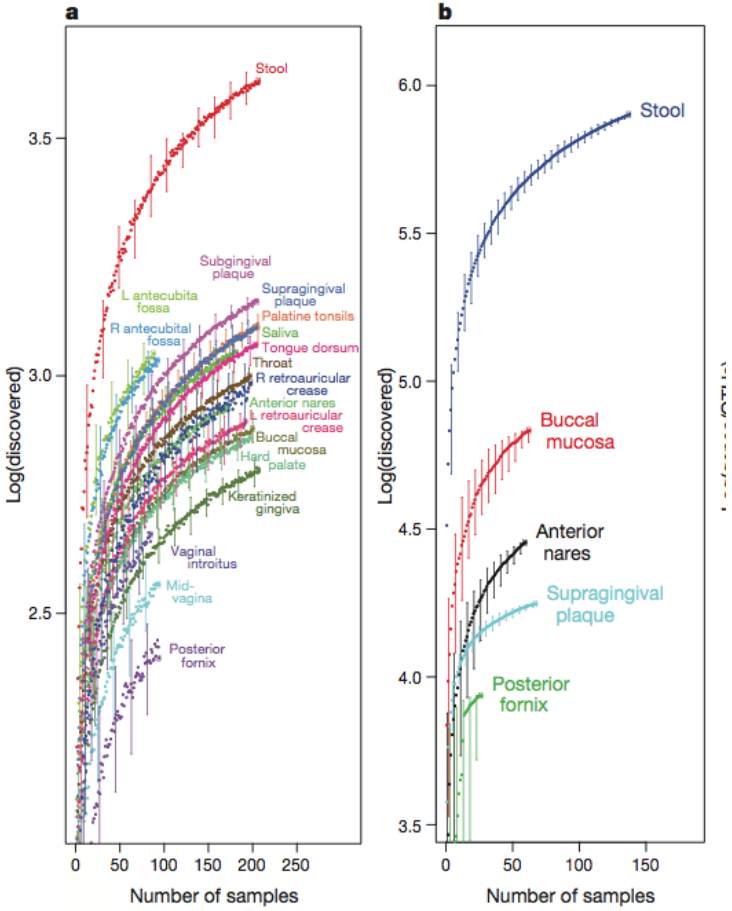
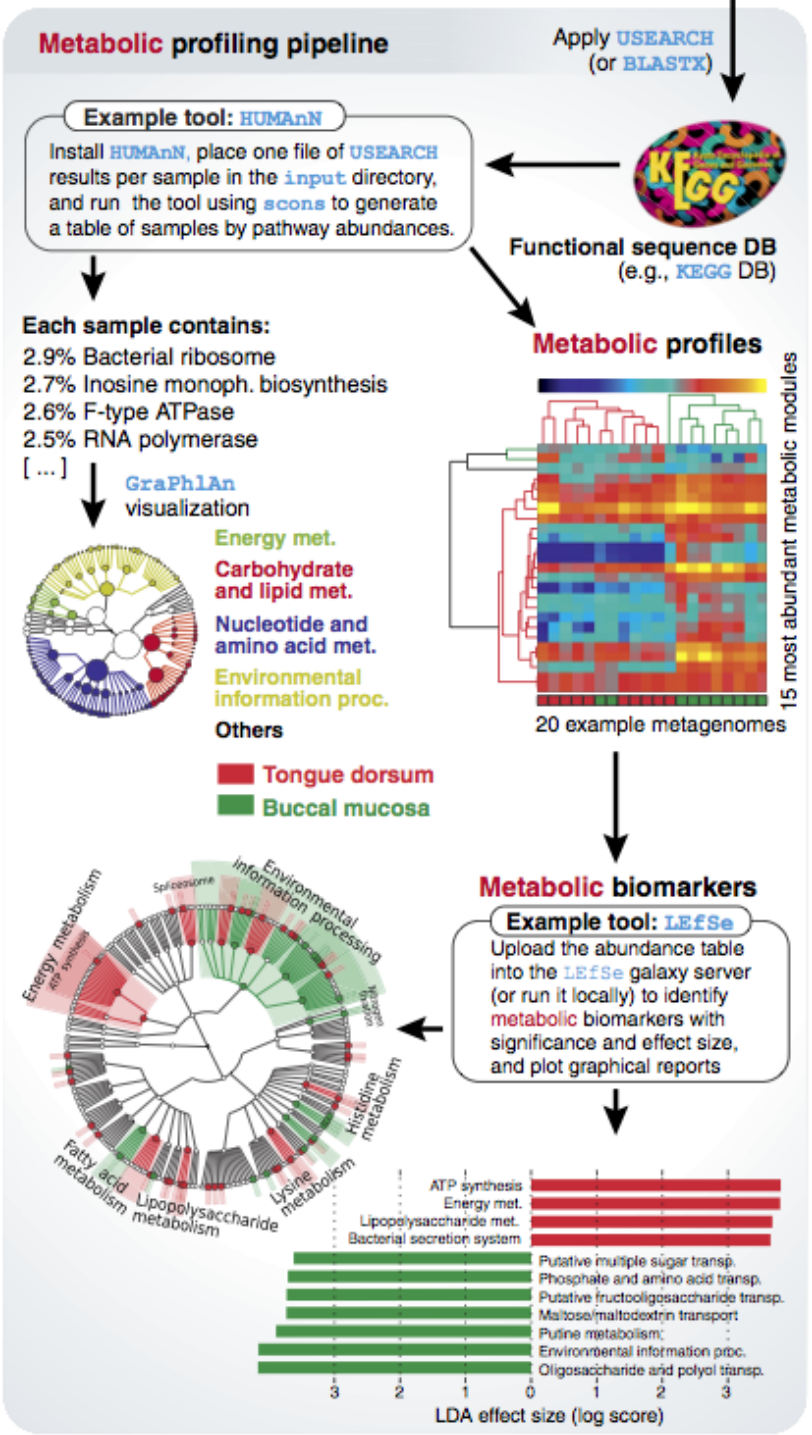


Figure 1 The Kraken sequence classification algorithm. To classify a sequence, each *k*-mer in the sequence is mapped to the lowest common ancestor (LCA) of the genomes that contain that *k*-mer in a database. The taxa associated with the sequence's *k*-mers, as well as the taxa's ancestors, form a pruned subtree of the general taxonomy tree, which is used for classification. In the classification tree, each node has a weight equal to the number of *k*-mers in the sequence associated with the node's taxon. Each root-to-leaf (RTL) path in the classification tree is scored by adding all weights in the path, and the maximal RTL path in the classification tree is the classification path (nodes highlighted in yellow). The leaf of this classification path (the orange, leftmost leaf in the classification tree) is the classification used for the query sequence.

