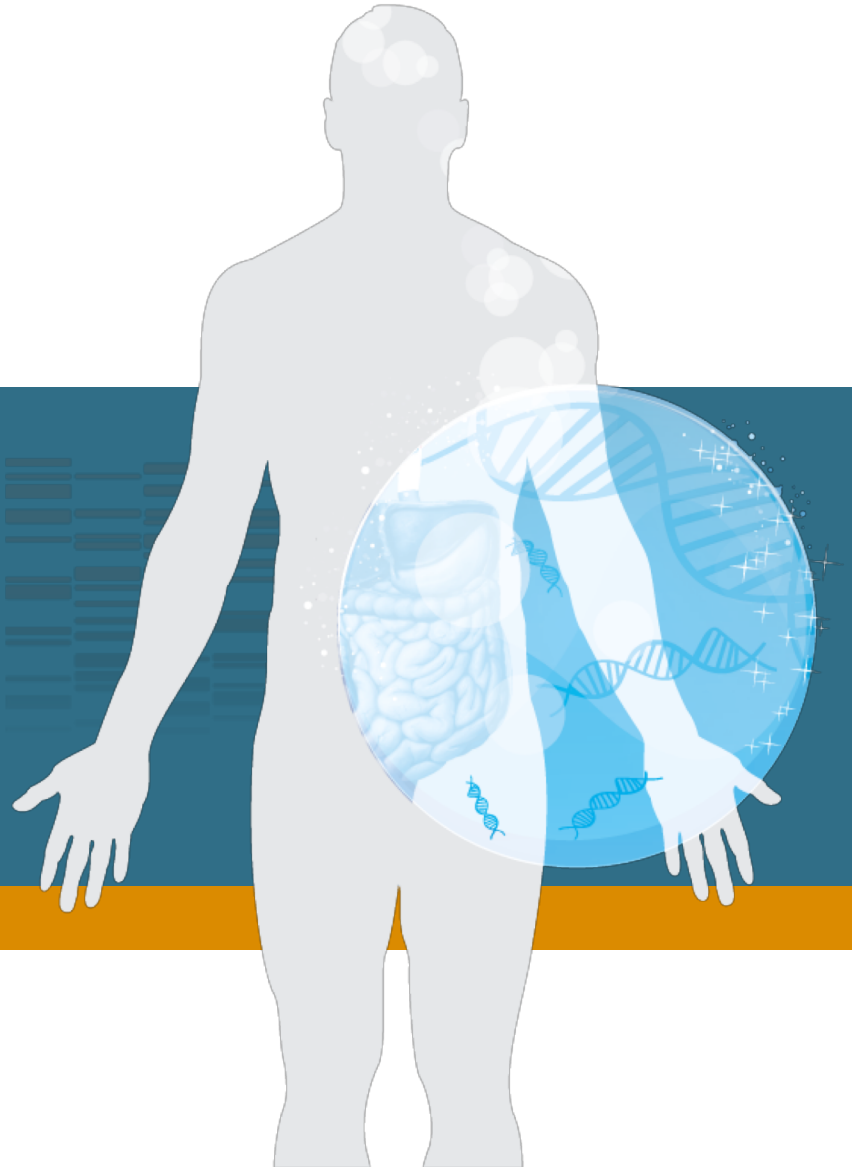


Module 3 - Analyse statistique avec R

Séance 3 : Analyse statistique des données métagénomiques

DUBii 2019

Boîte à outils R

Mise en œuvre pratique des fonctions indispensables pour manipuler les jeux de données

- 1. Fonctionalités de RStudio (Environnement, Historique, Console, Terminal, Scripts)
- 2. Travailler en mode Projet
- 3. Créer un notebook et générer des rapports html, pdf, word
- 4. Les types de données en R (les conversions entre différents types)
- 5. Manipuler les facteurs (réordonner les niveaux, gérer les variables ordinales)
- 6. Convertir une variable quantitative en facteur (fonction `cut()`)
- 7. Les chaines de caractères (fonctions `paste()` et `gsub()`)
- 8. Les vecteurs (fonctions `seq()`, `rep()` et `names()`)
- 9. Les matrices (fonctions `colnames()`, `rownames()`, `t()`, `apply()`)
- 10. Les listes (fonctions `summary()`, `lapply()`, `sapply()`, scope des variables)
- 11. Les data.frame (fonctions `dim()`, `ncol()`, `nrow()`, `head()`, `View()`, `cbind()`, `rbind()`)
- 12. Les tibbles
- 13. Gérer les valeurs manquantes (fonctions `anyNA()`, `is.na()`, `colSums()`)
- 14. Outils de manipulation des données (fonctions `select()`, `filter()`, opérateur `%>%`)
- 15. Visualisation des données (fonction `ggplot()`)

Statistiques pour les données « omiques »

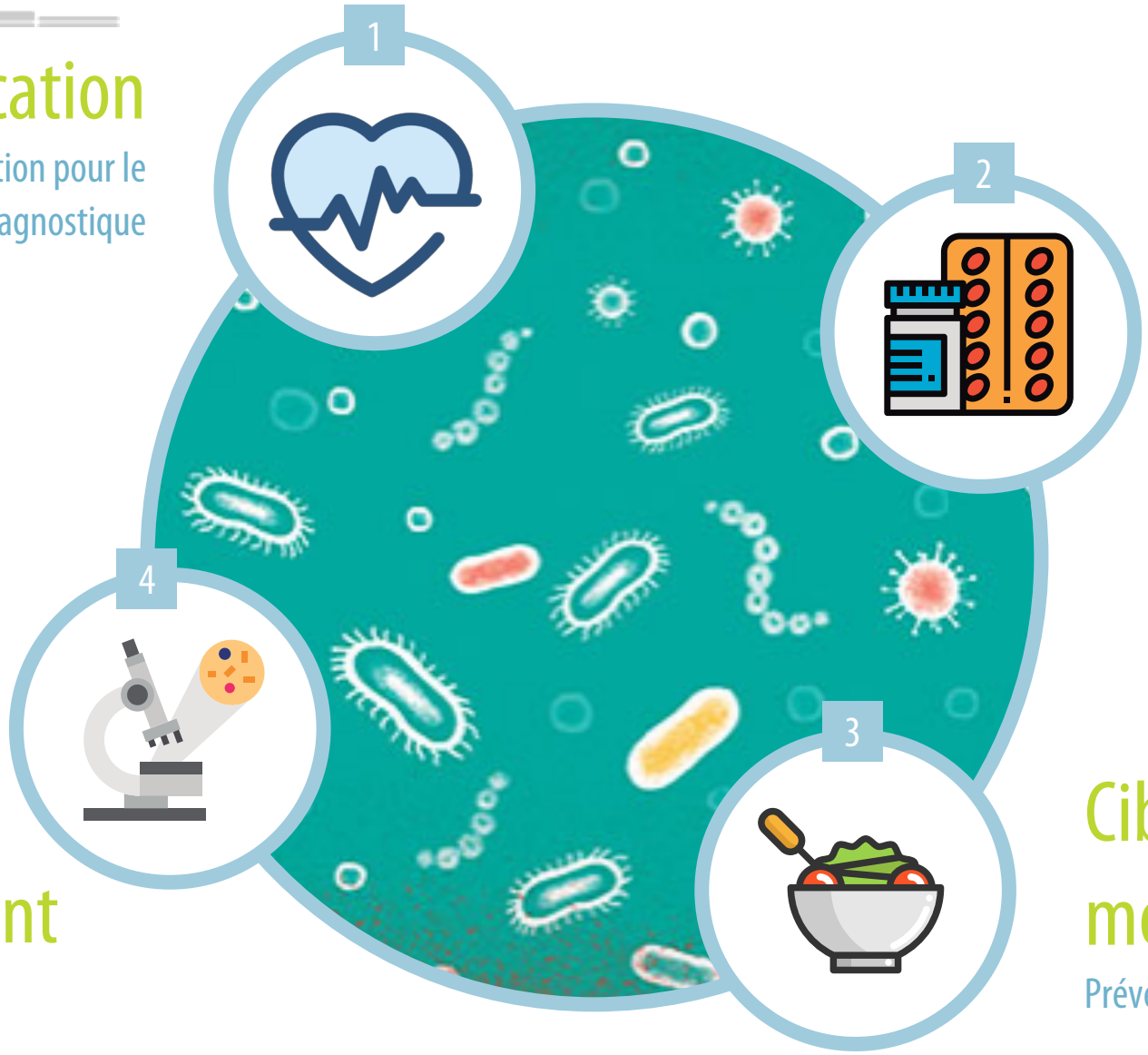
Exemple de l'analyse de données métagénomique dans l'étude du microbiote intestinal humain

Vers la découverte de l'immense diversité du microbiote

Applications en nutrition, santé et bien-être

Stratification

Personnalisation pour le diagnostique



Nouveaux médicaments

Nouvelles cibles thérapeutique

Le microbiote, organe médicament

Transplantation de microbiote

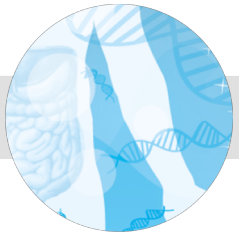
Cible de modulation

Préventive ou curative

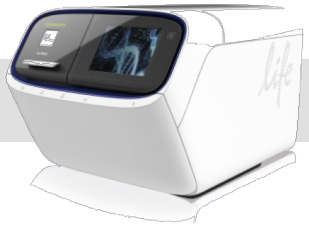
Acquisition des données pour une étude métagénomique

Objectifs et plan d'expérience

Collecte des échantillons



Séquençage



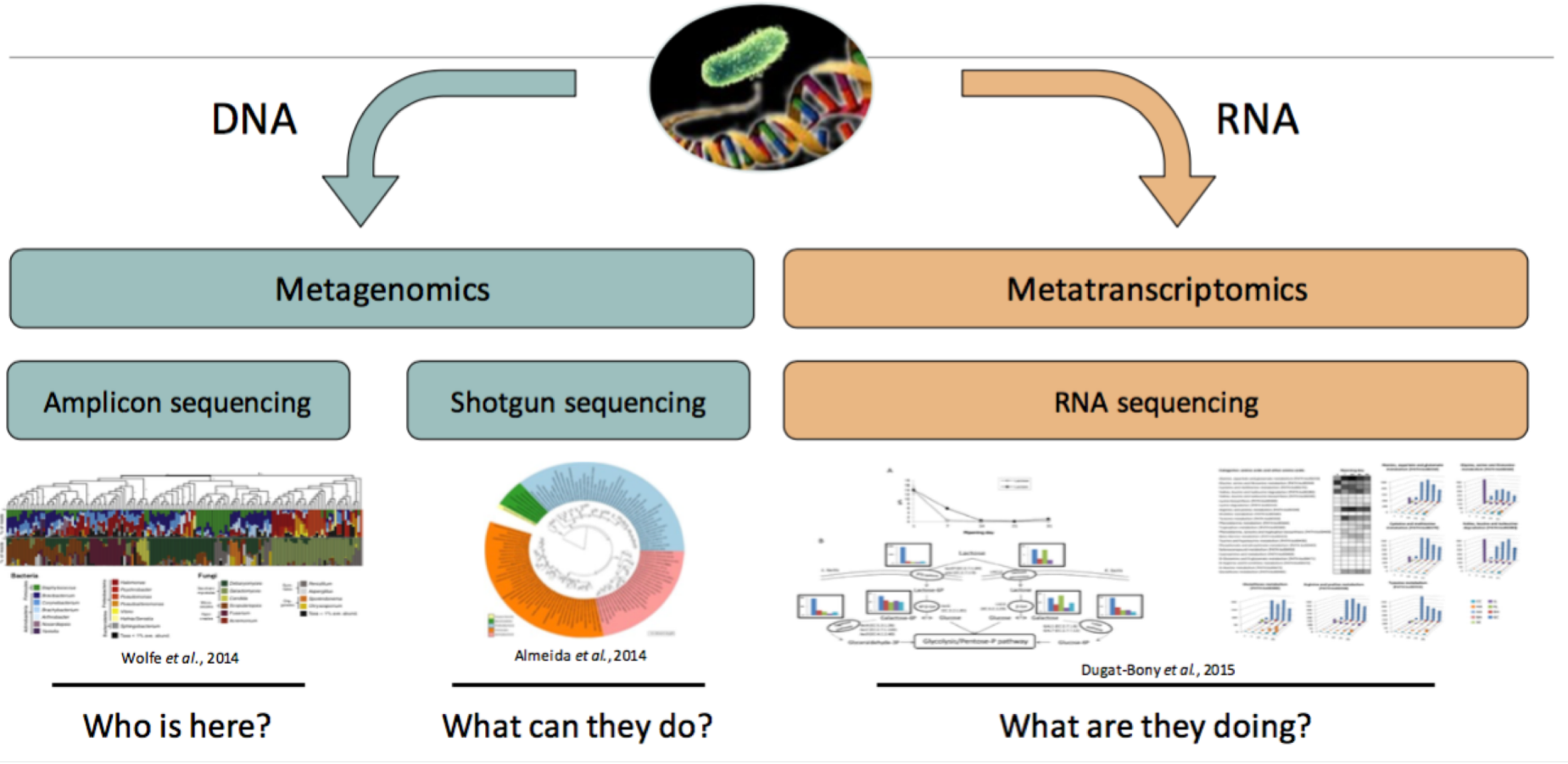
Traitement bioinformatique des séquences

