Rapport de Projet : ANOVA Phylogénétique

Alizée Geffroy Louis Lacoste

20 mars 2024

Table des matières

1 Introduction

Avec l'avènement des données massives de génomiques, transcriptomiques, protéomiques, il y a besoin de techniques statistiques robustes et passant à l'échelle permettant de mener à bien les analyses.

Ces données de génétiques proposent bien souvent deux informations, les mesures et l'arbre phylogénétique. Et pour certaines, l'arbre est ramifié au bout en proposant des répétitions intraspécifique.

C'est par exemple le cas pour les données de Chen et al. 2019 dont la figure 1 présente l'arbre phylogénétique : La problématique qui se pose souvent est celle de l'analyse de dif-

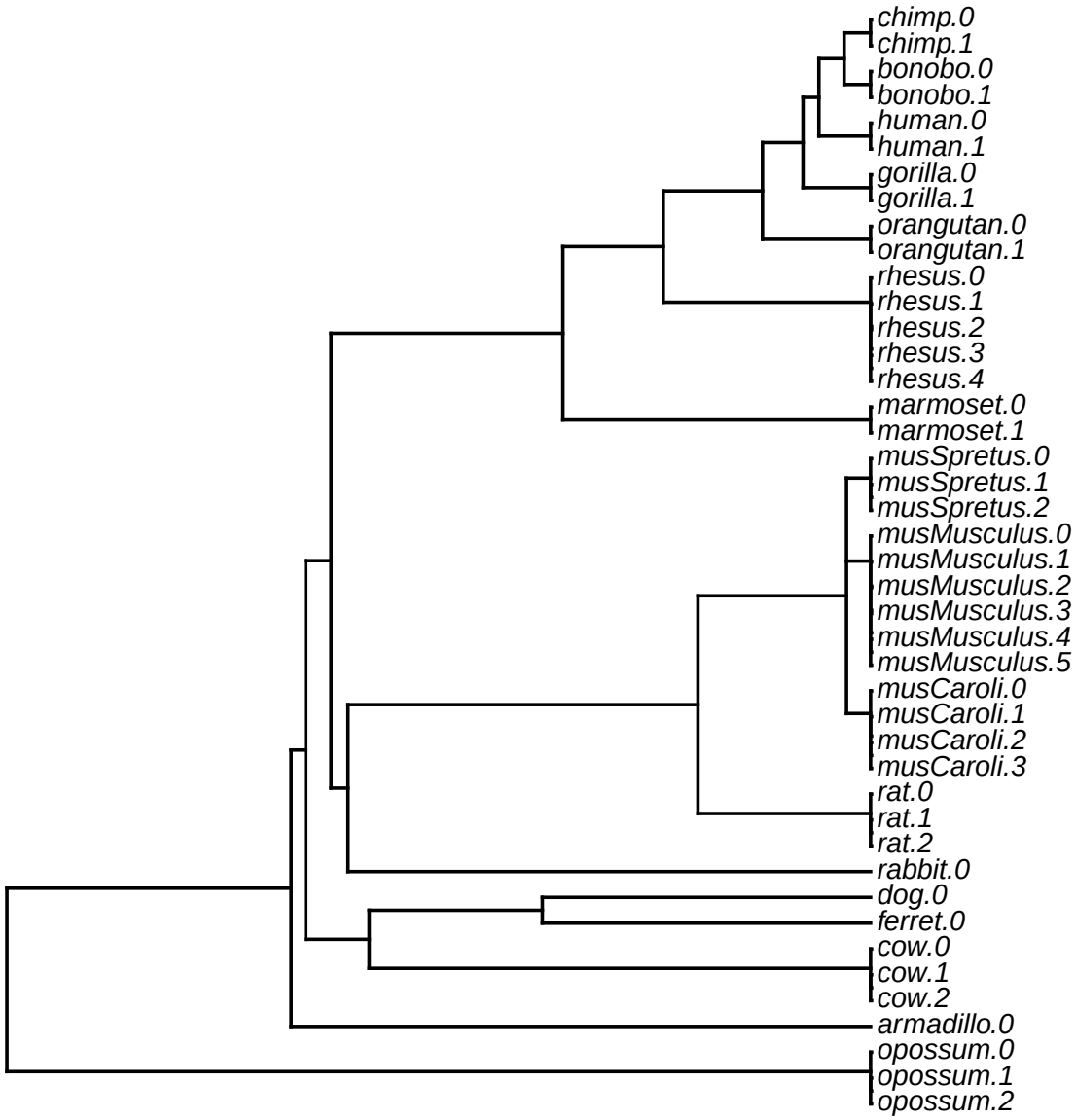


Figure 1 – Arbre phylogénétique de Chen et al. 2019

férents gènes qu'on mesure chez plusieurs espèces. On cherche alors d'abord à trouver quels gènes pourraient être différentiels chez certaines espèces, en détectant un changement d'expression dans certains groupes. En considérant l'arbre precédent 1, on pourra