Comparing abundance, diversity etc

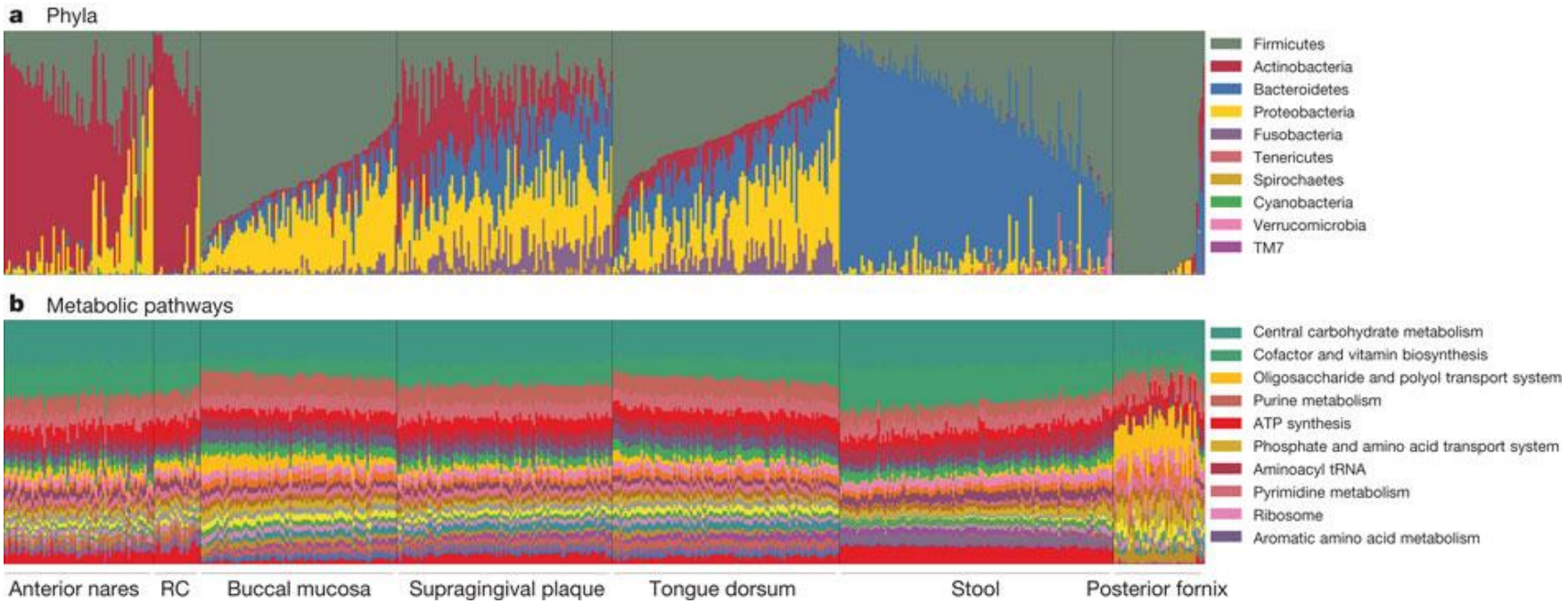


The HMP Consortium,. *Nature* **486**, 215–221 (2012).


Metabolic profiling pipeline
Apply USEARCH (or BLASTX)
Example tool: HUMAnN
Install HUMAnN, place one file of USEARCH results per sample in the input directory, and run the tool using scons to generate a table of samples by pathway abundances.
Functional sequence DB (e.g., KEGG DB)
Each sample contains:
2.9% Bacterial ribosome
2.7% Inosine monoph. biosynthesis
2.6% F-type ATPase
2.5% RNA polymerase
[...]
GraPhlAn visualization
Energy met.
Carbohydrate and lipid met.
Nucleotide and amino acid met.
Environmental information proc.
Others
Metabolic profiles
15 most abundant metabolic modules
20 example metagenomes
Tongue dorsum
Buccal mucosa
Metabolic biomarkers
Example tool: LEfSe
Upload the abundance table into the LEfSe galaxy server (or run it locally) to identify metabolic biomarkers with significance and effect size, and plot graphical reports
Energy metabolism
Environmental information processing
Histidine metabolism
Lysine metabolism
Lipopolysaccharide metabolism
Fatty acid metabolism
ATP synthesis
Energy met.
Lipopolysaccharide met.
Bacterial secretion system
Putative multiple sugar transp.
Phosphate and amino acid transp.
Putative fructooligosaccharide transp.
Maltose/maltodextrin transport
Purine metabolism
Environmental information proc.
Oligosaccharide and polyol transp.
3 2 1 0 1 2 3
LDA effect size (log score)

Functional annotation of metagenomic sequences

- Compare metagenomic sequences to large databases of metabolic annotations



Carriage of microbial taxa varies while metabolic pathways remain stable within a healthy population.

C Huttenhower *et al. Nature* **486**, 207-214 (2012) doi:10.1038/nature11234

Structure, function and diversity of the healthy human microbiome

Summary: Metagenomics pros and cons

Pros

- Samples whole genomes
- Can detect viruses, bacteria, fungus etc
- Novel gene discovery
- Much more scope for downstream analysis

Cons

- More expensive
- Human DNA present (small percent in stool but needs to be filtered out)
- Data analysis more complex
- IT storage and processing more demanding
- Poorer ascertainment of rare species

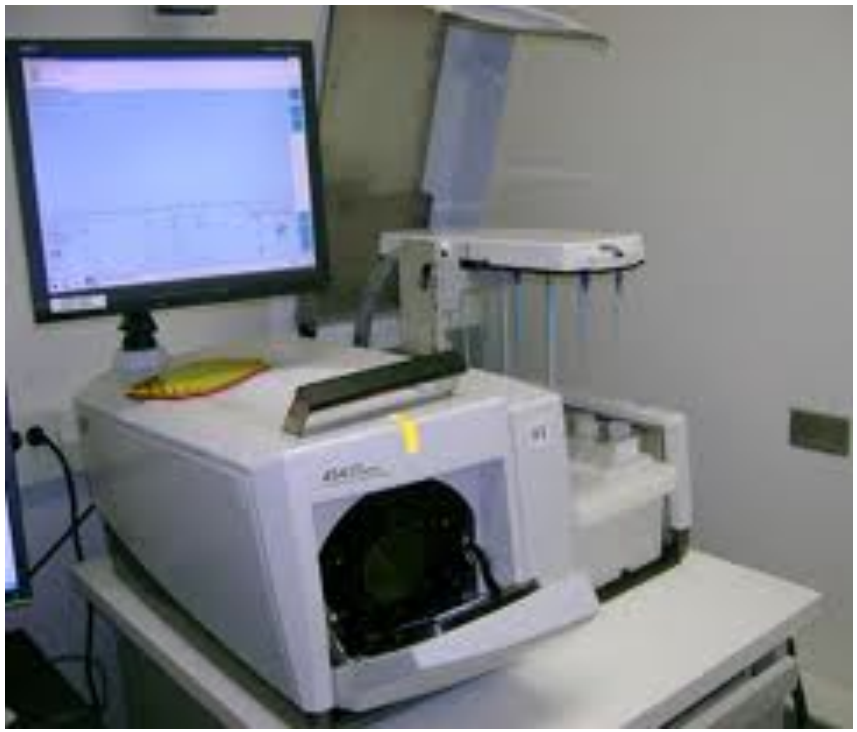
Sequencing technologies – pros and cons

Sanger sequencing – previous state of the art

- Dominant platform until 2005
- Long reads (500-1000 bp)
- Accurate
- High cost per base
- Cloning issues

454 – dominant from 2005-2011

- First nextgen technology
- High throughput pyrosequencing
- Long reads (500-1000 bp)
- Accurate
- High cost per base
- Frameshift error



Illumina MiSeq, HiSeq– current workhorse of microbiome studies

- Introduced in 2006
- Solid phase chain-extension
- Initially very short read (36 bp), now 250bp+ paired end
- Lowest cost per base
- VERY deep coverage possible
- Limited by short read length

PacBio, (Oxford Nanopore) – pretenders to the throne?

- Single molecule sequencing – no PCR amplification
- Medium cost per base
- Long reads
- Small number of reads (1000s rather than 1000000s) mean reduced depth

David et al. Genome Biology 2014, 15:R89
http://genomebiology.com/2014/15/7/R89

Genome Biology

RESEARCH Open Access

Host lifestyle affects human microbiota on daily timescales

Lawrence A David[1,2,11], Arne C Materna[3], Jonathan Friedman[4], Maria I Campos-Baptista[5], Matthew C Blackburn[6], Allison Perrotta[7], Susan E Erdman[8] and Eric J Alm[4,7,9,10*]

Abstract

Background: Disturbance to human microbiota may underlie several pathologies. Yet, we lack a comprehensive understanding of how lifestyle affects the dynamics of human-associated microbial communities.