Analyse biostatisque des données

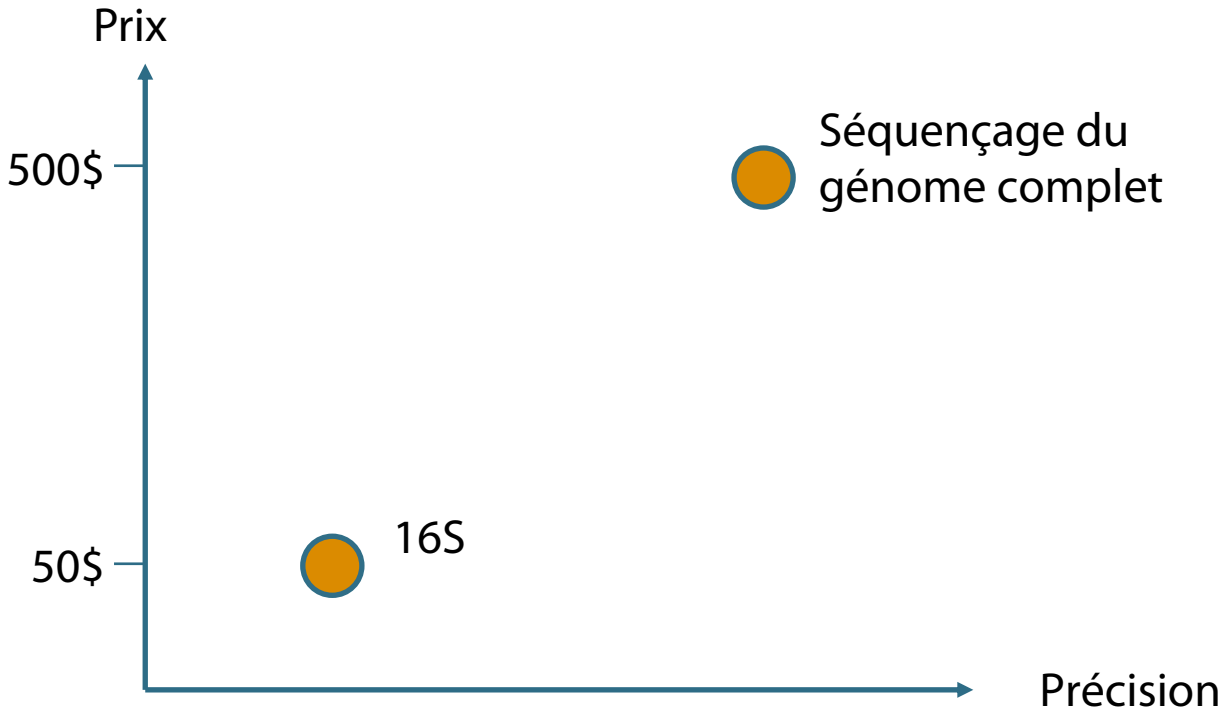
Catalogues de gènes

Catalogues d'espèces bactériennes

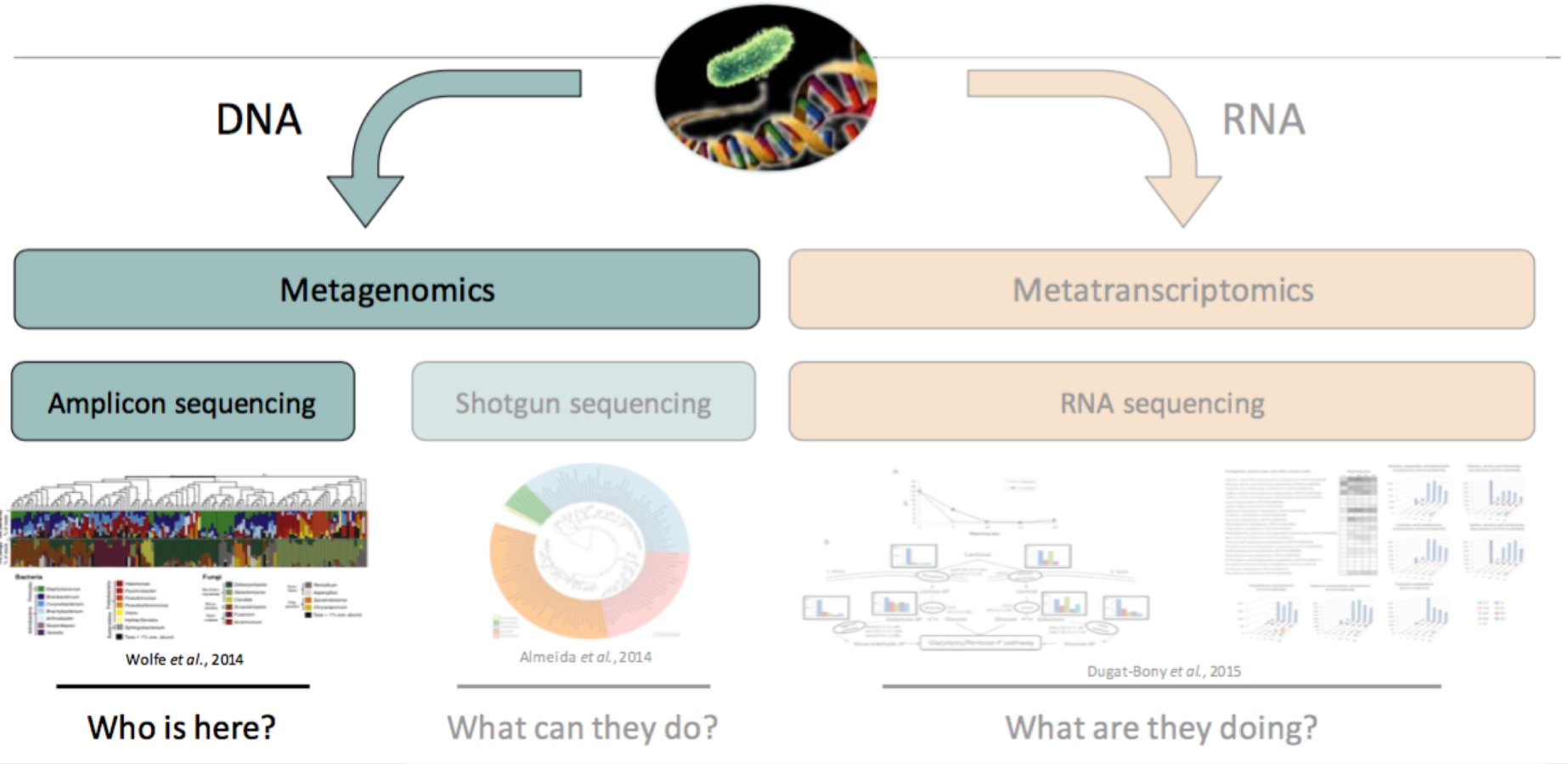
« Meta-omics » avec les NGS



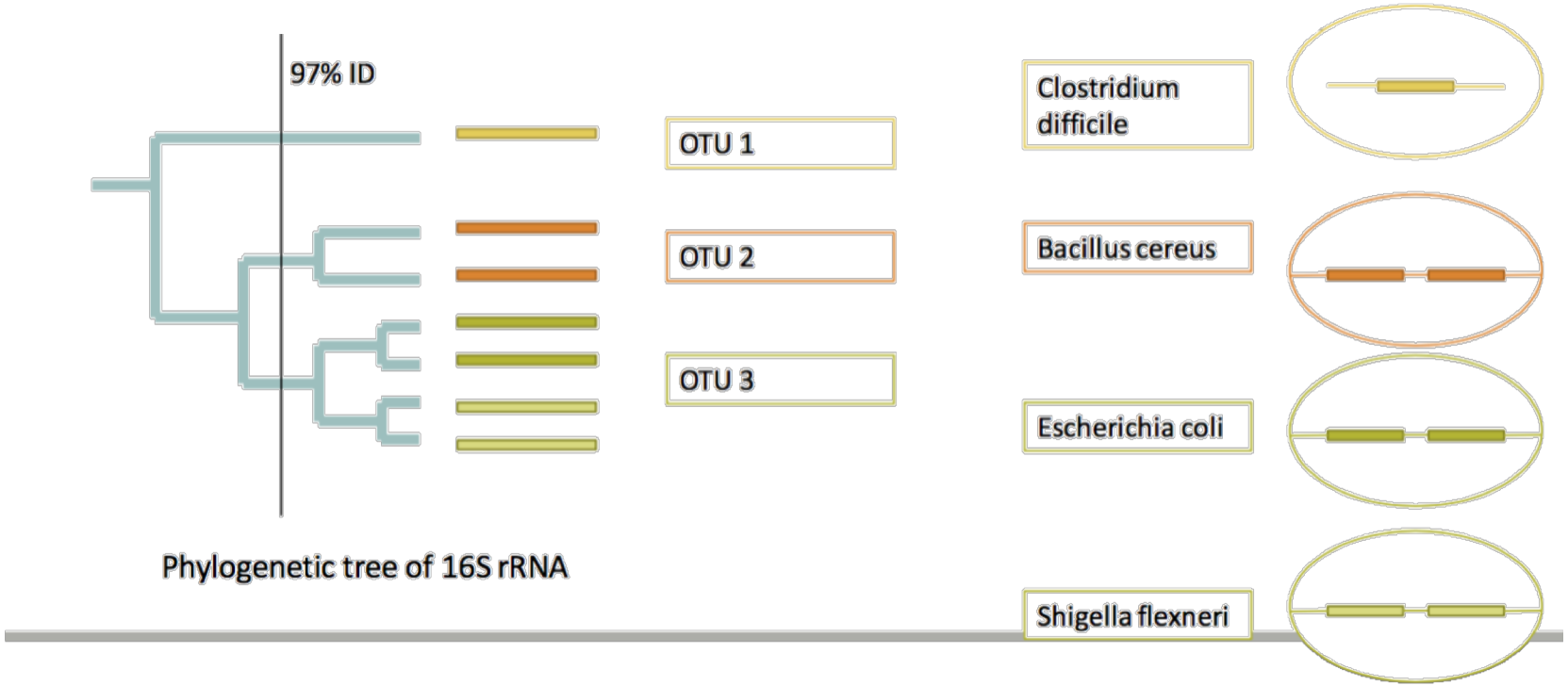
Séquençage 16S ou séquençage shotgun ?



Le séquençage16S (métabarcoding)

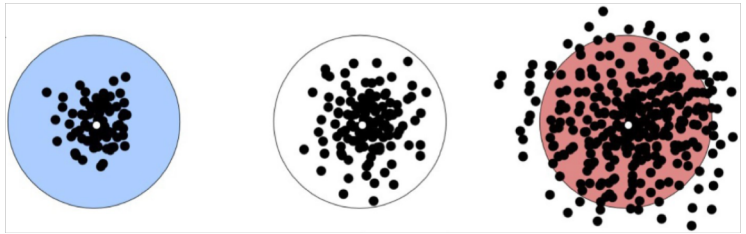


Les OTUs : un proxy pour les bactéries



OTU : Operational Taxonomic Units

- Le seuil de 97% est arbitraire



Les OTUs : un proxy pour les bactéries

SCIENTIFIC REPORTS

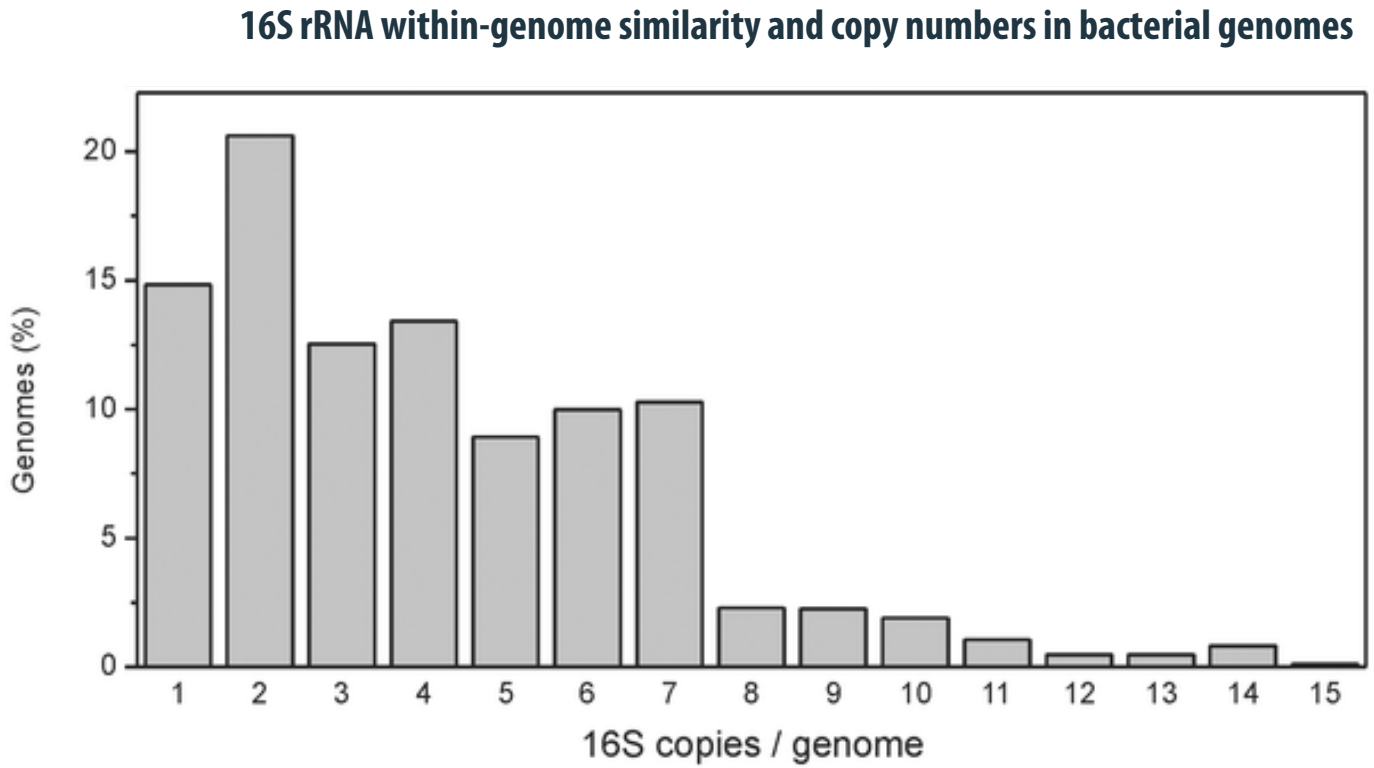
OPEN

Multiple *Streptomyces* species with distinct secondary metabolomes have identical 16S rRNA gene sequences

