Rapport de Projet : ANOVA Phylogénétique

Alizée Geffroy Louis Lacoste

11 mars 2024

Table des matières

1 Introduction

Ici contexte biologique, les données de Gomez-Mestre, Pyron et Wiens 2012, les données de Paul et Mélina, etc.

Avec l'avènement des données massives de génomiques, transcriptomiques, protéomiques etc, il y a besoin de techniques statistiques robustes et passant à l'échelle permettant de mener à bien l'anal Format des données : arbres phylogénétiques, données génétiques Arbres avec des petites branche : plusieurs individus par espèces avec chacun leurs données –> problème biologique

Deux sujets différents écologie et transcriptomique mais une même méthode.

Pour données Chen et al. 2019 la figure 1 présente l'arbre phylogénétique :

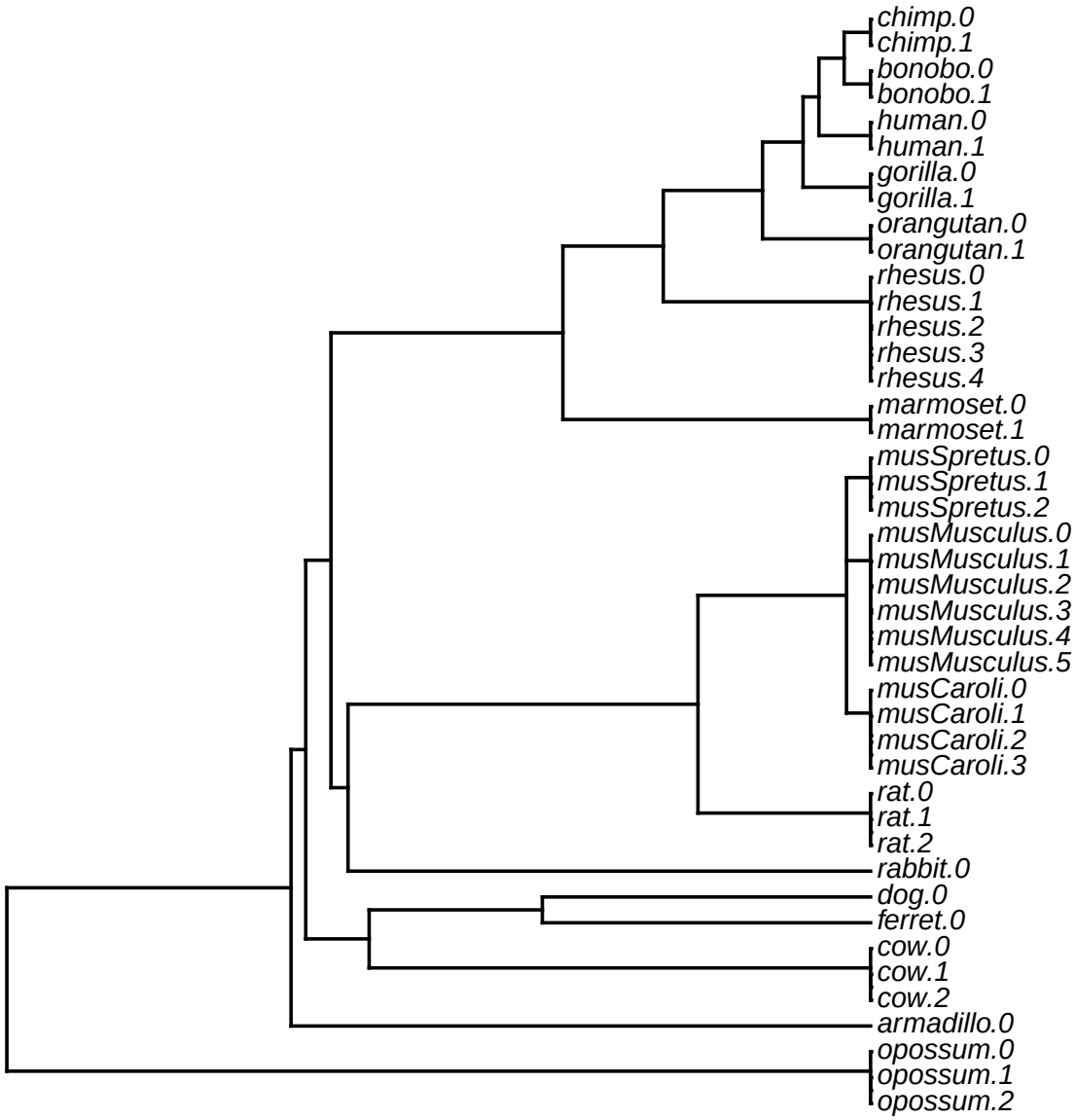

Figure 1 – Arbre phylogénétique de Chen et al. 2019

Transition, c'est pourquoi on va tester la méthode d'ANOVA phylogénétique avec cette forme de données. But ? Etudier cette méthode et les résultats

Un gène, comparer les moyennes d'expression d'un gène On connait les groupes exemple individus malade/sain

Comparaison non pas sur individus malades/pas malades mais sur espèces différentes. Pas possible de supposer iid, existe relations entre les individus et les groupes que l'on compare donc besoin de les prendre en compte.

Modele mixte la matrice des temps de divergences, BM simple sans erreurs, avec erreur (ajustement du ratio) avec OU...

