

# **Projet: ANOVA Phylogénétique**

Présentation du Mardi 26 Mars. 2024

---

Alizée Geffroy, Louis Lacoste, encadrés par Mélina Gallopin et Paul Bastide

M2 MathSV Université Paris-Saclay

1. Introduction
2. État de l'art
3. Calculs
4. Simulations
5. Application aux données réelles
6. Références et appendices



TODO Supprimer cette slide temporaire intro/contexte: biologique avec l'exemple de Chen (mettre l'arbre) + figure de l'article ? -> trouver les gènes différentiellement exprimés

Il existe déjà des méthodes statistiques pour cette problématique (EVEmodel ? State of the Art)

Transition avec le pourquoi du projet, trouver d'autres méthodes statistiques, adaptées de méthodes classiques qui pourraient bien marcher

Méthode pas par nous : 1 slide par tiret - Reprendre la forme matricielle de l'ANOVA phylo (mettre en rouge les diffs) - Présenter le MB qui évolue sur l'arbre + lien matrice K - Mettre la statistique de test (mettre en rouge la projection (donc diffs))

Transition vers notre travail - Mettre la formule avec erreur de mesure avec justification de l'ajout de l'erreur de mesure, formule transfo  $V_\lambda$ ,



pointer la limite qui est l'erreur due à l'estimation du  $\lambda$  Méthode par nous :

- Satterthwaite : préciser que c'est nos calculs à partir de résultats sur modèle mixte (faire slide en appendice) + stat approximée + df formule une méthode possible parmi tant d'autres: Kenward Roger classique

Simulations : - les 2 arbres avec les groupes - Modalités de simulations, bien préciser que l'idée de simuler c'est pour voir erreur de type I et puissance - Les résultats de simulations: pour les résultats Mettre ANOVA , ANOVA phylo Satterthwaite LRT

Applications aux données réelles : - Rappel du type de données, RNA-seq sur pleins de gènes (éventuellement un extrait du tableau ?) - Mentionner toutes les méthodes rapidement et présenter l'UpSet diagramme avec son analyse et la remarque sur Satterthwaite ML qui sur-sélectionne

Conclusions/Ouvertures: Conclusions: - Récap du projet sur son contenu scientifique

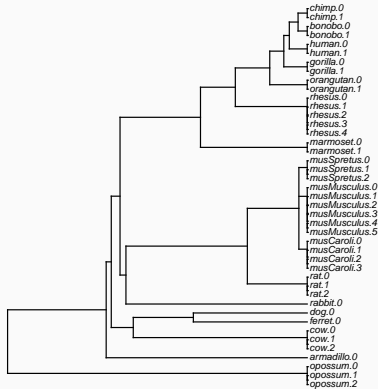


Ouvertures : - Utiliser un autre processus stochastique  
Ornstein-Uhlenbeck - Comprendre pourquoi Satterthwaite a  
sur-sélectionné dans l'application: mauvaise implémentation ? évaluer  
l'impact de l'approx - Prendre un autre arbre ou ré-échantillonner les  
groupes dans les simus - Agrandir le cadre de simulations - Appliquer les  
méthodes à d'autres données - modèle qui fait gène par gène: imaginer en  
prenant tous les gènes : Limma



# Introduction

---



**Figure 1:** Arbre phylogénétique de [Chen et al. 2019](#)

intro/contexte: biologique avec l'exemple de Chen + figure de l'article ?  
trouver les gènes différentiellement exprimés



Présenter le MB qui évolue sur l'arbre + lien matrice  $K$





## État de l'art

---

$$Y = X\beta + u, u \sim \mathcal{N}_n(0, \sigma_{phy}^2 K) \quad (1)$$

$$\text{où } \mathbf{Y} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1}_{n_1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, n = n_1 + n_2$$

Pour  $I = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$  :

$H_0 : \beta_2 = 0 \Leftrightarrow I^T \beta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ , les 2 groupes ont la même moyenne

$H_1 : \beta_2 \neq 0$ , les 2 groupes ont des moyennes différentes

On a alors la statistique de test suivante venant de [Bastide and Clavel 2022](#) :

$$F_{ANOVA_{phylo}} = \frac{\|\hat{Y} - \bar{Y}\|_{K-1}^2 (n-2)}{\|Y - \hat{Y}\|_{K-1}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}_{\text{isher}}(1, n-2) \quad (2)$$



$$Y = X\beta + u + \epsilon, \quad u \sim \mathcal{N}_n(0, \sigma_{phy}^2 K), \quad \epsilon \sim \mathcal{N}_n(0, \sigma_{err}^2 I_n) \quad (3)$$

avec justification de l'ajout de l'erreur de mesure, formule transfo  $V_\lambda$ ,  
pointer la limite qui est l'erreur dûe à l'estimation du  $\lambda$



# Calculus

---

# Calcul avec approximation de Satterthwaite

Satterthwaite : préciser que c'est nos calculs à partir de résultats sur modèle mixte (faire slide en appendice) + stat approximée + df formule



# Simulations

---

Afin d'avoir une idée des performances des méthodes, nous avons choisis de les comparer dans un contexte proche des cas d'application réels.





