Réseaux d’associations au sein de communautés bactériennes

Inférence de réseaux et distribution en loi de puissance de l’abondance des espèces

Arnaud Cougoul, Doctorant
Directrice : Gwenaël Vourc’h - Responsables : Patrick Gasqui et Xavier Bailly

13 / 10 / 2016

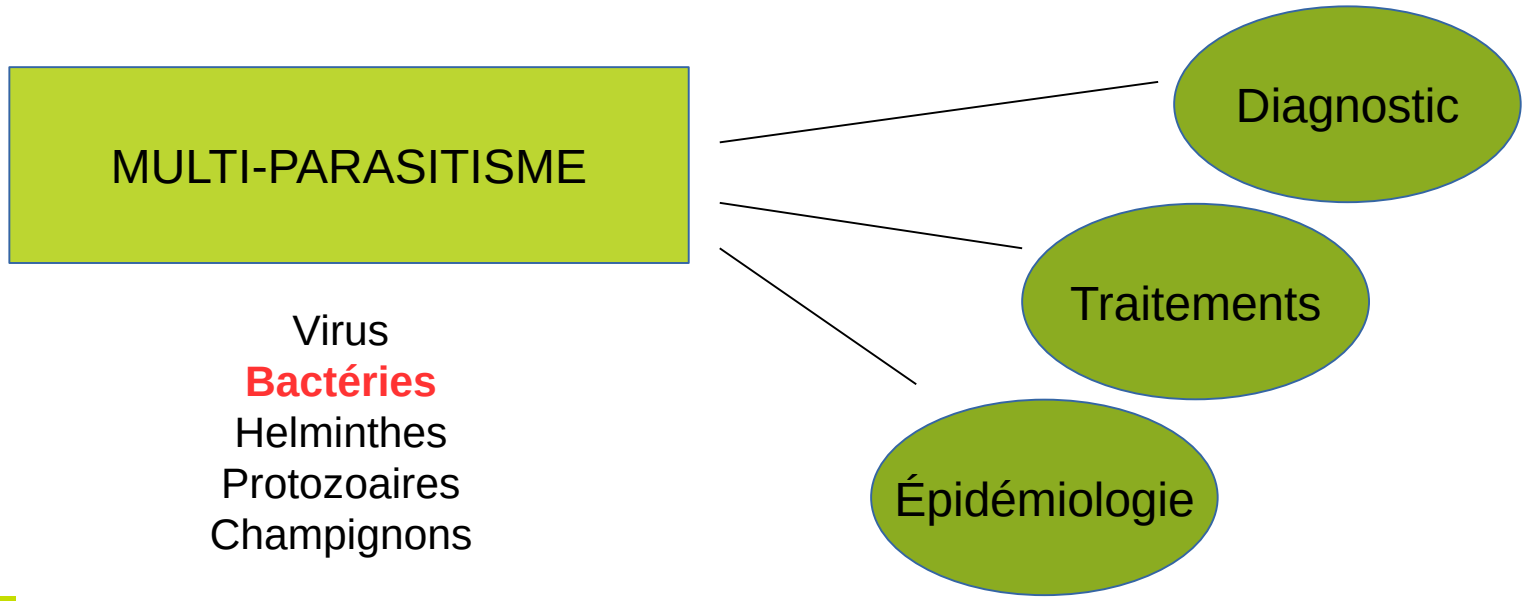
Multi-parasitisme

Contexte

Mes travaux vont à la suite de la thèse d'Élise Vaumourin (septembre 2014) sur la modélisation des associations entre parasites.

Lorsqu'il y a une infection, les organismes sont dans leur majorité infectés simultanément par plusieurs parasites.

Les infections multiples ont lieu dans 30 % voire 80 % des cas dans certaines populations humaines.



Interactions bactériennes et structure d'hôtes

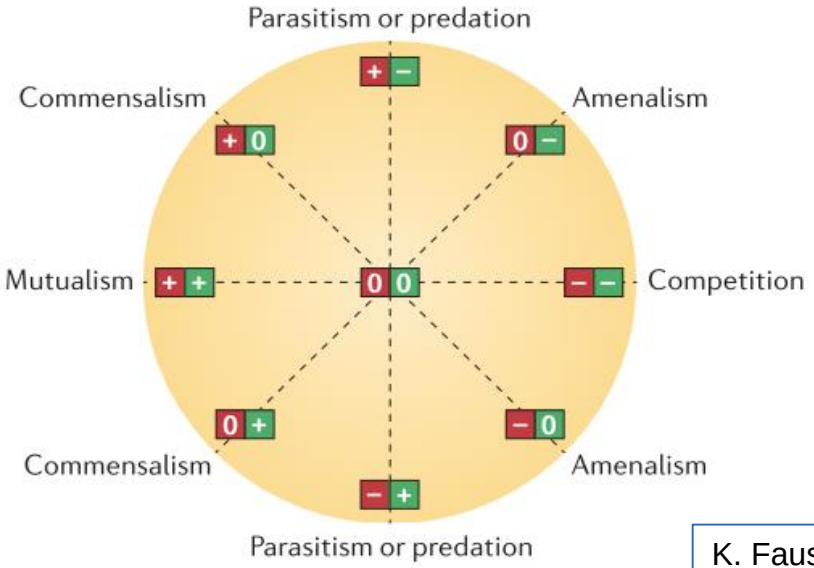
Les co-occurrences peuvent résulter du hasard.
Les co-occurrences nous intéressent quand elles sont **plus ou moins fréquentes que celles attendues au hasard.** On parle alors d'**associations.**

Ces associations peuvent résulter :

- d'un phénomène **d'interactions biologiques** entre les parasites.

Co-occurrence
i.e. interactions positives

mutualisme et commensalisme



Co-exclusion
i.e. interactions négatives

compétition, antagonisme et amensalisme

K. Faust and J. Raes, 2012

Interactions bactériennes et structure d'hôtes

Les co-occurrences peuvent résulter du hasard.
Les co-occurrences nous intéressent quand elles sont **plus ou moins fréquentes que celles attendues au hasard.** On parle alors d'**associations.**

Ces associations peuvent résulter :