2 Méthodes

Ici les rappels sur l'ANOVA, l'explication de l'ANOVA phylogénétique. La démonstration des limites de l'ANOVA phylogénétique par des simulations Méthode : la partie maths anova, anova phylo, satterthwaite,

2.1 L'ANOVA

L'ANOVA est un cas classique du modèle linéaire, nous utilisons ici les notations et le formalisme de Bel et al. s. d.

Le principe de l'ANOVA est d'expliciter le lien entre une variable quantitative et une ou plusieurs variables qualitatives.

La forme usuelle de l'ANOVA à 1 facteur est la suivante :

$$Y_{ik} = \mu_i + E_{ik}, \qquad i = 1, \dots I, k = 1, \dots n_i, E_{ik} \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

où dans cette équation, reprise du livre (Bel et al. s. d.), i représente le niveau du facteur et k indique le numéro de l'observation dans ce niveau. I est le nombre total de niveaux du facteur, n_i le nombre d'observation du niveau i.

L'ANOVA se généralise à deux facteurs, plus facilement compréhensible avec cette forme, non identifiable :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}, \quad i = 1, \dots I, j = 1, \dots J, k = 1, \dots n_i j, E_{ijk} \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

où μ représente un effet moyen de la population (*intercept*), α_i l'effet du premier facteur de niveau i, β_j l'effet du second facteur de niveau j.

Les paramètres de l'ANOVA sont estimables, grâce par exemple à la méthode du maximum de vraisemblance et ont des formules bien connues.

2.2 L'ANOVA phylogénétique

parler du BM ? PUis de la matrice V ou K qui donne la structure phylogénétique

Etre assez concis sur l'histoire de la projection et le modèle et les différences avec l'ANOVA.

2.3 Approximation de Satterthwaite

Pourquoi vouloir l'utiliser ? Réduire nbre de degrés de liberté utilisés dans la stat de test. Le but est d'approximé le nbre de degré de Liberté. On se basera sur la documentation du package lmer Kuznetsova, Brockhoff et Christensen 2017 pour calculer

les formules explicites de l'approximation dans notre cadre et ensuite implémenter l'implémenter..

$$Y = X\beta + u + \epsilon \tag{3}$$

$$\text{où} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad u \sim \mathcal{N}_n(0, \sigma^2_{phy} K), \quad \epsilon \sim \mathcal{N}_n(0, \sigma^2_{err} I_n)$$

$$\text{Alors} \quad Y \sim \mathcal{N}_n(X\beta, \sigma^2_{phy} K + \sigma^2_{err} I_n) \quad \text{et} \quad Var_\theta(Y) = V(\theta) = \sigma^2_{phy} K + \sigma^2_{err} I_n$$

De là on obtient :

$$C(\theta) = (Cov(\beta_i, \beta_j))_{i,j} = (X^T V(\theta)^{-1} X)^{-1} = (X^T (\sigma^2_{phy} K + \sigma^2_{err} I_n)^{-1} X)^{-1} \tag{4}$$

Toujours en suivant la documentation Kuznetsova, Brockhoff et Christensen 2017 on part de l'expression pour les degrés de liberté *df* et de l'approximation. Ce qui nous donne :

$$df = \frac{2(l^T \hat{C} l)^2}{[Var(l^T \hat{C} l)]} = \frac{2(f(\hat{\theta}))^2}{[Var(f(\hat{\theta}))]} \approx \frac{2(f(\hat{\theta}))^2}{[\nabla f(\hat{\theta})]^T A [\nabla f(\hat{\theta})]} \tag{5}$$

$$\text{où} \quad \hat{C} = C(\hat{\theta}) \quad \text{et} \quad f(\theta) = l^T C(\theta) l$$

A partir de cette expression, on calcule $\nabla f(\theta)$ qu'on appliquera en $\hat{\theta}$ et A la matrice de variance-covariance de $\hat{\theta} = (\hat{\sigma}^2_{phy}, \hat{\sigma}^2_{err})$

Calcul du gradient. Nous voulons calculer les dérivées partielles $\partial_{\sigma^2_{phy}} f(\theta)$ et $\partial_{\sigma^2_{err}} f(\theta)$. Pour les premières étapes de calculs, on écrira seulement ∂ sans distinction car ce sont les mêmes expressions pour les 2 dérivées. On utilisera dans la suite les formules de Petersen et Pedersen 2012 pour les dérivées de matrice

$$\partial f(\theta) = l^T \partial C(\theta) l$$

$$\partial C(\theta) = \partial (X^T V(\theta)^{-1} X)^{-1} = -C(\theta) \partial (X^T V(\theta)^{-1} X) C(\theta)$$

$$\partial (X^T V(\theta)^{-1} X) = \partial (X^T V(\theta)^{-1}) X + \cancel{X^T V(\theta)^{-1} \partial (X)} \quad (\partial_{\sigma^2_{phy}}(X) \text{ et } \partial_{\sigma^2_{err}}(X) \text{ sont nulles})$$

$$\partial (X^T V(\theta)^{-1}) = \partial (X^T) V(\theta)^{-1} + X^T \partial (V(\theta)^{-1}) = \cancel{\partial (X)^T V(\theta)^{-1}} + X^T \partial (V(\theta)^{-1})$$

$$\partial (V(\theta)^{-1}) = -V(\theta)^{-1} \partial (V(\theta)) V(\theta)^{-1}$$

$$\partial (V(\theta)) = \partial (\sigma^2_{phy} K + \sigma^2_{err} I_n)$$

Ce qui donne :

$$\partial_{\sigma^2_{phy}} (V(\theta)) = K, \quad \text{et} \quad \partial_{\sigma^2_{err}} (V(\theta)) = I_n$$

