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

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

Idée structure i

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))

Idée structure ii

- Transition vers notre travail
  - Mettre la formule avec erreur de mesure avec justification de l'ajout de l'erreur de mesure, formule transfo V_λ, pointer la limite qui est l'erreur dûe à l'estimation du λ
- **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** :

Idée structure iii

- 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

Contexte biologique

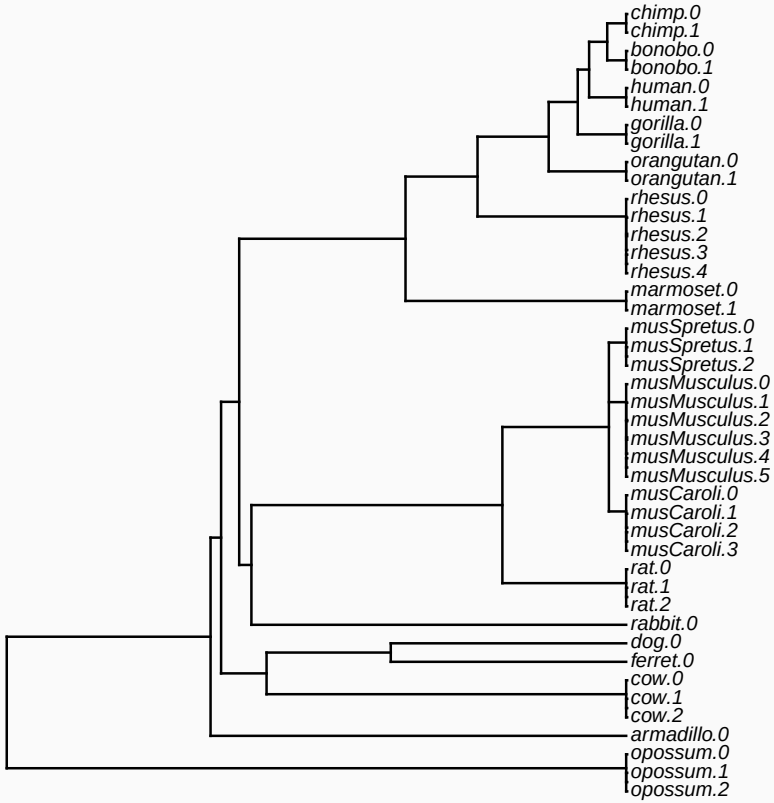
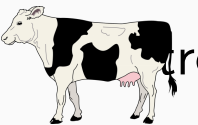


Figure 1: Arbre phylogénétique de Chen et al. 2019

intro/contexte: biologique avec l'exemple de Chen + figure de l'article ?

