

# 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

2024-03-26

Projet: ANOVA Phylogénétique

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. Conclusions et ouvertures
- 7. Références et appendices



2024-03-26

Projet: ANOVA Phylogénétique

└─Sommaire

Sommaire

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

# Introduction

---

2024-03-26

Projet: ANOVA Phylogénétique  
└ Introduction

Introduction

---

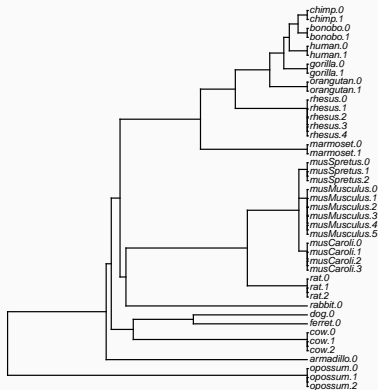


Figure 1: Arbre phylogénétique de [Chen et al. 2019](#)

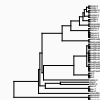
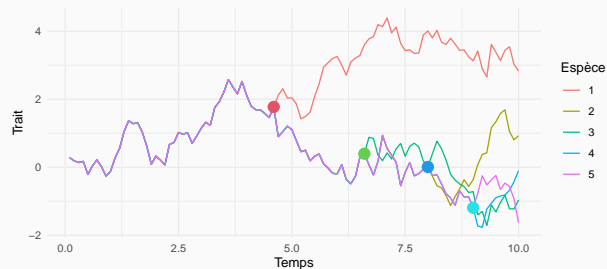
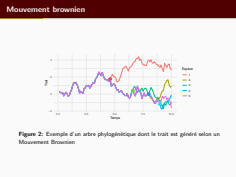


Figure 1: Arbre phylogénétique de [Chen et al. 2019](#)

- Avec l'avènement des données omiques de masses, on a accès à des données quantitatives très nombreuses pour plusieurs espèces.
- Le but derrière est de trouver les voies métaboliques et les gènes impliqués dans leur fonctionnement.
- Problème : cela coûte cher de regarder tous les gènes.
- On veut alors trouver les gènes qui s'expriment différemment en utilisant des méthodes statistiques en contrôlant l'erreur de type I, c'est-à-dire les faux positifs.

article de Chen:

- En compilant plusieurs jeux de données pour plusieurs espèces avec parfois plusieurs individus cf l'arbre les auteurs regardent par exemple entre les espèces les gènes qui sont différemment exprimés dans le foie.
- Pour voir s'il y a une différence entre 2 groupes d'espèces (*RatMus* vs *Autres*).  $\mu_1$ ,  $\mu_2$  etc.



**Figure 2:** Exemple d'un arbre phylogénétique dont le trait est généré selon un Mouvement Brownien

Pour un arbre phylo donné ça nous renseigne sur les instants de spéciation, donc moment de divergence entre 2 espèces représenté ici par les ronds.

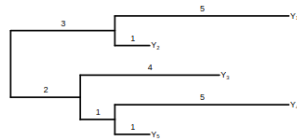
IMPORTANT : l'arbre phylogénétique est supposé connu, calibré<sup>1</sup> en temps et on n'y touche pas, nous.

- Ici représenté l'évolution d'un trait cad d'une valeur quantitative qu'on considère : ex comptage du nombre d'ARN exprimé pour un gène donné.
- La valeur du trait peut diverger pour chaque espèce à partir du moment de spéciation.
- Processus stochastique utilisé comme support de modélisation le mouvement brownien.
- Finalement, on a juste accès aux observations aux feuilles, jamais à ce qu'il se passe jusqu'alors



Pour un trait  $Y$  mesuré chez des espèces  $i$  et  $j$ ,  $Cov(Y_i, Y_j) = \sigma_{phylo}^2 t_{i,j}$  où  $t_{i,j}$  est le temps d'évolution commune.