De là en remettant les formules explicite les unes dans les autres, on obtient :

$$[\nabla f(\hat{\theta})] = \begin{bmatrix} \partial_{\sigma^2_{phy}} f(\hat{\theta}) \\ \partial_{\sigma^2_{err}} f(\hat{\theta}) \end{bmatrix} = \begin{bmatrix} l^T C(\hat{\theta}) X^T V(\hat{\theta})^{-1} K V(\hat{\theta})^{-1} X C(\hat{\theta}) l \\ l^T C(\hat{\theta}) X^T V(\hat{\theta})^{-1} I_n V(\hat{\theta})^{-1} X C(\hat{\theta}) l \end{bmatrix}$$

□

Calcul de A. A est la matrice variance-covariance de $\hat{\theta}$, c'est à dire l'inverse de la Hessienne H de la vraissemblance de $\hat{\theta}$.

$$A = H^{-1}$$

Dans ce cadre on peut obtenir une formule explicite de la Hessienne, même si dans la plupart des cas il est plus simple d'estimer cette matrice par des méthodes numériques. On va d'abord calculer la log-vraissemblance du vecteur Y défini précédemment :

$$\begin{aligned}\mathcal{L}(\mathbf{Y},\theta) &= \log(\frac{1}{(2\pi)^{n/2}|V(\theta)|^{1/2}}\exp\left(-\frac{1}{2}(Y-X\beta)^T V(\theta)^{-1}(Y-X\beta)\right)) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|V(\theta)|) - \frac{1}{2}(Y-X\beta)^T V(\theta)^{-1}(Y-X\beta)\end{aligned}$$

On calcule les dérivées premières de la log-vraissemblance

$$\begin{aligned}\partial_{\sigma^2_{phy}}\mathcal{L} &= -\frac{1}{2}\partial_{\sigma^2_{phy}}(\log(|V(\theta)|)) - \frac{1}{2}\partial_{\sigma^2_{phy}}((Y-X\beta)^T V(\theta)^{-1}(Y-X\beta)) \\ &= -\frac{1}{2}\frac{K}{|V(\theta)|} - \frac{1}{2}(Y-X\beta)^T V(\theta)^{-1} K V(\theta)^{-1}(Y-X\beta)\end{aligned}$$

$$\begin{aligned}\partial_{\sigma^2_{err}}\mathcal{L} &= -\frac{1}{2}\partial_{\sigma^2_{err}}(\log(|V(\theta)|)) - \frac{1}{2}\partial_{\sigma^2_{err}}((Y-X\beta)^T V(\theta)^{-1}(Y-X\beta)) \\ &= -\frac{1}{2}\frac{I_n}{|V(\theta)|} - \frac{1}{2}(Y-X\beta)^T V(\theta)^{-1} I_n V(\theta)^{-1}(Y-X\beta)\end{aligned}$$

Puis les dérivées secondes :

$$\begin{aligned}\partial_{\sigma^2_{\mathbf{phy}}\sigma^2_{\mathbf{phy}}}\mathcal{L} &= -\frac{1}{2}\partial_{\sigma^2_{phy}\sigma^2_{phy}}\left(\frac{K}{|V(\theta)|}\right) - \frac{1}{2}\partial_{\sigma^2_{phy}\sigma^2_{phy}}\left((Y-X\beta)^T V(\theta)^{-1} K V(\theta)^{-1}(Y-X\beta)\right) \\ &= \frac{1}{2}\frac{K^2}{V(\theta)^2} + (Y-X\beta)^T V(\theta)^{-1} K V(\theta)^{-1} K V(\theta)^{-1}(Y-X\beta)\end{aligned}$$

car

$$\partial\left((Y-X\beta)^T V(\theta)^{-1} K V(\theta)^{-1}(Y-X\beta)\right) = (Y-X\beta)^T \partial\left(V(\theta)^{-1} K V(\theta)^{-1}\right)(Y-X\beta)$$

et

$$\partial\left(V(\theta)^{-1} K V(\theta)^{-1}\right) = -V(\theta)^{-1}\partial V(\theta) V(\theta)^{-1} K V(\theta)^{-1} - V(\theta)^{-1} K V(\theta)^{-1}\partial V(\theta) V(\theta)^{-1}$$

ce qui donne

$$\partial_{\sigma^2_{phy}\sigma^2_{phy}}\left(V(\theta)^{-1} K V(\theta)^{-1}\right) = -2V(\theta)^{-1} K V(\theta)^{-1} K V(\theta)^{-1}$$

$$
\begin{aligned}
&\partial_{\sigma^2_{\mathbf{err}}\sigma^2_{\mathbf{phylo}}}\mathcal{L} = \partial_{\sigma^2_{\mathbf{phy}}\sigma^2_{\mathbf{err}}}\mathcal{L} \\
&= -\frac{1}{2}\partial_{\sigma^2_{phy}\sigma^2_{err}}\left(\frac{K}{|V(\theta)|}\right) - \frac{1}{2}\partial_{\sigma^2_{phy}\sigma^2_{err}}\left((Y-X\beta)^T V(\theta)^{-1} K V(\theta)^{-1}(Y-X\beta)\right) \\
&= \frac{1}{2}\frac{K}{V(\theta)^2} + \frac{1}{2}(Y-X\beta)^T(V(\theta)^{-1}V(\theta)^{-1}KV(\theta)^{-1} + V(\theta)^{-1}KV(\theta)^{-1}V(\theta)^{-1})(Y-X\beta)
\end{aligned}
$$

