

Introduction to metagenomics

Beatriz García Jiménez

Biological Informatics Group
Center for Plant Biotechnology and Genomics (CBGP), UPM-INIA
beatriz.garcia@upm.es



CENTRO DE BIOTECNOLOGÍA
Y GENÓMICA DE PLANTAS



CAMPUS
DE EXCELENCIA
INTERNACIONAL



Computational Biology Master UPM, Programming Challenges
13 December 2018

Outline

- 1 Introduction
 - Generalities
- 2 Basic 16S analysis
- 3 Important considerations
 - Pre-processing

Metagenomics: a new interesting and growing field

- Numbers:
 - 1 trillion of microbes in/on our body
 - 3 times more microbial cells than human cells (new proposal 1.3)
 - 100-1000 microbial genes per 1 human gene
 - > 1000 different types of microbes in our body
 - total weight = 2 * human brain weight
- Genome (human genes) vs meta-genome (microbial genes):
 - 99,9% of genome is the same
 - (only) 10% meta-genome is the same
- Microbiomes of the same cavity in different individuals are more similar between themselves than microbiomes of two different cavities in the same individual
- Functions:
 - Training the immune system
 - Digesting and processing nutrients
 - Defending against pathogens
 - Triggering signaling mechanisms

Bioinformatics allows to study all these microbial genes, thanks to the advances in the massive sequencing!!!

Metagenomics studies

- Many new manuscripts published every day about metagenomics
- Humans (distinct cavities), animals, plants-root-soils, water, ...
- Different kind of microbiome analysis about:
 - One cavity microbiome under specific conditions
 - Enviromental microbiomes: New York underground, hospitals, etc
 - Comparison of microbiomes under two different conditions (control vs disease, lean vs obese, C-section vs vaginal delivery, omnivorous vs vegetarian, etc)
 - Comparison of changes in microbiomes due to perturbations (food, drugs, ...) and time
- Studies to improve and standardize the techniques and their configuration (sequencing, filters, removing noise, variance stabilization, clustering of sequences, OTU taxonomic identification, ...)
- Research crowd-funding: American Gut Microbiome Project

Examples of metagenomics studies

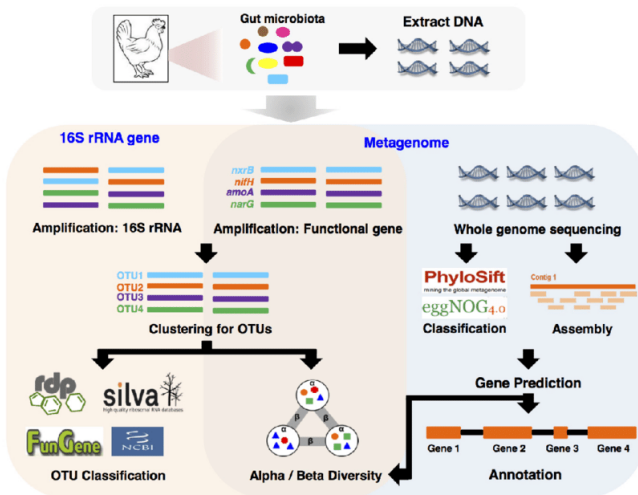
- Red wine and coffee modulate the microbiome
- The anorexia nervosa gut microbiome differs from healthy controls and is related to mental healthy
- The gut microbiome of a pre-Columbian Andean mummy looks much different than our own
- Fish oil may be important to altering the microbiome, reducing anxiety
- Gut bacteria may help prevent asthma in children
- We all emit our own 'microbial cloud'
- Elective vs. acute c-section deliveries: does it make a difference?
- The oral microbiome can predict childhood caries
- ...

Source: American Microbiome Institute blog

Outline

- 1 Introduction
 - Generalities
- 2 Basic 16S analysis
- 3 Important considerations
 - Pre-processing

Metagenomics workflow



(Source: *Hesed.info*)

Basic 16S rRNA analysis: Operational Taxonomic Unit (OTU) table

seqs_otu_table.txt

#Full OTU Counts-												
#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage-		
wf_otu_0	0	0	0	0	0	1	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_1	0	0	0	0	0	1	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_10	0	1	0	0	0	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales-		
wf_otu_100	0	0	0	1	0	0	0			Root;Bacteria-		
wf_otu_101	0	0	0	3	0	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_102	0	1	0	0	0	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_103	0	1	0	0	0	1	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_104	0	0	0	0	1	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales-		
wf_otu_105	0	1	0	0	0	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_106	0	0	0	0	1	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales-		
wf_otu_107	0	0	0	0	1	0	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"		
wf_otu_108	0	0	0	0	0	1	0			Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax-		
wf_otu_109	0	0	1	0	0	1	5	2		Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes-		
wf_otu_11	0	0	0	0	0	1	0			Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides-		
wf_otu_110	0	0	0	0	2	0	0			Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact		
wf_otu_111	0	0	0	0	0	1	0			Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides-		
wf_otu_112	0	0	0	0	0	1	0			Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides-		
wf_otu_113	0	0	0	1	0	0	0			Root;Bacteria-		

