Challenges in quantitative metagenomics: Accurately counting genes and genera

Manimozhiyan Arumugam

EMBL-Heidelberg

09.03.2011

EMBL

Outline

- Quantitative metagenomics
- Challenges
  - counting genes
  - counting genera
  - prokaryotic taxonomy
- Applications to human gut metagenomics
  - enterotypes of the human gut microbiome
  - functional biomarkers of host properties

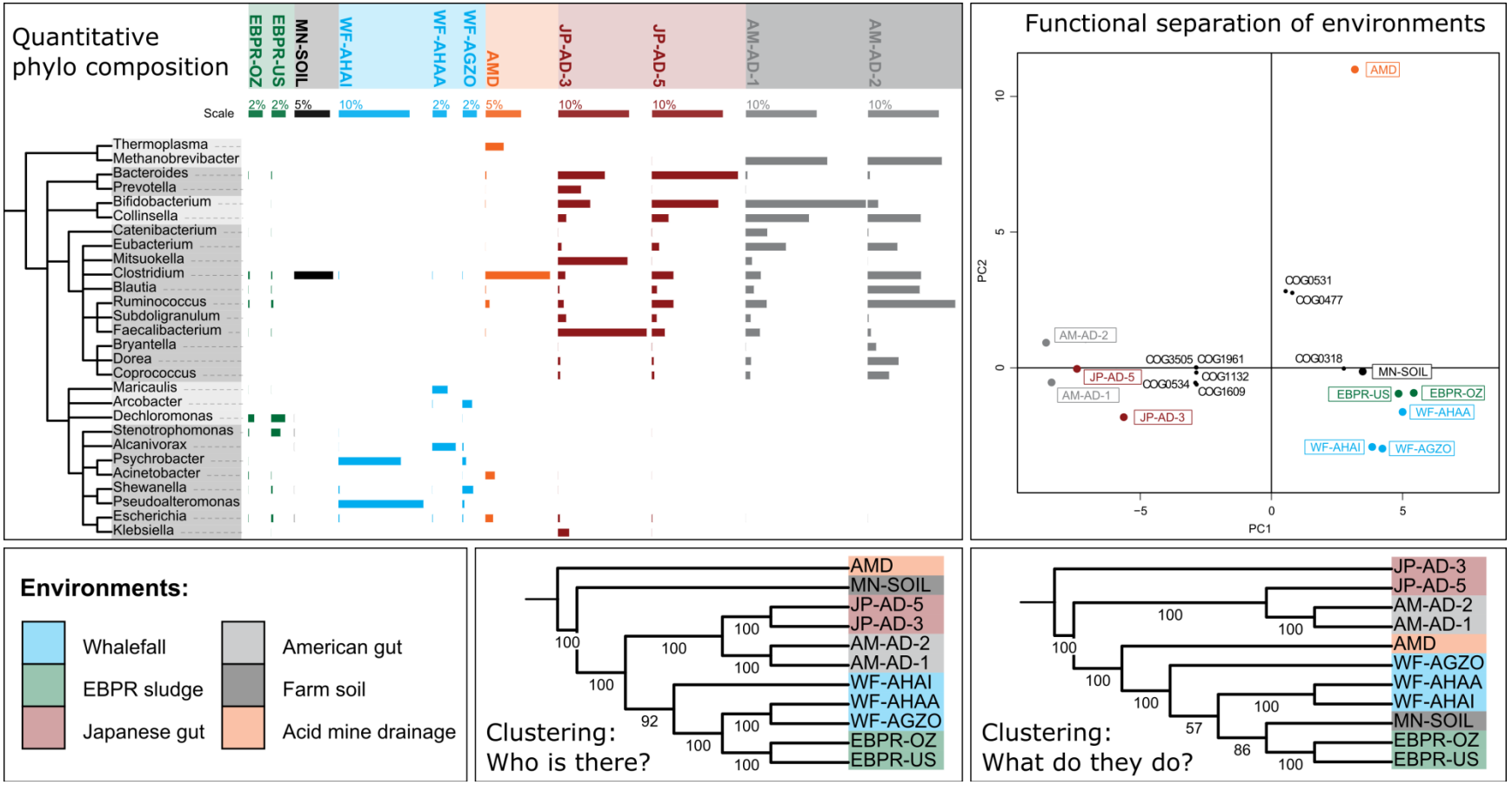
Introduction

What is (quantitative) metagenomics?

- Culture-free characterization of communities through a genomic snapshot
- Enables us to ask the following questions
- Who is there?
  - quantitative: how many cells of each genus/species/strain are in the environment?
- What are they doing?*
  - quantitative: how many instances (copies) of each gene are in the environment?
- Can we compare communities?
  - quantitative: evaluate per amount of sequence/sample (normalization)

*functional potential rather than expression levels

Visualizing quantitative results



Arumugam, *et al*. 2010. *Bioinformatics*. 26(23):2977

Challenges in Quantitative Metagenomics

Counting genes – premises

- Quantitative functional composition
  - number of instances of each functional unit (gene, orthologous group, pathway, etc)
- Representational bias
  - amount of DNA from a gene is proportional to gene length
  - functions are frequently restricted to regions/domains
  - irrelevant promiscuous domains may inflate the counts

Counting genes – our solution

- Identify genes
  - eggNOG orthologous groups
  - KEGG orthologous groups/modules/pathways
- Removing bias
  - normalize for gene length
  - consider the relevant regions/domains only