BM on a tree:



$$\mathbf{V} = \begin{matrix} & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \\ \begin{matrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{matrix} & \begin{pmatrix} 8 & 3 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 \\ 0 & 0 & 6 & 2 & 2 \\ 0 & 0 & 2 & 8 & 3 \\ 0 & 0 & 2 & 3 & 4 \end{pmatrix} \end{matrix}$$

2

- attention pas l'arbre correspondant au trait simulé dans la slide d'avant
- Une fois qu'on a les observations pour chaque, à partir du MB, on a cette covariance là
- expliciter avec l'exemple
- La matrice  $V$  ici porte alors l'information phylogénétique de l'arbre

Pour un trait  $Y$  mesuré chez des espèces  $i$  et  $j$ ,  $Cov(Y_i, Y_j) = \sigma_{phylo}^2 t_{i,j}$  où  $t_{i,j}$  est le temps d'évolution commune.



<sup>2</sup>Bastide and Clavel 2022



## État de l'art

---

2024-03-26

Projet: ANOVA Phylogénétique

└─ État de l'art

État de l'art

---

- On dispose des observations aux feuilles et de l'arbre, ou du moins de l'information phylo
- Une méthode classique c'est l'ANOVA mais ça ne s'applique pas à nos données car elles ne sont pas indépendantes
- Alors dans ce projet on va étudier et utiliser une méthode d'ANOVA phylogénétique que l'on va présenter

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

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

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

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

## Rappel de l'ANOVA

- Pour rappel ici formule de 'ANOVA classique matricielle -> écrire au tableau
- Au tableau pas oublier de dire  $\beta_1 = \mu_1, \beta_2 = \mu_1 - \mu_2$
- ou  $\mu_1$  et  $\mu_2$  sont les moyennes du trait pour les groupes 1 et 2 que l'on considèrera

## ANOVA phylogénétique

- L'anova phylo consiste à inclure l'information de l'arbre
- Remarque, ici  $K$  correspond à la matrice  $V$  présentée quand on parlait de l'arbre, la covariance des  $Y_i, Y_j$





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

$H_0 : \beta_2 = 0 \Leftrightarrow \ell^T \beta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ , 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 :

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

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

$H_0 : \beta_2 = 0 \Leftrightarrow \ell^T \beta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ , 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 :

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

- $H_0$  : les 2 groupes ont la même moyenne, c'est à dire pour notre exemple que le gène n'est pas diff exprimés le gène a le même niveau d'expression
- On peut noter ici par rapport à la stat de test pour l'anova classique, que cela revient à projeter l'écart à la moyenne et les erreurs sur l'inverse de K la matrice des temps de coévolution.
- Pour la démo voir les slides en appendices. (Mettre slide 39 et 46 [Bastide and Clavel 2022](#))
- Donc c'est la projection orthogonale par rapport au produit scalaire  $\langle u, v \rangle_{V^{-1}} = u^T V^{-1} v$ .
- Pourquoi les dl de la loi de Fisher :  $1 = K - 1$  ici  $K = 2$  et  $n - 2 = n - K$ .



On ajoute une erreur de mesure qui correspond mieux à la réalité des données: erreur intraspécifique

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

En posant  $\lambda = \frac{\sigma_{phy}^2}{\sigma_{err}^2}$  et  $E = u + \epsilon$ , on peut obtenir une nouvelle forme pour  $Y$

$$Y = X\beta + E, \quad \text{où } \text{Var}(E) = V(\theta) = \sigma_{phy}^2 (K - \lambda I_n) = \sigma_{phy}^2 V_\lambda \quad (4)$$

$$E \sim \mathcal{N}_n(0, \sigma_{phy}^2 V_\lambda)$$

Problème:  $\lambda$  n'est en général pas connu et il faut l'estimer. Dans ce cas, le test n'est pas exact et  $F$  ne suit plus la même loi de Fisher.

- Erreur intra-spécifique : variabilité entre les observations
- Donc on ne sait pas comment l'estimation de  $\lambda$  fait évoluer les degrés de libertés et l'idée est donc ici d'utiliser l'approximation de Satterthwaite pour estimer les degrés de liberté.
- Remarque : il est toujours possible de réaliser le test avec cette statistique mais l'on s'attend à ce que le test puisse se tromper.

## Calculs

---

2024-03-26

Projet: ANOVA Phylogénétique  
└─ Calculs

Calculs

---

- Jusqu'ici nous avons étudié le modèle d'ANOVA phylo, ça a été un apprentissage. À partir d'ici ce sont nos calculs avec pour but leur implémentation.