Results: Here, we link over 10,000 longitudinal measurements of human wellness and action to the daily gut and salivary microbiota dynamics of two individuals over the course of one year. These time series show overall microbial communities to be stable for months. However, rare events in each subjects' life rapidly and broadly impacted microbiota dynamics. Travel from the developed to the developing world in one subject led to a nearly two-fold increase in the Bacteroidetes to Firmicutes ratio, which reversed upon return. Enteric infection in the other subject resulted in the permanent decline of most gut bacterial taxa, which were replaced by genetically similar species. Still, even during periods of overall community stability, the dynamics of select microbial taxa could be associated with specific host behaviors. Most prominently, changes in host fiber intake positively correlated with next-day abundance changes among 15% of gut microbiota members.

Conclusions: Our findings suggest that although human-associated microbial communities are generally stable, they can be quickly and profoundly altered by common human actions and experiences.

Here, we address the dearth of coupled longitudinal datasets of human lifestyle and microbiota by tracking individuals and their commensal microbial communities each day over the course of 1 year. To let subjects comprehensively record their daily lives, we equipped them with iOS devices and a diary app that we configured to simplify personal record keeping. Paired with a simple diet record parsing algorithm that we wrote, this app allowed subjects to record data each day across 349 health and lifestyle variables spanning fitness, diet, exercise, bowel movements, mood, and illness (see Additional file 1 for a full list of measured variables). Even with our streamlined diary tools, we anticipated self-tracking to be inconvenient, and so we screened for study participants who would reliably collect daily records. Our screening yielded a small cohort of two healthy, unrelated male volunteers (Subjects A and B; see Additional file 2 for more demographic information). Yet, despite this small cohort size, the 10,124 measurements of subjects' daily activity collected over the course of 1 year provides an unprecedented window into the health and lifestyle factors potentially regulating human-associated microbial environments.

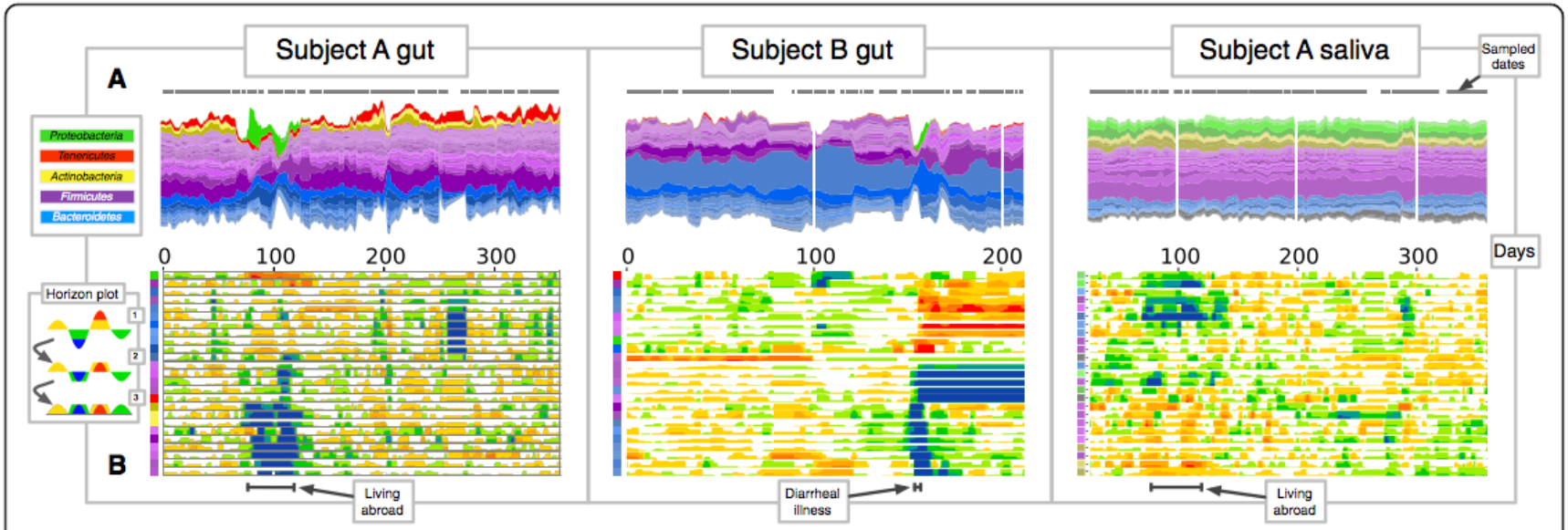


Figure 1 Gut and salivary microbiota dynamics in two subjects over 1 year. (A) Stream plots showing OTU fractional abundances over time. Each stream represents an OTU and streams are grouped by phylum: Firmicutes (purple), Bacteroidetes (blue), Proteobacteria (green), Actinobacteria (yellow), and Tenericutes (red). Stream widths reflect relative OTU abundances at a given time point. Sampled time points are indicated with gray dots over each stream plot. **(B)** Horizon graphs of most common OTUs' abundance over time. Horizon graphs enable rapid visual comparisons between numerous time series [21]. Graphs are made by first median-centering each OTU time series and dividing the curve into colored bands whose width is the median absolute deviation (Inset, step 1). Next, the colored bands are overlaid (step 2) and negative values are mirrored upwards (step 3). Thus, warmer regions indicate date ranges where a taxon exceeds its median abundance, and cooler regions denote ranges where a taxon falls below its median abundance. Colored squares on the vertical axis correspond to stream colors in **(A)**. Time series in both the stream plots and horizon graphs were smoothed using Tukey's running median. Lower black bars span Subject A's travel abroad (days 71 to 122) and Subject B's *Salmonella* infection (days 151 to 159).