trouver les gènes différentiellement exprimés



Mouvement brownien

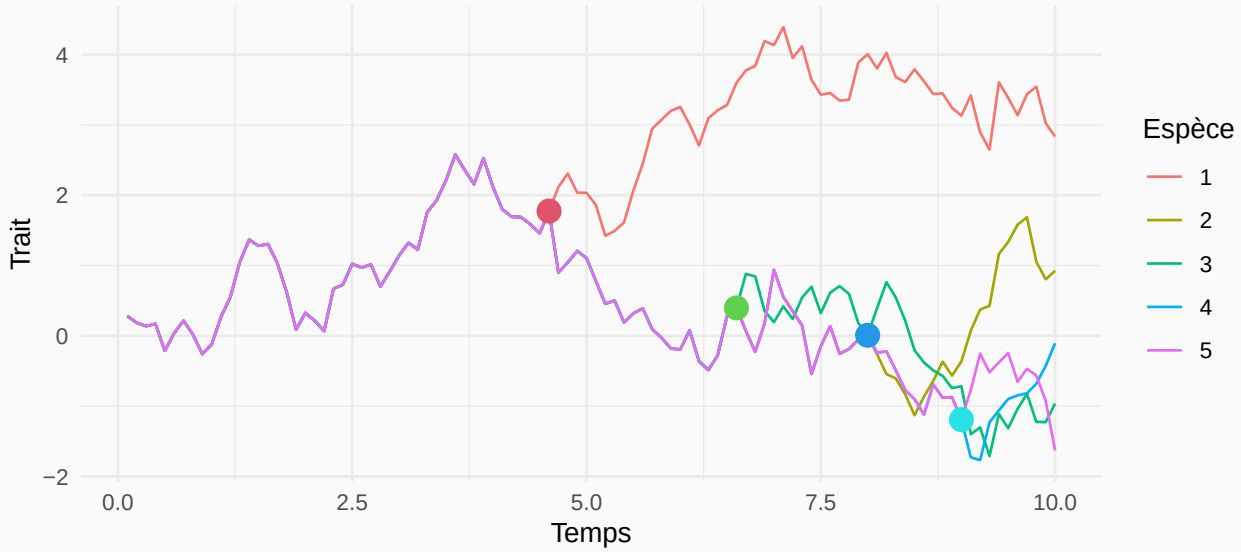
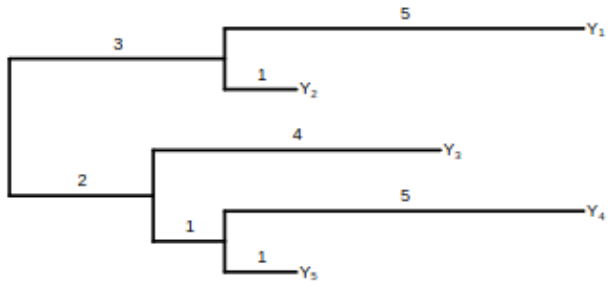


Figure 2: Exemple d'un arbre phylogénétique dont le trait est généré selon un Mouvement Brownien

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

BM on a tree:



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

1

[1] Bastide and Clavel 2022.

État de l'art

ANOVA phylogénétique

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

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

Statistique de test

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

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

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

On a alors la statistique de test suivante venant de Bastide and Clavel 2022 :

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

ANOVA phylogénétique et erreur de mesure

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

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

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

$$Y = X\beta + E, \text{ où } Var(E) = V(\theta) = \sigma^2_{phy}(K - \lambda I_n) = \sigma^2_{phy} V_\lambda \tag{4}$$
$$E \sim \mathcal{N}_n(0, \sigma^2_{phy} V_\lambda)$$

Problème: λ n'est souvent pas connu et il faut l'estimer. Dans ce cas, l'approximation de la distribution de F par une distribution de Fisher ne tient plus

Calculs

Calcul avec approximation de Satterthwaite

Méthode pour approximer les véritables degrés de liberté quand λ inconnu Pour cela on peur voir le modèle comme un modèle mixte

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

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

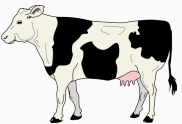
où $f(\theta) = I^T C(\theta) I$ et A matrice de variance-covariance de $\hat{\theta} = (\hat{\sigma}^2_{phy}, \hat{\sigma}^2_{err})$