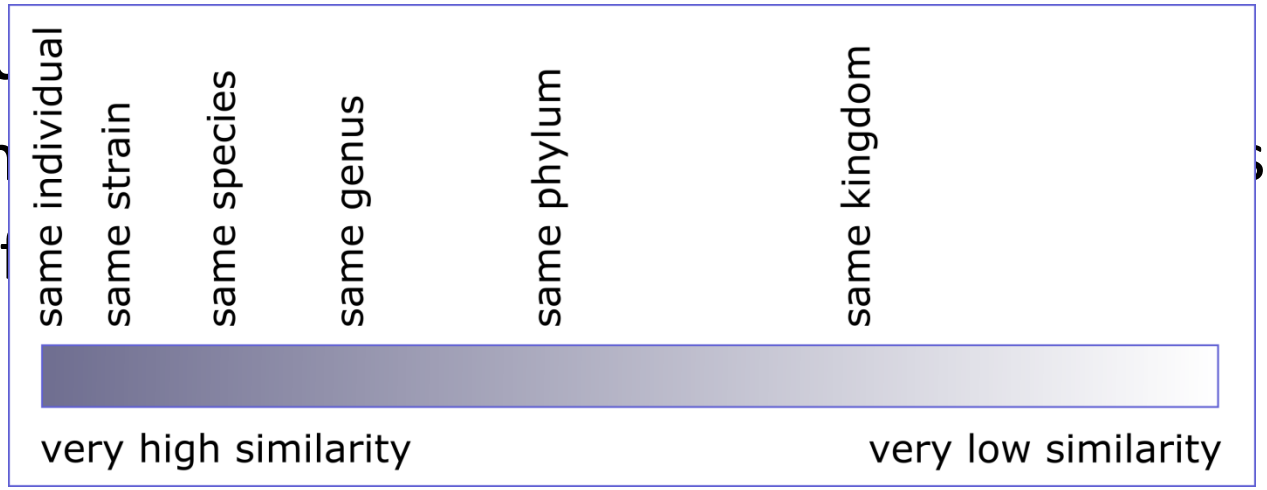
Tringe*, von Mering*, *et al*. 2005. *Science.* 308(5721):554
Arumugam, *et al*. 2010. *Bioinformatics*. 26(23):2977

Counting genera – premises

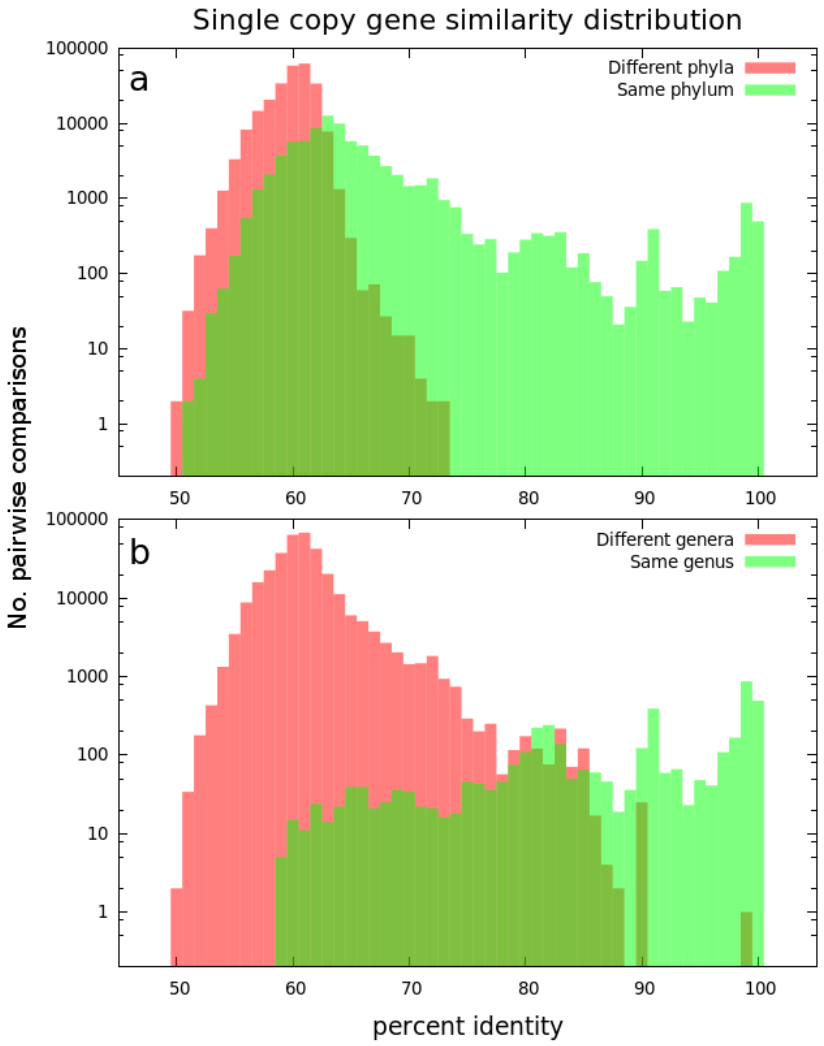
- Quantitative phylogenetic composition
  - number of cells (individuals) of each taxon
  - convert to relative abundance
- Representational bias
  - amount of DNA in metagenome proportional to genome size
  - (number of 16S rDNA reads proportional to 16S copy number)

Counting genera – our solution

- Identify genus of each metagenomic read
  - BLAST against known reference genomes
  - assign genus of best hit
- Best BLAST hit will produce many spurious hits
  - best hit at 50% similarity is most likely NOT the same genus
- Sequ
  - an
  - di