Received: 14 September 2016
Accepted: 23 August 2017
Published online: 11 September 2017

Sanjay Antony-Babu[1,4], Didier Stien[1], Véronique Eparvier[2], Delphine Parrot[3], Sophie Tomasi[3] & Marcelino T. Suzuki[1]

Le nombre de copies varie de 1 à 15

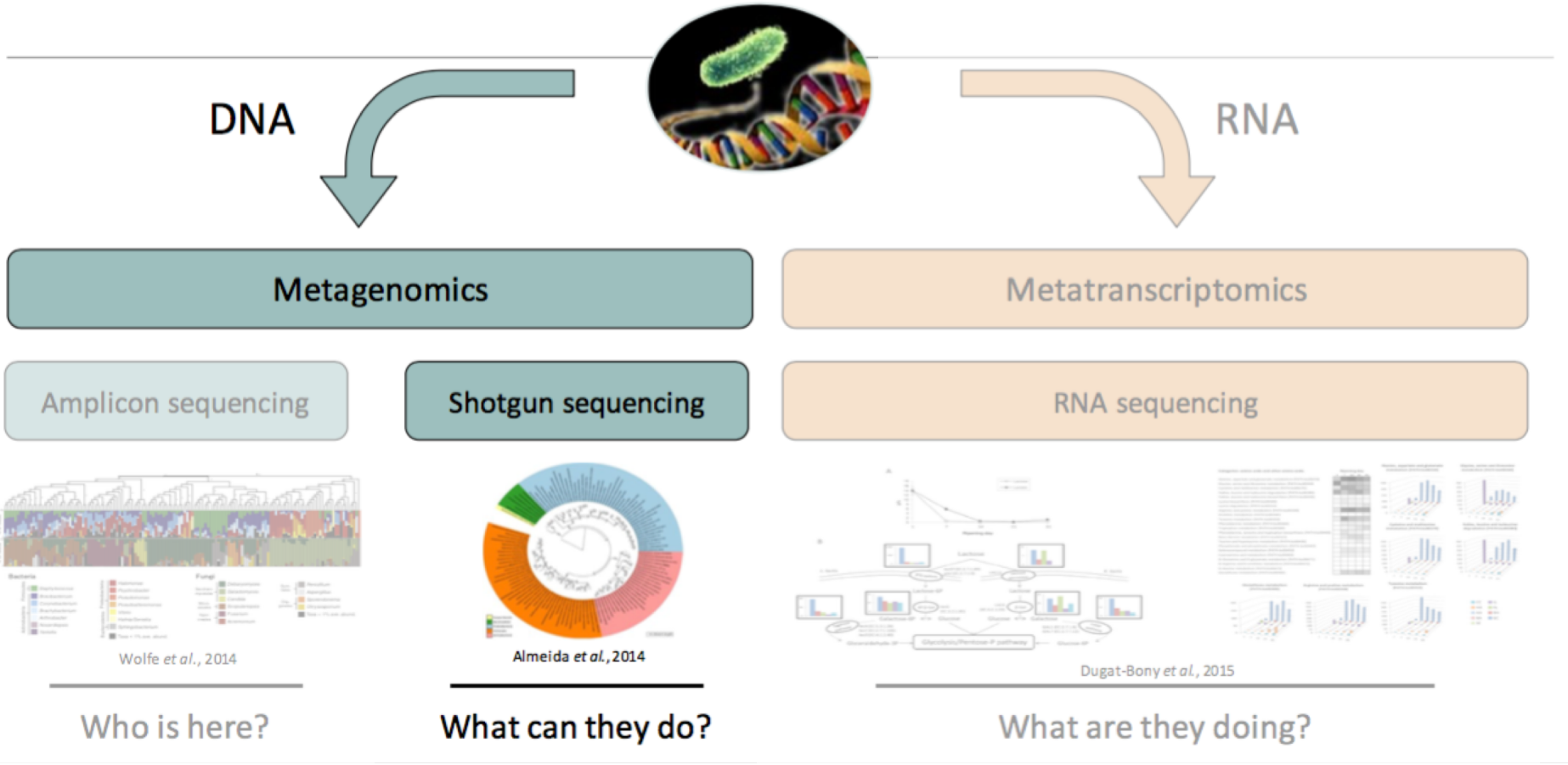


Větrovský T, Baldrian P (2013) The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLOS ONE 8(2): e57923.

Forces et faiblesses du séquençage 16S

- Forces
  - Détection des microorganismes dominants
  - Approximation de l'abondance relative
  - Profil taxonomique des communautés (généralement au niveau du genre, parfois de l'espèce)
- Faiblesses
  - Pas de quantification absolue possible
  - Pas d'identification exacte des organismes
  - Pas de vue fonctionnelle de l'écosystème

Le séquençage métagénomique shotgun

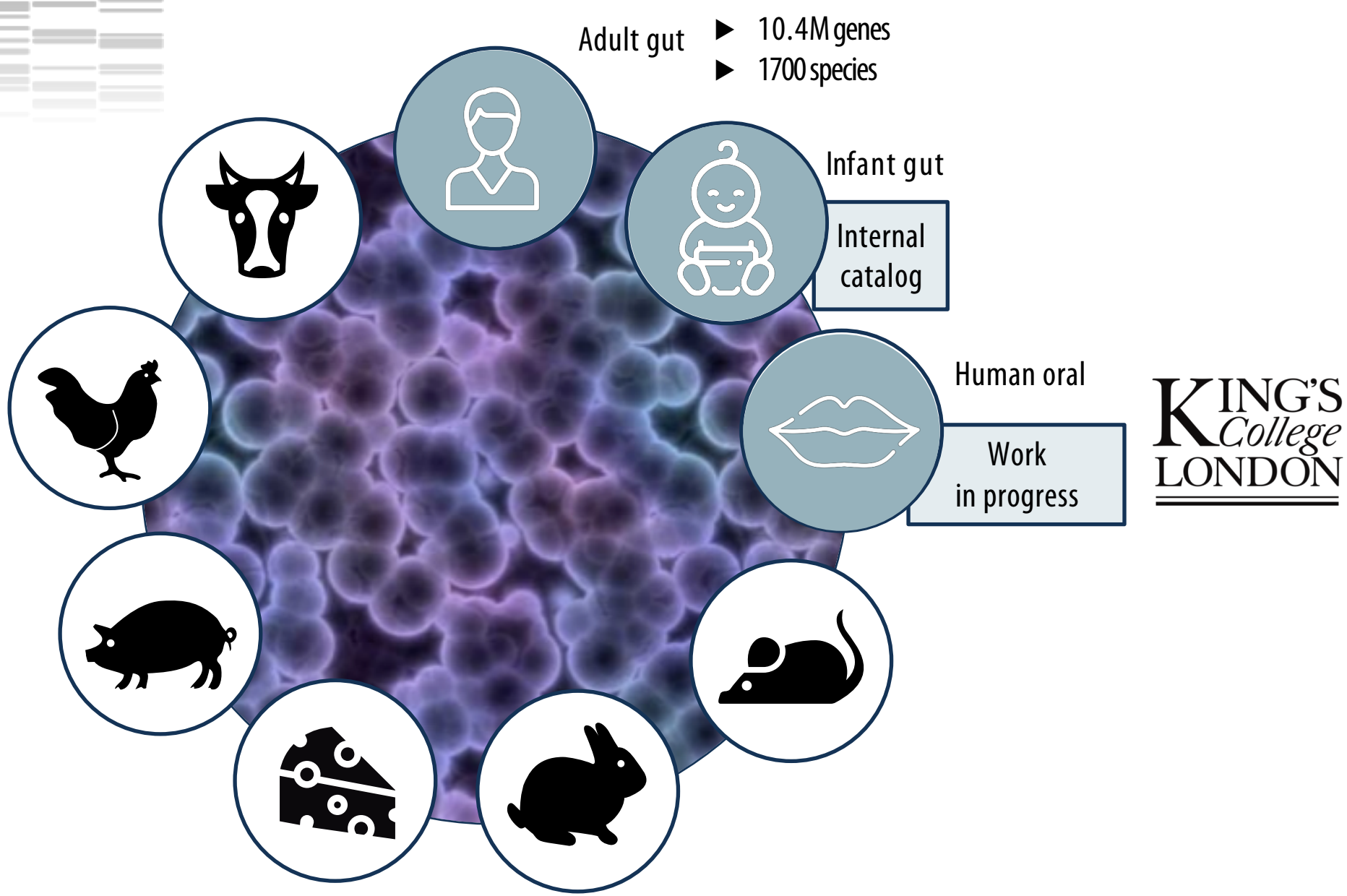


Le défi de la métagénomique shotgun

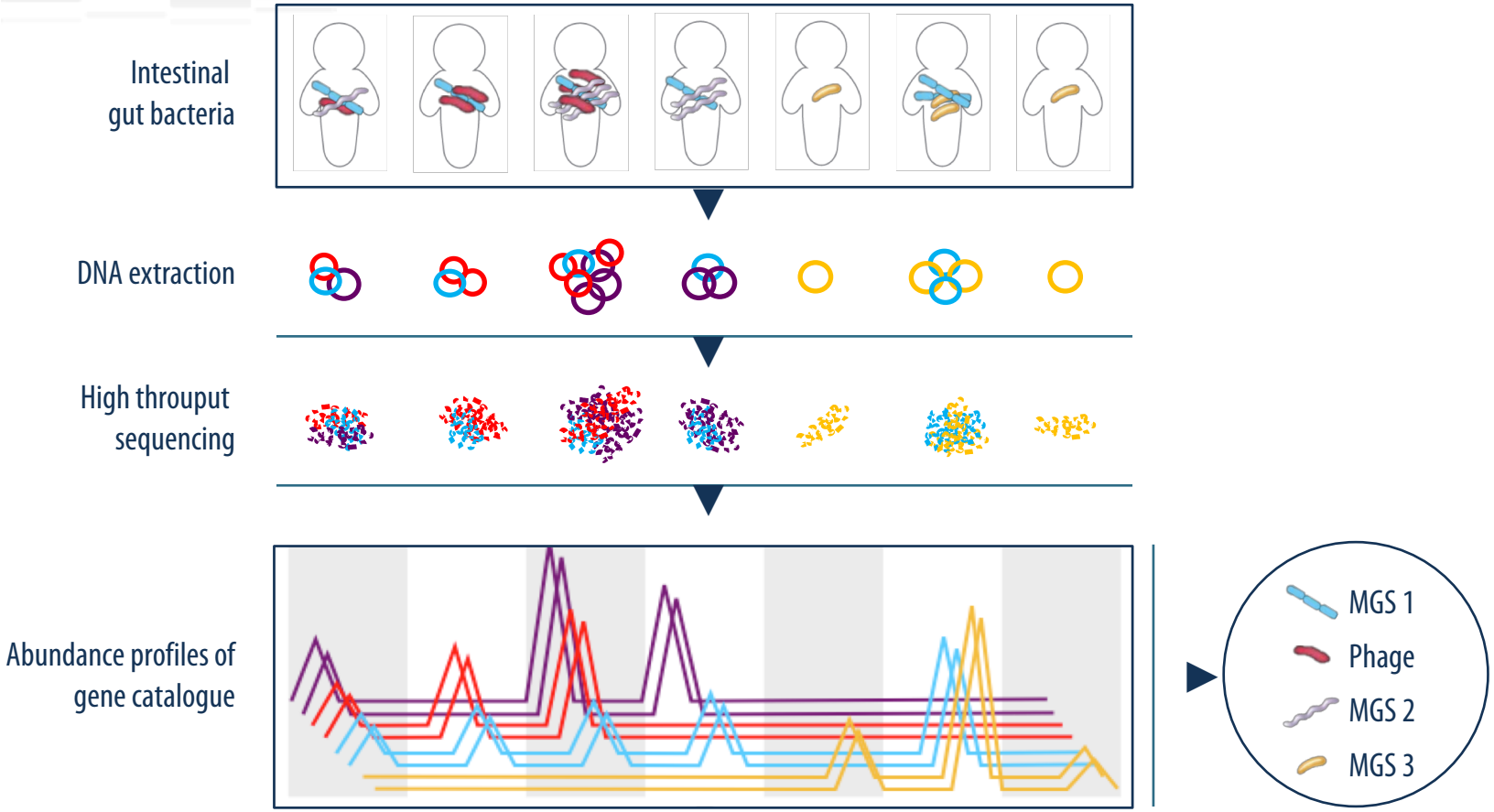
Traiter et analyser les données métagénomique