car

$$
\partial_{\sigma^2_{phy}\sigma^2_{err}}\left(V(\theta)^{-1}KV(\theta)^{-1}\right) = -V(\theta)^{-1}V(\theta)^{-1}KV(\theta)^{-1} + V(\theta)^{-1}KV(\theta)^{-1}V(\theta)^{-1}
$$

$$
\begin{aligned}
\partial_{\sigma^2_{\mathbf{err}}\sigma^2_{\mathbf{err}}}\mathcal{L} &= -\frac{1}{2}\partial_{\sigma^2_{err}\sigma^2_{err}}\left(\frac{I_n}{|V(\theta)|}\right) - \frac{1}{2}\partial_{\sigma^2_{err}\sigma^2_{err}}\left((Y-X\beta)^T V(\theta)^{-1} V(\theta)^{-1}(Y-X\beta)\right) \\
&= \frac{1}{2}\frac{I_n}{V(\theta)^2} + (Y-X\beta)^T V(\theta)^{-1}V(\theta)^{-1}V(\theta)^{-1}(Y-X\beta)
\end{aligned}
$$

□

3 Méthodologie

lrt ANOVA normale VANILLA = ANOVA phylo sans correction des degrés de liberté $df1 = K-1, df2 = n-K$ ANOVA phylo (avec REML)

test sur arbre quelconque puis sur arbre avec petites branches ?

3 parties : - théo - méthodo par simu - appli aux données réelles

3.1 Simulations

Simu : Plusieurs design, tailles etc On sait la vérité, on peut connaitre les vrais positifs etc Qu'est ce qu'on prend en entrées qu'est ce qu'on veut en sortie

Bien insister sur l'arbre d'entrée et l'objectif de la simu : quelle approche pour mieux détecter les gènes différentiellement exprimés.

Simulations :

— soit selon l'arbre des données

— soit partir sur regarder l'impact de la taille de l'arbre etc.

ANOVA vs ANOVA Phylogénétique

Dans cette partie nous souhaitons comparer les résultats de l'ANOVA et de l'ANOVA phylogénétique. Pour cela nous allons simuler des données selon plusieurs modalités et évaluer l'*erreur de première espèce* et la *puissance* obtenue.

— Des données réparties en deux groupes indépendants de la phylogénie.

— Des données réparties en deux groupes cohérents avec la phylogénie.

Pour les simulations qui ne se font pas selon la phylogénie, nous nous attendons à ce que l'ANOVA classique obtienne de bons résultats puisque c'est une situation correspondant à l'application du modèle.

Pour les simulations qui se font selon l'information de l'arbre phylogénétique, nous nous attendons à ce que l'ANOVA phylogénétique parvienne à mieux prendre en compte l'information apportée par la phylogénie et à démêler son effet.

Pour faire nos simulations dans un contexte proche du cas réel nous allons utiliser l'arbre présenté sur la figure 1.

Nous choisissons de diviser les espèces en deux groupes. Pour le groupe respectant la phylogénie, on a d'un côté les espèces du genre *Mus* avec les rats et les autres espèces dans un autre groupe (voir la figure 2).

Et pour le groupe ne respectant pas la phylogénie on attribue aléatoirement les individus à l'un des deux groupes en respectant les proportions du groupe défini avant afin de rendre les résultats comparables (voir la figure 3). Enfin pour que notre analyse soit reproductible nous fixons la graine à 1234.