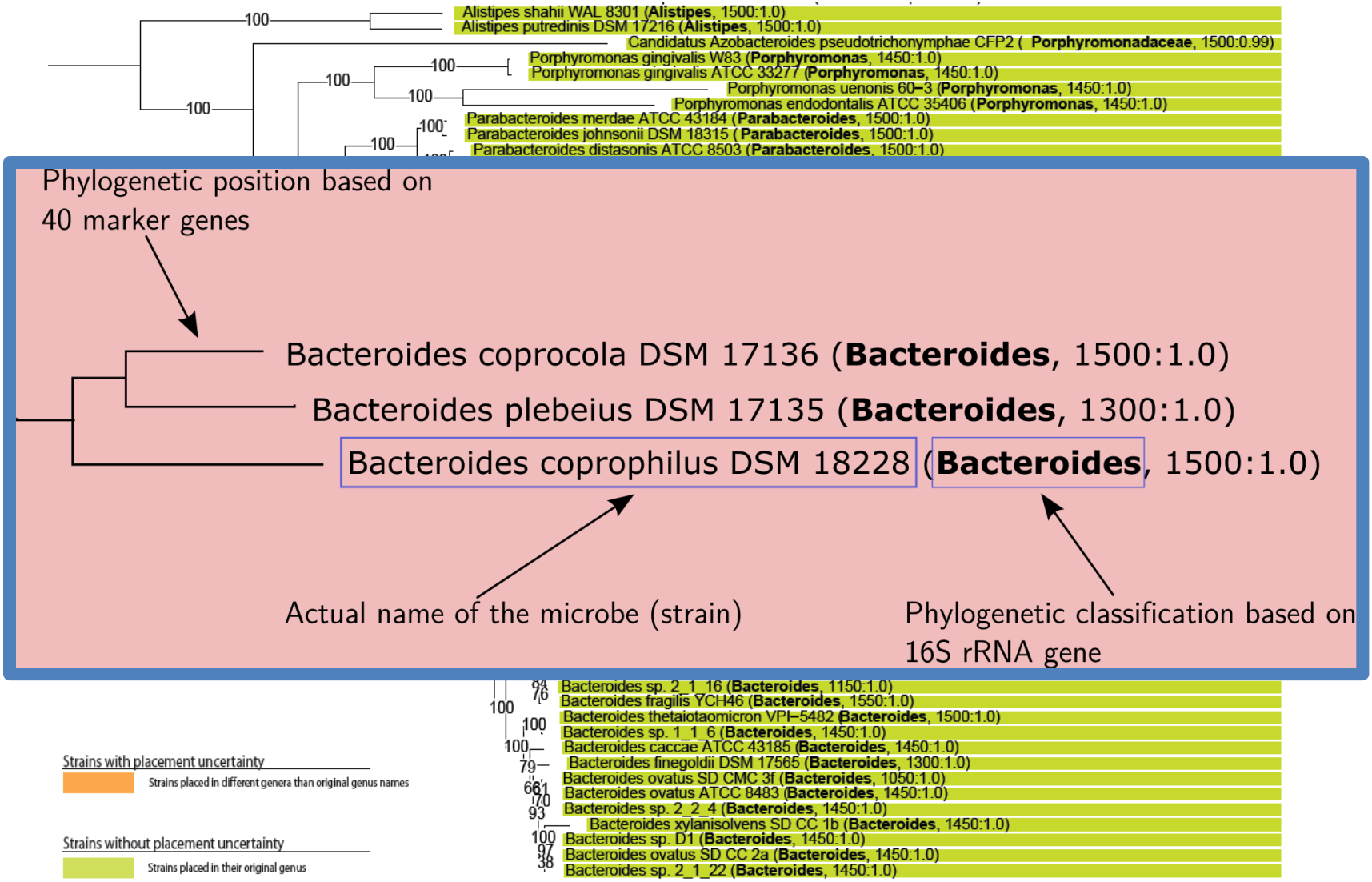
Rank-specific similarity thresholds



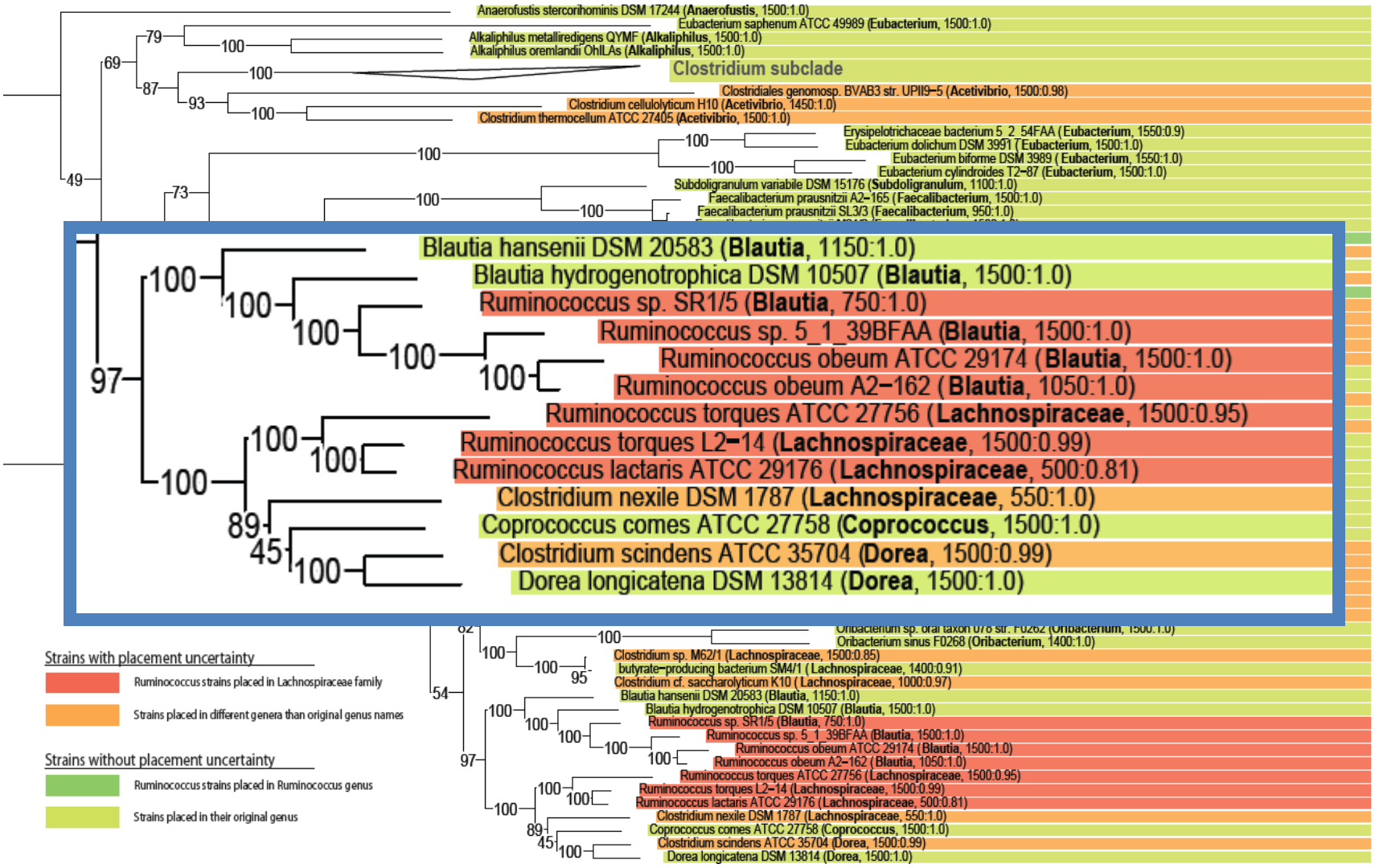
- Estimating similarity thresholds
  - 40 universal single copy marker genes
  - >65% match → phylum assignment
  - >85% match → genus assignment
- Currently exploring:
  - robust validation of thresholds
  - map more reads without compromising accuracy

Challenges in prokaryotic taxonomy through an example

Bacteroidales: a good clade



Clostridiales: a difficult clade



Prokaryotic taxonomy: solution?