chercher les gènes qui sont différents entre les groupes des *mus* et *rat* par rapport aux autres espèces.
Le modèle le plus couramment utilisé est actuellement l'Expression Variance and Evolution modèle (EVE) présenté dans Rohlfs et Nielsen 2015. L'EVE modèle est basé sur un Likelihood Ratio Test (LRT), une méthode statistique classique. Ce projet s'inscrit alors dans un questionnement plus large qui cherche à se demander si d'autres modèles classiques comme l'ANOVA, en les adaptant, pourrait produire des résultats similaires voire meilleurs que l'EVE modèle. En effet, avoir un bon modèle qui, en particulier, donne peu de faux positifs est important. On peut ensuite étudier les gènes potentiellement intéressants selon une problématique et des groupes d'espèces données.
Au vu de la forme des données étudiées, le projet s'est tourné vers une méthode d'ANOVA phylogénétique. Celle-ci sera d'abord décrite ainsi que d'autres outils mathématiques utilisés pour affiner la fiabilité du test dans une première partie. Certains auront fait l'objet de calculs explicites en vue de leur implémentation. A partir de ces résultats, nous avons implémenter ces méthodes en R d'abord appliquées à des simulations destinées à comparer et étudier la méthode d'ANOVA phylogénétique sur des données d'arbre simulés. Enfin, on a testé sur des données réelles.

Au cours de ce projet nous avons donc eu une partie d'étude théoriques et mathématiques des modèles de l'ANOVA et de l'ANOVA phylogénétique afin de bien l'adapter à nos données. A partir de la formulation mathématiques des modèles

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.

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

Contrairement à une comparaison basée sur la santé des individus, cette approche se focalise sur les espèces. La non-indépendance et les relations complexes entre individus et groupes comparés nécessitent l'utilisation d'un modèle mixte, impliquant la matrice des temps de divergences, ainsi que l'intégration de processus stochastiques tels que le Mouvement Brownien sans erreurs, avec ajustement du ratio erreur de mesure / erreur dûe à la phylogénie.

2 Méthodes

Dans cette partie nous présentons les modèles statistiques d'ANOVA et sa dérivée phylogénétique. Après avoir posé le cadre mathématique à partir des recherches bibliographiques, nous développerons les outils mathématiques. En particulier pour l'approximation nous avons calculé une forme explicite afin de l'implémenter. Finalement, nous faisons une présentation succinte des méthodes REML et du modèle LRT.

2.1 L'ANOVA

L'ANOVA est un cas classique du modèle linéaire, nous utilison ici une forme matricielle.

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

La forme matricielle usuelle de l'ANOVA à 1 facteur et 2 groupes de taille respectivement n_1 et n_2 est la suivante :

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

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

On noter qu'ici $\beta_1 = \mu_1$ la moyenne du groupe 1, et $\beta_2 = \mu_2 - \mu_1$ la différence des moyennes entre les groupes dans cette paramétrisation.

Les paramètres $(\beta_1, \beta_2, \sigma^2)$ de l'ANOVA sont estimables, grâce par exemple à la méthode du maximum de vraisemblance et ont des formules bien connues.

Test statistique

Dans le cadre d'ANOVA classique nous allons rappeler les hypothèses du test et la statistique de test. On fait un test sur les moyennes de chaque groupe. Ce peut être la moyenne de la valeur d'un trait génétique ou bien de la valeur de la fréquance d'une séquence ou allèle. On testera alors les hypothèses suivantes avec $l = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$:

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

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

On a alors la statistique de test suivante :

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

$$\text{Où } \bar{Y} = \frac{1}{n}\sum_{i,j=1}^{n_1,n_2} Y_{i,j} \text{ et } \hat{Y} = X\hat{\beta}$$