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

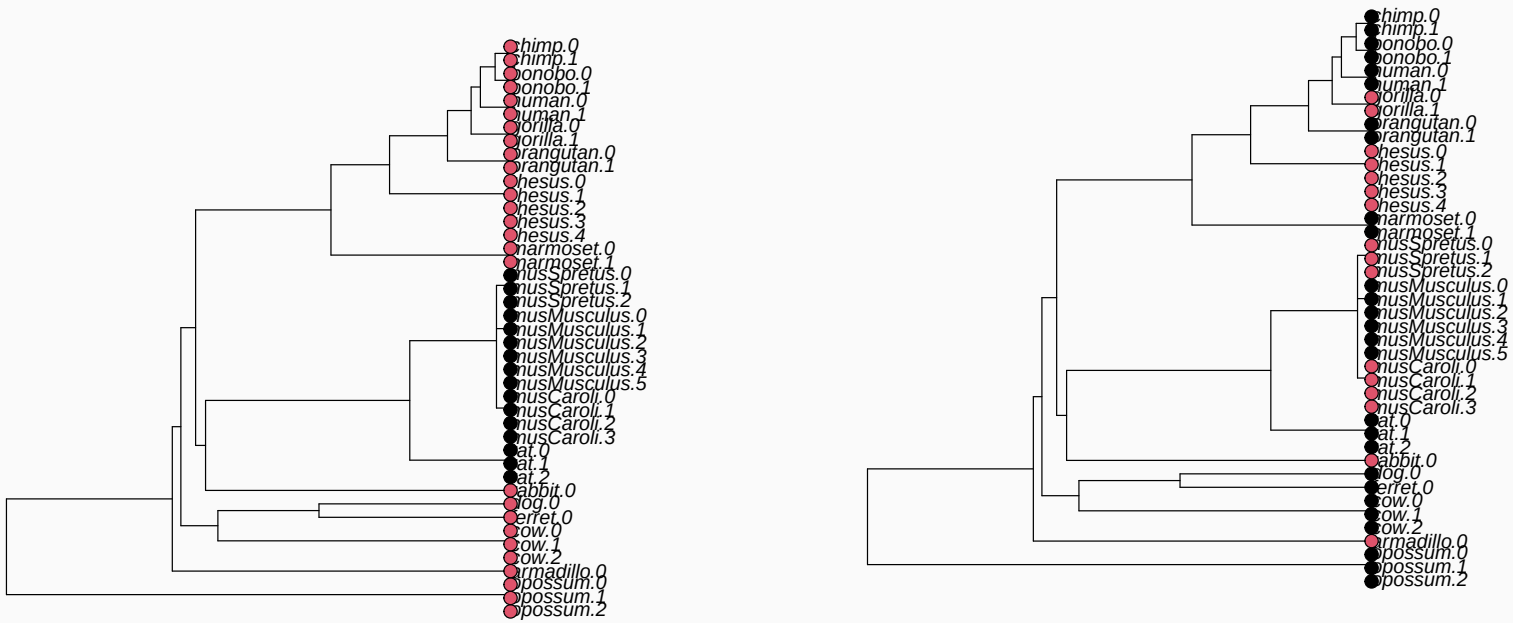
Simulations

Modalités de simulations i

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



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

(b) Groupes sélectionnés sans respect de la phylogénie.

Figure 3: Arbre et groupes pour les simulations

Modalités de simulations iii

Nous re-paramétrisons :

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

Et alors les paramètres du modèles sont :

$$h \in (0.3, 0.5, 0.7, 0.9),$$
$$\sigma^2_{phylo} = h \times v_{tot},$$
$$\sigma^2_{measure} = (1 - h) \times v_{tot}$$

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

Résultats: Erreur de type I

Add erreur de type 1 pour comparer que ça avec Satterthwaite REML, ANOVA et ANOVA phylo REML