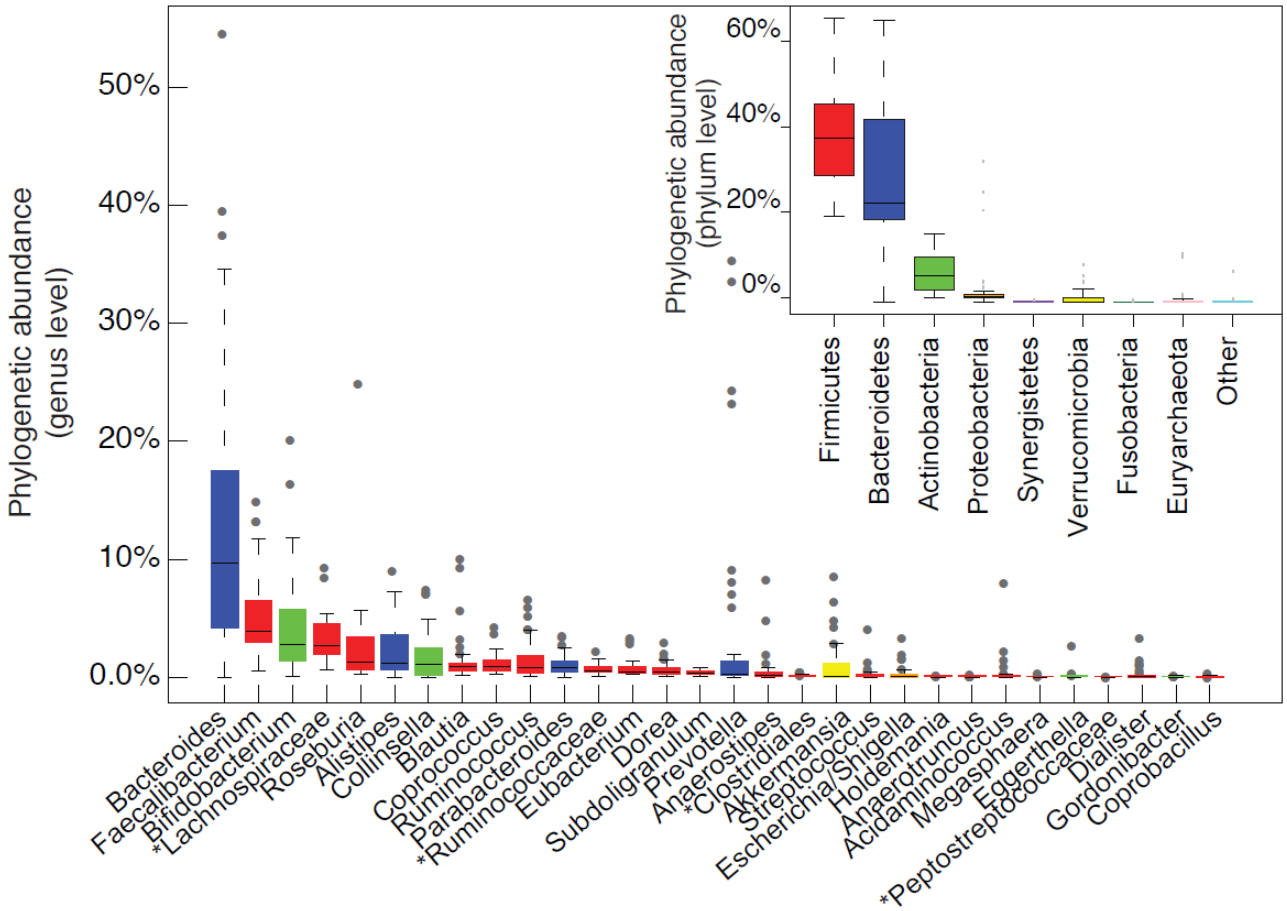
- How to phylogenetically classify a reference genome?
- unreliable: name based classification
- reasonable: 16S rRNA based classification
  - feasible, but limited resolution
  - goes down to genus level
- ideal goal: multiple phylogenetic markers
  - hard, but higher resolution