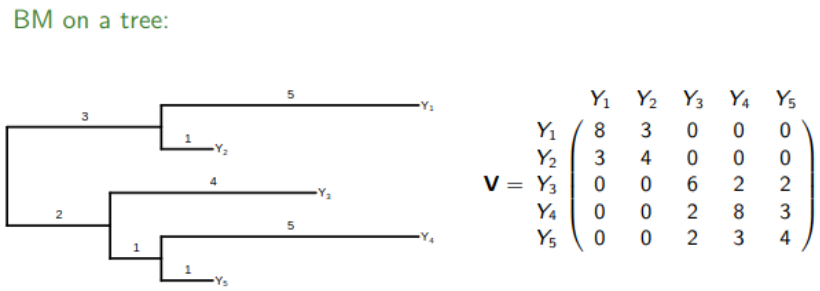
2.2 L'ANOVA phylogénétique

Dans la méthode d'ANOVA classique l'information portée par l'arbre phylogénétique n'est pas prise en compte. Le but de cette nouvelle méthode est de ne plus mettre cette information de côté et peut être obtenir de meilleurs résultats. En effet on peut imaginer, en considérant des traits évolutifs ou des séquences d'ADN, que des individus d'une même espèce ou bien d'espèces proche phylogénétiquement, pourraient avoir des valeurs proches. Il s'agira alors de modéliser l'arbre et les informations évolutives qu'ils contient de manière à l'incorporer.

Comme décrit dans Bastide, Mariadassou et Robin 2022 l'évolution d'un trait nécessite de décrire ses fluctuations le long de l'arbre et ses branches. C'est pour cela que souvent cela est le résultat d'un processus stochastique à temps continu branchant sur un arbre phylogénétique, supposé connu et fixé. Le processus classique est le mouvement brownien et c'est celui que nous avons utilisé. Il a cependant quelques limites qui ne font pas l'objet de ce rapport mais qui peuvent alors justifier le choix d'autres types de processus comme celui d'Ornstein-Uhlenbecks. Le modèle de mouvement brownien va alors induire que les feuilles des arbres (nos observations) auront une distribution gausienne que l'on écrira sous la forme suivante :

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

Les notations correspondent toujours à celles utilisées pour (1). La seule différence se trouvant dans la distribution de u et la présence d'une matrice K. Dans le cadre du mouvement brownien $K = (K_{i,j})_{1 \leq i,j \leq n} = (t_{i,j})_{1 \leq i,j \leq n}$ où $t_{i,j}$ représente le temps d'évolution commun aux espèces i et j. Comme on peut le voir dans l'exemple suivant, cette matrice a bien la forme attendue : deux espèces proches dans l'arbre ont un coefficient de covariance élevé (leur temps d'évolution commun est grand), alors que deux espèces éloignées sont plus faiblement corrélées. On peut voir un exemple utilisé dans les slides de cours Bastide et Clavel 2022, où la matrice V correspond à notre matrice K :



De plus l'arbre présenté dans la figure du dessus n'a pas de lien avec la figure suivante.

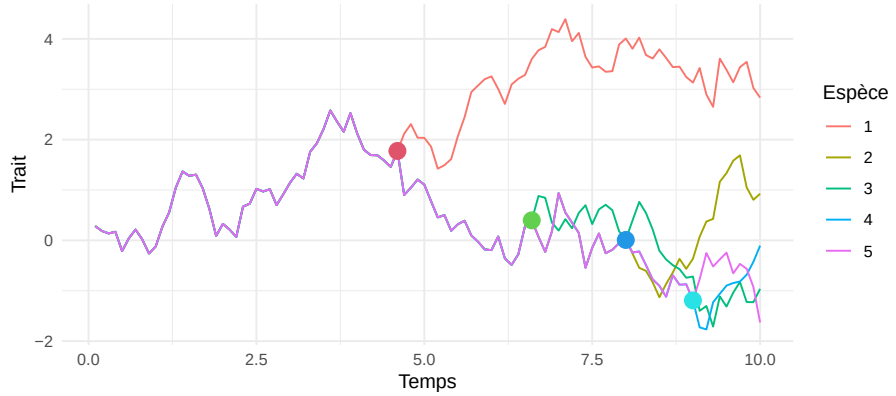


Figure 2 – Exemple d'un arbre phylogénétique dont le trait est généré selon un Mouvement Brownien

A note que seules les réalisation du processus aux feuilles de l'abre (ici, à t = 10) sont observées. Le reste du processus est latent.

Test statistique

En considérant le même test et les mêmes hypothèses que 2.1 mais en prenant en compte la nouvelle formule 3, on obtient une nouvelle statistique de test.