Résultats: puissance

Comparer les puissances des méthodes principales

Application aux données réelles

Rappel des données

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

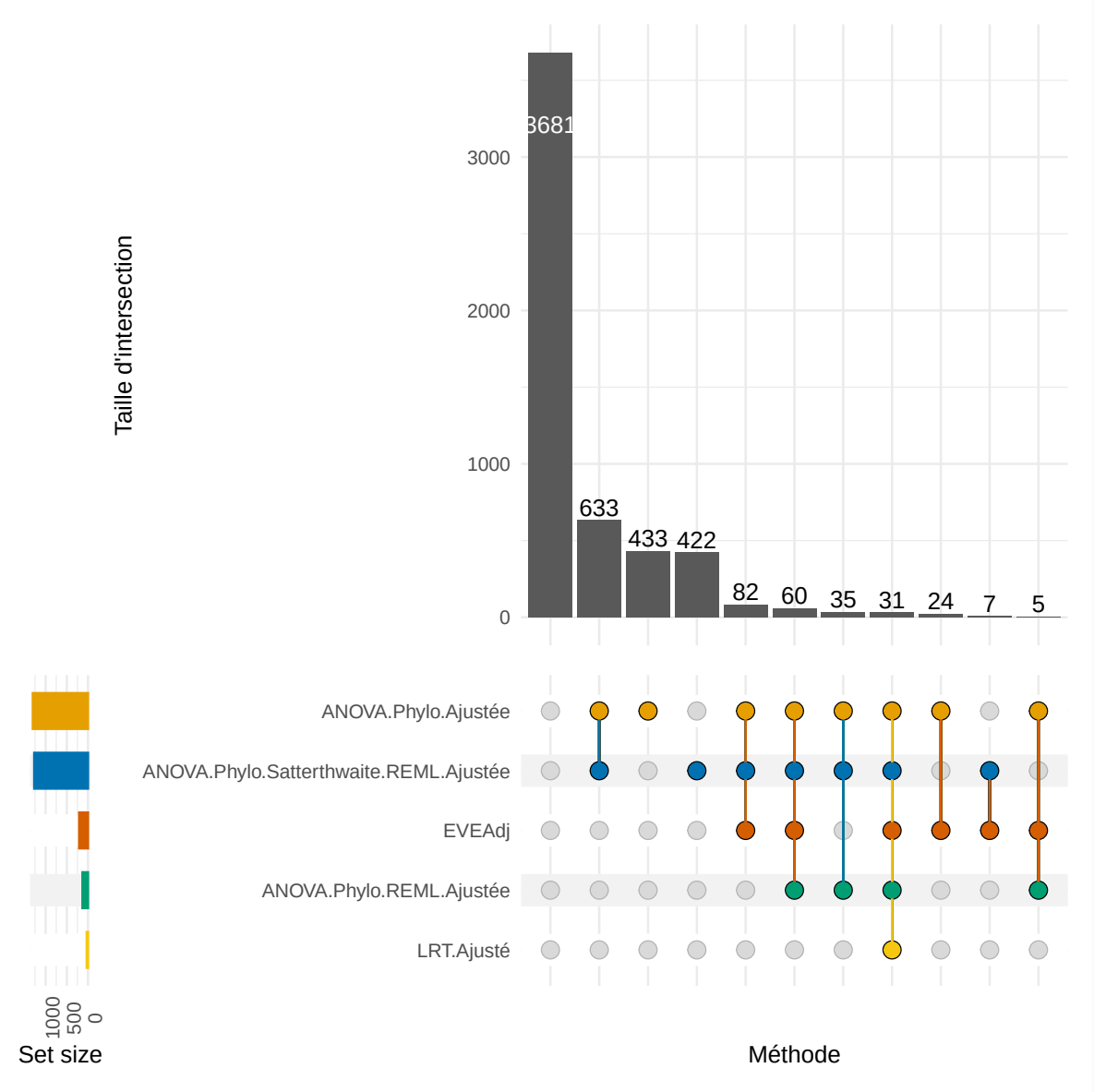


Figure 4: *UpSet diagram* de toutes les méthodes en incluant la méthode EVE

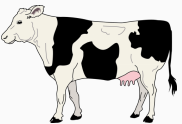
Conclusions et ouvertures

Conclusions scientifiques

Récap du projet sur son contenu scientifique

Problèmes rencontrés

- Utilisation de l'approx de la Hessienne -¿ expression analytique



Ouvertures

- Utilisation du processus d'Ornstein-Uhlenbeck
- Pourquoi Satterthwaite a surselectionné ? Creuser
- Prendre un autre arbre, autres données, ou ré-échantillonner les groupes dans les simus
- Modèle qui fait gène par gène: imaginer en prenant tous les gènes -¿ méthode LIMMA

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.

Code pour les simulations i

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

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

Concernant les fautes d'orthographes

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 emboité 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_α, ok pour satterthwaite masi prendre α en compte aussi Modifie la structure de variance et ajoute un paramètre α, $K(\alpha)$, 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

questions posables iii

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)