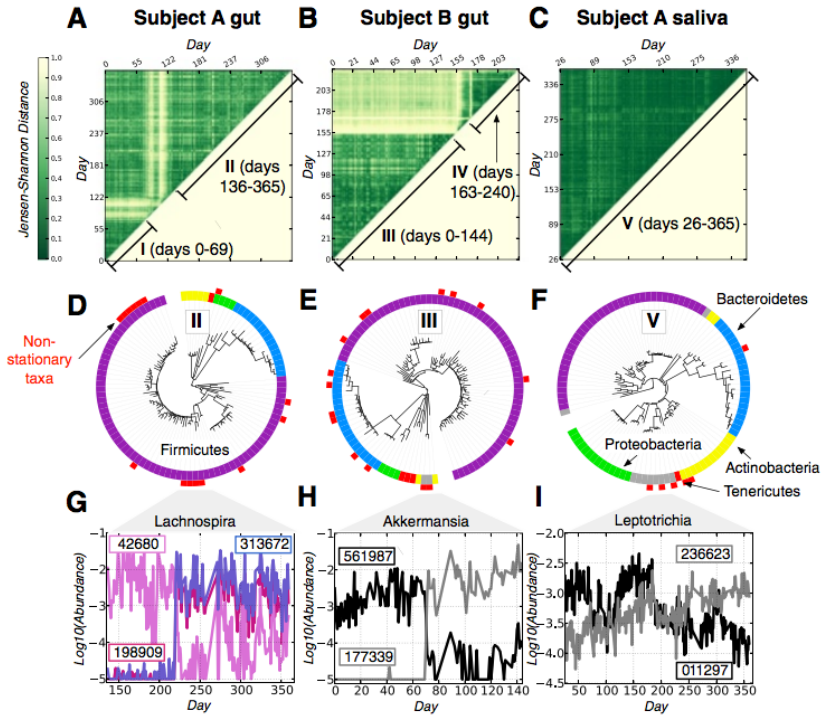


Figure 2 Stability testing of gut and saliva microbiota time series. (A-C) Pairwise Jensen Shannon Distances between samples from Subject A's gut **(A)**, Subject B's gut **(B)**, and Subject A's saliva **(C)**. Dark green regions indicate date ranges with similar microbiota. To quantify how stable individual microbial taxa were across the labeled date ranges, we performed the Augmented Dickey Fuller (ADF) test, which evaluated the null hypothesis that a given OTU is non-stationary (that is, the OTU tends to return to an equilibrium value). The majority of tested OTUs were stationary according to the ADF test (88%, 85%, 84%, 79%, and 94% for date ranges I-V, $P < 0.05$). **(D-F)** Phylogeny of stationary and non-stationary OTUs. Inner rings denote phyla (the Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria and Tenericutes are colored purple, blue, green, yellow, and red, respectively). Outer rings are white for stationary OTUs and red for non-stationary ones. Non-stationary taxa clustered phylogenetically for date ranges II **(D)**, III **(E)**, and V **(F)** ($P < 0.05$, P-test), supporting the hypothesis that closely-related taxa are more likely to be in competition. **(G-I)** Time series of closely-related, non-stationary OTUs (Greengenes prokMSA ids given in boxes). An artificial abundance floor of 1e-5 was added to improve visibility. Shown are members of the genus *Lachnospira* over date range II **(G)**, the genus *Akkermansia* over date range III **(H)**, and the genus *Leptotrichia* over date range V **(I)**. The summed abundances of the selected *Lachnospira* and *Leptotrichia* are stationary over the given date ranges ($P < 0.05$, ADF test).

Augmented Dickey–Fuller test

From Wikipedia, the free encyclopedia

In statistics and econometrics, an **augmented Dickey–Fuller test (ADF)** is a test for a unit root in a time series sample. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models. The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.[1]

Testing procedure [edit]

The testing procedure for the ADF test is the same as for the Dickey–Fuller test but it is applied to the model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

where α is a constant, β the coefficient on a time trend and p the lag order of the autoregressive process. Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modelling a random walk and using the constraint $\beta = 0$ corresponds to modelling a random walk with a drift. Consequently, there are three main versions of the test, analogous to the ones discussed on Dickey–Fuller test (see that page for a discussion on dealing with uncertainty about including the intercept and deterministic time trend terms in the test equation.)

By including lags of the order *p* the ADF formulation allows for higher-order autoregressive processes. This means that the lag length *p* has to be determined when applying the test. One possible approach is to test down from high orders and examine the *t*-values on coefficients. An alternative approach is to examine information criteria such as the Akaike information criterion, Bayesian information criterion or the Hannan–Quinn information criterion.

The unit root test is then carried out under the null hypothesis $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. Once a value for the test statistic

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

is computed it can be compared to the relevant critical value for the Dickey–Fuller Test. If the test statistic is less (this test is non symmetrical so we do not consider an absolute value) than the (larger negative) critical value, then the null hypothesis of $\gamma = 0$ is rejected and no unit root is present.

Intuition [edit]

The intuition behind the test is that if the series is not integrated then the lagged level of the series (y_{t-1}) will provide no relevant information in predicting the change in y_t besides the one obtained in the lagged changes (Δy_{t-k}). In that case the $\gamma = 0$ null hypothesis is not rejected.

Examples [edit]

A model that includes a constant and a time trend is estimated using sample of 50 observations and yields the DF_τ statistic of –4.57. This is more negative than the tabulated critical value of –3.50, so at the 95 per cent level the null hypothesis of a unit root will be rejected.

Critical values for Dickey–Fuller *t*-distribution.

	Without trend		With trend	
Sample size	1%	5%	1%	5%
T = 25	–3.75	–3.00	–4.38	–3.60
T = 50	–3.58	–2.93	–4.15	–3.50
T = 100	–3.51	–2.89	–4.04	–3.45
T = 250	–3.46	–2.88	–3.99	–3.43
T = 500	–3.44	–2.87	–3.98	–3.42
T = ∞	–3.43	–2.86	–3.96	–3.41

Source[2]:373

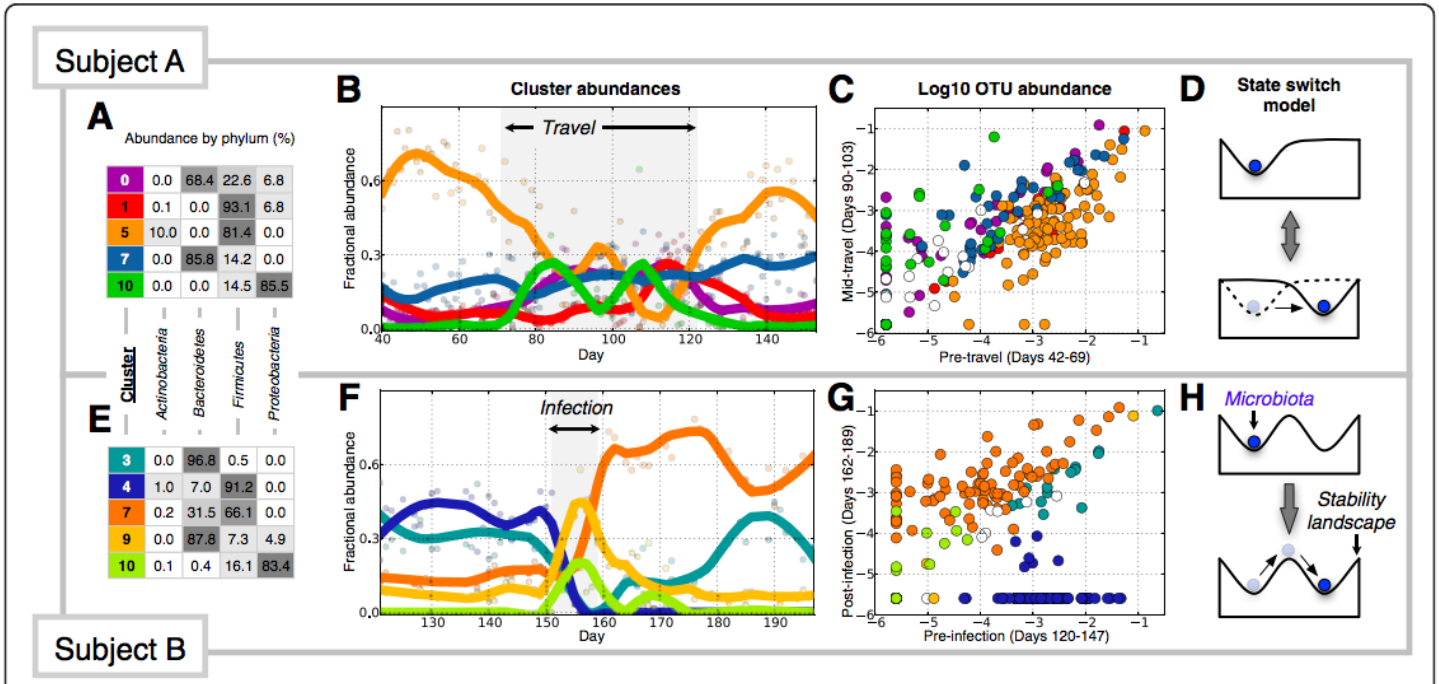


Figure 3 Dynamics of major OTU clusters across major perturbations. Highly abundant OTUs were clustered by their dynamics across Subject A's travel period **(A-D)** and Subject B's acute enteric infection **(E-H)**. Clusters were produced separately for the two environments; see Methods for more details. **(A,E)** Taxonomic composition of major clusters (fractional abundance exceeds 10% for more than 3 days). **(B,F)** Cluster abundances over time (shaded points) and trend lines (solid) fit using LOESS smoothing, colored using the same scheme as in **(A,E)**. Subject A's travel abroad (days 71 to 122) and Subject B's enteric infection (days 151 to 159) are shaded in gray. **(C,G)** Median $\log_{10}$(abundance) of OTUs in each cluster before and after perturbation. OTUs are colored by cluster membership, except for uncolored OTUs belonging to clusters not plotted in **(A,E)**. OTU detection limits were set to the minimum fractional abundance observed in each subject's time series (1e-5.8 for Subject A and 1e-5.6 for Subject B). **(D,H)** Cartoon of microbiome state models, in which microbiota are considered to be balls in a landscape shaped by environmental factors [24]. Subject A's travel-related microbiota shift is consistent with a model where environmental disturbances cause state changes **(D)**, while Subject B's infection-related shift is consistent with state transitions caused by direct community perturbations **(H)**.