# Calcul avec approximation de Satterthwaite

Méthode pour approximer les véritables degrés de liberté quand  $\lambda$  inconnu

Pour cela on peut voir le modèle comme un modèle mixte<sup>3</sup> :

$$F_{approx} = \frac{\|\hat{Y} - \bar{Y}\|_{V_{\lambda}^{-1}}^2 df_{approx}}{\|Y - \hat{Y}\|_{V_{\lambda}^{-1}}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{Fisher}(1, df_{approx}) \quad (5)$$

$$\text{Avec } df_{approx} = \frac{2(f(\hat{\theta}))^2}{[\nabla f(\hat{\theta})]^T A [\nabla f(\hat{\theta})]} \quad (6)$$

où  $f(\theta) = \ell^T C(\theta) \ell$  et A matrice de variance-covariance de  $\hat{\theta} = (\hat{\sigma}_{phy}^2, \hat{\sigma}_{err}^2)$

$$\begin{aligned} C(\theta) &= (Cov(\beta_i, \beta_j))_{i,j} \\ &= (X^T V(\theta)^{-1} X)^{-1} = (X^T (\sigma_{phy}^2 K + \sigma_{err}^2 I_n)^{-1} X)^{-1} \end{aligned}$$

Projet: ANOVA Phylogénétique

└─ Calculs

└─ Calcul avec approximation de Satterthwaite

2024-03-26

Méthode pour approximer les véritables degrés de liberté quand  $\lambda$  inconnu  
Pour cela on peut voir le modèle comme un modèle mixte<sup>3</sup> :

$$F_{approx} = \frac{\|\hat{Y} - \bar{Y}\|_{V_{\lambda}^{-1}}^2 df_{approx}}{\|Y - \hat{Y}\|_{V_{\lambda}^{-1}}^2} \underset{\mathcal{H}_0}{\sim} \mathcal{Fisher}(1, df_{approx}) \quad (5)$$

$$\text{Avec } df_{approx} = \frac{2(f(\hat{\theta}))^2}{[\nabla f(\hat{\theta})]^T A [\nabla f(\hat{\theta})]} \quad (6)$$

où  $f(\theta) = \ell^T C(\theta) \ell$  et A matrice de variance-covariance de  $\hat{\theta} = (\hat{\sigma}_{phy}^2, \hat{\sigma}_{err}^2)$

$$C(\theta) = (Cov(\beta_i, \beta_j))_{i,j}$$

$$= (X^T V(\theta)^{-1} X)^{-1} = (X^T (\sigma_{phy}^2 K + \sigma_{err}^2 I_n)^{-1} X)^{-1}$$

<sup>3</sup>Kuznetsova, Brockhoff, and Christensen 2017.

- On obtient alors des degrés de liberté approximés, qui nous permettent d'obtenir une stat de test elle aussi approximée.
- A partir de la doc du package lmerTest, en considérant le contexte de modèle mixte on a une formule approximée des degrés de liberté et donc nous avons calculé des formules explicites du gradient de f et de A. Et voilà.
- Et nous les avons implémentées pour pouvoir réaliser les tests à partir de cette nouvelle statistique.



## Simulations

---

2024-03-26

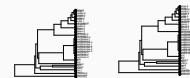
Projet: ANOVA Phylogénétique

└─ Simulations

Simulations

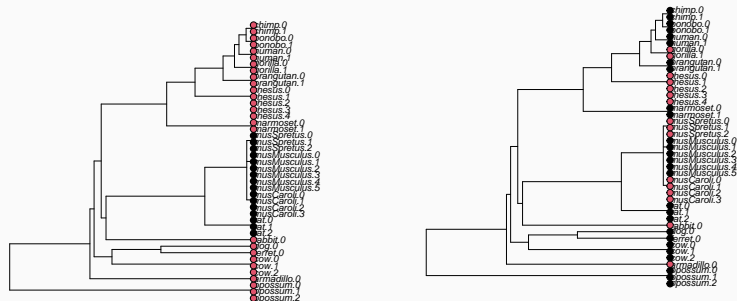
---

Après avoir implémenté, nous avons donc réalisé des simulations afin de regarder comment se comporte les méthodes : ANOVA classique, ANOVA phylogénétique et ANOVA phylogénétique avec approximation de Satterthwaite.



(a) Groupes *Mus* et rats contre les autres (b) Groupes sélectionnés sans respect de la phylogénie.