# Modalités de simulations ii

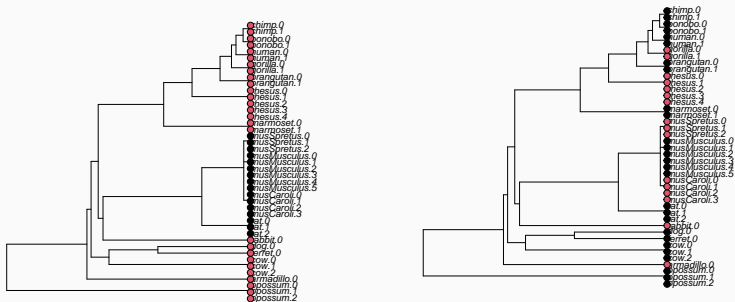


Figure 2: Arbre et groupes pour les simulations



Nous re-paramétrisons :

$$v_{tot} = \sigma_{phylo}^2 + \sigma_{measure}^2 = 1$$

Et alors les paramètres du modèles sont :

$$h \in (0.3, 0.5, 0.7, 0.9),$$

$$\sigma_{phylo}^2 = h \times v_{tot},$$

$$\sigma_{measure}^2 = (1 - h) \times v_{tot}$$

Ainsi,  $h = 0$  signifie qu'il y a seulement du bruit, et  $h = 1$  seulement de l'information phylogénétique.



# **Application aux données réelles**

---

Nous allons appliquer les différentes méthodes aux données compilées par [Chen et al. 2019](#). Il s'agit de données de RNA-seq chez 17 espèces et de l'arbre phylogénétique présenté figure 1.



# Remerciements

Merci pour votre attention.

Merci à nos encadrants pour leur accompagnement, leur disponibilité et leur gentillesse.



## Références et appendices

---

## References

---



Bastide, Paul and Julien Clavel (Dec. 2022). “Continuous Trait Evolution”.



Chen, Jenny et al. (Jan. 2019). “A Quantitative Framework for Characterizing the Evolutionary History of Mammalian Gene Expression”. In: *Genome Res* 29.1, pp. 53–63. ISSN: 1549-5469. DOI: 10.1101/gr.237636.118. pmid: 30552105.



Le code pour les simulations est disponible sur notre dépôt GitHub :

`https:  
//github.com/Polarolouis/anova-phylogenetique-projet-msv/`



## Concernant les fautes d'orthographe

Après relecture du rapport, nous avons pu constater que celui-ci contenait de nombreuses coquille. Nous vous présentons nos excuses.

## questions posables i

- comment obtenir la stat de test pour anova phylo (Cholesky) - en quoi c'est un modèle mixte pour Satterthwaite ? - calcul de la Hessienne optim vs formule analytique, mettre formule analytique - Le LRT un modèle emboîté blabla ? - sur quoi est basé EVEmodel ? - Mettre la démo du calcul de la Hessienne - Ornstein Uhleinbeck : qu'est ce que ça change par rapport au MB ? EVE dit optimum qui saute pas le processus qui saute Modélise deux niches différentes. Effet sur la moyenne masi ok , et sur la variance  $K_\alpha$ , ok pour satterthwaite masi prendre  $\alpha$  en compte aussi Modifie la structure de variance et ajoute un paramètre  $\alpha$ ,  $K(\alpha)$ , un saut sur l'optima. - données de comptage transformées donc ok de modéliser par MB

En écologie ne travaille pas sur autant de traits, spécificité de la RNA-seq des milliers de données.

LIMMA pour le cas non phylogénétique. Pour le cas phylogénétique `phylolimma`.

## questions posables ii

Méthodes d'amélioration essayer de faire quelque chose qui prennent en compte plusieurs gènes à la fois - Est ce q'on pourrait faire une méthode comme LIMMA et faire Satterthwaite ? - c'est bizarre d'utiliser des mesures

Questions Mélina : - Qu'est qu'une ANOVA phylogénétique ? En quoi différent l' ANOVA classique et l' ANOVA phylogénétique ? - Comment modéliser l'évolution d'un trait continu sur un arbre (choix du processus dans l'ANOVA phylogénétique : savoir qu'il existe différentes manières de faire, soit on prend un brownien, soit on prend un OU ... ) - Comment prendre en compte les erreurs de mesures dans l'anova phylogénétique ? (Car ici, dans le cadre de l'expression des gènes chez plusieurs espèces, on mesure plusieurs individus par espèce, on a donc une variabilité intra-espèce et une variabilité inter-espèces. . . il faut donc prendre en compte cela dans le modèle, et c'est d'ailleurs ce que fait EVE) - Quel test effectuer pour tester si on a une différence d'expression significative entre différent groupes d'espèces ? (LRT ou test basé sur la stat de

Fisher). - Qu'est ce qu'un modèle mixte ? Comment estimer les paramètres dans un modèle mixte ? Quels tests stats ? Quel est le lien entre une anova phylo et un modèle mixte ? - Pourquoi faire du REML au du ML classique ? Dans quel context? - Pour l'analyse de données réelles, vous avez également été confrontés à un problème de tests multiples : puisque vous faites un test par gènes, et que vous avez des milliers de gènes, alors vous devez "corriger les p-values" pour extraire votre sous liste de "gènes différentiellement exprimés" (deux approches classiques : Bonferroni / BH )