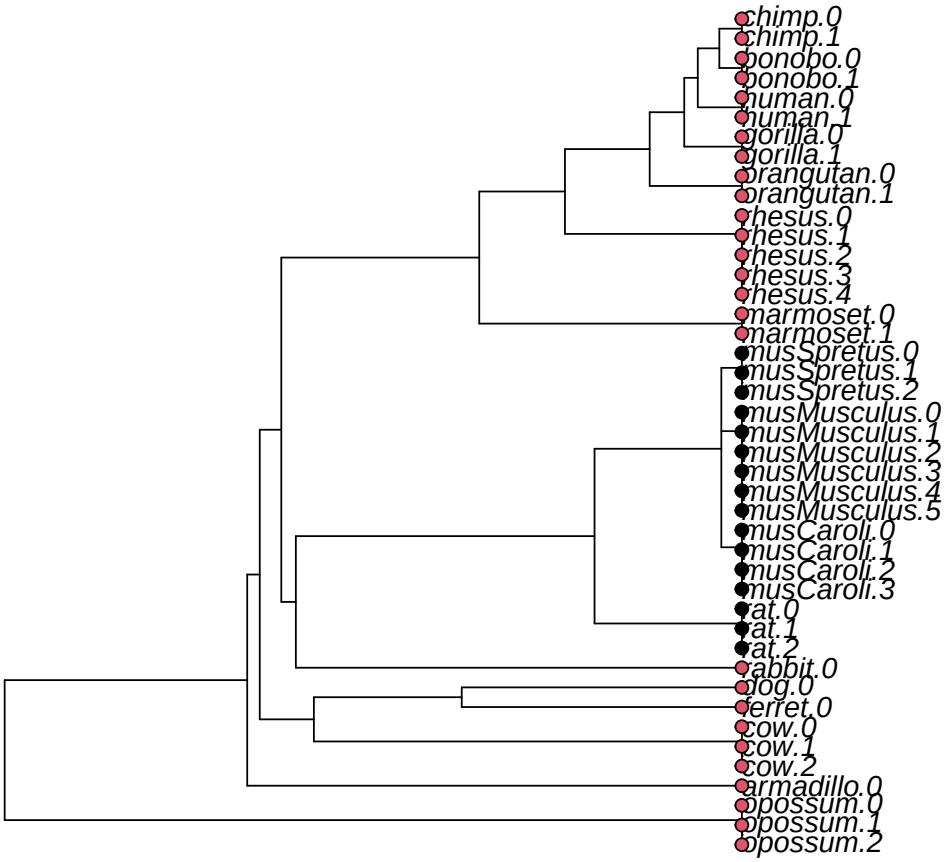

FIGURE 2 – Groupes *Mus* et rats contre les autres

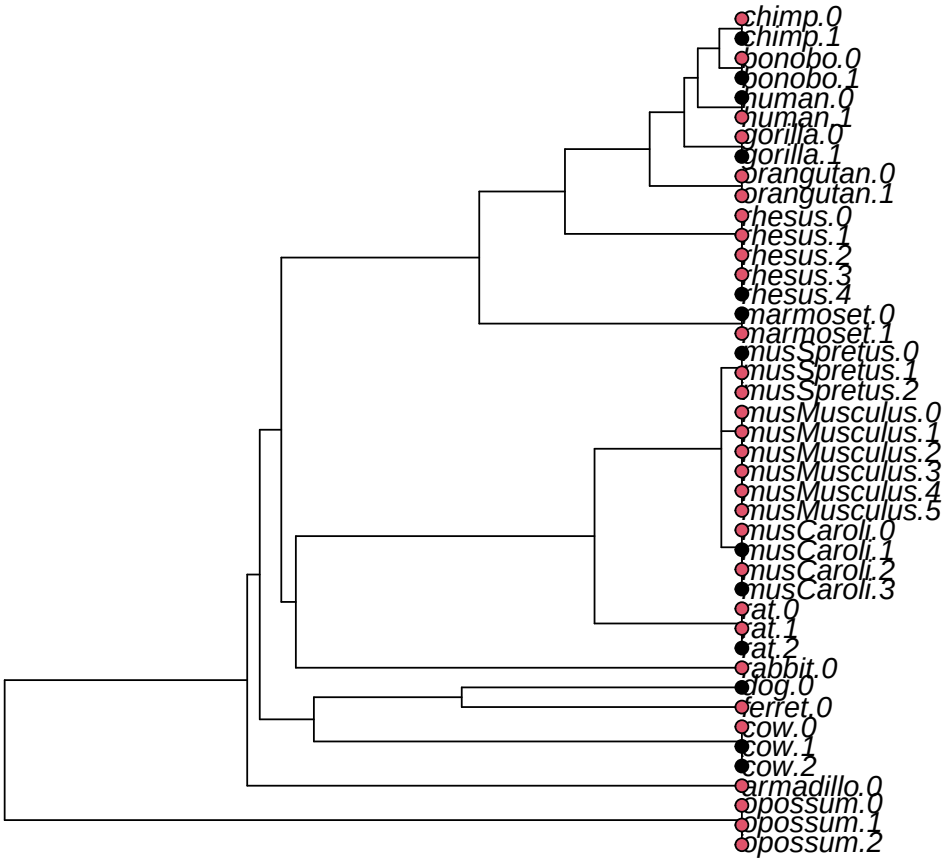

FIGURE 3 – Groupes aléatoires respectant les proportions

Pour les paramètres de la simulation nous allons faire prendre à h, défini comme l'héritabilité, les valeurs $h \in (0.3, 0.5, 0.7, 0.9)$.

Avec Satterthwaite et le *Likelihood Ratio Test* (LRT)

4 Données

Ici nous appliquons les méthodes implémentées sur l'arbre de CHEN et al. 2019.

Vanilla (ML et REML), Satterthwaite (ML et REML), LRT

Nous appliquons les différentes méthodes que nous avons implémentés dans le code.

Ci-dessous la figure 4 présente les p-values des différentes méthodes. Il est important de noter que ce graphique présente les p-values *non ajustées*.