Figure 3: Arbre et groupes pour les simulations



(a) Groupes *Mus* et rats contre les autres

(b) Groupes sélectionnés sans respect de la phylogénie.

Figure 3: Arbre et groupes pour les 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.
- Nous reprenons l'arbre de [Chen et al. 2019](#)
- Et nous allons tester deux situations
  1. RatMus contre les autres espèces Figure 3a, donc l'information phylogénétique joue un rôle.
  2. ET les espèces réparties en deux groupes sans lien avec la classification phylogénétique
- Ce à quoi on s'attend **LE DESSIN**: un trait évoluant selon un processus stochastique, aboutissant à 2 valeurs différentes, l'ANOVA dit qu'il y a un saut, mais l'ANOVA phylogénétique connaissant la matrice  $K$  dit que c'est normal c'est la dérive, on a divergé il y a longtemps.



Nous re-paramétrisons :

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

Et alors les paramètres du modèle s'expriment :

$$\begin{aligned}\sigma_{phylo}^2 &= h \times v_{tot}, \\ \sigma_{err}^2 &= (1 - h) \times v_{tot}\end{aligned}$$

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

Et nous avons réalisé des simulations pour  $h \in \{0.3, 0.5, 0.7, 0.9\}$ .

Nous re-paramétrisons :

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

Et alors les paramètres du modèle s'expriment :

$$\begin{aligned}\sigma_{phylo}^2 &= h \times v_{tot}, \\ \sigma_{err}^2 &= (1 - h) \times v_{tot}\end{aligned}$$

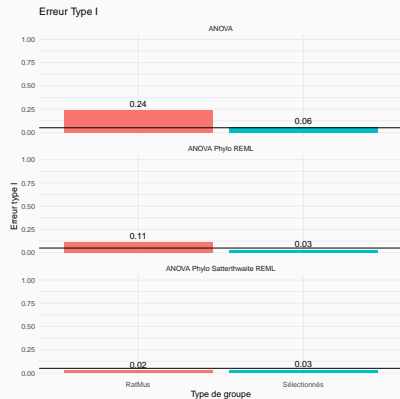
Ainsi,  $h = 0$  signifie qu'il y a seulement du bruit, et  $h = 1$  seulement de l'information phylogénétique.

Et nous avons réalisé des simulations pour  $h \in \{0.3, 0.5, 0.7, 0.9\}$ .

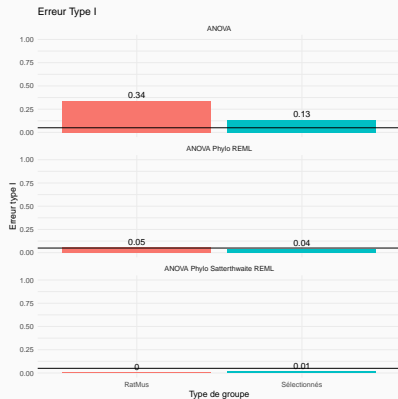
- A l'implémentation on ré-écrit le modèle pour n'avoir qu'un seul paramètre à faire varier,  $h$ , l'héritabilité. Qui se base sur le fait que la variance totale,  $v_{tot}$  **la formule**.



# Résultats: Erreur de type I



(a) Erreur de type I pour les simulations avec  $h = 0.3$



(b) Erreur de type I pour les simulations avec  $h = 0.9$

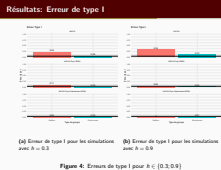
Figure 4: Erreurs de type I pour  $h \in \{0.3; 0.9\}$



## Projet: ANOVA Phylogénétique

### Simulations

### Résultats: Erreur de type I



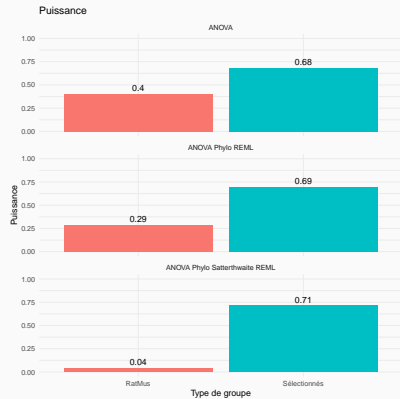
- Pour comparer les méthodes nous nous intéressons à deux métriques, l'erreur de type I et la puissance.
- L'erreur de type I est particulièrement importante à contrôler car comme nous en avons parlé plus haut, des faux-positifs impliqueraient donc des expériences inutiles et particulièrement coûteuses.
- A noter ici que nous ne regardons que les méthodes qui minimisent le critère REML (Restricted Maximum Likelihood) car ce sont elles qui fournissent une estimation de la variance non biaisée. Ce critère améliore sensiblement l'estimation de la variance.
- Remarque: l'ANOVA telle qu'implémentée dans R utilise directement le REML.