Bastide et Clavel 2022 nous donne la forme de la statistique pour la méthode d'ANOVA de cette forme.

$$F_{ANOVAphylo} = \frac{||\hat{Y} - \bar{Y}||^2_{K^{-1}}(n-2)}{||Y - \hat{Y}||^2_{K^{-1}}} \underset{\mathcal{H}_0}{\sim} \mathcal{F}\text{isher}(1, n-2) \tag{4}$$

Où $||Y - \hat{Y}||^2_{K^{-1}} = ||Y - X\hat{\beta}||^2_{K^{-1}} = Proj^K_X Y = (Y - \hat{Y})^T K^{-1}(Y - \hat{Y})$
et $||\hat{Y} - \bar{Y}||^2_{K^{-1}} = (\hat{Y} - \bar{Y})^T K^{-1}(\hat{Y} - \bar{Y})$

Concernant cette statitstique, on peut dire qu'elle est toujours exacte car on connait la matrice K.

2.3 ANOVA phylogénétique avec erreur de mesure

Dans la section précedente, on a supposé que la seule source de variabilité provenait du mouvement brownien sur l'arbre. On rajoute dans cette section une autre variabilité specifiée par σ^2_{err} qui à partir de la formule précédente (3), nous donne :

$$Y = X\beta + u + \epsilon, \quad u \sim \mathcal{N}_n(0, \sigma^2_{phy}K), \quad \epsilon \sim \mathcal{N}_n(0, \sigma^2_{err}I_n) \tag{5}$$

Alors $Y \sim \mathcal{N}_n(X\beta, \sigma^2_{phy}K + \sigma^2_{err}I_n)$
On pose $\theta = (\sigma^2_{phy}, \sigma^2_{err})$
On définit pour la suite $Var_\theta(Y) = V(\theta) = \sigma^2_{phy}K + \sigma^2_{err}I_n$ (6)

Comme décrit dans Bastide, Mariadassou et Robin 2022, l'ajout de cette variance résiduelle dans notre modèle est crucial pour mieux représenter la complexité des données que nous traitons. En effet, supposer que la seule source de variation entre les observations est le processus stochastique sur l'arbre phylogénétique (specifiée par $\sigma^2_{phy}K$) est souvent peu réaliste, surtout dans des contextes où les données sont hétérogènes ou comme on le verra plus tard, nous avons les données de plusieurs individus d'une même espèce. C'est d'ailleurs pour ça qu'on peut parler de variation intraspécifique. Cette hypothèse simplificatrice peut introduire des biais significatifs dans nos analyses, compromettant ainsi la validité des résultats obtenus. En intégrant la variance résiduelle, qui capture l'effet indépendant de l'environnement sur chaque mesure, notre modèle devient plus flexible et mieux adapté pour tenir compte de la variabilité observée dans les données. Le modèle mixte phylogénétique résultant, combinant à la fois la variance phylogénétique et la variance résiduelle, nous permet de distinguer les effets héritables des effets non héritables, offrant ainsi une approche plus nuancée et réaliste de l'analyse comparative des données évolutives.

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

$$Y = X\beta + E, \text{ où } Var(E) = V(\theta) = \sigma^2_{phy}(K - \lambda I_n) = \sigma^2_{phy}V_\lambda \tag{7}$$
$$E \sim \mathcal{N}_n(0, V_\lambda)$$

Test statistique

Le test statistique et ses hypothèses sont conservés et on obtient de la même façon que dans la section précédente une nouvelle statistique de test en lien avec l'équation 7.

$$F_{ANOVAphylo-error} = \frac{||\hat{Y} - \bar{Y}||^2_{V_\lambda^{-1}}(n-2)}{||Y - \hat{Y}||^2_{V_\lambda^{-1}}} \underset{\mathcal{H}_0}{\sim} \mathcal{F}\text{isher}(1, n-2) \tag{8}$$

Où $||Y - \hat{Y}||^2_{V_\lambda^{-1}} = ||Y - X\hat{\beta}||^2_{V_\lambda^{-1}} = Proj_X^{V_\lambda} Y = (Y - \hat{Y})^T V_\lambda^{-1}(Y - \hat{Y})$
et $||\hat{Y} - \bar{Y}||^2_{V_\lambda^{-1}} = (\hat{Y} - \bar{Y})^T V_\lambda^{-1}(\hat{Y} - \bar{Y})$

Il est important de noter que lorsque le paramètre λ est connu, l'ANOVA phylogénétique est exacte. Cependant, dans la pratique, λ est généralement inconnu et doit être estimé à partir des données. Dans ce cas, l'approximation de la distribution de F par une distribution de Fisher ne tient plus, et il est nécessaire d'utiliser des méthodes alternatives telles que la méthode de Satterthwaite pour estimer les degrés de liberté. Cette méthode tient compte de l'incertitude associée à l'estimation de λ et fournit une approximation plus précise de la distribution de la statistique de test.