1. Catalogue de gènes
2. Clustering des gènes en espèces métagénomiques (MGS)
3. Calcul de l'abondance des MGS
4. Analyses statistiques exploratoires
5. Statistiques avancées : intégration de données et machine learning

1. Catalogues de gènes et d'espèces pour divers microbiomes

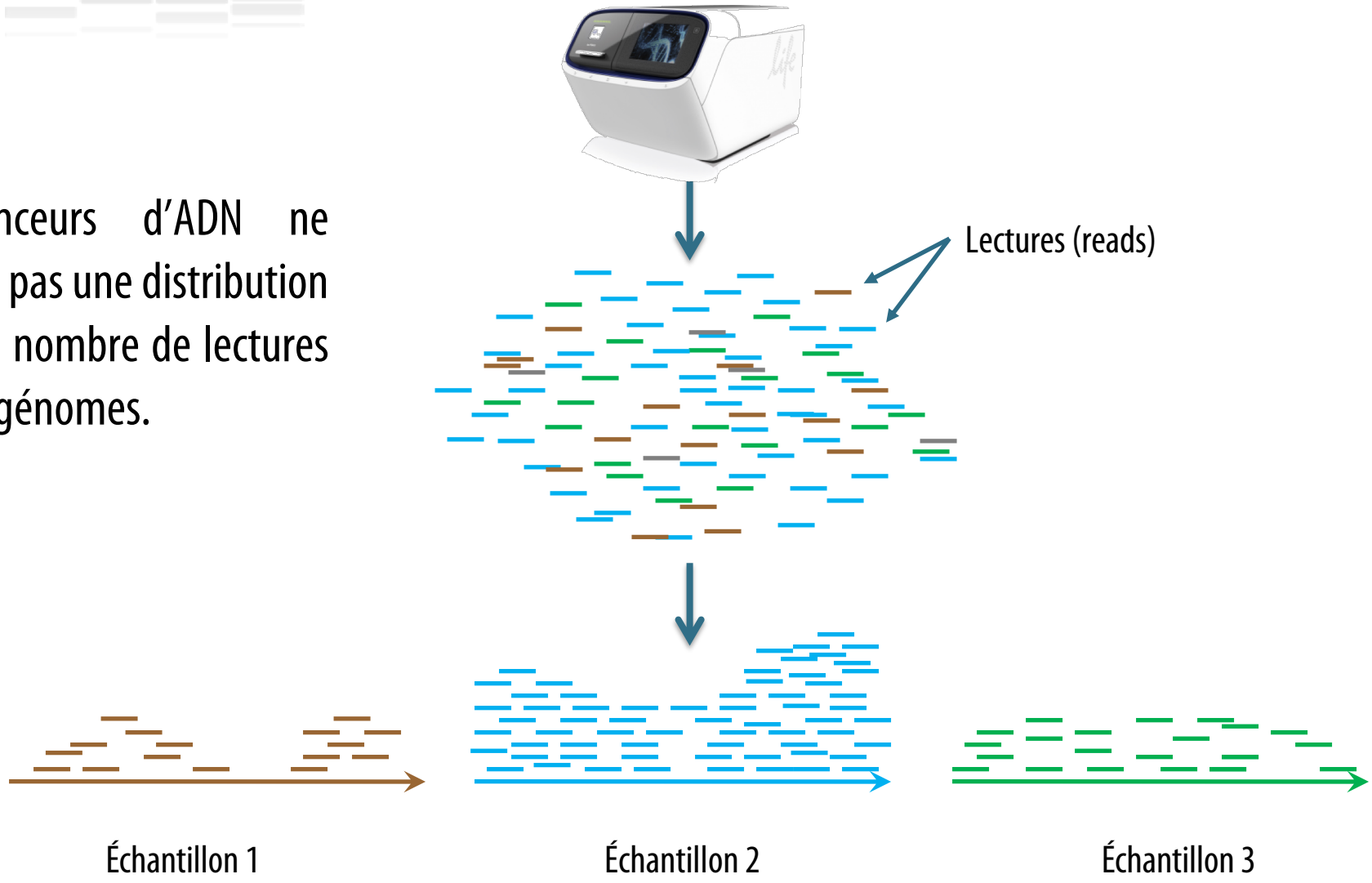


2. Clustering des gènes en espèces métagénomiques

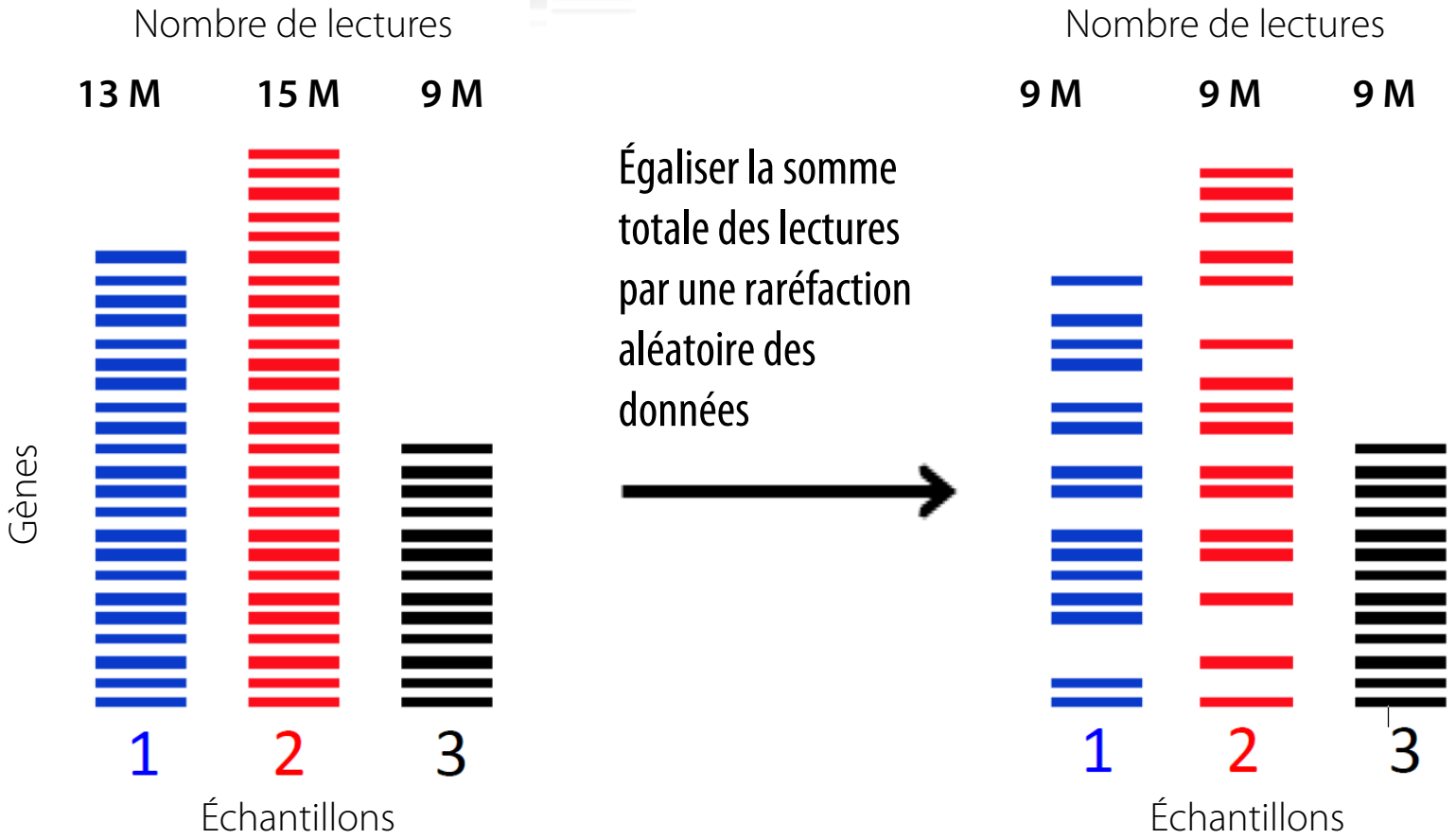


3. Calcul de l'abondance des MGS
Le nombre total de lectures varie entre les échantillons

Les séquenceurs d'ADN ne garantissent pas une distribution uniforme du nombre de lectures sur les métagénomes.



Une méthode pour résoudre le problème : le downsizing / raréfaction



Inconvénients :

- Une grande partie des données sont éliminées de manière peu rentable
- La précision de la mesure est réduite

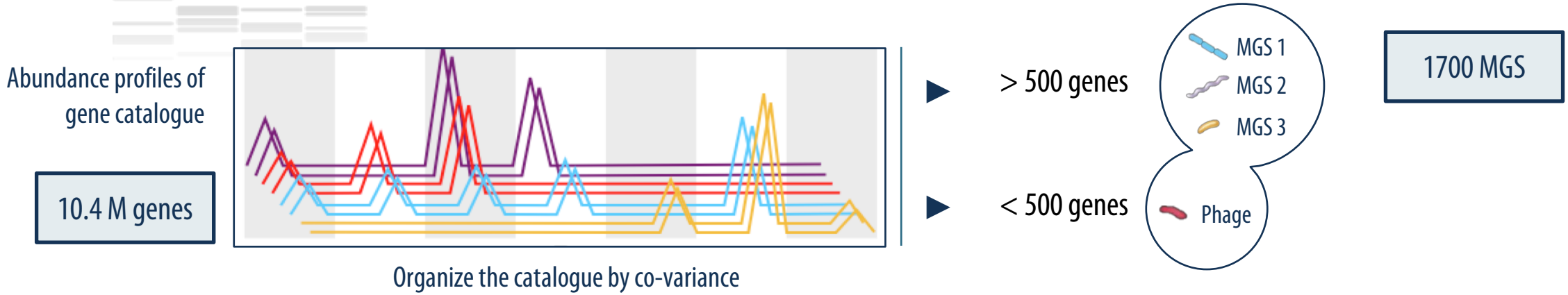
McMurdie et al., 2014

3. Calcul de l'abondance des MGS
Les gènes ne font pas tous la même taille

Étape de normalisation :

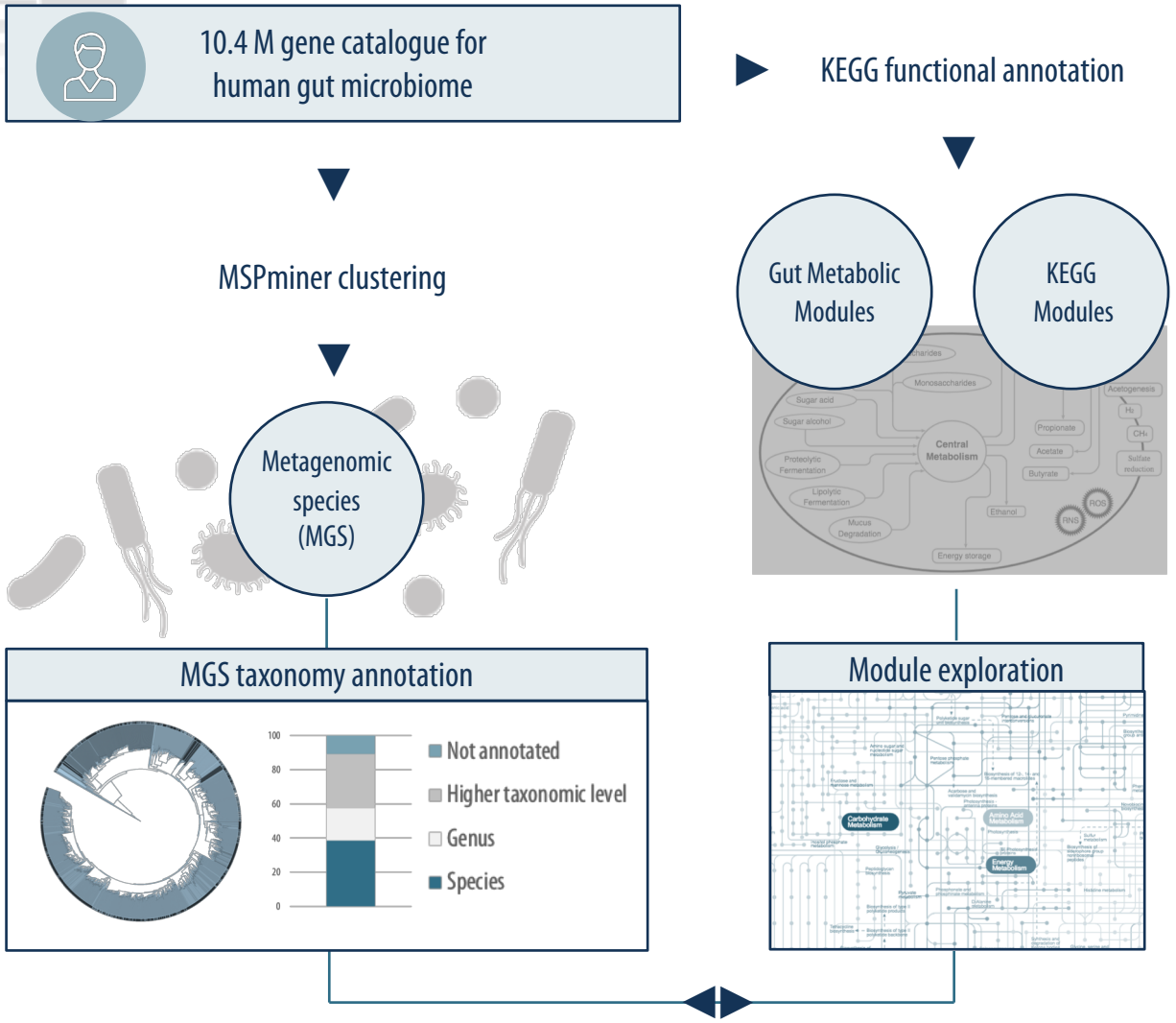
- Les comptages de chaque gène sont divisés par la taille du gène
- Le vecteur de comptage est divisé par la somme des comptages pour obtenir des fréquences
- La somme de l'abondance des gènes vaut 1 pour chaque individu.