Line: 1 Column: 1 Rest Tab Size: 4

(Source: Caparoso, 2012)


```
seqs_otu_table.txt
#Full OTU Counts-
#OTU ID:PC.354, PC.355, PC.356, PC.481, PC.593, PC.607, PC.634, PC.635, PC.636, Consensus Lineage-
wf_otu_0 0 0 0 0 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_1 0 0 0 0 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_10 0 1 0 0 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales-
wf_otu_100 0 0 0 1 0 0 0 0 0 Root;Bacteria-
wf_otu_101 0 0 0 3 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_102 0 1 0 0 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_103 0 1 0 0 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_104 0 0 0 0 0 1 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales-
wf_otu_105 0 1 0 0 0 0 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_106 0 0 0 0 0 1 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales-
wf_otu_107 0 0 0 0 0 1 0 0 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"-
wf_otu_108 0 0 0 0 0 0 0 1 0 Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax-
wf_otu_109 0 0 1 0 0 0 1 5 2 Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes-
wf_otu_11 0 0 0 0 0 0 0 1 0 Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides-
wf_otu_110 0 0 0 0 0 2 0 0 0 Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact-
wf_otu_111 0 0 0 0 0 0 0 1 0 Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides-
wf_otu_112 0 0 0 0 0 0 1 0 0 Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides-
wf_otu_113 0 0 0 0 0 1 0 0 0 Root;Bacteria-
```

(Source: Caparoso, 2012)

Basic 16S rRNA analysis: Operational Taxonomic Unit (OTU) table

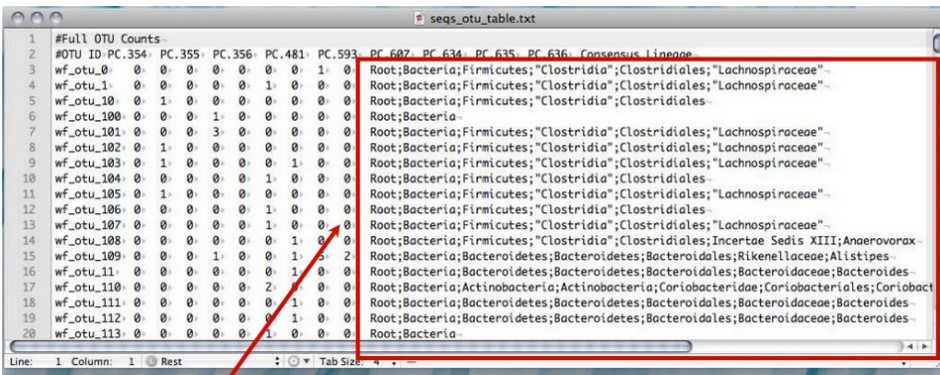
The screenshot shows a text file named 'seqs_otu_table.txt' containing an OTU table. The first row is a header: '#Full OTU Counts'. The second row lists the OTU IDs and their counts for various samples: '#OTU ID PC.354 PC.355 PC.356 PC.481 PC.593 PC.607 PC.634 PC.635 PC.636 Consensus Lineage'. The subsequent rows show the counts for each OTU across the samples, followed by the consensus lineage for each OTU. A red box highlights the header row, and a red arrow points to the sample identifiers in the first column.

