Metagenomics with R

R Users Grenoble

Antoine Bichat

November 23, 2017

Metagenomics

What is this?

- Mentagenomics is the study of the genomes of all species living in a given environment

[*] This presentation is simplified to make the topic understable in 5 minutes.

What is this?

- Mentagenomics is the study of the genomes of all species living in a given environment
- One wants to know the composition in micro-organisms of different samples

[*] This presentation is simplified to make the topic understable in 5 minutes.

What is this?

- Mentagenomics is the study of the genomes of all species living in a given environment
- One wants to know the composition in micro-organisms of different samples
- Principally *Bacteria*, *Archea* and *Fungi*

[*] This presentation is simplified to make the topic understable in 5 minutes.

What is this?

- Mentagenomics is the study of the genomes of all species living in a given environment
- One wants to know the composition in micro-organisms of different samples
- Principally *Bacteria*, *Archea* and *Fungi*
- More and more companies are interested in metagenomics

[*] This presentation is simplified to make the topic understable in 5 minutes.

Abundance table

- Output from bioinformatic pipeline
- Input for statistical analysis

Taxon	Sample1	Sample2	Sample3
Escherichia coli	27.8	22.1	19.0
Enterobacter cloacae	23.8	16.5	24.2
Bifidobacterium longum	7.9	21.2	27.9
Klebsiella sp	3.6	10.5	11.5
Staphylococcus aureus	1.7	13.4	9.0
Bacteroidetes fragilis	19.3	0.0	0.8
Other	15.9	16.3	7.6

R Pipeline

Biological process

- One gene (16S) present in all bacteria with variation is isolated and sequenced

Biological process

- One gene (16S) present in all bacteria with variation is isolated and sequenced

- Output of the sequencing: One FASTA file per sample

```
> GTCGATCGATGCCCTAGCCGATAGATCCCGATATAGCCGATAGAAAATATACGA...
> GTCGATCGATGCCCTAGCCGATAGATCGCGATATAGCCGATAGAAAATATACGT...
> GTCGATCGATGCCCTAGCCGATAGATCGCGATATAGCCGATAGAAAATATACGA...
> GTCGATCGATGCCATAGCCGATAGATCCCGATATAGCCGATAGAAAATATACGA...
...
```

Clustering

Similar sequences are grouped

- Similarity threshold in genomes
- Error correction of sequences

```
> GTCGATCGATGCCCTAGCCGATAGATCCCGATATAGCCGATAGAAAATATACGA... -> Group 1
> GTCGATCGATGCCCTAGCCGATAGATCGCGATATAGCCGATAGAAAATATACGT... -> Group 2
> GTCGATCGATGCCCTAGCCGATAGATCGCGATATAGCCGATAGAAAATATACGA... -> Group 1
> GTCGATCGATGCCATAGCCGATAGATCCCGATATAGCCGATAGAAAATATACGA... -> Group 3
...
```

Annotation

A species is assignated to each group

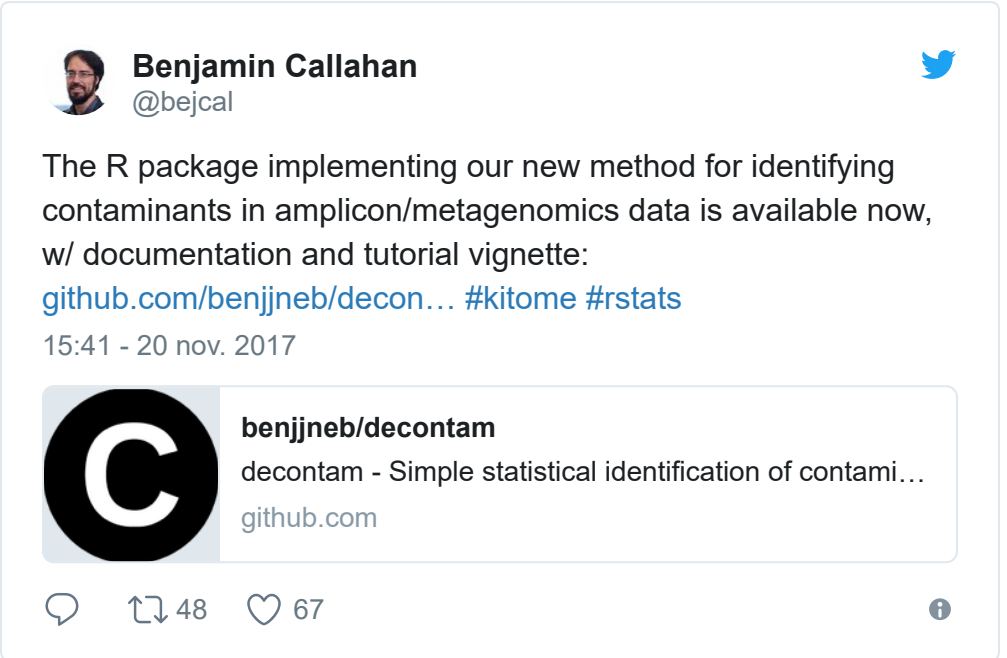
- NBC (Naive Bayesian Classifier)
- BLAST (Basic Local Alignment Search Tool)

```
Group 1 -> Escherichia coli
Group 2 -> Enterobacter cloacae
Group 3 -> Bifidobacterium longum
...
```