3. Calcul de l'abondance des MGS



- Dans chaque MGS, l'ensemble des gènes « essentiels », ou gènes marqueurs, est défini comme celui des 50 gènes les plus corrélés entre eux.
- Au sein d'un échantillon, l'abondance des MGS est calculée en faisant la moyenne des comptages des 50 gènes marqueurs qui la composent.
- Si moins de 10 % de ces gènes sont présents, le comptage de la MGS est mis à 0.
- On obtient ainsi une matrice d'abondance des MGS

4. Analyses statistiques exploratoires
Composition taxonomique et fonctionnelle



4. Analyses statistiques exploratoires

Metagenomic species (MGS)

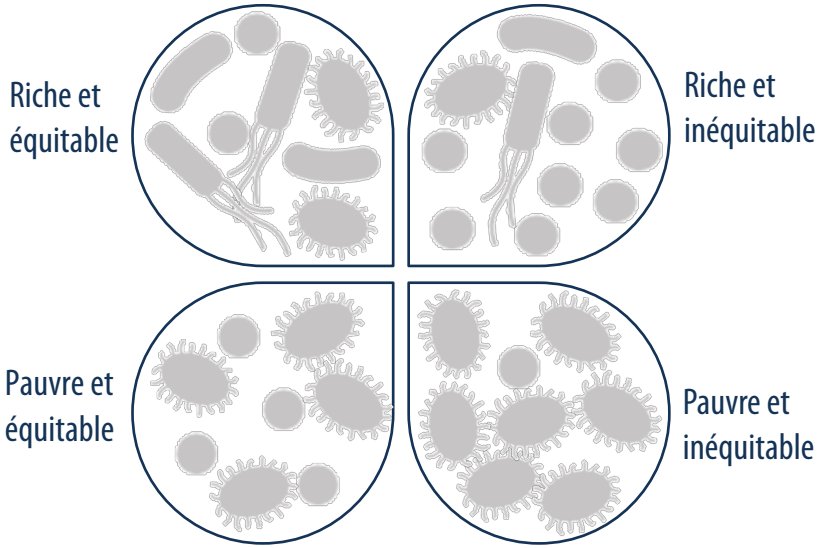
Gut Metabolic Modules

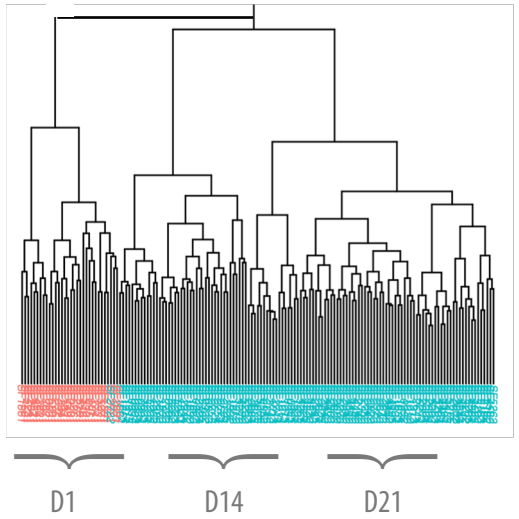
Analyse statistique exploratoire

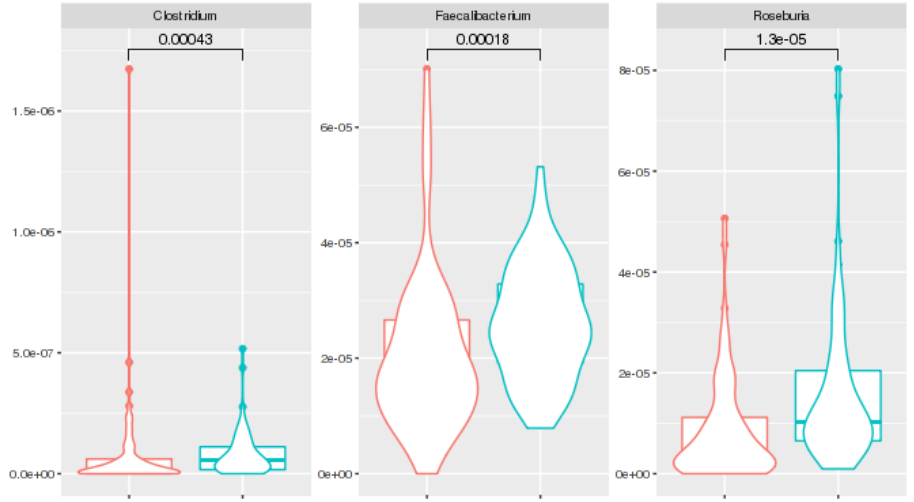
Diversité alpha

Diversité beta

Identifier les facteurs de modification du microbiote





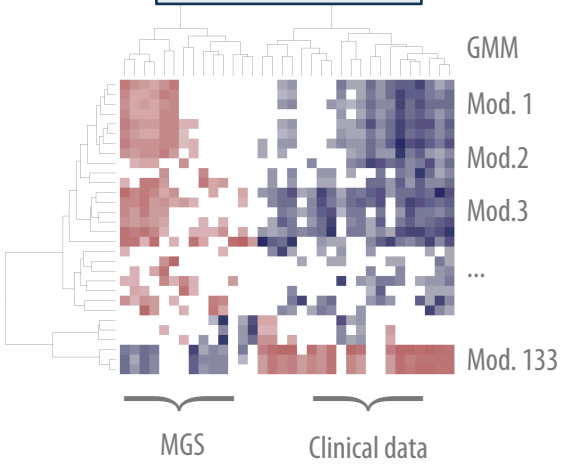


5. Statistiques avancées : intégration de données et machine learning

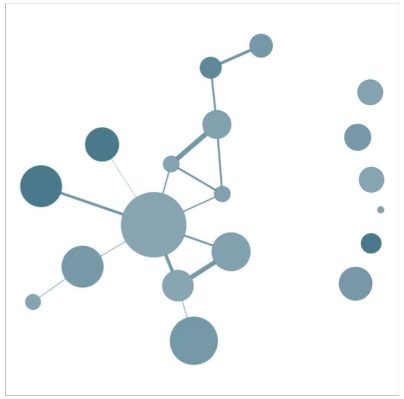
Metadata | Serum/fecal metabolites | MGS abundance | GMM abundance