2.4 Approximation de Satterthwaite

On s'est basé sur la documentation du package `lmerTest` Kuznetsova, Brockhoff et Christensen 2017 pour calculer les formules explicites de l'approximation dans notre cadre. En effet il existe des formules explicite dans le cadre du modèle mixte. Dans notre cas, on peut voir l'équation (5) comme l'équation d'un modèle linéaire mixte où β représente tous les paramètres à effets fixes, avec sa matrice de design associée X, u les effets aléatoires et ϵ les résidus. Dans l'optique de l'implémenter, nous avons calculé la formule explicite de l'approximation de Satterthwaite. A partir de 5 on rappelle les valeurs suivantes :

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

De la documentation on obtient alors la covariance suivante :

$$C(\theta) = (Cov(\beta_i, \beta_j))_{i,j} = (X^T V(\theta)^{-1} X)^{-1} = (X^T(\sigma^2_{phy}K + \sigma^2_{err}I_n)^{-1}X)^{-1} \tag{9}$$

Approximation (F-statistique et approximation de Satterthwaite)**.**

$$F_{approx} = \frac{||\hat{Y} - \bar{Y}||^2_{V_\lambda^{-1}} df_{approx}}{||Y - \hat{Y}||^2_{V_\lambda^{-1}}} \underset{\mathcal{H}_0}{\sim} \mathcal{F}isher(1, df_{approx}) \tag{10}$$

$$Avec\ df_{approx} = \frac{2(f(\hat{\theta}))^2}{[\nabla f(\hat{\theta})]^T A [\nabla f(\hat{\theta})]} \tag{11}$$

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

Calcul explicite de l'approximation. 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 :

$$df = \frac{2(l^T \hat{C} l)^2}{[Var(l^T \hat{C} l)]} = \frac{2(f(\hat{\theta}))^2}{[Var(f(\hat{\theta}))]} \approx \frac{2(f(\hat{\theta}))^2}{[\nabla f(\hat{\theta})]^T A [\nabla f(\hat{\theta})]} \tag{12}$$

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

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

Étape 1 : 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 :

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

Étape 2 : Calcul de A

Par Cramer-Rao on sait que pour les estimateurs non biaisés, $Var(\hat{\theta}) \geqslant I(\theta)^{-1}$. $I(\theta)^{-1}$ est l'inverse de l'information de Fisher. Lorsque la fonction de vraissemblance est assez régulière comme ici, on a $I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} l(Y, \theta)\right]$ La matrice de variance-covariance $Var(\hat{\theta})$ peut alors être approximée par $I(\hat{\theta})^{-1} = A$.
Dans la plupart des cas il est plus simple d'estimer cette matrice par des méthodes numériques, pour autant une formule explicite de la Hessienne rend le calcul plus rapide et plus robuste quant aux erreurs numériques.
On va donc abord calculer la log-vraissemblance du vecteur Y défini précédemment :

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

On calcule les dérivées premières de la log-vraissemblance

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

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

Puis les dérivées secondes :

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

car $\partial((Y-X\beta)^T V(\theta)^{-1}KV(\theta)^{-1}(Y-X\beta)) = (Y-X\beta)^T\partial(V(\theta)^{-1}KV(\theta)^{-1})(Y-X\beta)$
et $\partial\left(V(\theta)^{-1}KV(\theta)^{-1}\right) = -V(\theta)^{-1}\partial V(\theta)V(\theta)^{-1}KV(\theta)^{-1} - V(\theta)^{-1}KV(\theta)^{-1}\partial V(\theta)V(\theta)^{-1}$
ce qui donne $\partial_{\sigma^2_{phy}\sigma^2_{phy}}\left(V(\theta)^{-1}KV(\theta)^{-1}\right) = -2V(\theta)^{-1}KV(\theta)^{-1}KV(\theta)^{-1}$