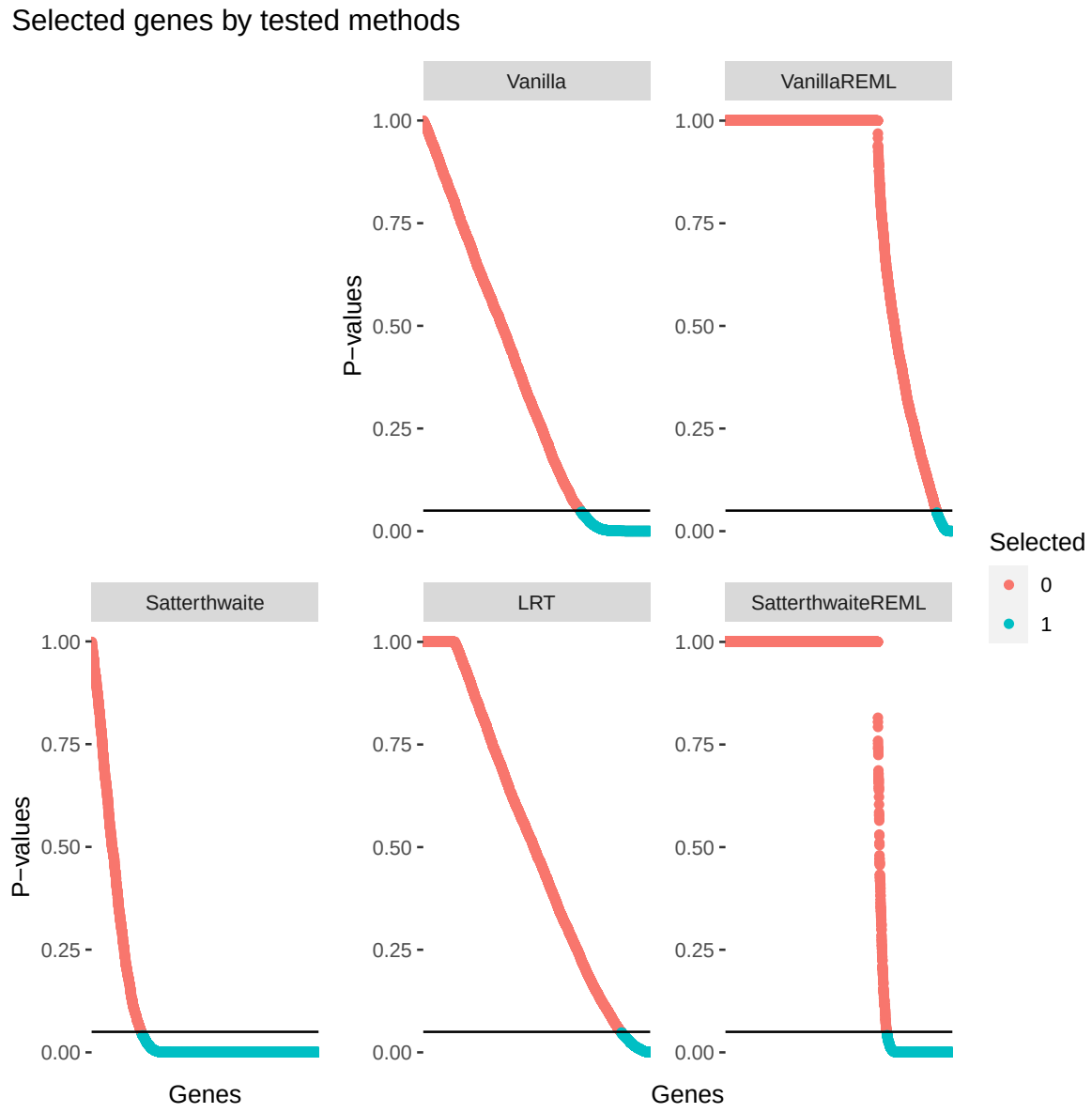

FIGURE 4 – *p-values* pour les différents tests

Pour la suite de cette analyse, nous allons appliquer un ajustement des p-values pour les test multiples, nommément la correction de Benjamini-Hochberg.

Une fois ces corrections appliquées, nous allons comparer les gènes sélectionnés, c'est-à-dire différentiellement exprimés.

Ces résultats sont présentés dans le diagramme de Venn (figure 5)