DADA2

The R package **dada2** uses error correction and NBC

Other bioinformatic tools

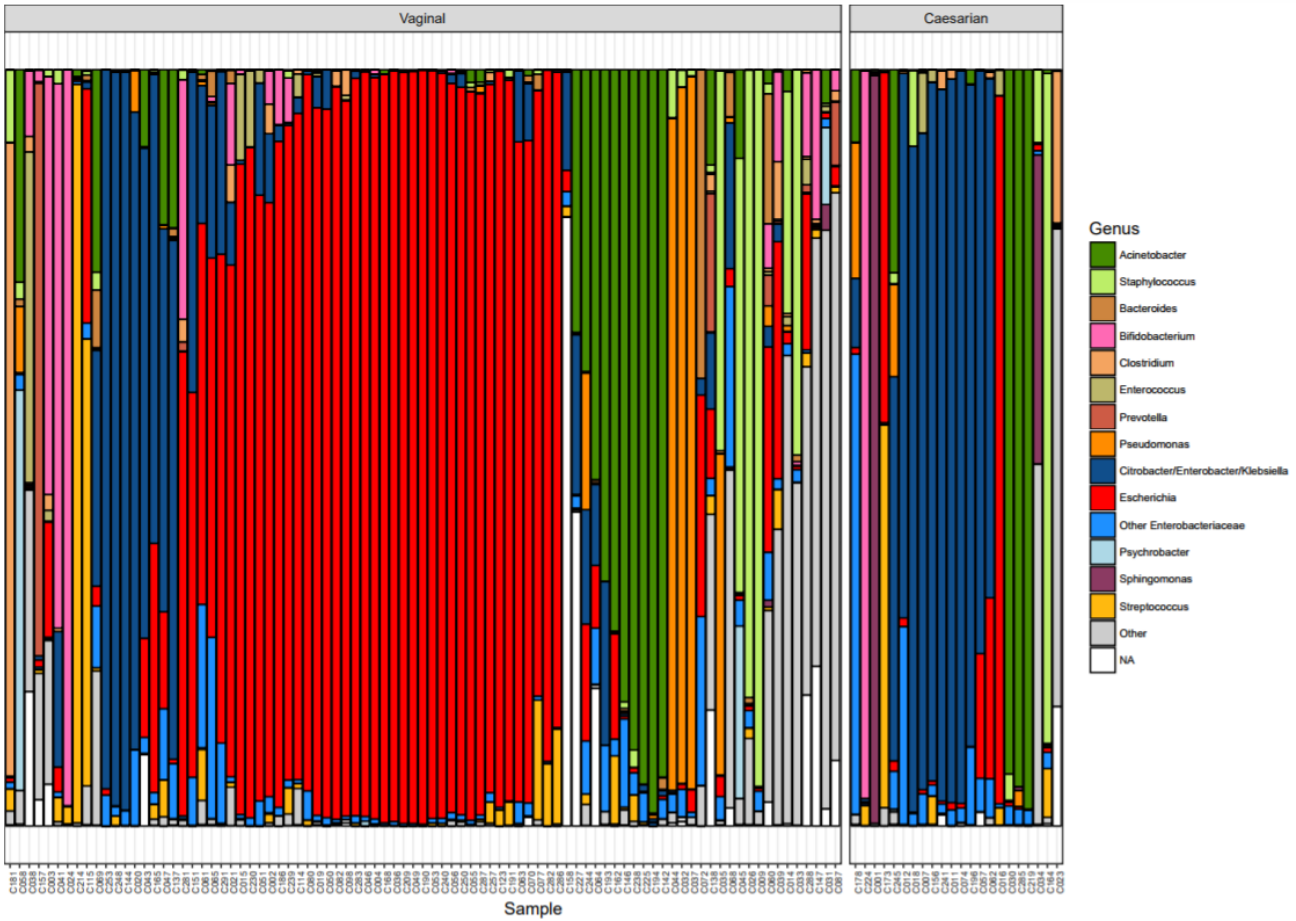


Data visualisation

With ggplot2

Sample composition...

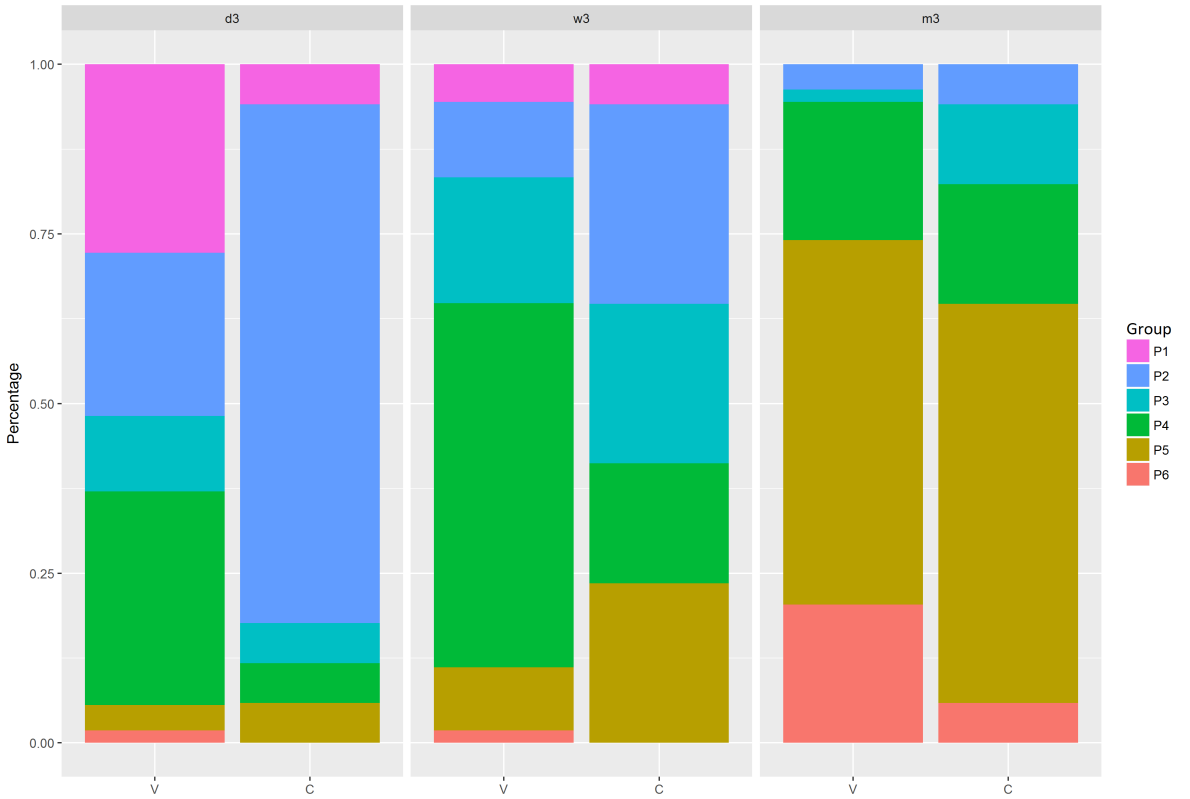
according to delivery mode



Repartition of groups ...

according to delivery mode and age

Each sample is assignated to one group according to its composition



Thanks for your attention :)

Slides created via the R package **xaringan**.