Enterotypes of the human gut microbiome

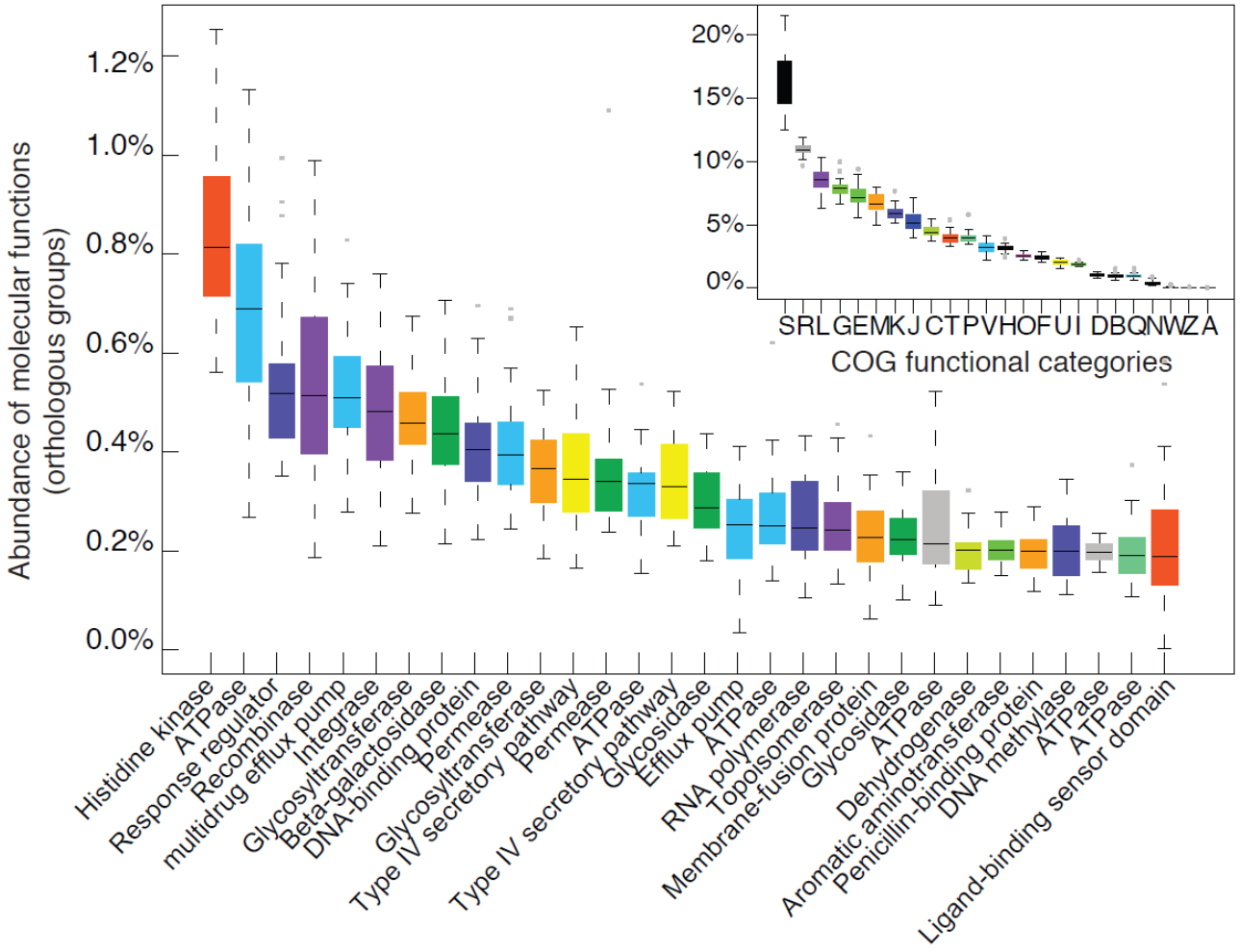
Summary of data

Nationality	Cohort	Technology	Count
American	Turnbaugh et al 09	454 Titanium	2
Japanese	Kurokawa et al 07	Sanger	9
Danish	MetaHIT (obesity)	Sanger	4
Spanish	MetaHIT (IBD)	Sanger	4
French	MicroObes (obesity)	Sanger	8
Italian	MicroAge (>70 yrs)	Sanger	6
6 countries	6 cohorts	2 technologies	33 samples

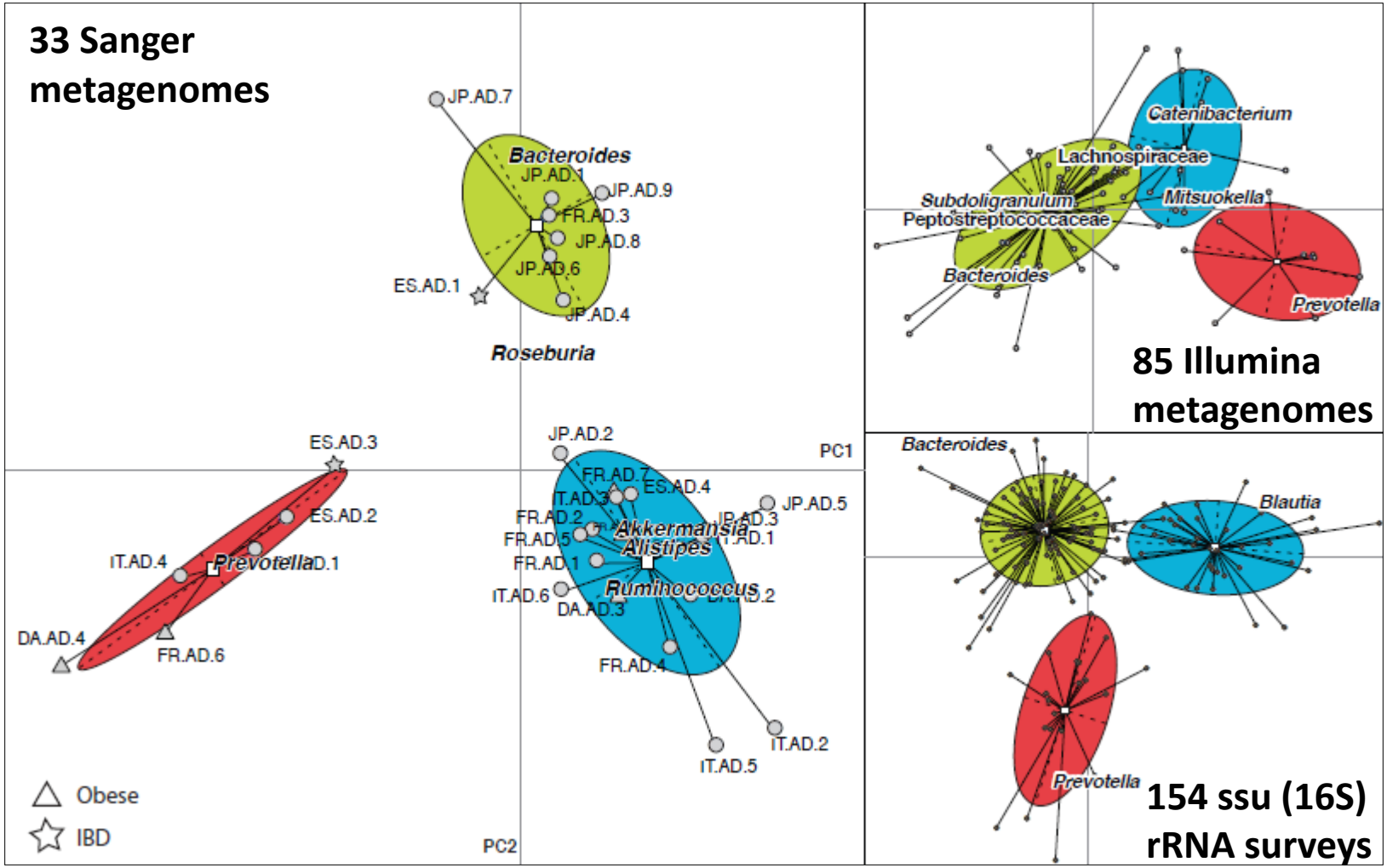
Who is there?
Bacteroidetes & Firmicutes