#### Analyse :

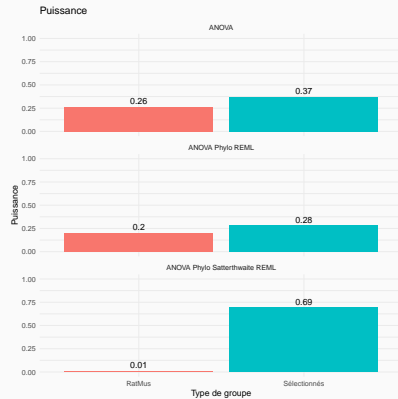
- erreur type I pour groupe phylogénétique:  $h = 0.3$  l'ANOVA classiques trompe bcp, l'ANOVA phylo se trompe, seule ANOVA phylo Satter est sous la barre des 0.5 = contrôle bien les faux positifs
- $h = 0.9$  : ANOVA classique se trompe toujours et plus, ANOVA phylo est au seuil 5 % pas étonnant il y a plus d'info phylogénétique, avec Satterthwaite on a aucune erreur
- au global l'ANOVA phylo avec Satterthwaite contrôle dans les 2 cas l'erreur de type I, et comme on s'y attend l'ANOVA phylo fait mieux que l'ANOVA classique
- groupe pas phylo:  $h = 0.3$  l'ANOVA se trompe légèrement, elle dépasse le seuil, les autres sont en dessous à 0.03
- pour  $h = 0.9$  l'ANOVA se trompe plus, elle dépasse le seuil, les autres sont en dessous
- touk, avec faible héritabilité on est dans un résultat proche de l'attendu : l'ANOVA se trompe à peine, avec forte héritabilité l'erreur est plus marquée ce qui est étonnant au vu des groupes sélectionnés
- Tout d'abord nos données ne respectent les hypothèses de l'ANOVA. On suspecte que la manière dont on a constitué les groupes n'a pas suffisamment cassé la phylogénie.



# Résultats: puissance



(a) Puissances pour les simulations avec  $h = 0.3$



(b) Puissances pour les simulations avec  $h = 0.9$

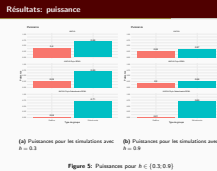
Figure 5: Puissances pour  $h \in \{0.3; 0.9\}$

Projet: ANOVA Phylogénétique

2024-03-26

Simulations

Résultats: puissance



- puissance statistique: quantifie les vraies positifs il faut qu'elle soit grande
- Situation groupe avec info phylogénétique: Revers de la médaille mauvaises puissances voire très mauvaises: CATASTROPHIQUE
- $h = 0.3$  + groupes selectionnes: on a des puissance correctes ce qui nous rassure sur l'implémentation des Méthodes
- $h = 0.9$  + groupes selectionnes : ANOVA et ANOVA phylo pas très bonne puissances, c'est inquiétant : toujours meme hypothèse il reste ude l'info phylo
- pour Satterthwaite ca parait bien c'est la meme chose



## Application aux données réelles

---

2024-03-26

Projet: ANOVA Phylogénétique  
└ Application aux données réelles

Application aux données réelles

- On a des données pour environ 5400 genes pour 17 especes toujours selon le meme arbre
- On passe a un test multiple sur tous les genes : on fait un test par gene et on corrige par la technique de Benjamini et Hochberg
- 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.