Integrative data analysis

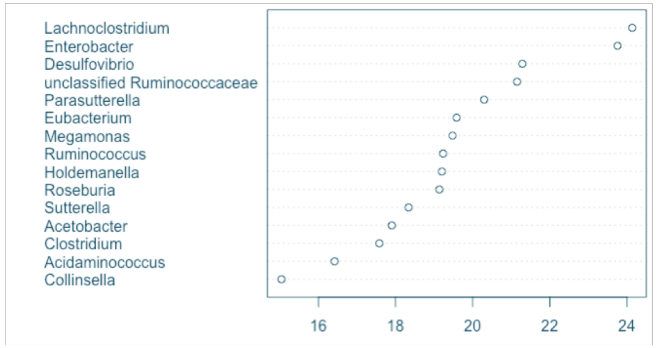
Correlation analysis



Network inference

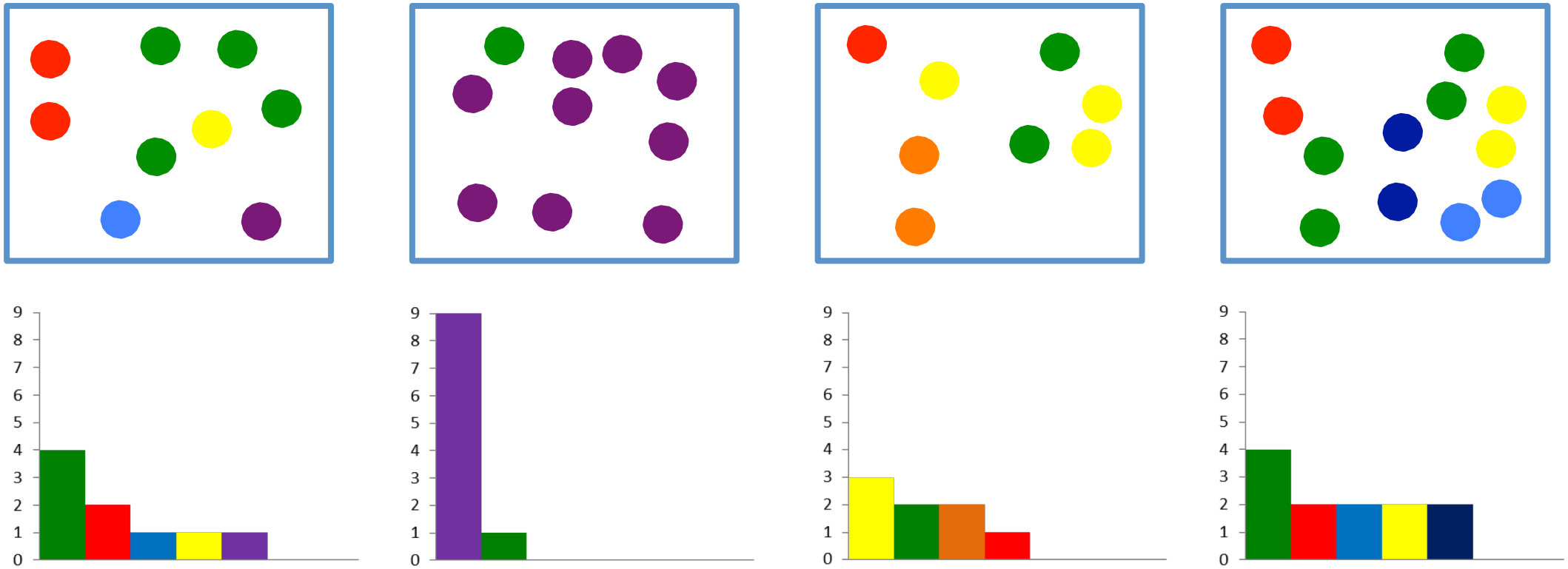


Variable selection and modelling

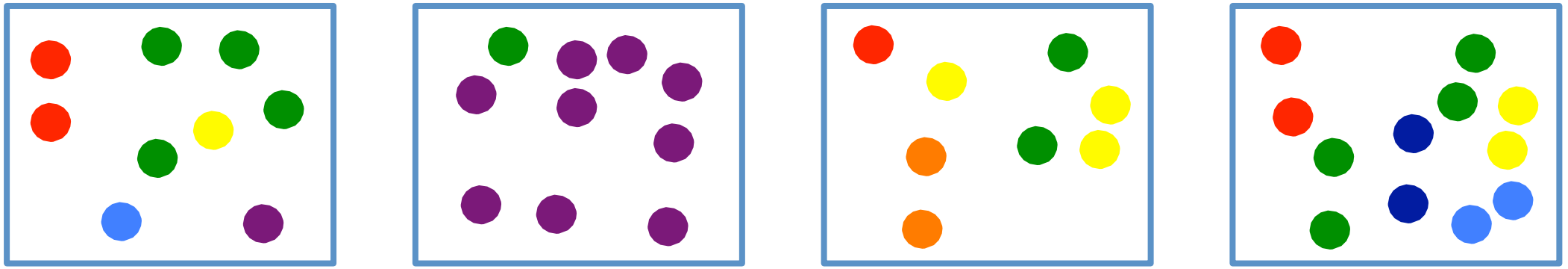


Analyses statistiques exploratoires

Diversité alpha : diversité au sein d'un échantillon

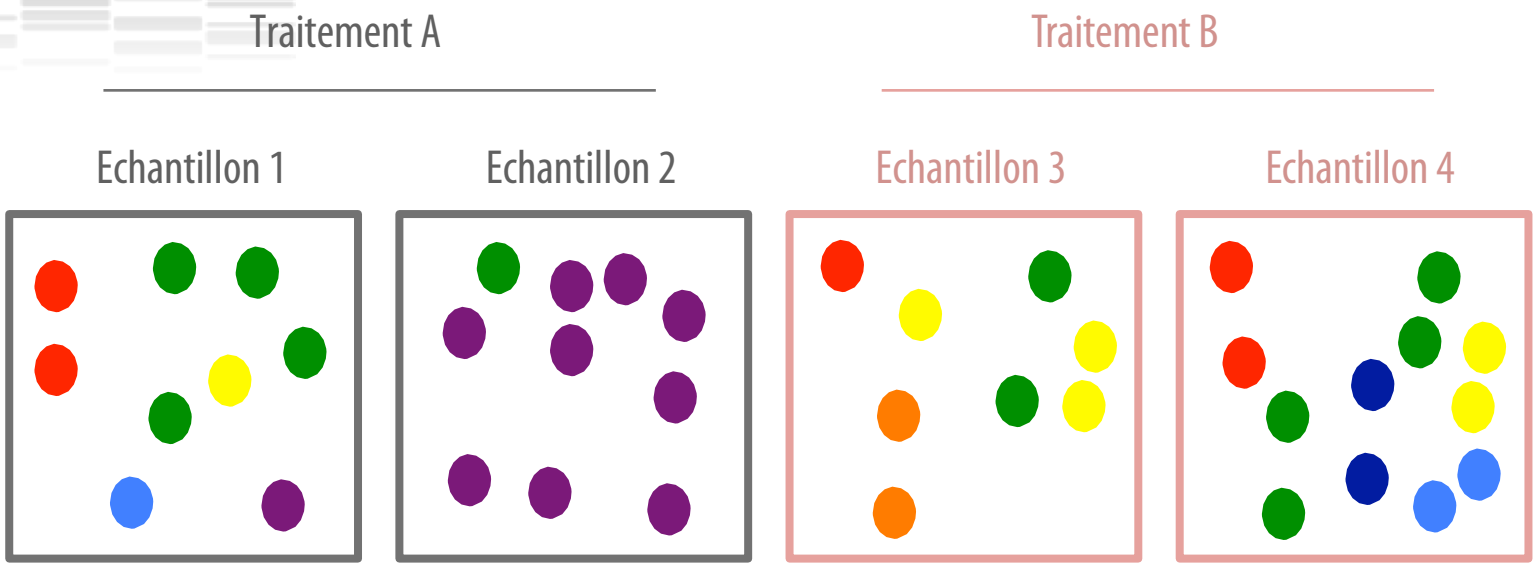


Diversité alpha : diversité au sein d'un échantillon



- Estimateurs de la richesse et de la diversité
  - Nombre d'espèces différentes : *Richesse* (R)
  - Nombre d'espèces estimées : *Indice de Chao1*
  - Information statistique : *Indice de Shannon*
  - Dominances : *Indices de Simpson inverse*

Diversité beta : diversité entre les échantillons



- Quelle est l'influence des traitements A et B ?
- L'échantillon 1 est-il plus similaire à l'échantillon 2 qu'aux échantillons 3 et 4 ?

Diversité beta : diversité entre les échantillons

- Calcul des distances/dissimilarités entre les échantillons x_i

  $\mathbf{x}_i = (x_{1i}, \ldots, x_{Si})$: vecteur de fréquence / abondance / présence

  - Euclidienne $E(i,j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$

  - Rapproche les échantillons dont les profils de comptages sont proches
  - Influencé par les espèces les plus abondantes
  - Min : 0 ssi les échantillons sont identiques
  - Max : non borné

Diversité beta : diversité entre les échantillons

- Calcul des distances/dissimilarités entre les échantillons x_i

  $\mathbf{x}_i = (x_{1i}, \ldots, x_{Si})$: vecteur de fréquence / abondance / présence

  - Euclidienne
  - Bray-Curtis

$$B(i,j) = \frac{\sum_{s=1}^{S} |x_{si} - x_{sj}|}{\sum_{s=1}^{S} x_{si} + x_{sj}}$$

  - Prend en compte l'abondance des espèces
  - Min : 0 si les échantillons sont identiques
  - Max : 1 si les échantillons sont totalement dissemblables

Diversité beta : diversité entre les échantillons

- Calcul des distances/dissimilarités entre les échantillons x_i

  $\mathbf{x}_i = (x_{1i}, \ldots, x_{Si})$: vecteur de fréquence / abondance / présence

  - Euclidienne
  - Bray-Curtis
  - Jaccard
  - Jensen-Shannon
  - 1 – |corrélation Spearman / Pearson|
  - …

En pratique dans R

1. Calcul de la distance (dist, vegdist, as.dist)

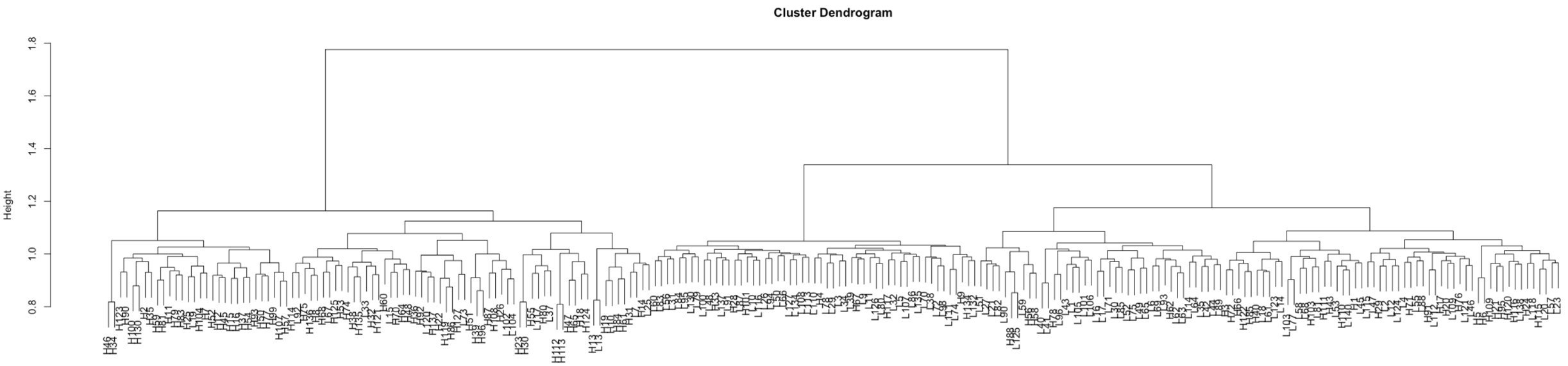
```
library(vegan)  # Pour le calcul des distances
library(ade4)   # Pour la calcul de la PCoA
```

```
profile.dist = dist(x = microbiota.abundance, method = "euclidean")
profile.dist = vegdist(exp(microbiota.abundance), method = "bray")
profile.dist = as.dist(sqrt(1 - cor(t(microbiota.abundance), method = "pearson")**2 ))
profile.dist = as.dist(sqrt(1 - cor(t(microbiota.abundance), method = "spearman")**2 ))
```

En pratique dans R

1. Calcul de la distance (dist, vegdist, as.dist)
2. Représentation graphique
   - Sous forme de dendogramme

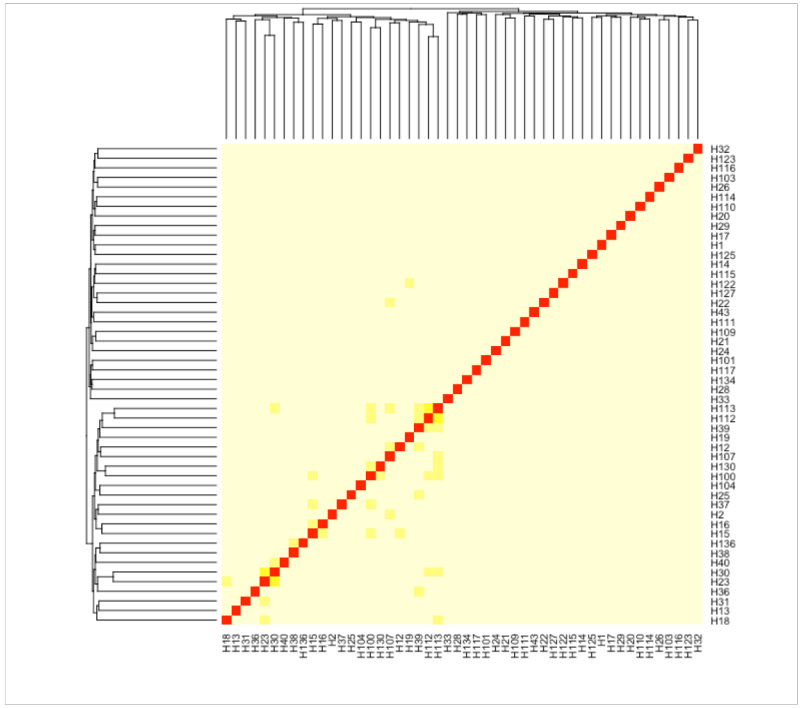
```
profile.hclust = hclust(profile.dist, method = "ward.D2")
plot(profile.hclust, xlab = "Distance linked to the Spearman correlation")
```



En pratique dans R

1. Calcul de la distance (dist, vegdist, as.dist)
2. Représentation graphique
   - Sous forme de heatmap

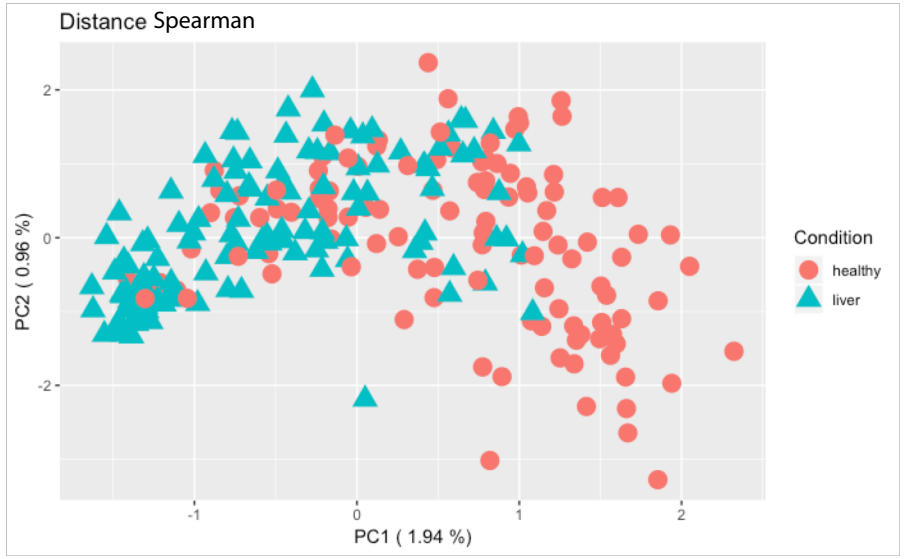
```
heatmap(as.matrix(profile.dist))
```

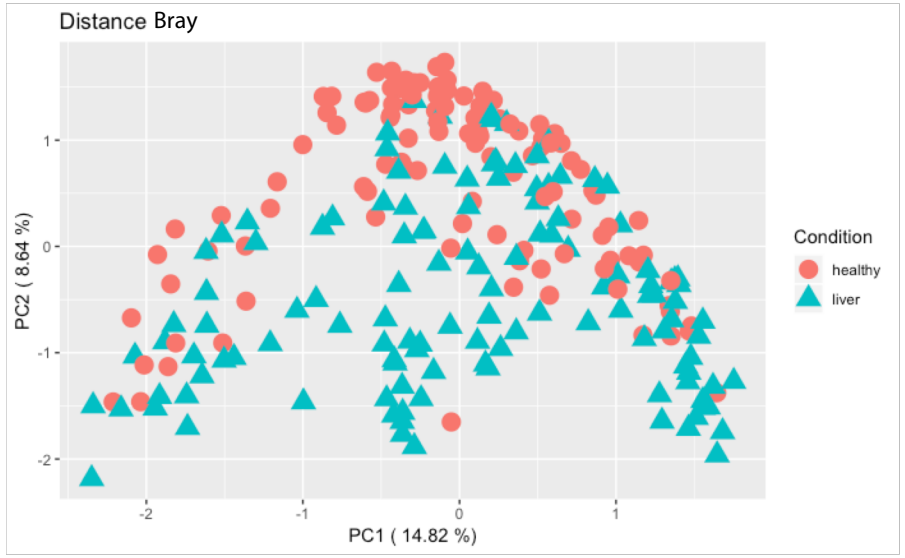


En pratique dans R

1. Calcul de la distance (dist, vegdist, as.dist)
2. Représentation graphique
   - Sous forme de PCoA

```
profile.pco = dudi.pco(d = profile.dist, scannf = FALSE, nf = 7)
```