- **de facteurs de risques communs** des hôtes

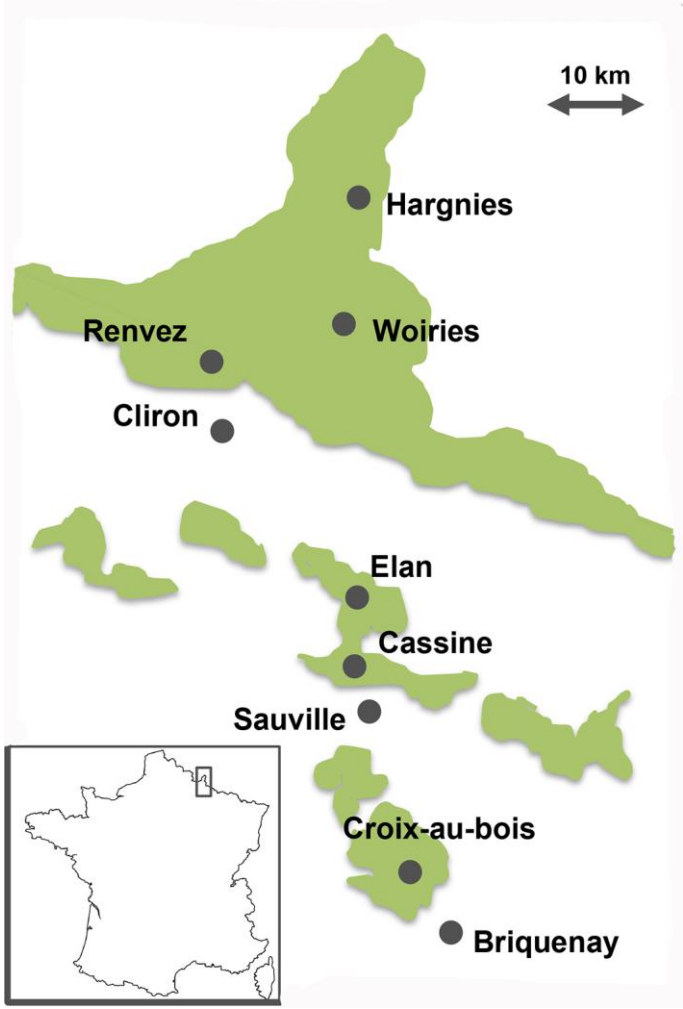
Par exemple :

Répartition spatiale,
Climat,
Comportement,
Susceptibilité physiologique

Données microbiote tiques Ardennes

JF Cosson et al.

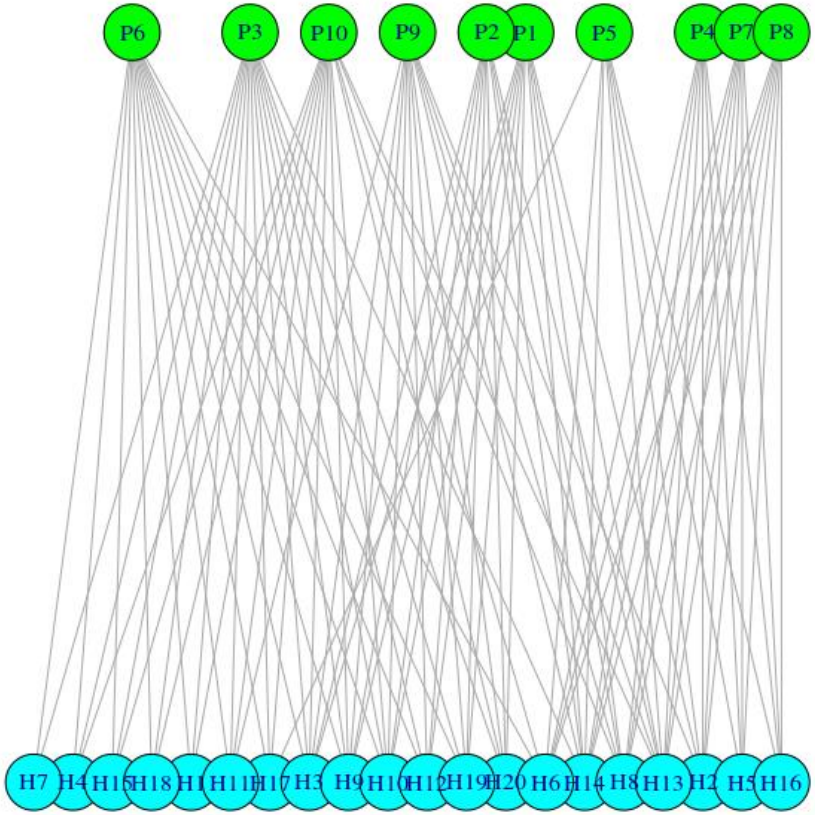
- Les tiques sont vecteurs de pathogènes
- 267 tiques *ixodes ricinus* femelles récoltées
- 9 sites des Ardennes → facteur environnemental
- Composition du microbiote obtenue par séquençage de la région 16S du génome
- 629 genres de bactéries observés
- **Données transversales de présence/absence de bactéries**
  - Données illumina pas assez nombreuses pour avoir des données quantitative (faibles efforts de séquençage)
  - Pathogènes à faible densité
  - Echantillonnage destructif