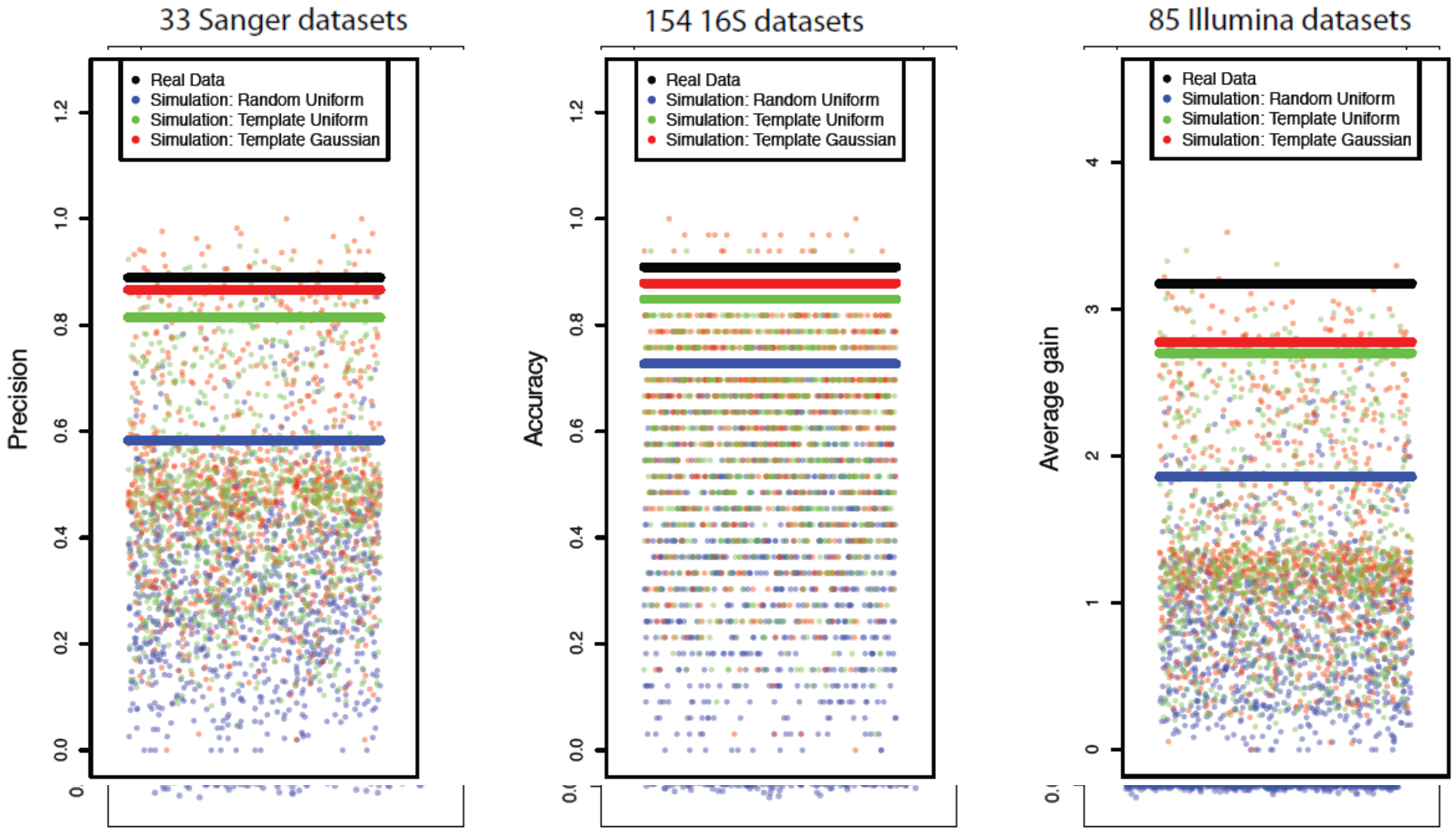
What do they do?
Signaling, drug resistance, sugar utilization