# UpSet diagram

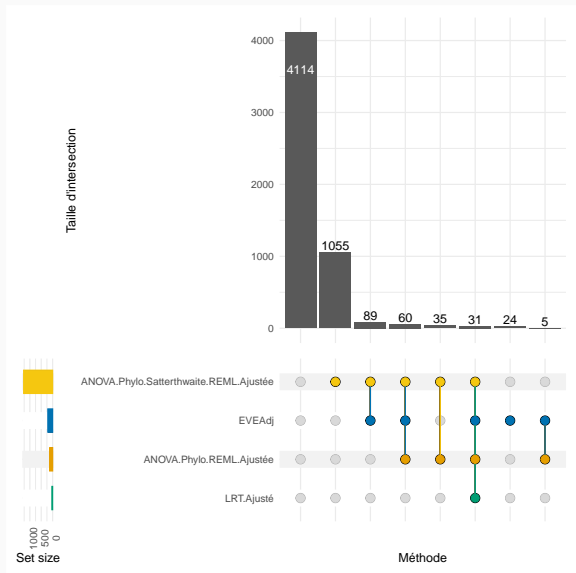


Figure 6: UpSet diagram de toutes les méthodes en incluant la méthode EVE

Projet: ANOVA Phylogénétique  
Application aux données réelles

UpSet diagram

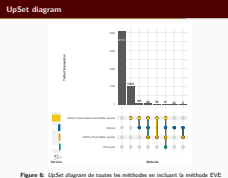


Figure 6: UpSet diagram de toutes les méthodes en incluant la méthode EVE

- ANOVA phylo Satterthwait REML surselectionne des genes
- mais il y a des recoupements
- Lrt sous selectionne
- la méthode EVE (Expression Variance and Evolution) c'est l'état de l'art basé sur Lrt

# Conclusions et ouvertures

---

2024-03-26

Projet: ANOVA Phylogénétique  
└─ Conclusions et ouvertures

Conclusions et ouvertures

---

- La méthode d'ANOVA phylogénétique avec Satterthwaite paraît intéressante, notamment pour le contrôle de l'erreur de type I. Mais il faudra creuser pour essayer de comprendre la dégradation de la puissance.
- Au début nous utilisons une approximation de l'hessienne, remplacée par la forme analytique une fois celle-ci obtenue.

- La méthode d'ANOVA phylogénétique avec Satterthwaite paraît intéressante, notamment pour le contrôle de l'erreur de type I. Mais il faudra creuser pour essayer de comprendre la dégradation de la puissance.
- Au début nous utilisons une approximation de l'hessienne, remplacée par la forme analytique une fois celle-ci obtenue.

- Approximation Hessienne : calculée par approximation numérique, méthode de Richardson.



- Changer de processus stochastique ? Le processus d'Ornstein-Uhlenbeck.
- Comprendre pourquoi avec l'approximation de Satterthwaite sur les données réelles il y a eu une sur-sélection.
- Changer les conditions de simulations : prendre un autre arbre, autres données, ou ré-échantillonner les groupes.
- Ces méthodes test gène par gène puis corrige pour faire un test multiple. On pourrait développer des méthodes qui font sur tous les gènes en même temps (adaptation phylogénétique de la méthode LIMMA).

- Changer de processus stochastique ? Le processus d'Ornstein-Uhlenbeck.
- Comprendre pourquoi avec l'approximation de Satterthwaite sur les données réelles il y a eu une sur-sélection.
- Changer les conditions de simulations : prendre un autre arbre, autres données, ou ré-échantillonner les groupes.
- Ces méthodes test gène par gène puis corrige pour faire un test multiple. On pourrait développer des méthodes qui font sur tous les gènes en même temps (adaptation phylogénétique de la méthode LIMMA).



Merci pour votre attention.

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



# Références et appendices

---

2024-03-26

Projet: ANOVA Phylogénétique  
└ Références et appendices




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.
-  Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). “**lmerTest** Package: Tests in Linear Mixed Effects Models”. In: *J. Stat. Soft.* 82.13. ISSN: 1548-7660. DOI: 10.18637/jss.v082.i13. URL: <http://www.jstatsoft.org/v82/i13/> (visited on 03/01/2024).

-  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.
-  Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). “**lmerTest** Package: Tests in Linear Mixed Effects Models”. In: *J. Stat. Soft.* 82.13. ISSN: 1548-7660. DOI: 10.18637/jss.v082.i13. URL: <http://www.jstatsoft.org/v82/i13/> (visited on 03/01/2024).