Exemple de graphe biparti pour représenter un réseau hôtes-bactéries

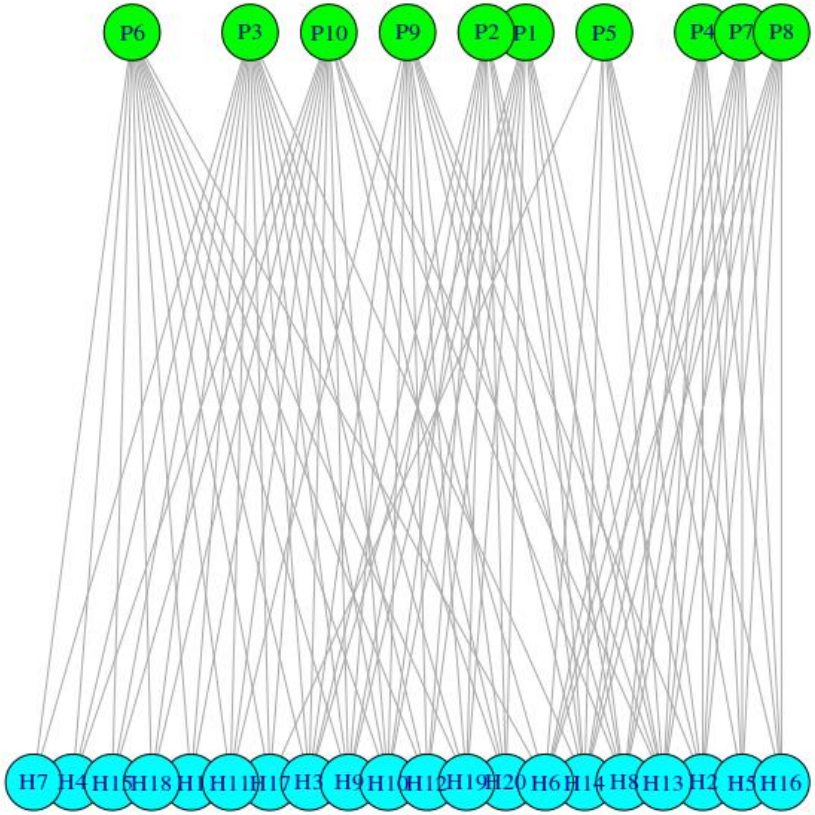
Couche parasites

Couche hôtes



Exemple de graphe biparti pour représenter un réseau hôtes-bactéries

Couche parasites

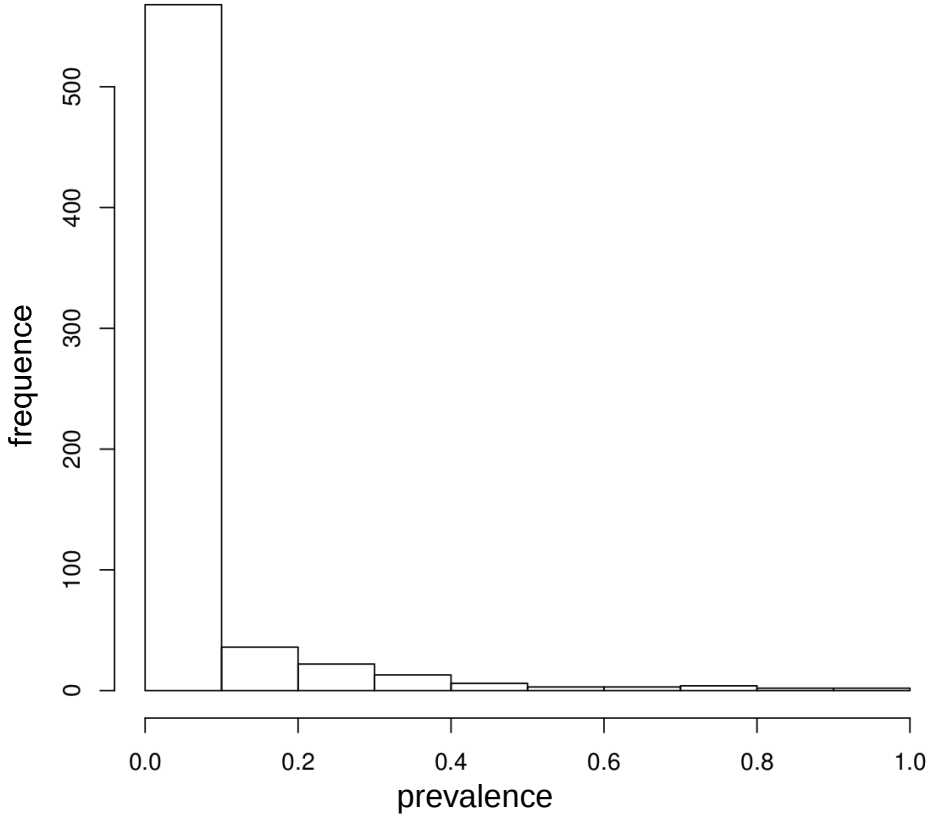


Couche hôtes

degré / nb_ech = prévalence d'un pathogène dans une population

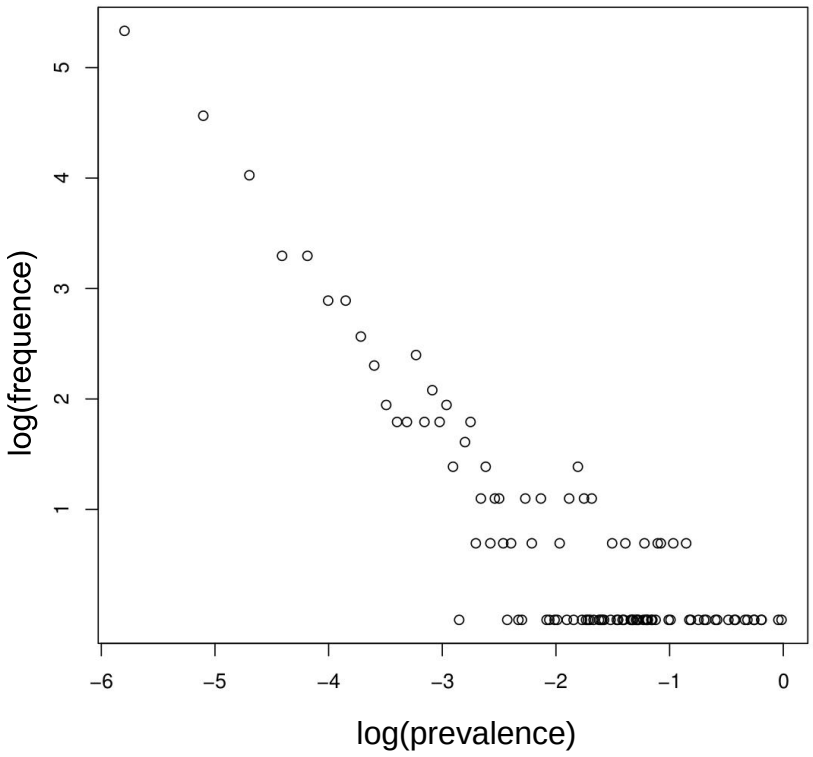
Distribution des prévalences des espèces en loi de puissance

Histogramme des valeurs de prévalences



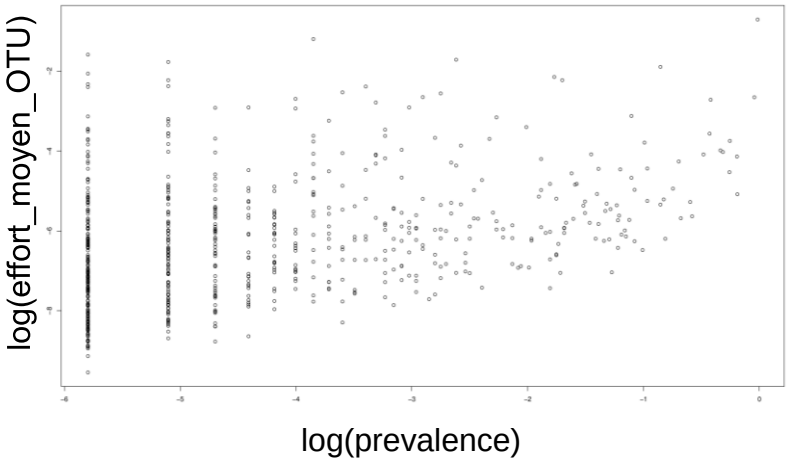
Nombre de bactéries par valeurs de prévalences