#Full OTU Counts	#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage
1	wf_otu_0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
2	wf_otu_1	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
3	wf_otu_10	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales"
4	wf_otu_100	0	0	0	1	0	0	0	0	0	Root;Bacteria"
5	wf_otu_101	0	0	0	3	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
6	wf_otu_102	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
7	wf_otu_103	0	1	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
8	wf_otu_104	0	0	0	0	1	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales"
9	wf_otu_105	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
10	wf_otu_106	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales"
11	wf_otu_107	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
12	wf_otu_108	0	0	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax"
13	wf_otu_109	0	0	0	1	0	0	1	5	2	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes"
14	wf_otu_11	0	0	0	0	0	1	0	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides"
15	wf_otu_110	0	0	0	0	2	0	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact"
16	wf_otu_111	0	0	0	0	0	1	0	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides"
17	wf_otu_112	0	0	0	0	0	1	0	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides"
18	wf_otu_113	0	0	0	0	1	0	0	0	0	Root;Bacteria"

Sample identifiers

(Source: Caparoso, 2012)

Basic 16S rRNA analysis: Operational Taxonomic Unit (OTU) table

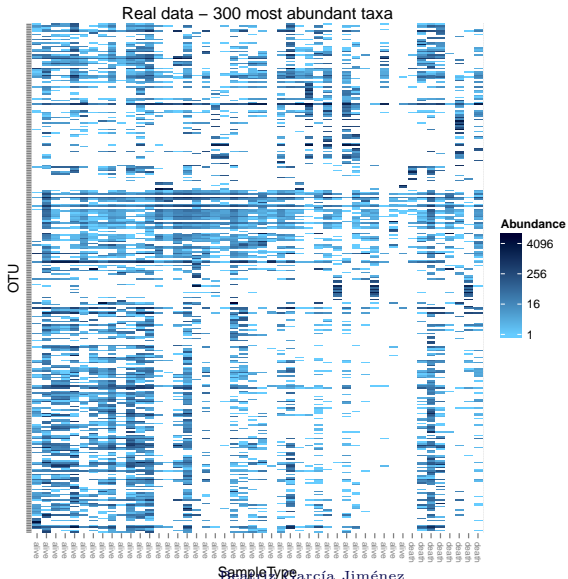


#Full OTU Counts	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage
wf_otu_0	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_1	0	0	0	0	1	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_10	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales"
wf_otu_100	0	0	0	1	0	0	0	0	0	Root;Bacteria"
wf_otu_101	0	0	0	3	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_102	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_103	0	1	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_104	0	0	0	0	1	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales"
wf_otu_105	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_106	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales"
wf_otu_107	0	0	0	0	1	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_108	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax"
wf_otu_109	0	0	0	1	0	1	5	2	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes"
wf_otu_11	0	0	0	0	0	1	0	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides"
wf_otu_110	0	0	0	0	2	0	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobacter"
wf_otu_111	0	0	0	0	0	1	0	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides"
wf_otu_112	0	0	0	0	0	1	0	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides"
wf_otu_113	0	0	0	0	1	0	0	0	0	Root;Bacteria"

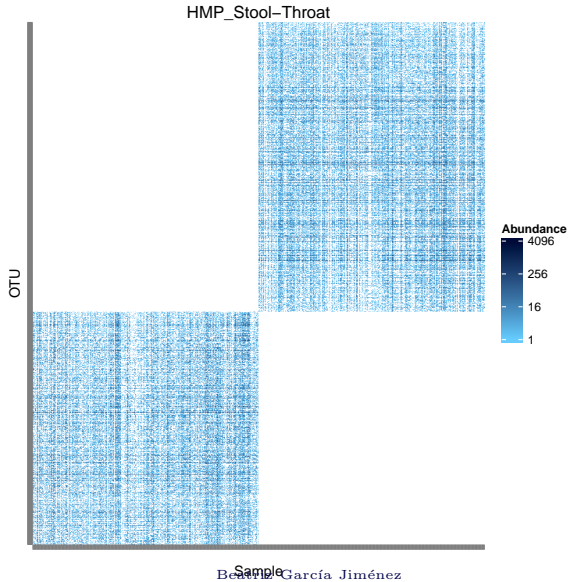
Optional per OTU taxonomic information

(Source: Caparoso,2012)

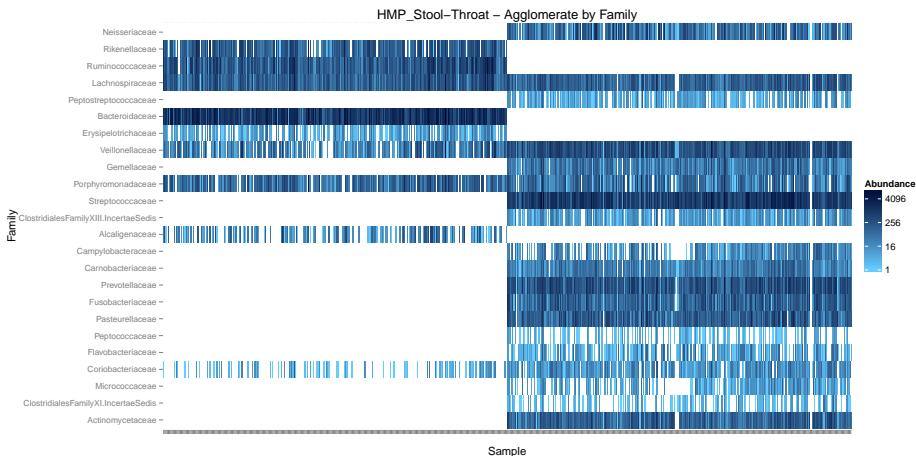
Basic 16S analysis: Graphical representations