Can we compare gut communities? Enterotypes in the gut microbiota

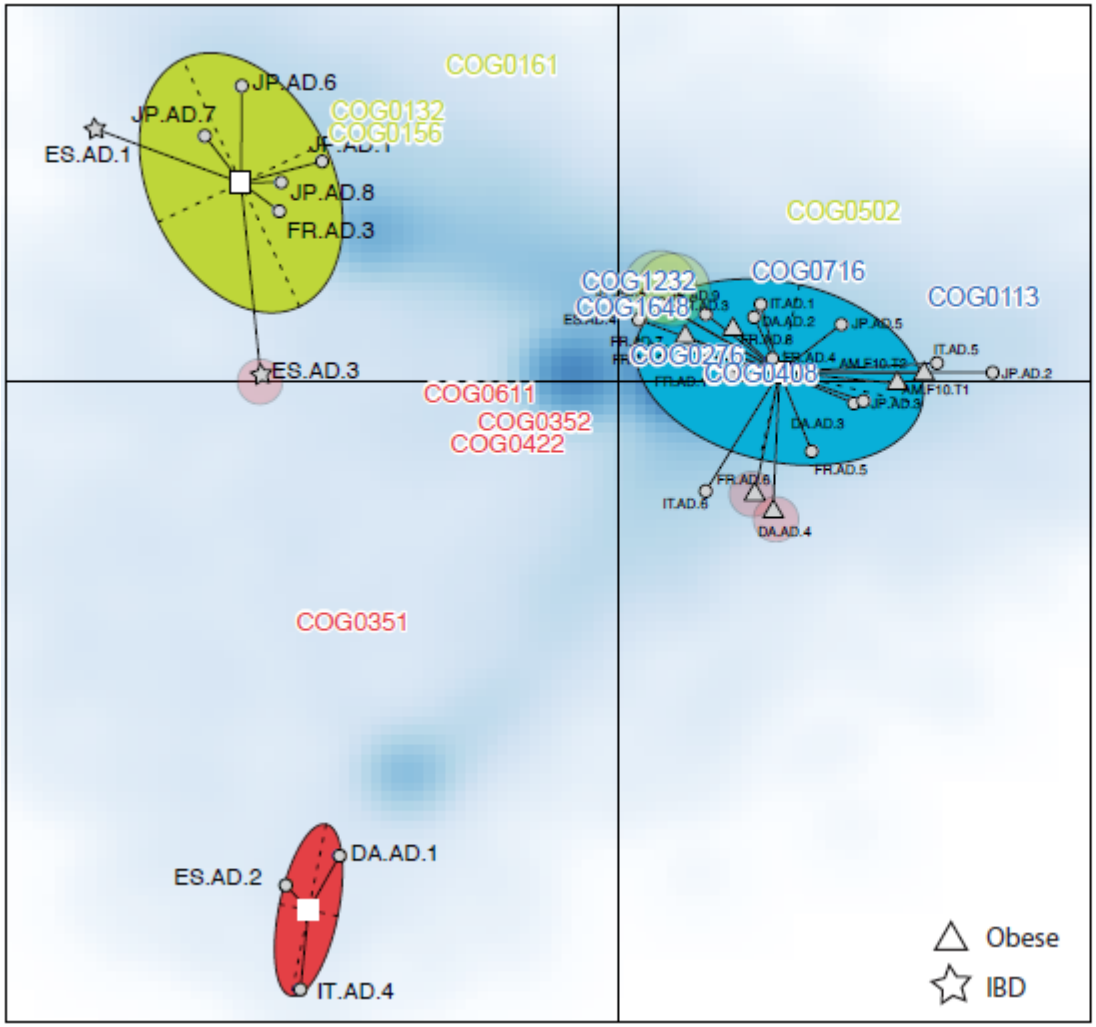


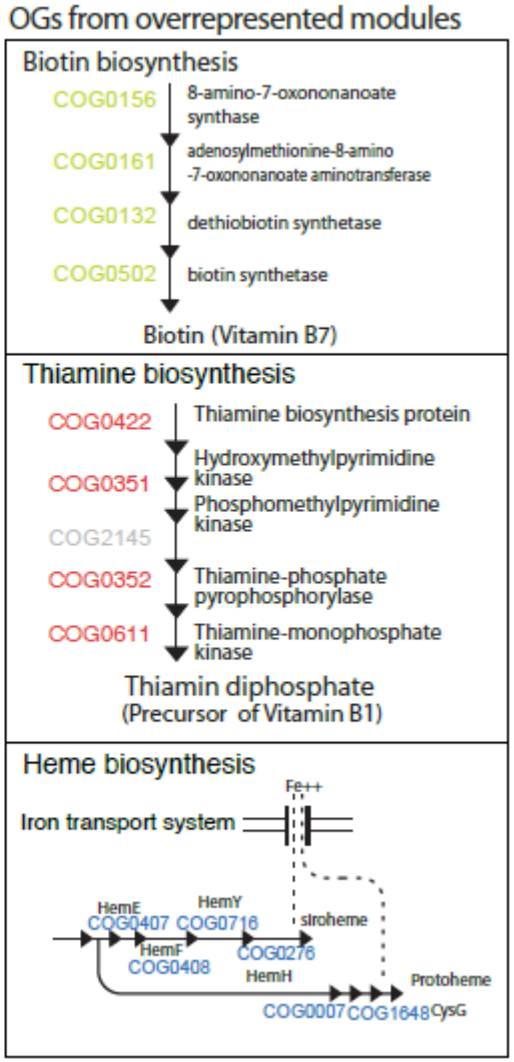
Validating enterotypes



3 clusters, 3 types of simulations

Functional separation of enterotypes

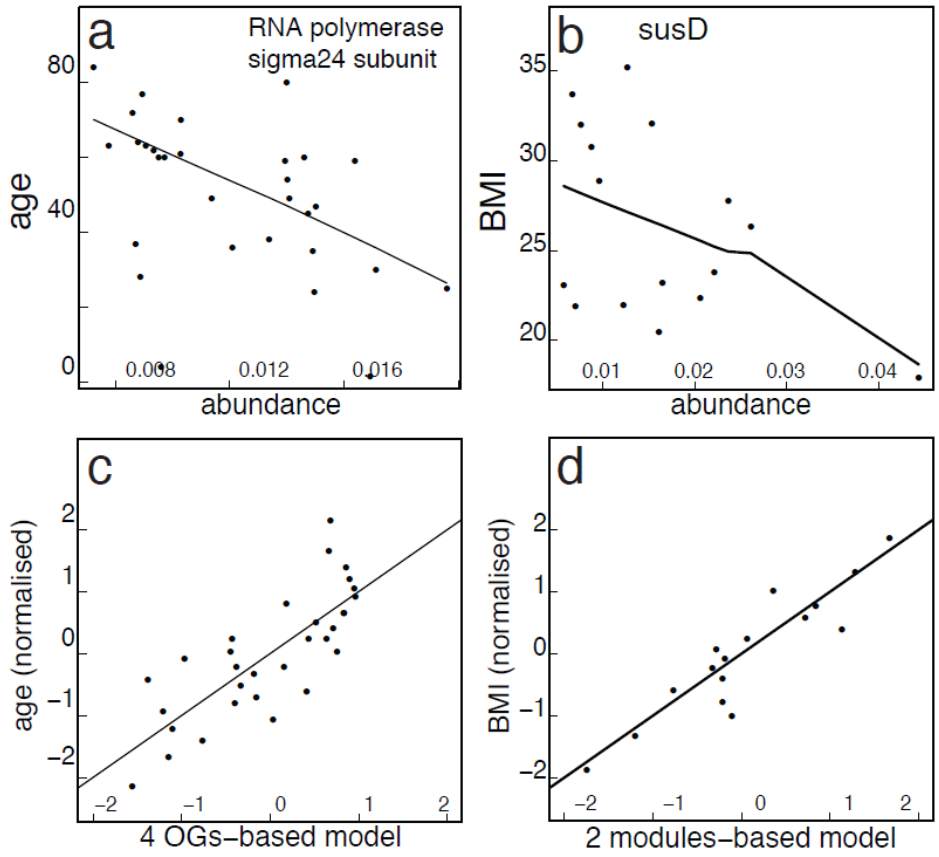




What are enterotypes?

- Stable symbiotic host-microbial interaction states?
- Potential classification of human groups that respond differently to diet or drug intake?
- Are there more such groups with some individuals in transition between them?