Table 1 Significant correlations between Subject A's metadata and microbiota

Body site	Lag (days)	Host factor	Representative OTUs (n)	ρ	*P* value	Abun. (%)	Cluster ID	Total OTUs
Subject A Gut	0	Stool: Hardness	*Eggerthella/Clostridium* (11)	-0.30	1.0E-06	0.2144	10	23
	0	Stool: Time Of Day	*Eggerthella/Clostridium* (11)	0.27	7.4E-06	0.2144	10	23
	1	Nutrition: Fiber	*Clostridium* (6)	-0.38	7.4E-06	0.0442	6	9
	1	Nutrition: Fiber	Ruminococcaceae/*F. prausnitzii* (4)	-0.44	1.1E-07	0.3745	8	8
	1	Nutrition: Fiber	*Eggerthella/Clostridium* (11)	-0.39	3.3E-06	0.2144	10	23
	1	Nutrition: Fiber	*Ruminococcus/R. gnavus/Clostridium* (4)	-0.51	2.9E-10	0.5479	51	12
	1	Nutrition: Fiber	*Ruminococcus/R. gnavus/Clostridium* (5)	-0.51	3.8E-10	0.3495	52	7
	1	Nutrition: Fiber	*Blautia* (3)	-0.38	7.6E-06	0.0346	53	3
	1	Nutrition: Fiber	Bifidobacteriales (13)	0.36	1.6E-05	6.0786	86	13
	1	Nutrition: Fiber	*Coprococcus* (8)	0.44	7.2E-08	4.2192	89	12
	1	Nutrition: Fiber	*Clostridium* (1)	-0.42	4.6E-07	0.0716	111	1
	1	Nutrition: Fiber	*Ruminococcus/R. gnavus/Clostridium* (6)	-0.44	1.2E-07	2.0690	118	14
	1	Nutrition: Fiber	*Roseburia/E. rectale* (30)	0.37	8.4E-06	5.0446	127	40
	1	Food: OrangeJuice	*Clostridium* (1)	0.28	4.7E-06	0.0457	106	2
	1	Food: BreakfastBar	*Ruminococcus/R. gnavus/Clostridium* (4)	-0.27	6.6E-06	0.5479	51	12
	1	Food: BreakfastBar	*Ruminococcus/R. gnavus/Clostridium* (5)	-0.40	2.9E-11	0.3495	52	7
	1	Food: BreakfastBar	Bifidobacteriales (13)	0.27	9.5E-06	6.0786	86	13
	1	Food: BreakfastBar	*Clostridium* (1)	-0.43	5.3E-13	0.0716	111	1
	1	Food: Yogurt	Bifidobacteriales (2)	0.45	2.7E-14	0.0069	85	2
	1	Food: Fruits: Fresh	Clostridiales (4)	-0.27	1.1E-05	0.1866	120	9
	1	Food: Fruits: Citrus	Ruminococcaceae/*F. prausnitzii* (4)	0.36	1.7E-09	1.7152	107	4
	1	Food: Soup	Clostridiales (1)	-0.25	3.3E-05	0.0014	62	2
	1	Food: Soup	*Blautia* (21)	-0.26	2.4E-05	3.8126	68	31
	1	Food: Soup: Other	Clostridiales (1)	-0.27	1.3E-05	0.0014	62	2
	1	Food: Soup: Other	*Blautia* (21)	-0.28	4.2E-06	3.8126	68	31
Subject A Saliva	-7	Exercise: TookPlace	*S. mutans/S. sanguinis* (2)	-0.28	1.6E-05	0.0142	21	2
	1	OralCare: Flossing	*S. mutans/S. sanguinis* (2)	-0.30	2.5E-06	0.0142	21	2
	1	Fitness: BodyFat	*Prevotella* (4)	-0.36	1.4E-06	1.5761	35	14

Non-parametric statistics were used to identify metadata variables significantly correlated with clustered OTUs lagged forward or backwards in time (q <0.05, Spearman correlation). Representative taxonomic names from each cluster are shown along with the percentage of overall reads accounted for by each OTU cluster ('Abundance'). Redundant correlations (for example, yogurt subtypes associated with the same OTU cluster) are shown in Additional file 9, and a full list of taxa associated with each cluster can be found in Additional files 10 and 11. We excluded Subject A's travel period and Subject B's post-infection period from metadata correlation analysis. No significant correlations were found among Subject B's microbiota and metadata.

ARTICLE

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1]*, Jeroen Raes[1,2]*, Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]

"To analyse the feasibility of comparative metagenomics of the human gut across cohorts and protocols and to obtain first insights into commonalities and differences between gut microbiomes across different populations…"

Methods

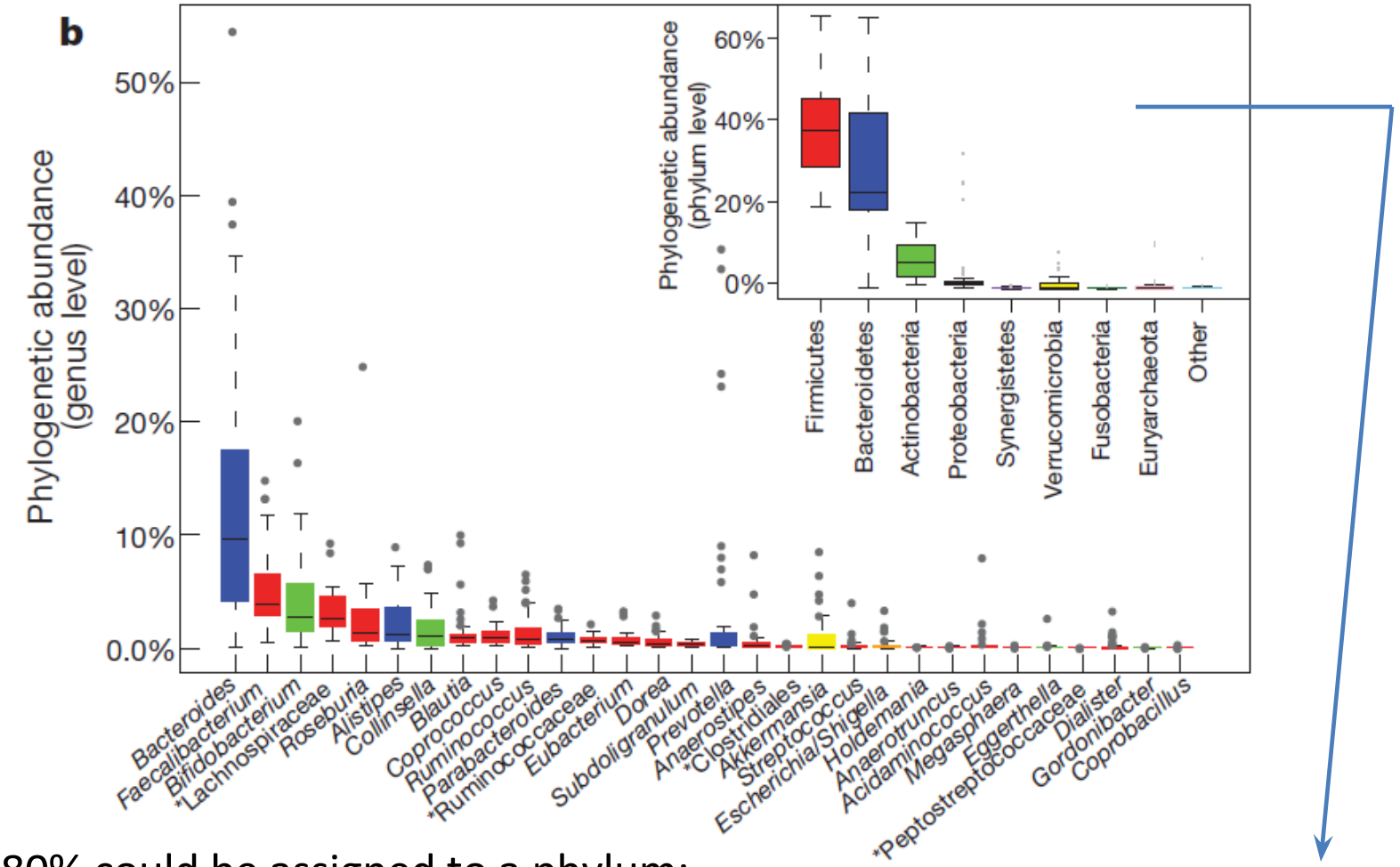
Samples:
Bacterial DNA extracted from fecal samples from Danish, French, Italian and Spanish individuals.
Plus sequence data from Japanese and Americans
(39 individuals in total; final assessment with 33 individuals)