Échelle log-log

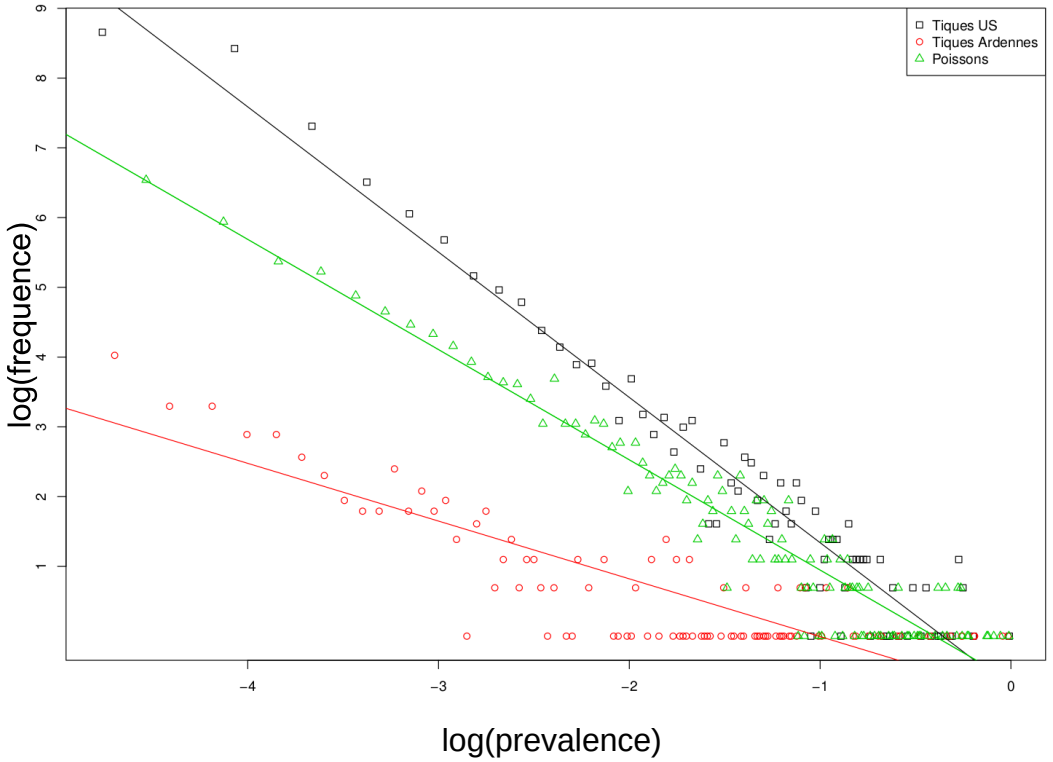


Généricité de la loi de puissance

1. Les OTUs à faible prévalence ne sont pas rares du fait de leur faible densité



2. Comparaison avec d'autres données de microbiotes.
Échelle log-log



3. Loi de puissance observée en écologie des communautés...

McGill, B. J. (2003)

et dans les communautés bactériennes
Ex : *Human microbiome project*

Microbiotes épinoches : Bolnick, D. I. et al. (2014)
Microbiotes tiques US : Van Treuren W. et al. (2015)

Questionnement

Plus de la moitié des espèces ont une prévalence inférieure à 1%
Plus de 500 espèces ont une prévalence inférieure à 10%

Quels effets ont les faibles prévalences
sur l'inférence d'un réseau biologique ?
sur le calcul des corrélations?

Corrélation entre deux variables (présence/absence)

On utilise le **coefficient phi** pour mesurer la **corrélation entre deux variables de présence/absence** de bactéries x et y.
Ce coefficient est une variation de la définition de Pearson de *r* pour deux variables égales à 0 ou à 1.

Tableau Croisé :

	$y = 1$	$y = 0$	total
$x = 1$	n_{11}	n_{10}	$n_{1\bullet}$
$x = 0$	n_{01}	n_{00}	$n_{0\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}$$

$$\phi^2 = \frac{\chi^2}{n}$$

- Le coefficient phi s'exprime aussi en fonction des prévalences :

$$\phi(x,y) = \frac{P(x = y = 1) - p_x \times p_y}{\sqrt[2]{p_x \times (1 - p_x) \times p_y \times (1 - p_y)}}$$

Fahrmeir and Tutz, 1994

Où

- $p_x = n_{1*} / n$ et $p_y = n_{*1} / n$ sont les fréquences marginales de x=1 et de y=1 respectivement

- P (x=y=1) est le taux de cooccurrences observé

Amplitude des corrélations dépendante des prévalences

Pour 2 valeurs de prévalences données p_x et p_y,
la corrélation dépend que du taux de cooccurrences P(x=y=1).

$$\phi(x,y) = \frac{P(x=y=1) - p_x \times p_y}{\sqrt[2]{p_x \times (1-p_x) \times p_y \times (1-p_y)}}$$

P (x=y=1) est borné par :

max (0 , $p_x + p_y - 1$) ≤ P(x=y=1) ≤ min (p_x , p_y)

Application pour $p_x = 0.05$ et $p_y = 0.6$,

P (x=y=1) minimum = 0
max = 0.05

$$\frac{-0.05 \times 0.6}{\sqrt[2]{0.05 \times 0.95 \times 0.6 \times 0.4}}$$

Corrélation min
- 0,28

$$\frac{0.05 - 0.05 \times 0.6}{\sqrt[2]{0.05 \times 0.95 \times 0.6 \times 0.4}}$$

Corrélation max
0,19

Des pathogènes rares...
Impossibilité de détecter une relation

Les prévalences induisent une borne minimale et maximale de la corrélation.

Plus la prévalence est petite plus ces bornes se resserrent.

Pour les faibles prévalences, il devient impossible de détecter une association.

La présence d'OTUs rares dégrade la performance des méthodes d'inférence de réseaux (S. Weiss et al, 2016).

Un seuillage des OTUs sur leur taux de présence (=prévalence) est donc nécessaire.

Filtering step

« Nettoyage » des données

Seuils existant :

Kurtz, Z. D. et al. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Comput. Biol. 11, 1–25 (2015).

Propose de filtrer les OTUs avec un taux de présence dans les échantillons inférieur à 37%.

Weiss, S. et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. Isme J 10, 1–13 (2016).

Propose d'enlever les OTUs les plus rares pour ne pas avoir plus de 50 % de zéros.

Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput. Biol. 8, 1–11 (2012).

Propose d'enlever les OTUs avec moins de 2 *reads* en moyenne par échantillons

Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front. Microbiol. 5, 1–14 (2014).