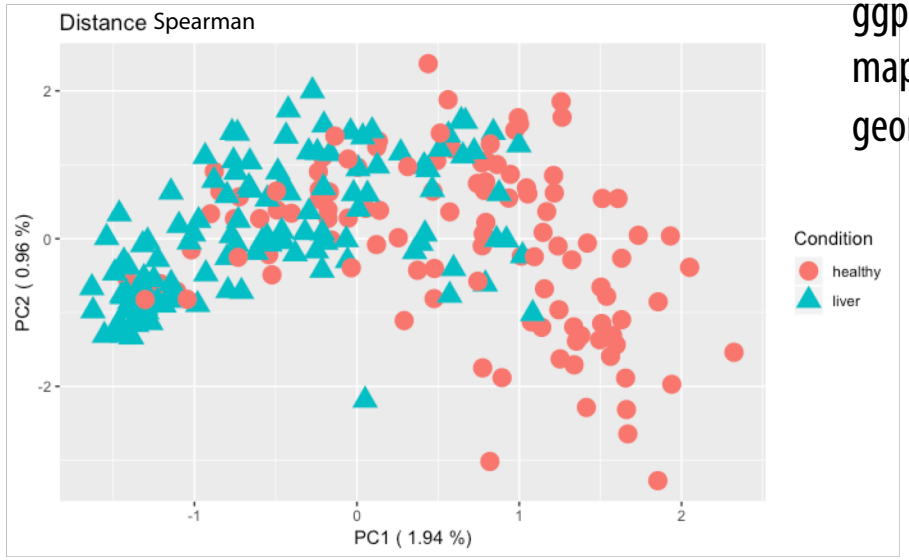




En pratique dans R

1. Calcul de la distance (dist, vegdist, as.dist)
2. Représentation graphique
   - Sous forme de PCoA

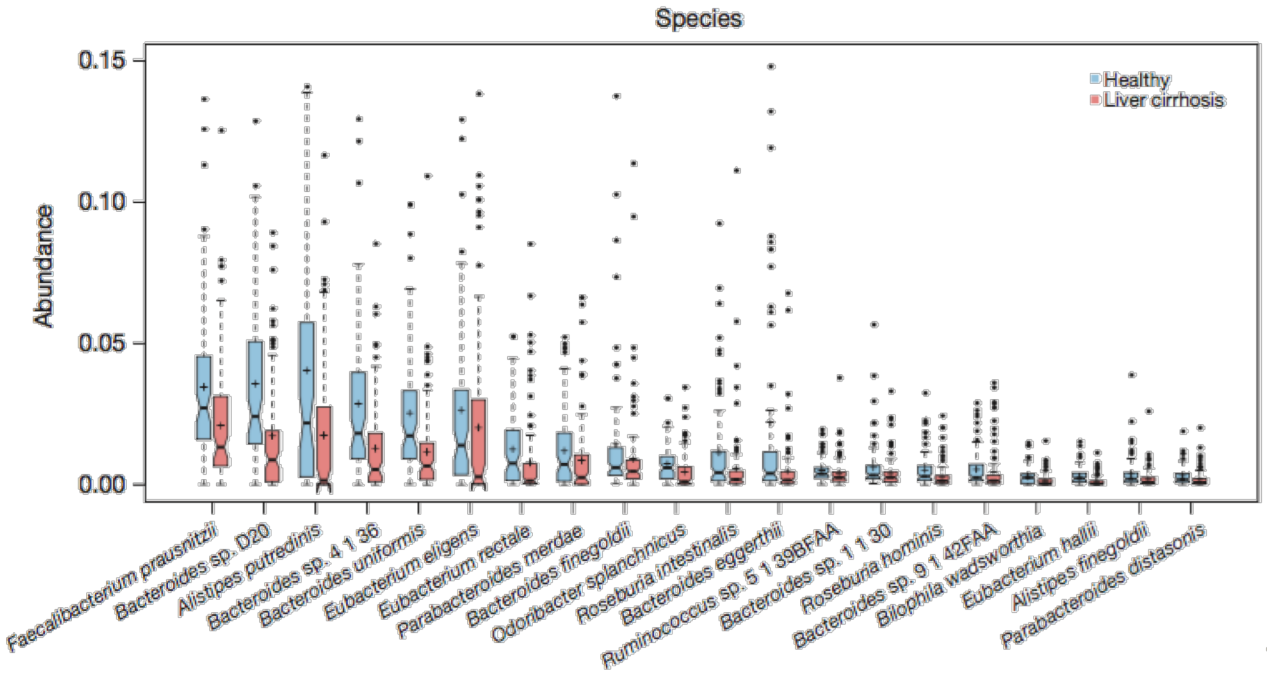
```
profile.pco  = dudi.pco(d = profile.dist, scannf = FALSE, nf = 7)
```

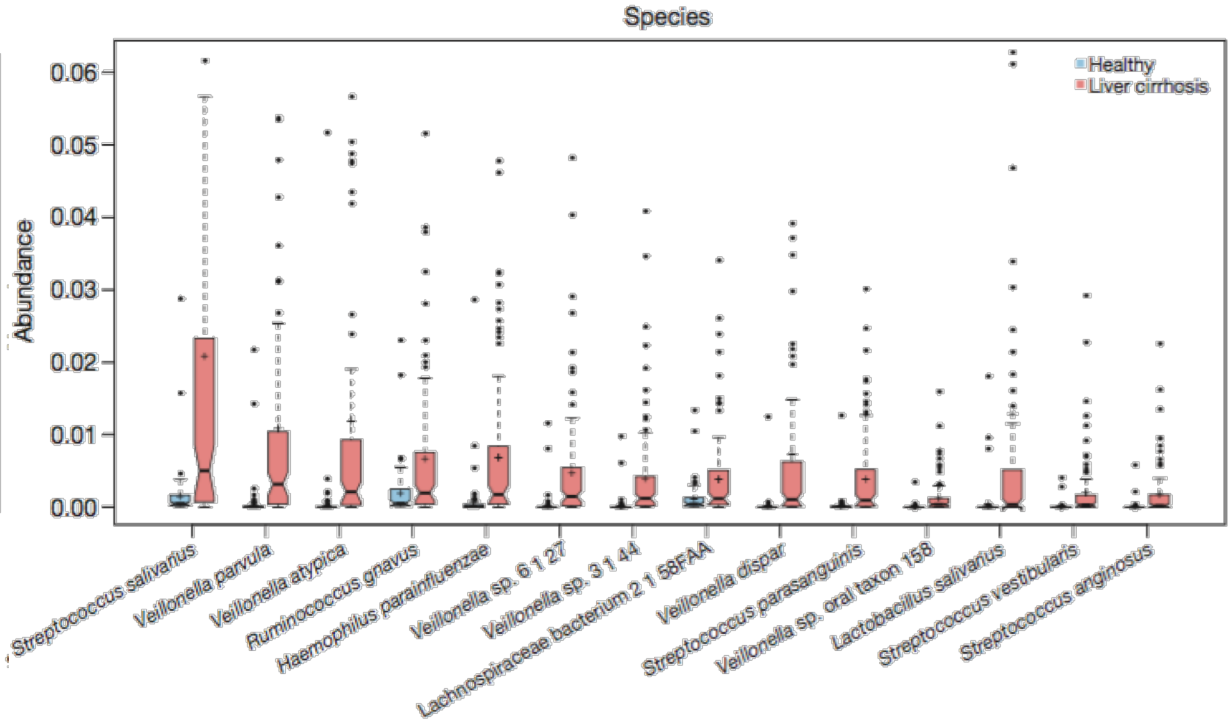



```
ggplot(data=profile.pco$l1,
mapping = aes(x = RS1, y = RS2, colour = metaFile$status, shape=metaFile$status)) +
geom_point(size = 5)
```

Identifier les facteurs de modification du microbiote

MGS différentiellement abondantes





Les tests statistiques

- L'analyse statistique pour identifier facteurs de modification du microbiote est basée sur un **échantillon représentatif**
- L'abondance relative d'une espèce peut être considérée comme une **variable aléatoire** réalisée dans chaque métagénome individuel
- L'hypothèse nulle (**H0**) postule que la distribution de cette variable aléatoire est indépendante du facteur examiné.

  (H0) : La différence observée dans l'abondance relative d'une espèce microbienne entre les sujets sains et les patients est simplement due au hasard

- Les différences entre les distributions sont évaluées, pour conclure ou non au **rejet de l'hypothèse nulle**

Ajustement des pvalues pour les tests multiples

- Dans le cas de **comparaisons multiples** il est nécessaire de contrôler le nombre de « fausses découvertes »
- En supposant un seuil de 0,05 pour chaque test, nous admettons une probabilité de 5 % de faire une « fausse découverte »
- Cependant, la probabilité de faire une « fausse découverte » est **beaucoup plus grande** dans le cas des tests multiples :

  Deux tests : 1 - (1 - 0,05) × (1 - 0,05) = **0,0975**
  Trois tests : **0,14**
  Quatre test : **0,4**

- Afin d'éviter de tirer des conclusions erronées, les pvalues doivent être ajustées par rapport au nombre de tests effectués

Les méthodes de correction des pvalues

- Contrôle du **FWER** (family-wise error rate) : probabilité d'avoir *au moins une erreur de type I*

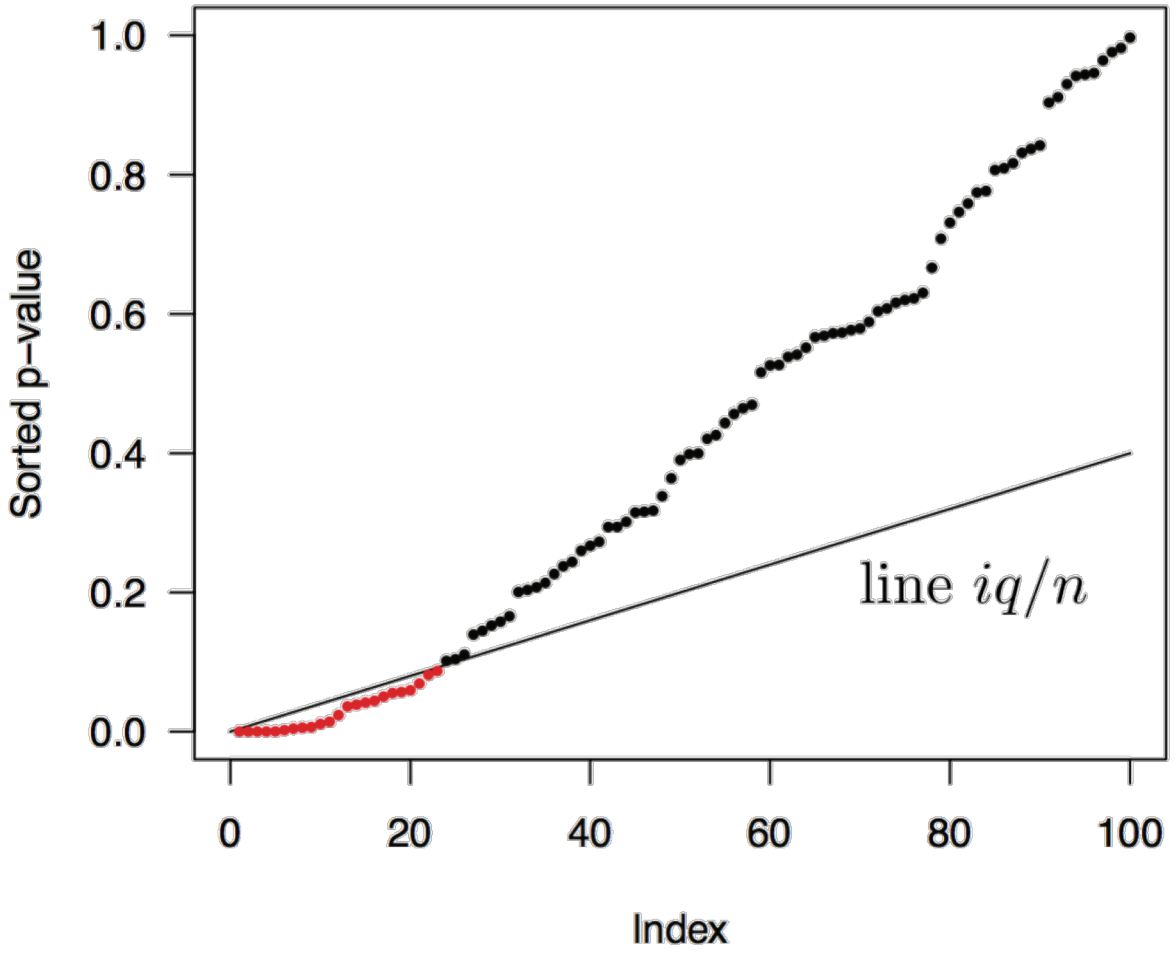
  Ex : Méthode de Bonferroni : diviser le seuil choisi par le nombre de test effectué

- Contrôle du **FDR** (false discovery rate) : *proportion* attendue de fausses découvertes
- Un FDR de 0,05 signifie qu'environ 5% des tests significatifs seront des «fausses découvertes».