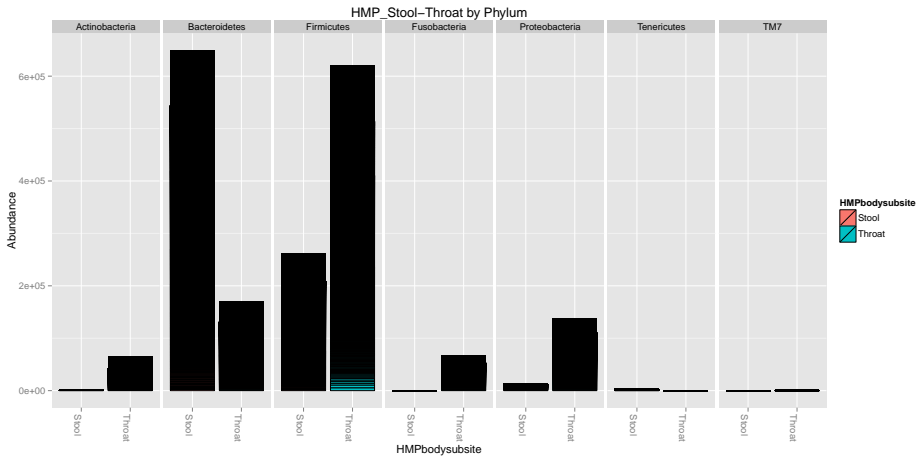
Basic 16S analysis: Graphical representations



Basic 16S analysis: Graphical representations



Basic 16S analysis: Graphical representations



Outline

- 1 Introduction
 - Generalities
- 2 Basic 16S analysis
- 3 Important considerations
 - Pre-processing

Important considerations with OTU matrices

- ❶ Number of samples \ll Number of features/OTUs (under-sampling)
- ❷ Sparsity and zero inflated data
- ❸ Compositional data [Aitchison,1982]
- ❹ Microbial data distributions: not follow common distributions (tricky)
- ❺ There are much less associations between different cavities than within the same cavity [CCREPE, Raes, 2012]
- ❻ How to compare different methods (not gold standard)
- ❼ Pre-procesing (with high variability, without standardization):
 - Filtering OTUs
 - Filtering samples
 - Normalizing
 - Abundance aggregation by taxonomic levels

Pre-processing steps

- Filtering OTUs (options):
 - Removing OTUs without samples
 - Removing OTUs without taxonomy category
 - Removing singletons
 - Removing OTUs not present in at least an 30-50% of samples
 - Removing OTUs with a high variance
- Filtering samples:
 - By sample type
 - Removing low-depth/low-read count (<10,000, first quartile, etc.)
- Normalizing (next slide)
- Abundance aggregation by taxonomic levels:
 - Aggregating OTUs with the same assigned taxonomy, although they are grouped in different OTUs
 - Aggregating all lower level OTUs corresponding to the same rank at a higher level

Parameters to tune in our controlled experimentation: no.samples, no.OTUs, normalization techniques, initial OTU and sample filtering,...

Pre-processing: Normalization

Normalization (or variance stabilization)

It is a pre-processing step with a high influence on the final results. However, it is not standardized. It depends on the following task

- Many different methods: rarefaction, scaling, log-ratio transformations, proportion, Total sum normalization per sample (TSS), quantile, etc.
- Not only different methods, also different parameters to tune inside each method: pseudocount, %number of samples to remove, %number of OTUs to remove, choosing the percentile, etc.

Introduction to metagenomics

Beatriz García Jiménez

Biological Informatics Group
Center for Plant Biotechnology and Genomics (CBGP), UPM-INIA
beatriz.garcia@upm.es



CENTRO DE BIOTECNOLOGÍA
Y GENÓMICA DE PLANTAS



CAMPUS
DE EXCELENCIA
INTERNACIONAL



Computational Biology Master UPM, Programming Challenges
13 December 2018