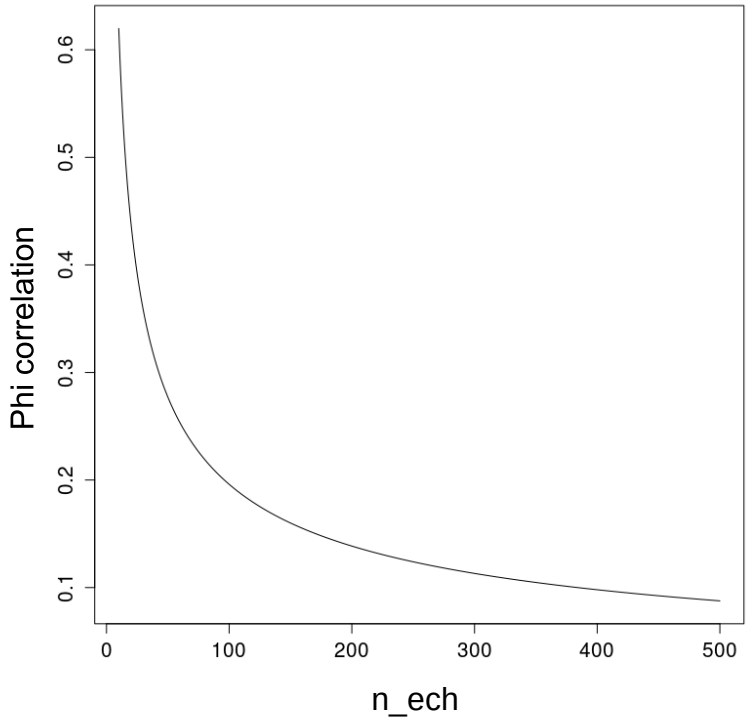
Propose d'enlever les OTUs les plus rares jusqu'à ce que la similarité moyenne des sites soit d'au moins 20%.

Notre proposition de seuillage sur les paires de prévalences

- Seuiller a partir des IC, la distribution de phi est connue.

  $phi^2 = chi^2 / n$

- Amplitude de phi dépend uniquement des valeurs de prévalences de la paire étudiée
- Si amplitude de phi est inclue dans l'IC, alors aucune détection sera possible



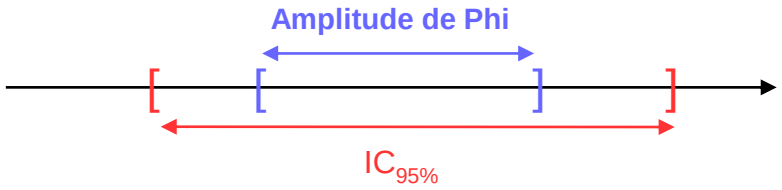
Valeur absolue minimale de phi pour être significativement différent de zéro en fonction du nombre d'échantillons

Application

Pour n_ech = 30, **$IC_{95\%}$ = [-0,36 ; 0,36]**

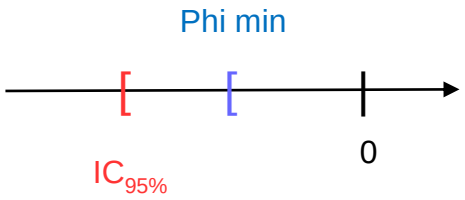
Avec **p_x = 0.05** et **p_y = 0.6,**

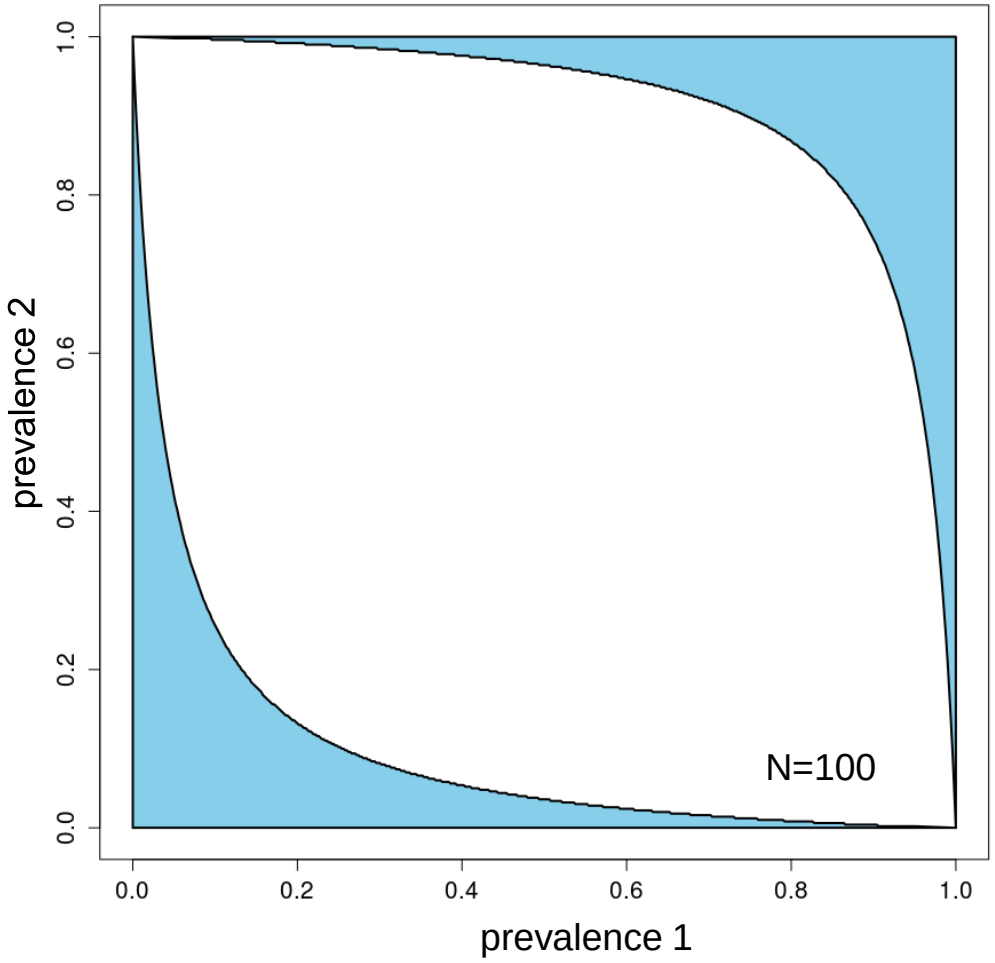
Phi_amp = [-0,28 ; 0,19]



Courbes de seuil de détection – corrélation négative

Dans la zone bleue, il est impossible de détecter une corrélation négative pour ces valeurs de paires de prévalences