Le code pour les simulations, nos implémentations, le rapport et cette présentation est disponible sur notre dépôt GitHub :

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

2024-03-26

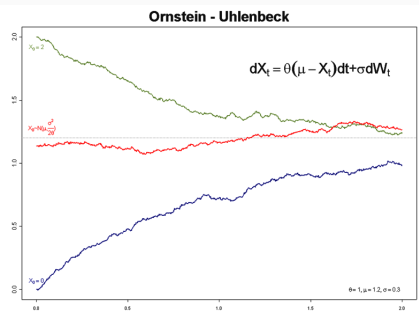
Concernant les fautes d'orthographe

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

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

Une autre possibilité de processus stochastique est le processus d'Ornstein-Uhlenbeck (*mean-reverting process*) décrit par l'EDS suivante:

$$dr_t = -\theta(r_t - \mu) + \sigma dW_t$$



**Figure 7:** Exemple de trajectoires de processus d'Ornstein-Uhlenbeck (tiré de Wikipédia)

## Ornstein-Uhlenbeck

- Ce processus tend avec le temps vers la valeur  $\mu$  et peut donc modéliser l'existence d'un optimum pour le trait considéré.
- C'est le processus qui sous-tend le modèle EVE
- Et l'idée derrière que ce n'est pas le processus qui saute, mais l'optimum et il modélise donc deux niches différentes.

$$dr_t = -\theta(r_t - \mu) + \sigma dW_t$$

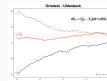


Figure 7: Exemple de trajectoires de processus d'Ornstein-Uhlenbeck (tiré de Wikipédia)

TODO Ajouter la formule canonique du modèle mixte

2024-03-26

Obtention des estimateurs

En connaissant l'arbre et le processus stochastique, le modèle s'écrit:

$$Y = X\beta + \sigma E, E \sim \mathcal{N}(0_n, K)$$

De là par les formules classiques on a les estimateurs :

$$K = LL^T$$

$$\text{Var}[L^{-1}E] = L^{-1}V[L^{-1}]^T = I$$

Et on peut alors écrire la régression décorrélée :

$$L^{-1}Y = (L^{-1}X)\beta + \sigma E', E' \sim \mathcal{N}(0_n, I)$$

$$\beta = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$$

$$\hat{\sigma}^2 = \frac{1}{n-p} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta})$$

$$= \frac{1}{n-p} \|Y - X\hat{\beta}\|_{V^{-1}}^2$$

En connaissant l'arbre et le processus stochastique, le modèle s'écrit:

$$Y = X\beta + \sigma E, E \sim \mathcal{N}(0_n, K)$$

$$K = LL^T$$

$$\text{Var}[L^{-1}E] = L^{-1}V[L^{-1}]^T = I$$

Et on peut alors écrire la régression décorrélée :

$$L^{-1}Y = (L^{-1}X)\beta + \sigma E', E' \sim \mathcal{N}(0_n, I)$$

De là par les formules classiques on a les estimateurs :

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-p} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta}) \\ &= \frac{1}{n-p} \|Y - X\hat{\beta}\|_{V^{-1}}^2 \end{aligned}$$

- 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 mais ok, et sur la variance  $K_\alpha$ , ok pour Satterthwaite mais 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 mais transformées donc ok de modéliser par MB

- 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 mais ok, et sur la variance  $K_\alpha$ , ok pour Satterthwaite mais 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 mais 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.
- Méthodes d'amélioration essayer de faire quelque chose qui prennent en compte plusieurs gènes à la fois
  - Est-ce qu'on pourrait faire une méthode comme LIMMA et faire Satterthwaite ?
  - C'est bizarre d'utiliser des mesures
- Questions Mélina :
  - Qu'est-ce qu'une ANOVA phylogénétique ? En quoi diffère 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 ... )

- 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.
- Méthodes d'amélioration essayer de faire quelque chose qui prennent en compte plusieurs gènes à la fois
  - Est-ce qu'on pourrait faire une méthode comme LIMMA et faire Satterthwaite ?
  - C'est bizarre d'utiliser des mesures
- Questions Mélina :
  - Qu'est-ce qu'une ANOVA phylogénétique ? En quoi diffère 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 ... )



## questions posables iii

- 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 différents 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 contexte?

## Projet: ANOVA Phylogénétique

2024-03-26

└ questions posables

- 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 différents 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 contexte?

- 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 )