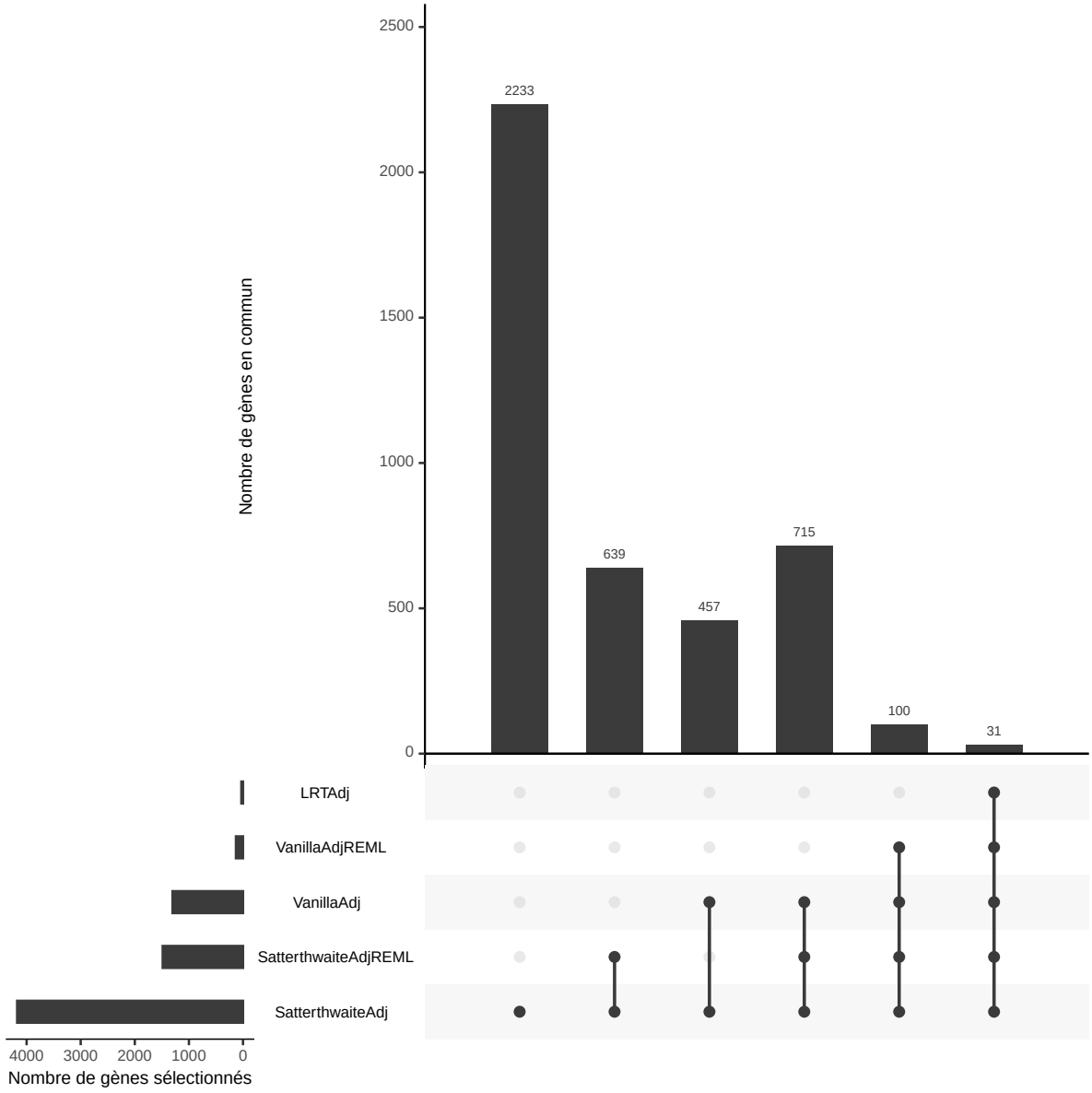

FIGURE 5 – Diagramme de Venn comparant les gènes sélectionnés selon les méthodes

EVEmodel

Dans l'article ROHLFS et NIELSEN 2015, les auteurs introduisent une méthode de détection des gènes différentiellement exprimés. Cette méthode est à l'heure actuelle très utilisée.

Remarque : La méthode a produit des `NA` pour certains gènes, d'après le message d'erreur, une optimisation n'a pas convergé. Ces gènes sont présentés dans le tableau 1.

Toutes les méthodes

Nous allons ici comparer toutes les méthodes dans un diagramme de Venn (figure 6) afin de voir les gènes sélectionnés en commun et les éventuelles différences entre les méthodes.

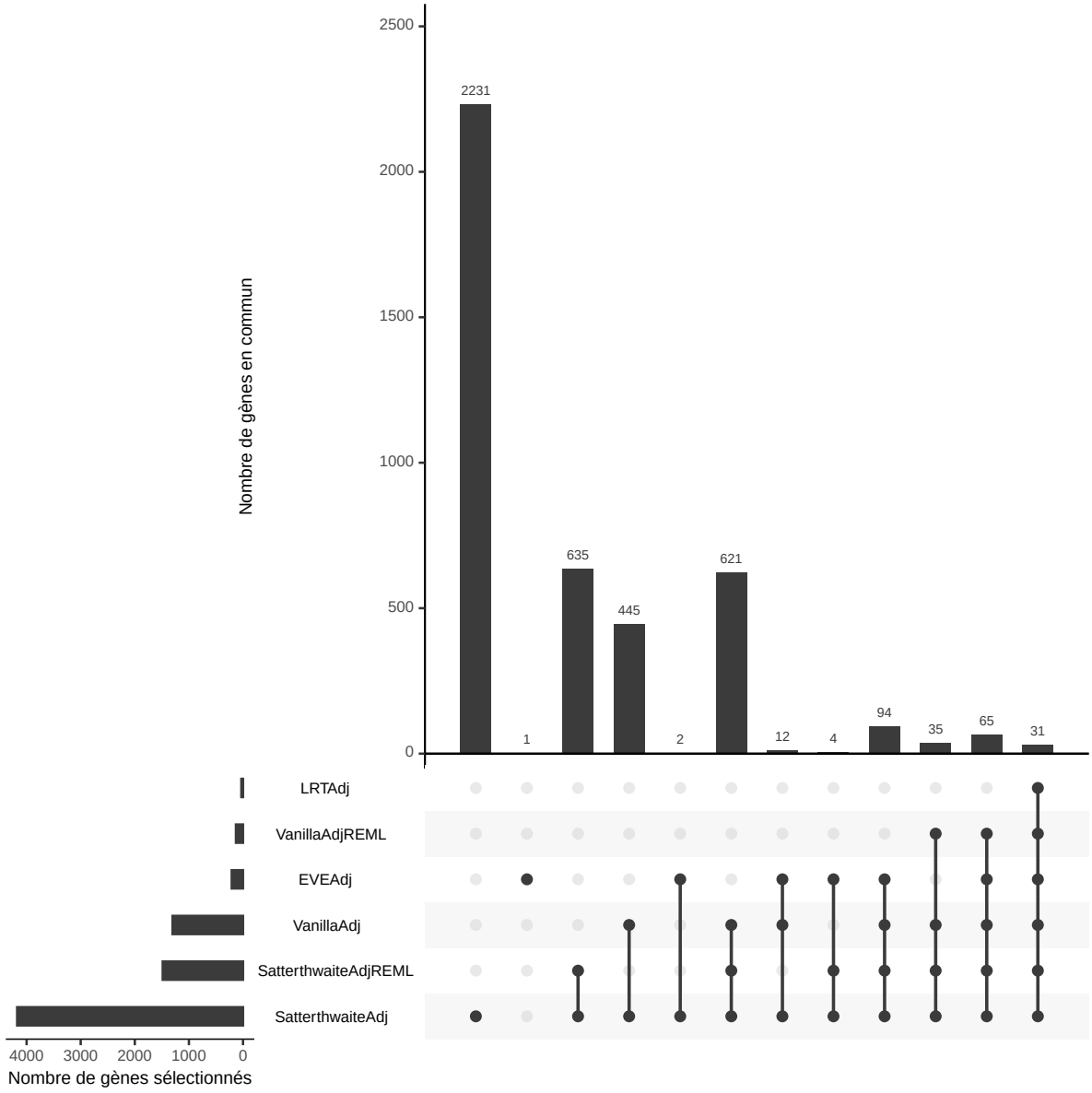

FIGURE 6 – Diagramme de Venn de toutes les méthodes en incluant la méthode EVE

Revenir sur explication de gènes différentiellement exprimées etc.