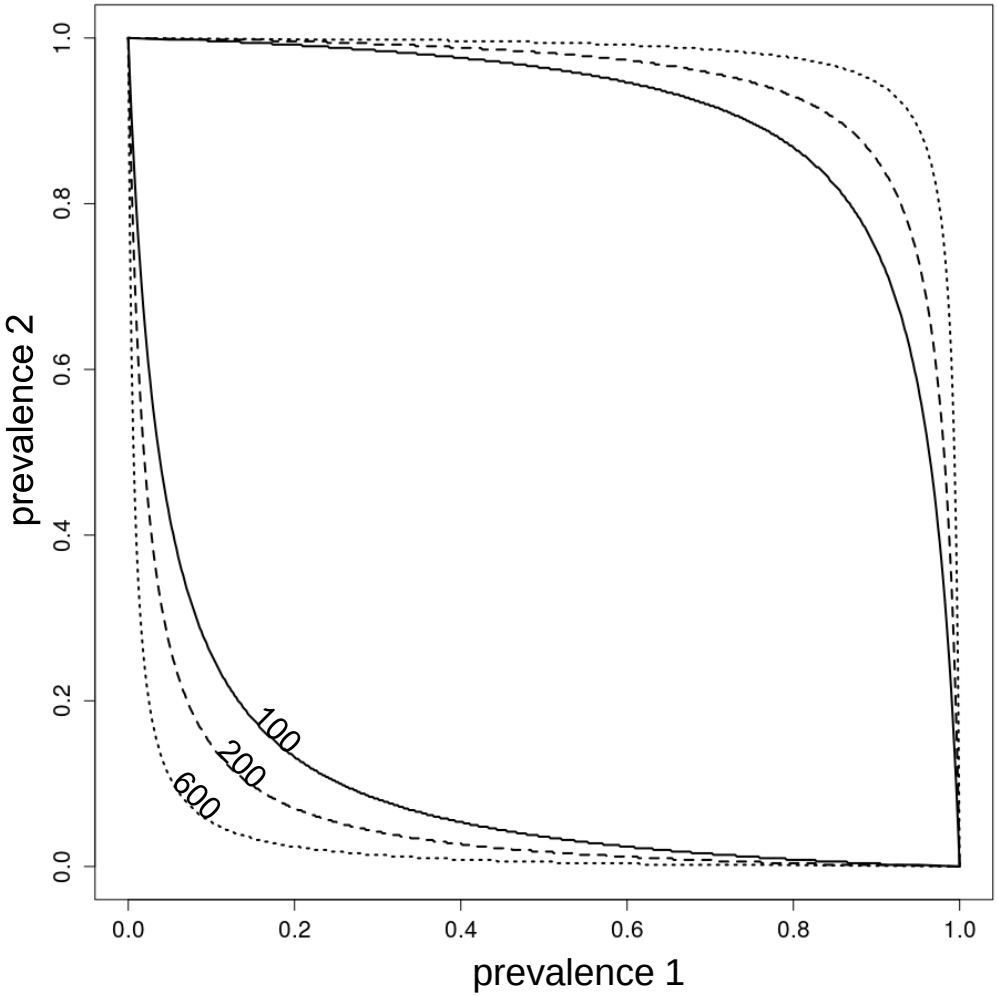




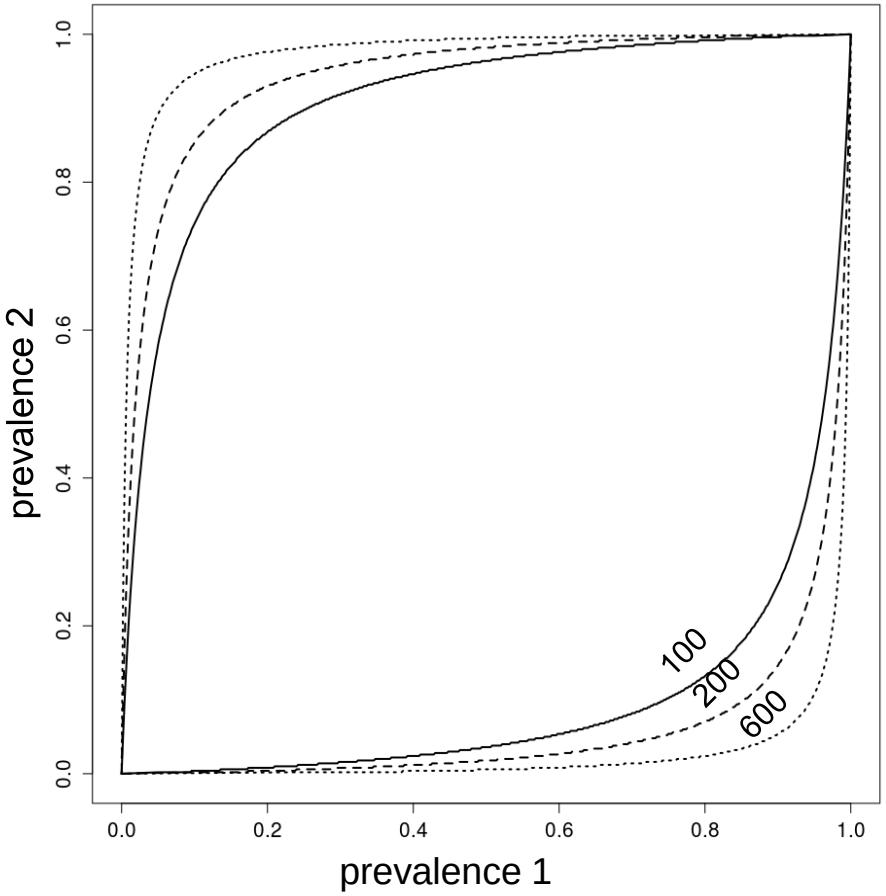
Courbes de seuil de détection dépendante du nombre d'échantillons

Plus le nombre d'échantillons augmente, moins les déficits de détections apparaissent.