  Ex : Méthode de Benjamini – Hochberg

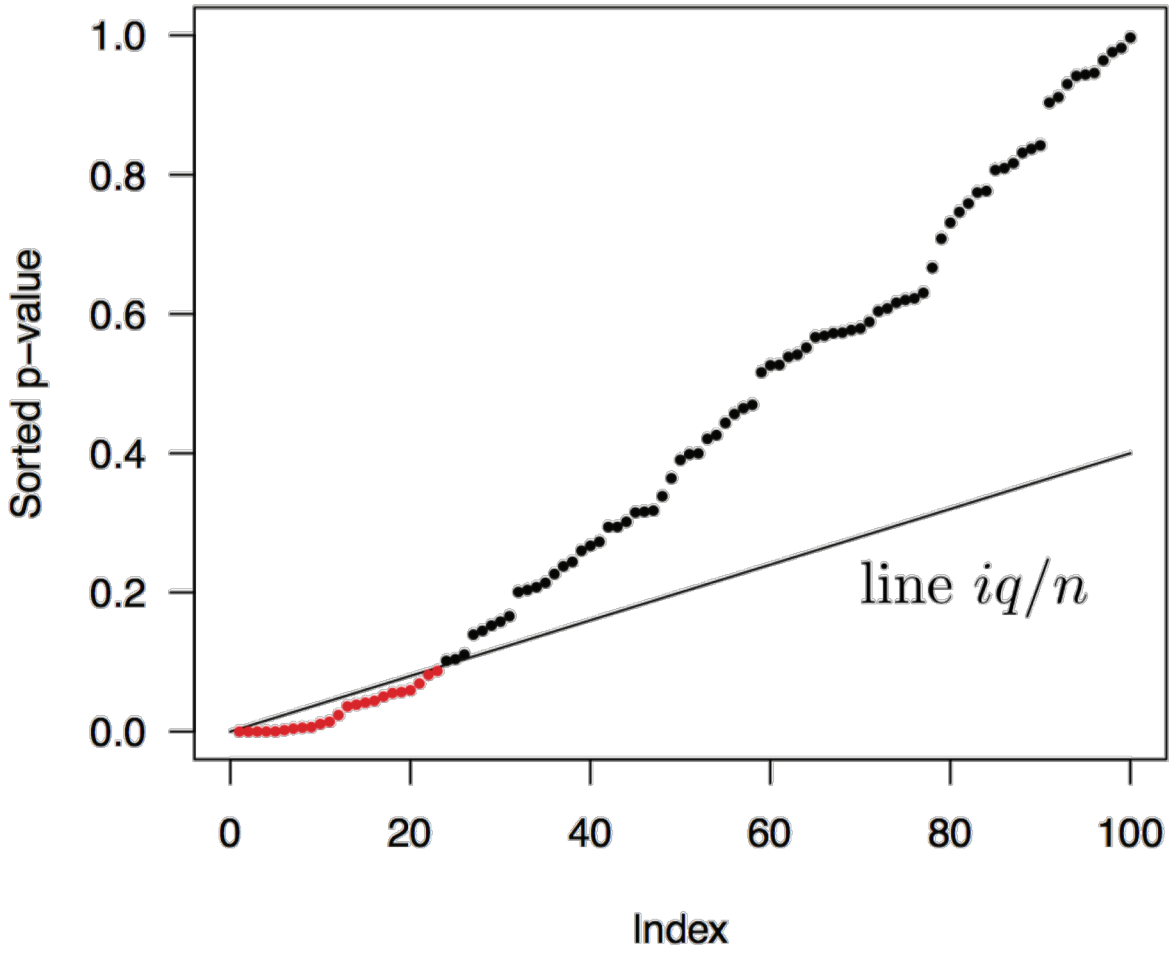
Le contrôle du FDR de BH (sous hypothèse d'indépendance)

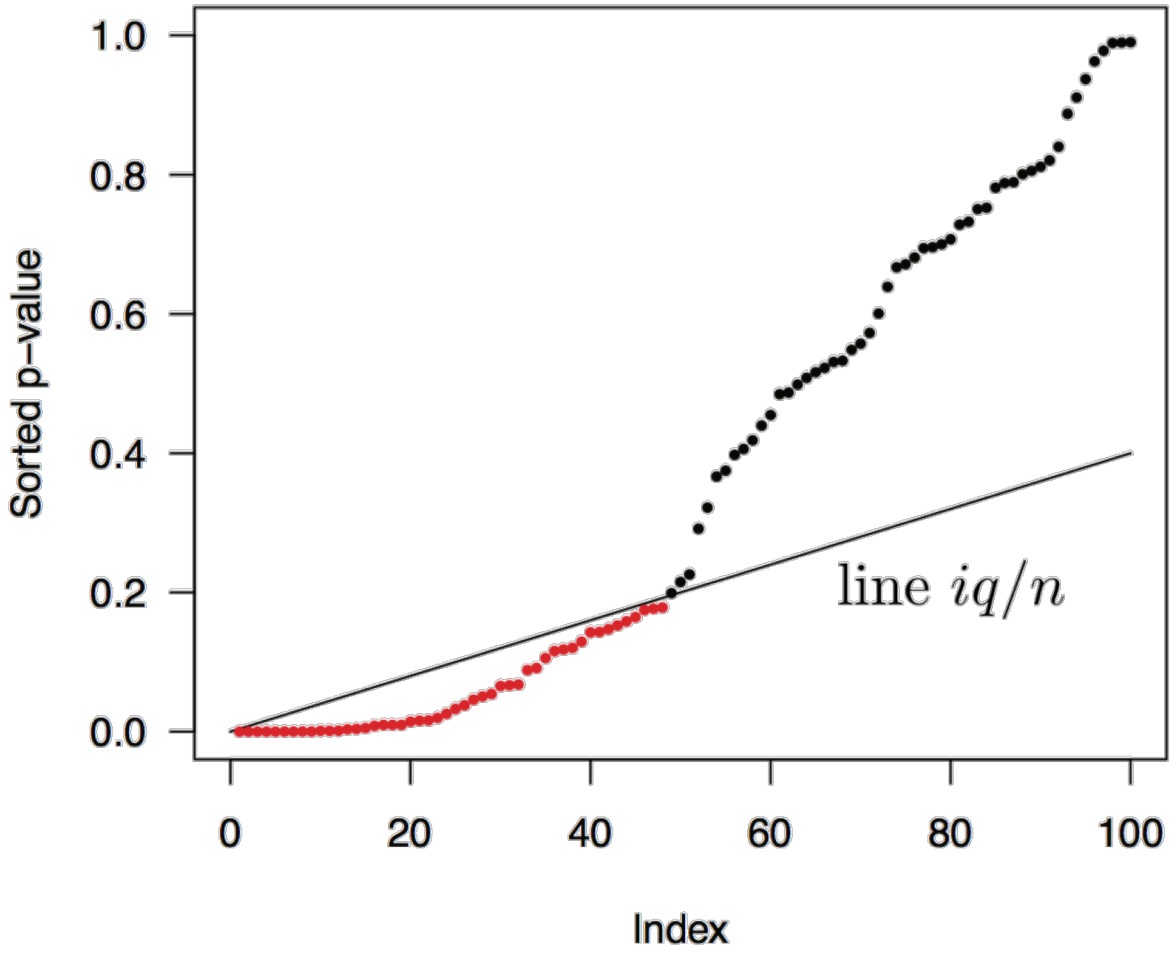
- Trier les pvalues: $p(1) \leq p(2) \leq \ldots \leq p(n)$ (de la plus à la moins significative)
- Choisir un seuil de FDR q



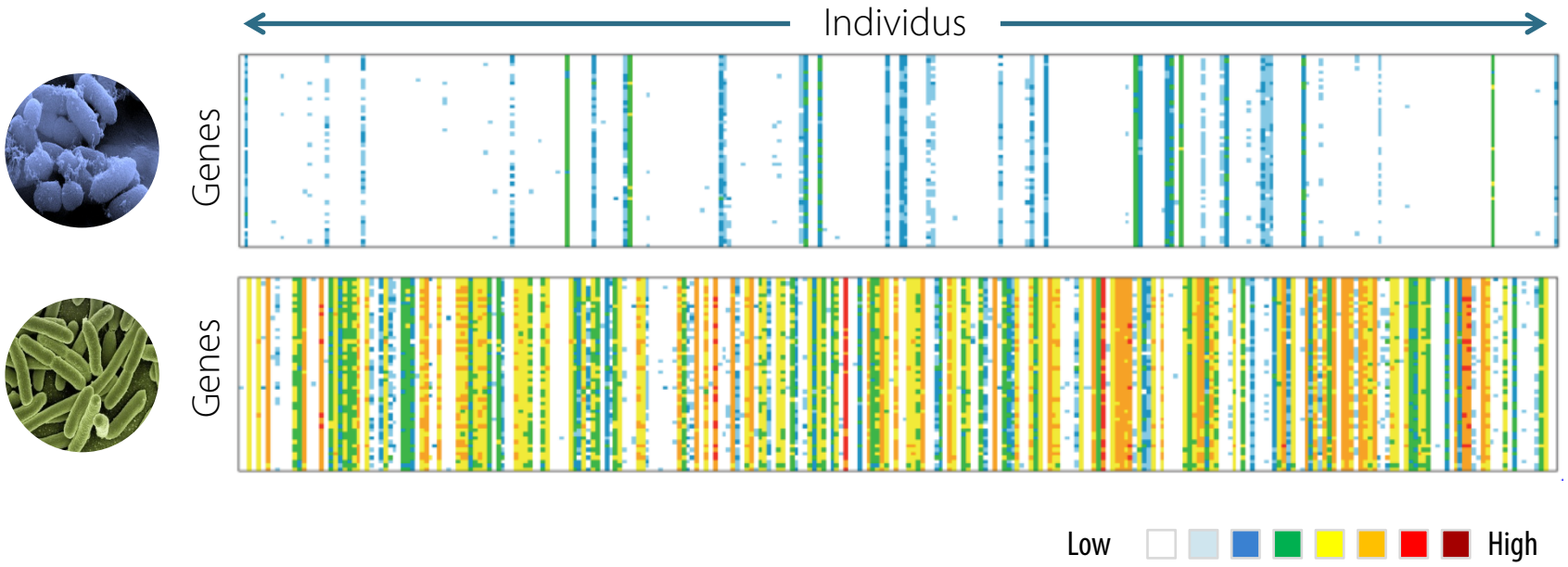
Le contrôle du FDR de BH (sous hypothèse d'indépendance)

- Trier les pvalues: $p(1) \leq p(2) \leq \ldots \leq p(n)$ (de la plus à la moins significative)
- Choisir un seuil de FDR q

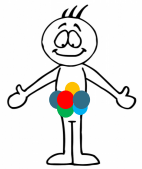




Validation de la pertinence des résultats : MGS barcode



Exemple : l'étude « liver cirrhosis »



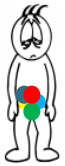
83 H

ranksum $q<10^{-4}$

75 245 gènes

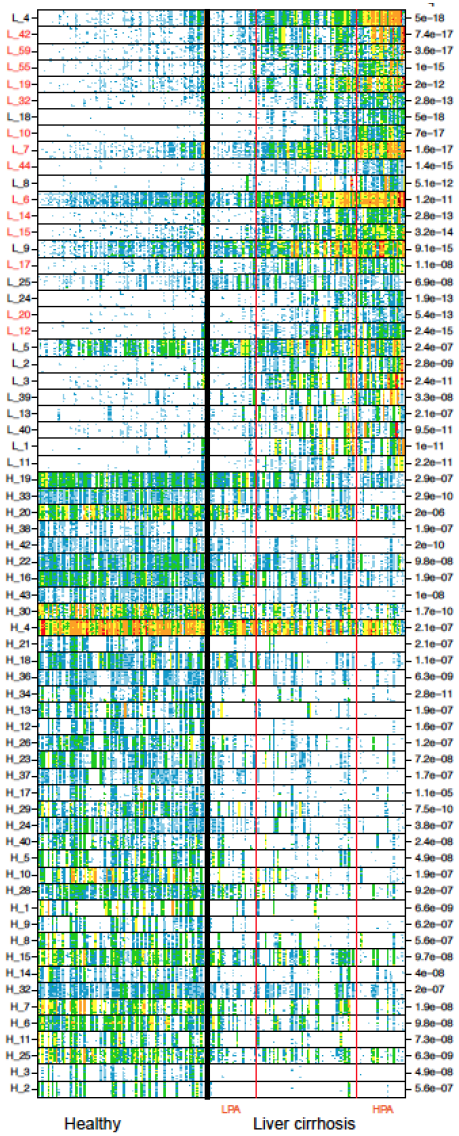
↓

66 MGS



98 LC

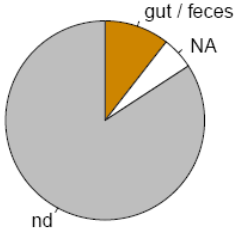
28 LC MGS



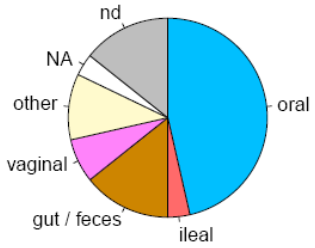
38 Healthy MGS

Taxonomie des MGS

38 Healthy MGS



28 LC MGS



Espèces bucales

Qin Nan et al. Nature 2014

Plan du TP

- I. Identification des MGS contrastantes (différentiellement abondantes)
- II. Traitement d'une matrice de comptage de gènes
  - 1. Importer la matrice
  - 2. Distribution du nombre de reads mappés
  - 3. Downsizing à 5000 reads
  - 4. Normalisation
  - 5. Calcul de l'abondance des MetaGenomic Species (MGS)