Sequence assembly and alignment:
Sanger sequenced
Assembled using SMASH
Aligned sequences to 1,511 reference genomes
(from NIH Human Microbiome Project and the European MetaHIT consortium) using BLASTN

Alignment Results

- On average 52.8% of sequence fragments could be assigned to a genus;

30.7% not aligned due to strict criteria; 16.5% of reads belong to unknown genera

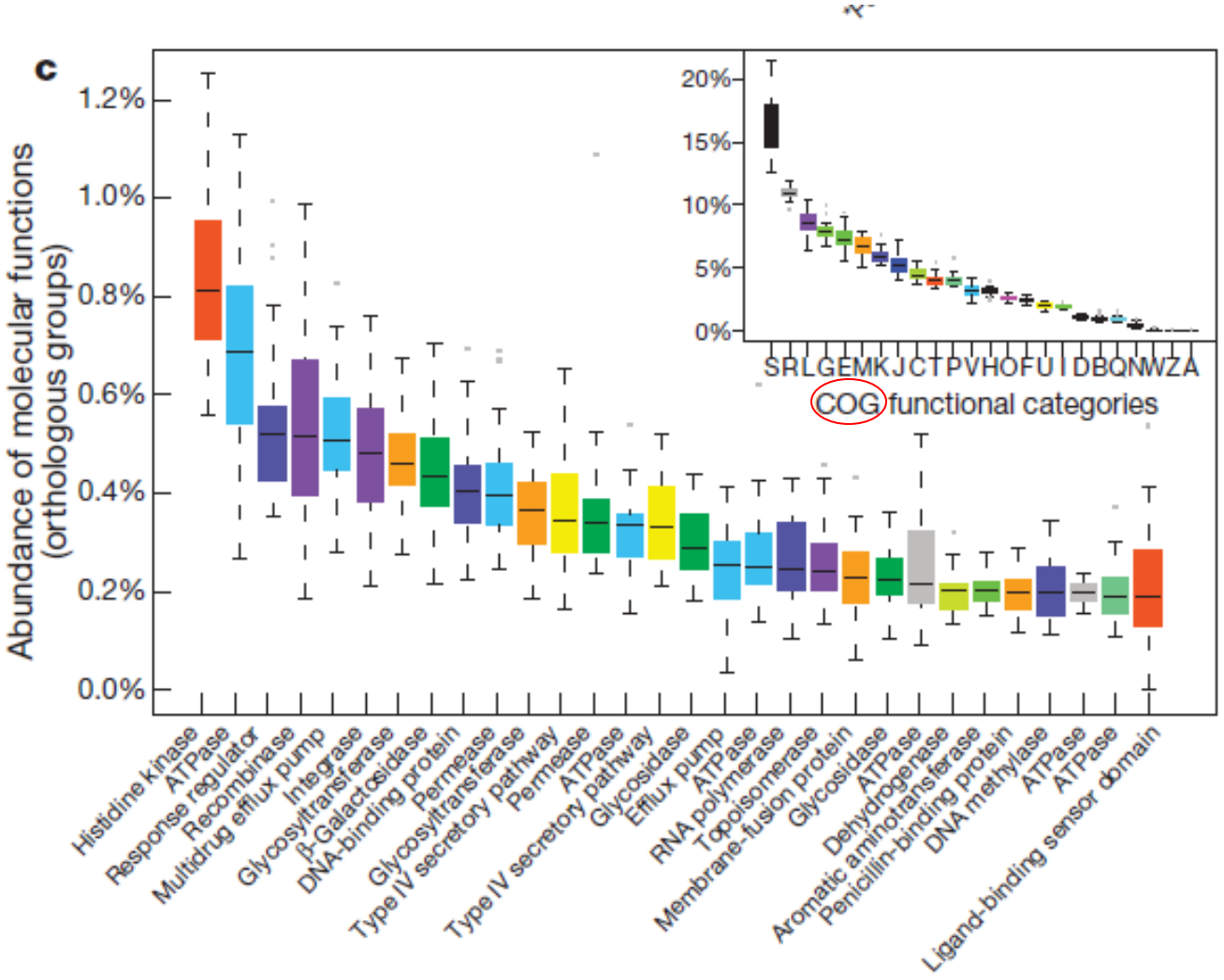


- 80% could be assigned to a phylum;

Firmicutes and Bacteriodetes dominant phyla in human gut microbiome

Alignment Results (2)

Reads aligned to gene orthologs predicted using eggNOG
(evolutionary genealogy of genes: Non-supervised Orthologous Groups)



COG=clusters of orthologous groups

COGs

Functional annotation

NCBI

Code	COGs	Domains	Description	Pathways and functional systems
Information storage and processing				
J	**217**	6449	Translation, ribosomal structure and biogenesis	4
K	**132**	5438	Transcription	3
L	**184**	5337	DNA replication, recombination and repair	2
Cellular processes				
D	**32**	842	Cell division and chromosome partitioning	-
O	**110**	3165	Posttranslational modification, protein turnover, chaperones	-
M	**155**	4079	Cell envelope biogenesis, outer membrane	1
N	**133**	3110	Cell motility and secretion	2
P	**160**	5112	Inorganic ion transport and metabolism	1

Methods cont…

Cluster analysis and PCA:
Clustered by using Jensen-Shannon distance and PAM (partitioning around medoid clustering) to identify microbiome groups by genus and function

Validation:
Results with other datasets, microarray (HITChip) and other clustering algorithms