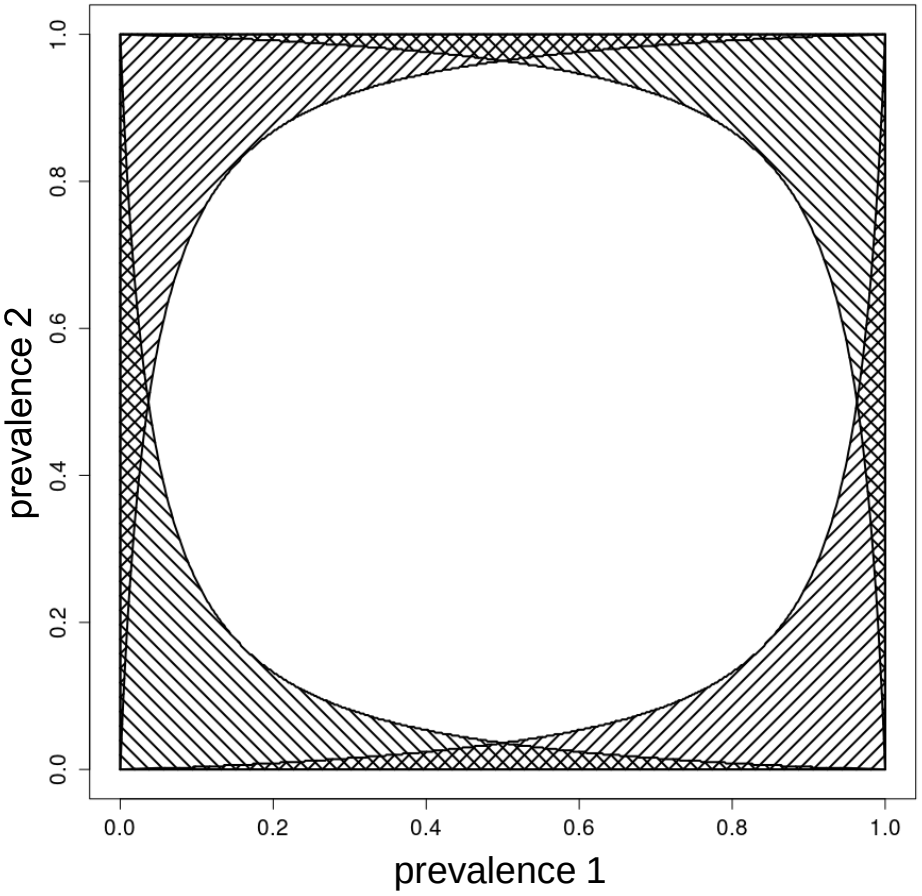
Sur les données de composition du microbiote de tiques, il est impossible de détecter une association négative sur plus de 90 % des paires de bactéries.



Courbes de seuil de détection – corrélation positive



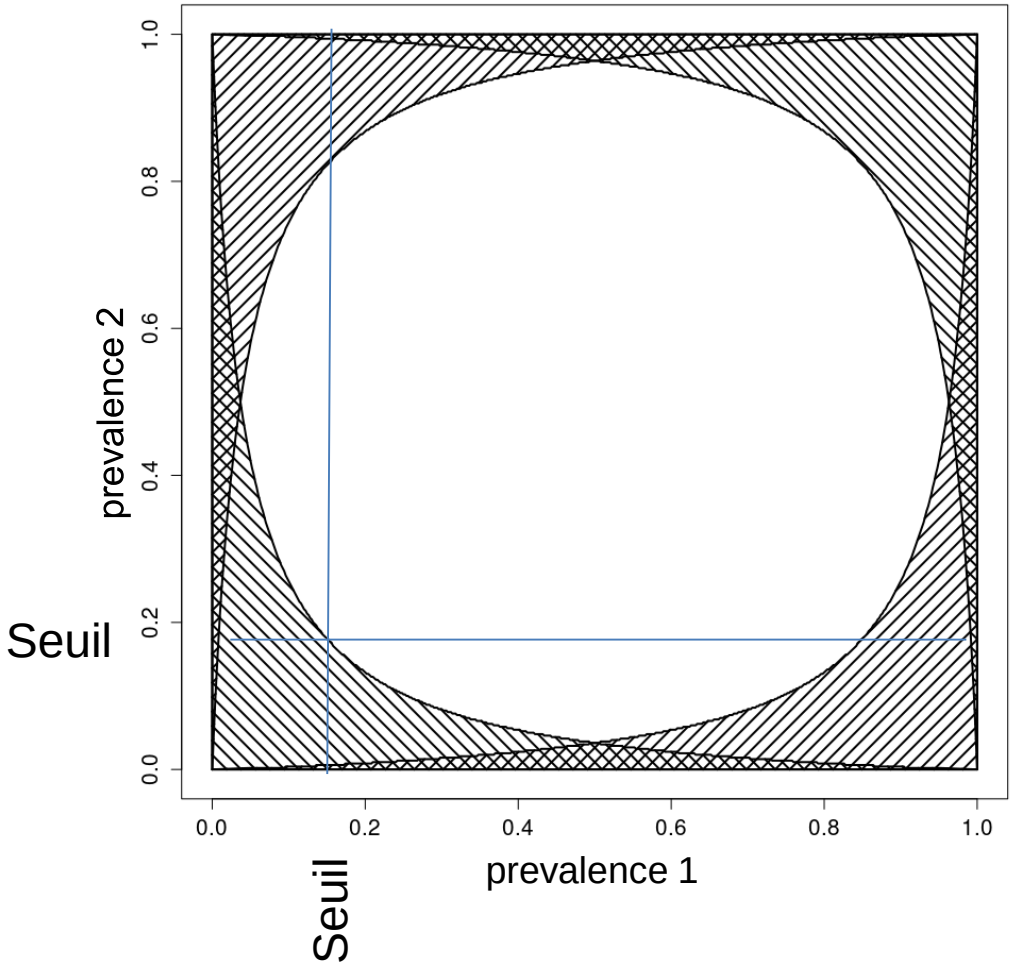
Les problèmes dans la détection d'une corrélation positive apparaissent quand une prévalence est faible et l'autre grande



Superposition des 2 courbes de limites de détection (corrélation positive et négative)