Applications aux données réelles de Chen mais ne pas perdre de temps à expliquer en détails EVEmodel (dire que c'est State of the art).

5 Résultats

6 Discussion et conclusion

Intro

Application/Résultats : décrire les données, vite fait normalisation avec vrai aebre, on ne connait pas Discussion/COnclusion ? Interprétation des résultats sinon la mettre dans les f-cicd : CI/CD to build Latex PDF ... CI/CD to build Latex pdf and create a release

in with GitHub Actions. The workflow triggers on push to the repository. Integrates with Overleaf.

Références

Bastide, Paul et Julien Clavel (s. d.). « Continuous Trait Evolution ».

Bastide, Paul, Mahendra Mariadassou et Stéphane Robin (juill. 2022). « Modèles d'évolution de caractères continus ». In : Didier, Gilles et Stéphane Guindon. *Modèles et méthodes pour l'évolution biologique*. ISTE Group, p. 47-85. isbn : 978-1-78948-069-6. doi : 10.51926/ISTE.9069.ch3. url : https://www.istegroup.com/fr/produit/modeles-et-methodes-pour-levolution-biologique/?/47495 (visité le 14/11/2023).

Bastide, Paul, Charlotte Soneson et al. (1er jan. 2023). « A Phylogenetic Framework to Simulate Synthetic Interspecies RNA-Seq Data ». In : *Molecular Biology and Evolution* 40.1, msac269. issn : 1537-1719. doi : 10.1093/molbev/msac269. url : https://doi.org/10.1093/molbev/msac269 (visité le 20/11/2023).

Bel, L et al. (s. d.). *Le Modèle Linéaire et ses Extensions*.

Bgee (2023). *Bgee : Gene Expression Data in Animals*. url : https://www.bgee.org/ (visité le 20/11/2023).

Chen, Jenny et al. (jan. 2019). « A Quantitative Framework for Characterizing the Evolutionary History of Mammalian Gene Expression ». In : *Genome Res* 29.1, p. 53-63. issn : 1549-5469. doi : 10.1101/gr.237636.118. pmid : 30552105.

Gomez-Mestre, Ivan, Robert Alexander Pyron et John J. Wiens (2012). « Phylogenetic Analyses Reveal Unexpected Patterns in the Evolution of Reproductive Modes in Frogs ». In : *Evolution* 66.12, p. 3687-3700. issn : 1558-5646. doi : 10.1111/j.1558-5646.2012.01715.x. url : https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2012.01715.x (visité le 13/11/2023).

Kuznetsova, Alexandra, Per B. Brockhoff et 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/ (visité le 01/03/2024).

Petersen, Kaare Brandt et Michael Syskind Pedersen (2012). *The Matrix Cookbook*. Version 20121115. url : http://matrixcookbook.com.

Rohlfs, Rori V. et Rasmus Nielsen (1er sept. 2015). « Phylogenetic ANOVA : The Expression Variance and Evolution Model for Quantitative Trait Evolution ». In : *Systematic Biology* 64.5, p. 695-708. issn : 1063-5157. doi : 10.1093/sysbio/syv042. url : https://doi.org/10.1093/sysbio/syv042 (visité le 06/03/2024).

Satterthwaite, F. E. (déc. 1946). « An Approximate Distribution of Estimates of Variance Components ». In : *Biometrics Bulletin* 2.6, p. 110. issn : 00994987. doi : 10.2307/3002019. JSTOR : 10.2307/3002019. url : https://www.jstor.org/stable/10.2307/3002019?origin=crossref (visité le 08/01/2024).

Wide Cross-species RNA-Seq Comparison Reveals Convergent Molecular Mechanisms Involved in Nickel Hyperaccumulation across Dicotyledons - García de La Torre - 2021 - New Phytologist - Wiley Online Library (2023). url : https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.16775 (visité le 20/11/2023).

A Application aux données réelles

Comme nous l'avons remarqué dans la section 4 l'application de la méthode EVEmodel a produit des valeurs manquantes pour les gènes présentés dans le tableau suivant.

Gènes ayant produits des NA
OG15121
OG3765
OG4072
OG412
OG4690
OG594
OG7272
OG7523
OG7564
OG8117
OG8343
OG9829

TABLE 1 – Table des gènes pour lesquels la méthode `EVEmodel` a produit des NA

Code du projet

Tout le code produit est disponible sur le dépôt GitHub suivant `https://github.com/Polarolouis/anova-phylogenetique-projet-msv/`. Ce dépôt contient le code pour implémenter la méthode, faire les simulations et compiler le rapport.

Nous avons au maximum indiqué le code qui n'a pas été écrit par nous, la plupart du temps dans les commentaires du code.