Functional biomarkers



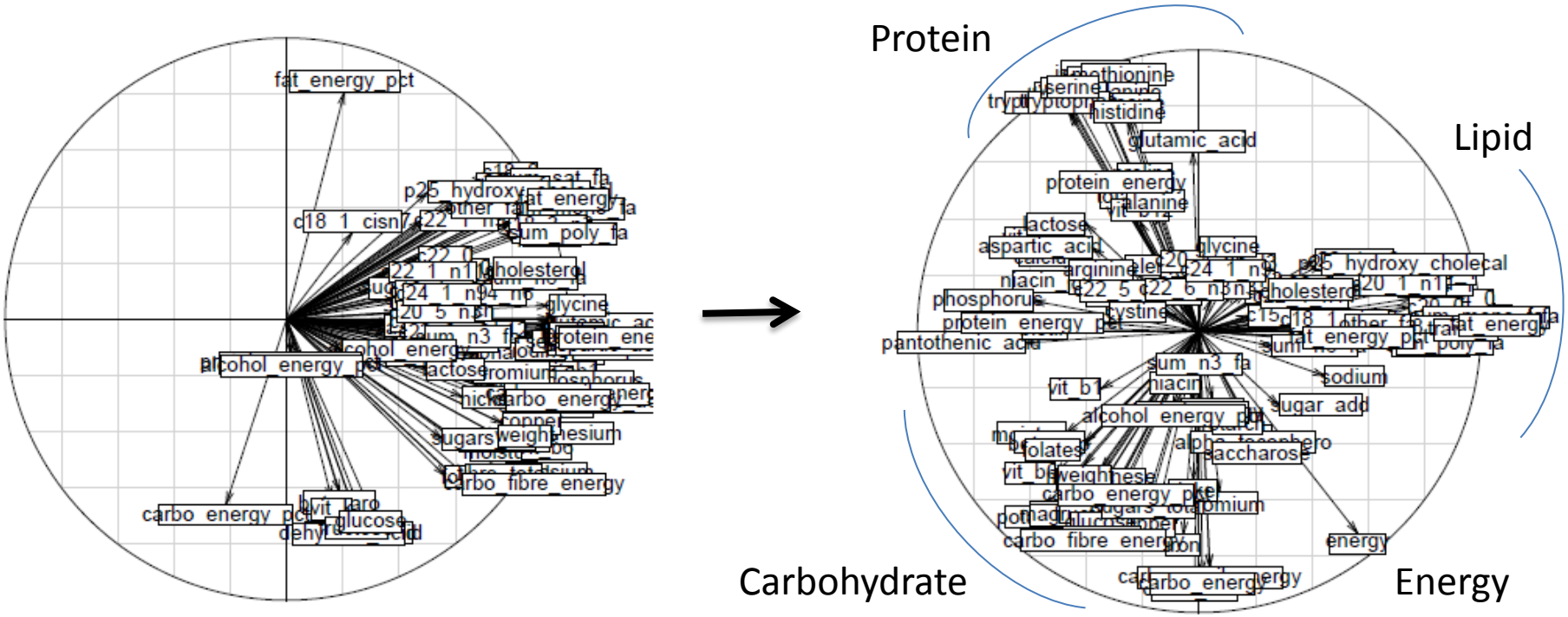
After stringent testing for multiple correction,
three functional modules correlate with bmi, and 12 genes with age.
susD is the glycan binding protein in *Bacteroides* species

Arumugam*, Raes*, *et al*. Nature (*in press*).

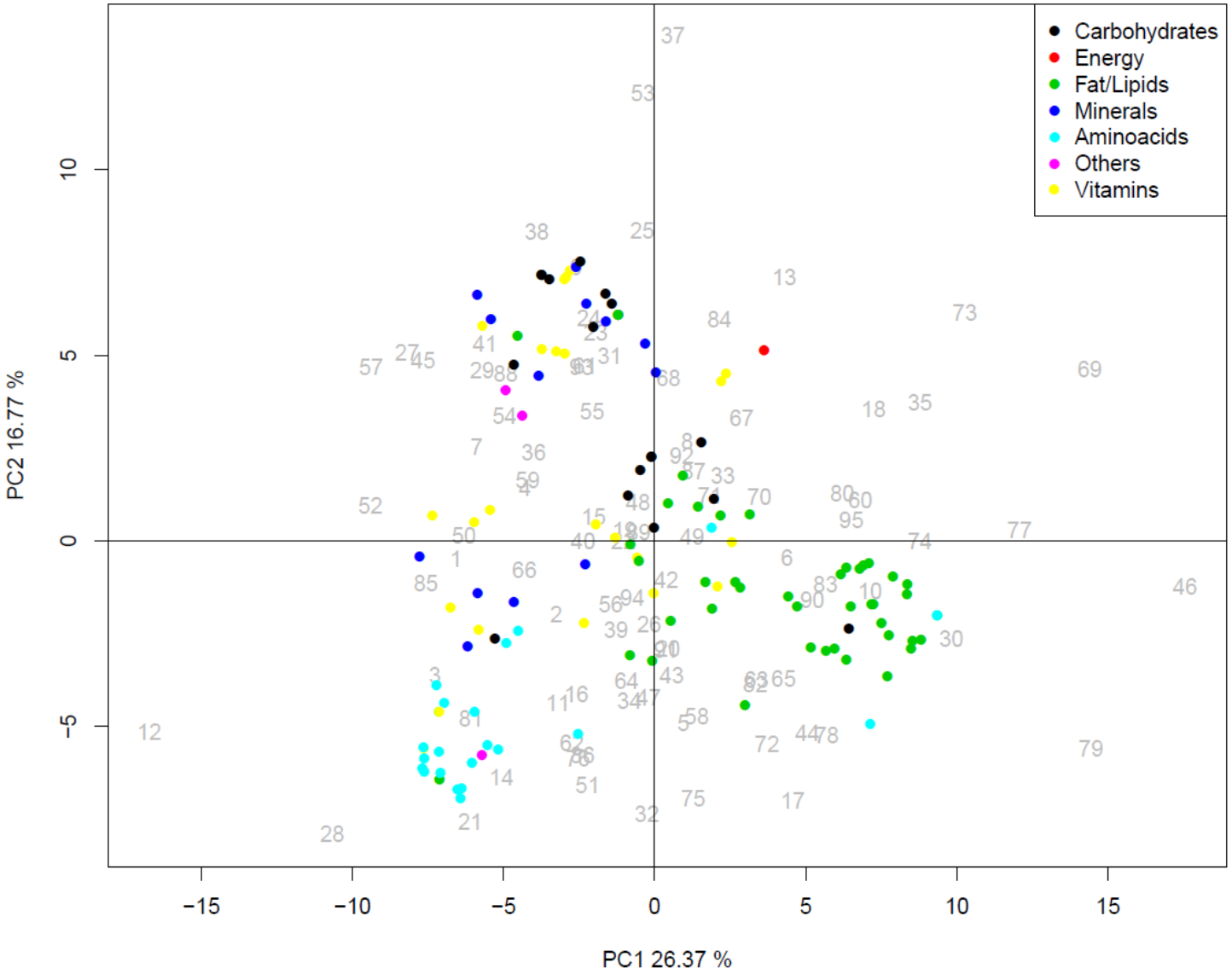
Mining the sample metadata: diet as an example

Dietary nutrients from a food questionnaire

- Dietary information contains redundancies!



Dietary nutrients as confounding factors



Acknowledgments

EMBL

- Peer Bork
- Jeroen Raes
- Takuji Yamada
- Daniel Mende
- Julien Tap
- Shini Sunagawa
- Gabriel Fernandes
- Bork Group

MetaHIT

- Dusko Ehrlich
- Joel Dore
- Francisco Guarner
- Oluf Pederson
- Eric Pelletier

Thank You