Courbes de seuil de détection – corrélation positive

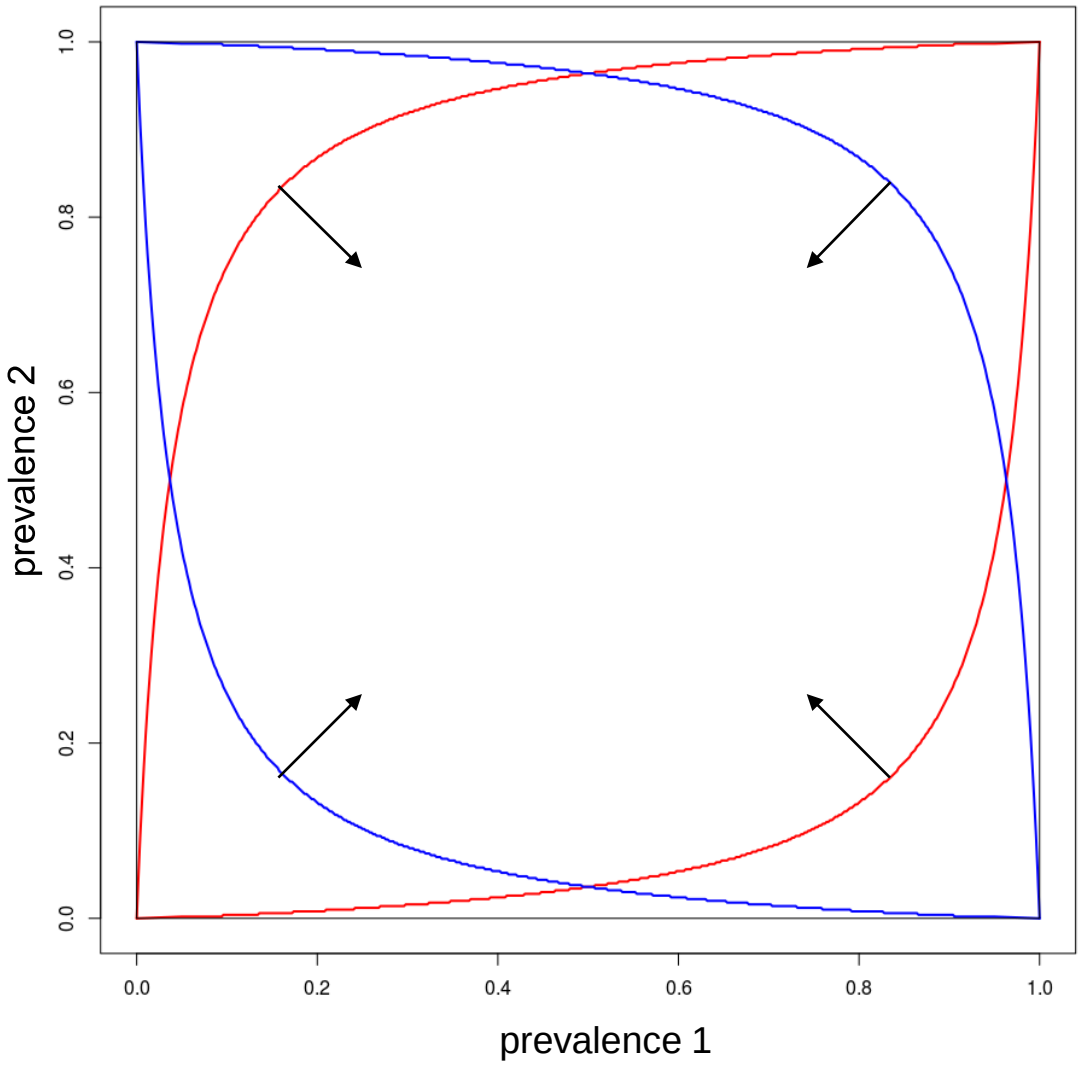
Difficulté d'avoir « une » valeur de seuil



Prise en compte des tests multiples

Lorsque l'on construit un réseau de corrélaton,
un grand nombre de test est réalisé

On veut contrôler le alpha globale



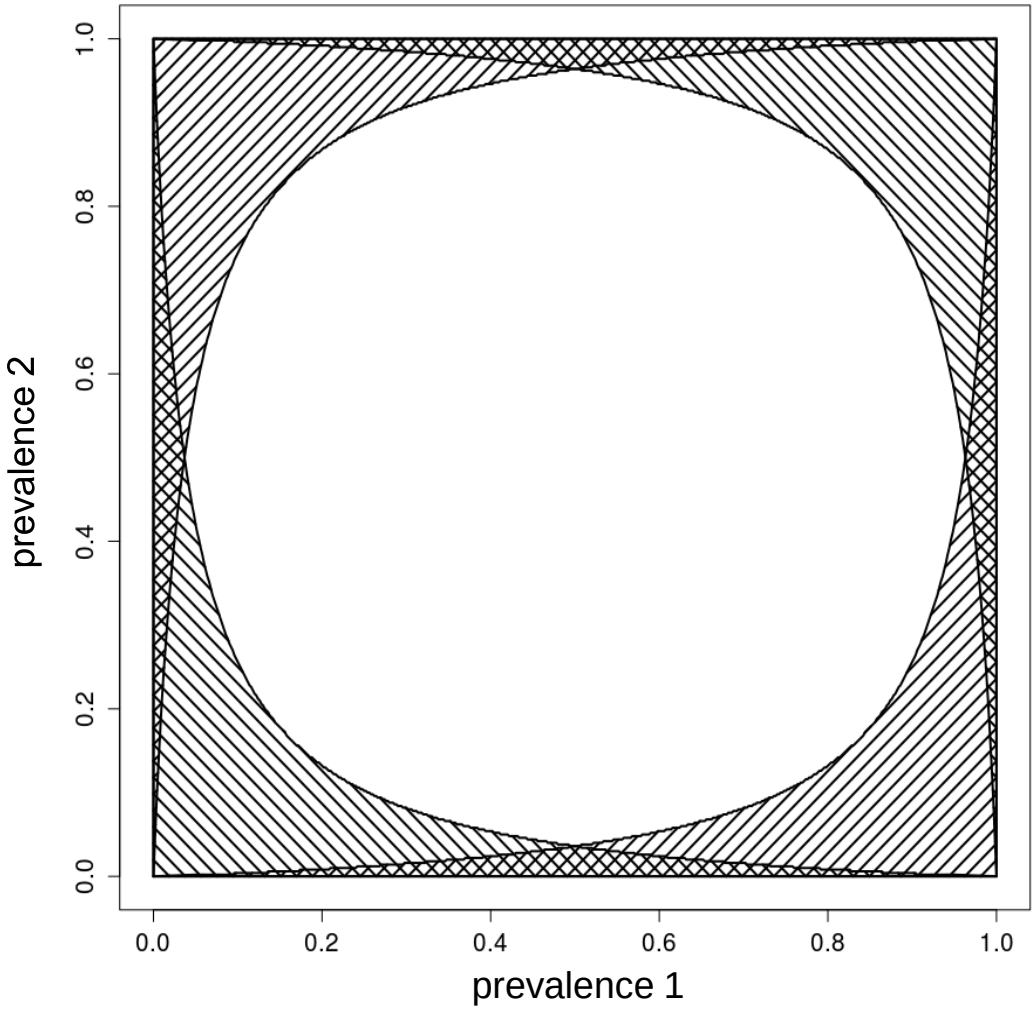
Conclusion

Nous proposons une méthode originale de seuillage des prévalences pour limiter le nombres de paires de bactéries à tester
Diminuer le nombre de tests et contrôler le risque alpha

Extension à d'autre mesure d'associations

Seuil sur des pairs de prévalences et pas sur une valeur précise

Permet de garder des pathogènes rares d'importance médicale que l'on aurait du supprimer avec un seuil



Ma thèse

Développer des méthodes statistiques basées sur la théorie des réseaux afin d'identifier des associations au sein d'une communauté bactérienne tout en tenant compte d'une potentielle structure dans la population hôtes

- 3 grands axes
  - Indicateurs réseaux
  - Comparaison réseaux
  - Données longitudinales

- Financement metaprogramme INRA GISA « Gestion intégrée de la santé des animaux » et MEM « Méta-omiques des écosystèmes microbiens »

Remerciement

Merci pour votre attention

Je tiens particulièrement à remercier

- Mes encadrants & l'unité d'Epidémiologie Animale
- L'équipe organisatrice du colloque CARTABLE