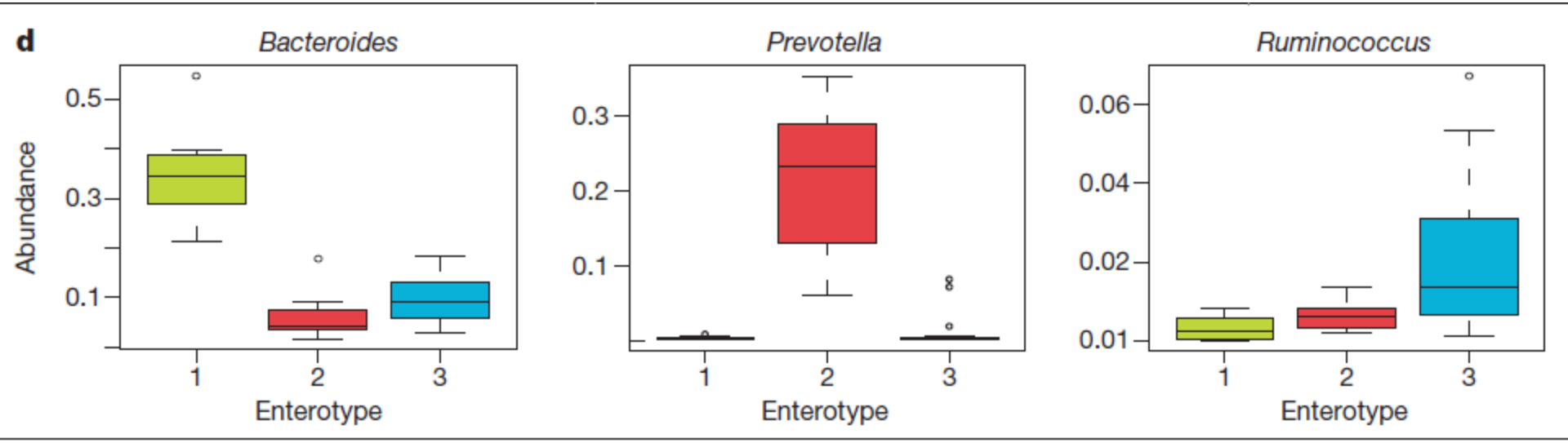
Interpretation of clustering results:
Correlations of phylogenetic or functional classifications by host characteristics (age, gender, nationality, BMI)

Enterotypes

“Multidimensional cluster analysis and principal component analysis (PCA) revealed that the remaining 33 samples formed three distinct clusters that we designate as enterotypes …. (that) .. are identifiable by the variation in the levels of one of three genera: Bacteroides (enterotype 1), Prevotella (enterotype 2) and Ruminococcus (enterotype 3) (Fig. 2a, d) “

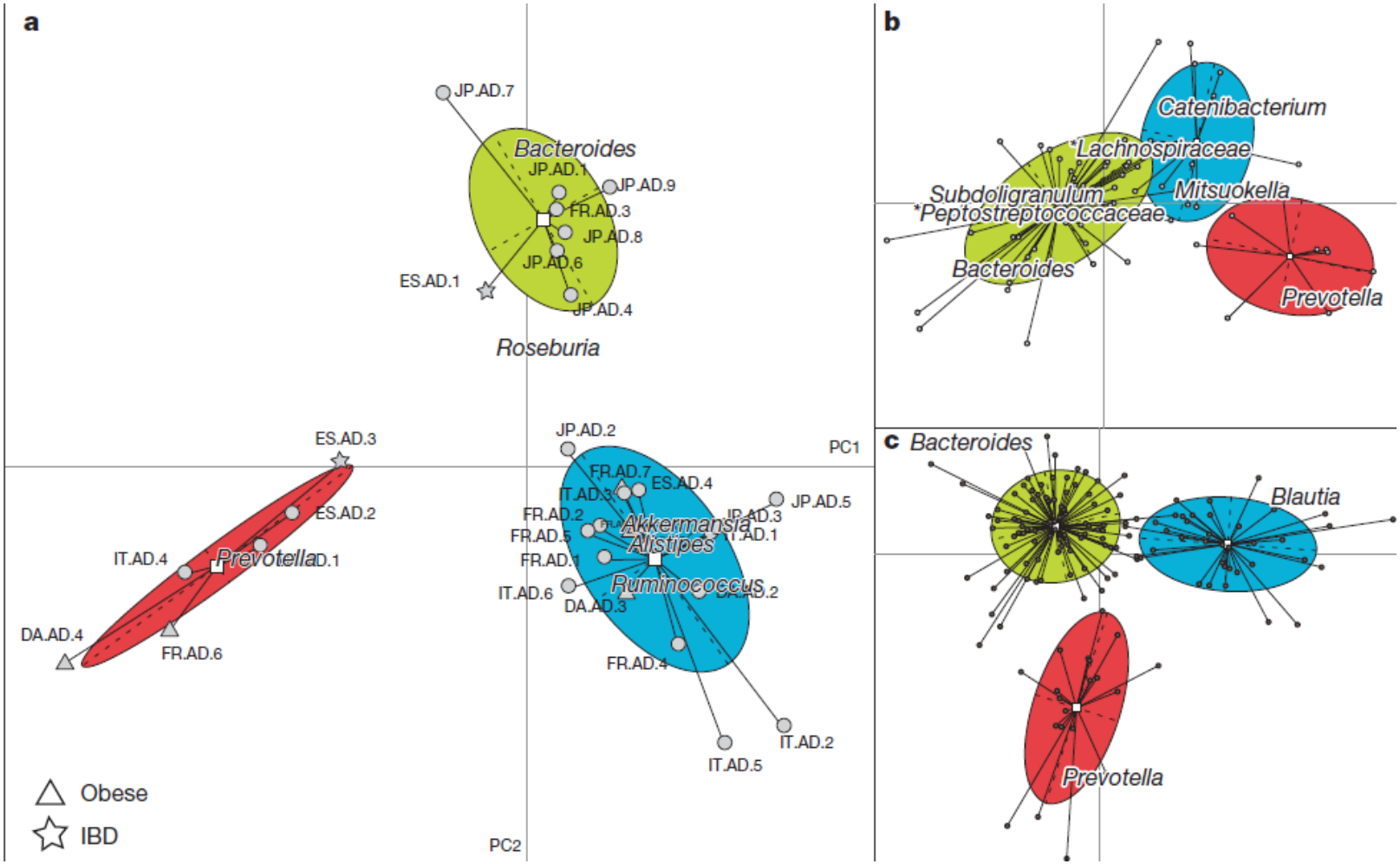
Cluster Analysis Results

The human gut is comprised of 3 enterotypes!!
(identified by prevalence of 1 of 3 genera)



Validation

- 3 enterotypes (clusters) were also observed using two other datasets of 16s pyrosequencing data from 154 Americans and Illumina metagenomic sequence from 85 Danish individuals
- Also validated using HITChip (human intestinal tract microarray)

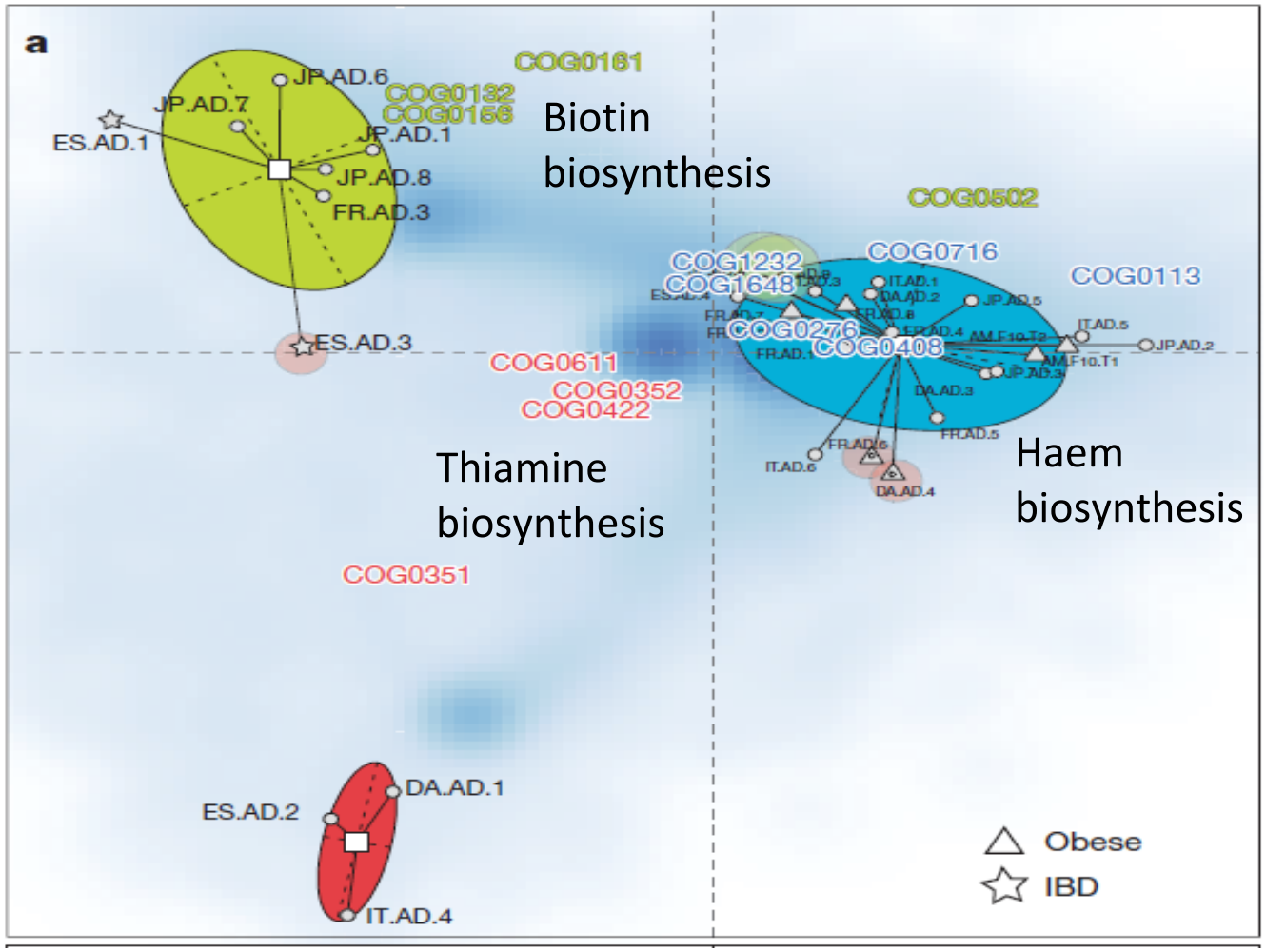


More Cluster Analysis Results

Like phylogenetic clustering, there are also 3 cluster of orthologous groups

Individuals in each cluster are similar to those in the phylogenetic clusters

Function and species composition coincide



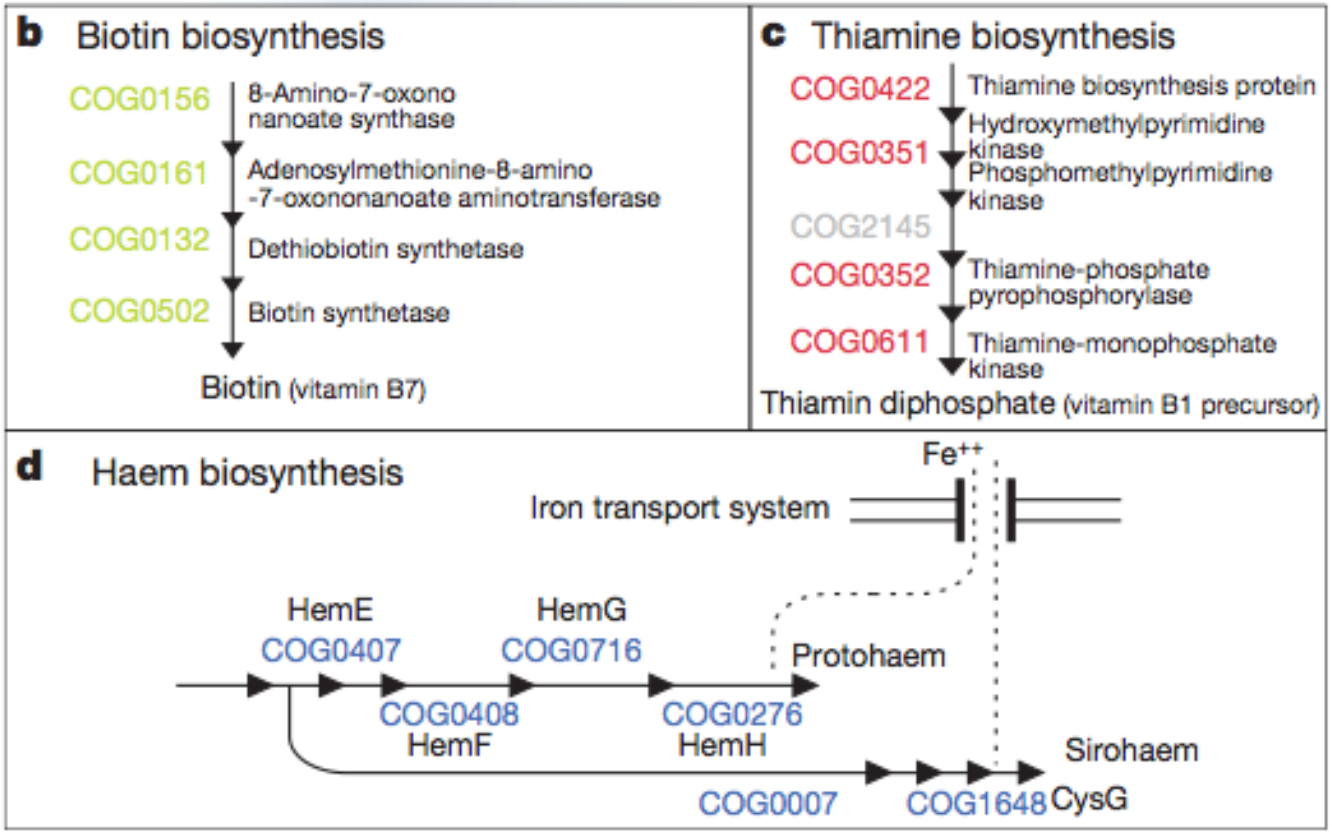


Figure 3 | Functional differences between enterotypes. **a**, Between-class analysis (see Fig. 2) of orthologous group abundances showing only minor disagreements with enterotypes (unfilled circles indicate the differing samples). The blue cloud represents the local density estimated from the coordinates of orthologous groups; positions of selected orthologous groups are highlighted. **b**, Four enzymes in the biotin biosynthesis pathway (COG0132, COG0156, COG0161 and COG0502) are overrepresented in enterotype 1. **c**, Four enzymes in the thiamine biosynthesis pathway (COG0422, COG0351, COG0352 and COG0611) are overrepresented in enterotype 2. **d**, Six enzymes in the haem biosynthesis pathway (COG0007, COG0276, COG407, COG0408, COG0716 and COG1648) are overrepresented in enterotype 3.

Can these data be used as biomarkers?

Phylogenetic based enterotypes do not appear to correlate with nationality, gender, age or body mass index (BMI)

Found correlations between functional groups with age, gender and BMI (not nationality)

Will the microbiome reveal the health status of an individual?

Perhaps….larger more controlled cohorts are needed

The ability to identify enterotypes means the future looks bright!