$$\begin{aligned}\partial_{\sigma^2_{err}\sigma^2_{phylo}} l &= \partial_{\sigma^2_{phy}\sigma^2_{err}} l \\ &= -\frac{1}{2}Tr(\partial_{\sigma^2_{phy}\sigma^2_{err}}(V(\theta)^{-1})K) + \frac{1}{2}(Y-X\beta)^T\partial_{\sigma^2_{phy}\sigma^2_{err}}(V(\theta)^{-1}KV(\theta)^{-1})(Y-X\beta) \\ &= \frac{1}{2}Tr(V(\theta)^{-1}I_nV(\theta)^{-1}K) - \frac{1}{2}(Y-X\beta)^T(V(\theta)^{-1}V(\theta)^{-1}KV(\theta)^{-1} + \\ &V(\theta)^{-1}KV(\theta)^{-1}V(\theta)^{-1})(Y-X\beta)\end{aligned}$$

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

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

De là on obtient la Hessienne $\begin{pmatrix}\partial_{\sigma^2_{phy}\sigma^2_{phy}} l & \partial_{\sigma^2_{phy}\sigma^2_{err}} l \\ \partial_{\sigma^2_{err}\sigma^2_{phy}} l & \partial_{\sigma^2_{err}\sigma^2_{err}} l\end{pmatrix}$ puis A en inversant la matrice de Fisher liée, ce qui peut se faire par des méthodes numériques ou analytiques ici car cela signifie inverser une matrice 2 x 2 ce qui est facile.

□

2.5 REML

Le REML, ou Maximum de Vraisemblance Restreint (Restricted Maximum Likelihood en anglais), est une méthode statistique utilisée dans l'estimation des paramètres de modèles linéaires mixtes (ou modèles à effets mixtes) et dans l'analyse de la variance (ANOVA). Il s'agit d'une approche alternative à la méthode de maximum de vraisemblance (ML) standard, notamment lorsque l'on travaille avec des modèles à effets aléatoires.

L'une des formules pour la vraissemblance restreinte consite à regarder la vraissemblances des observations en intégrant sur les effets fixes, ici β sur lesquels on a mis un prior impropre uniforme entre moins l'infini et plus l'infini. C'est à dire sous l'hypothèse que les estimations des effets fixes sont éliminées (conditionnées)

$$L_{REML}(Y;\theta) = \int_{\beta\in\mathbb{R}^2} \frac{1}{(2\pi)^{n/2}|V(\theta)|^{1/2}} \exp\left(-\frac{1}{2}(Y-X\beta)^T V(\theta)^{-1}(Y-X\beta)\right) d\beta$$

Lorsque dans un estimateur on a la division par $(n-p)$ au lieu de n, on fait sans le dire un REML dans le cas de la régression linéaire classique.
Satterthewaite maximum likelihood ne fait pas l'astuce de diviser par $n-p$, au contraire de l'anova classique ou phylogénétique. Il est donc particulièrement mauvais. Le REML permet de mieux estimer la variance, et aussi le λ (ratio des variances), ce qui est crucial dans la dérivation des degrés de libertés approchés.

2.6 Méthode Likelihood Ratio Test

La méthode du test de rapport de vraisemblance (LRT) est une technique statistique utilisée pour comparer deux modèles statistiques et déterminer s'ils diffèrent significativement en termes d'ajustement aux données. Le rapport de vraisemblance est calculé comme le rapport des vraisemblances maximales de deux modèles (sous les deux hypothèses) emboîtés :

$$\text{LRT} = -2\log\left(\frac{L(\theta_{\text{H}_0})}{L(\theta_{\text{H}_1})}\right)$$

Sous l'hypothèse nulle que le modèle plus simple est correct, ce rapport suit approximativement une distribution du chi-deux avec un nombre de degrés de liberté égal à la différence dans le nombre de paramètres entre les deux modèles. Ainsi, en comparant la valeur observée du rapport de vraisemblance à la distribution du chi-deux, on peut décider si l'ajout de paramètres dans le modèle conduit à une amélioration significative de l'ajustement aux données.
Il est à noter que cette méthode est plus coûteuse car il est nécessaire d'ajuster 2 modèles au lieu d'un seul dans les autres méthodes.

3 Simulations

Dans cette partie nous souhaitons comparer les résultats de l'ANOVA et de l'ANOVA phylogénétique classique, avec approximation de Satterthwaite et avec le *Likelihood ratio test*. 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 au hasard par rapport à la phylogénie.
— Des données réparties en deux groupes cohérents avec la phylogénie.

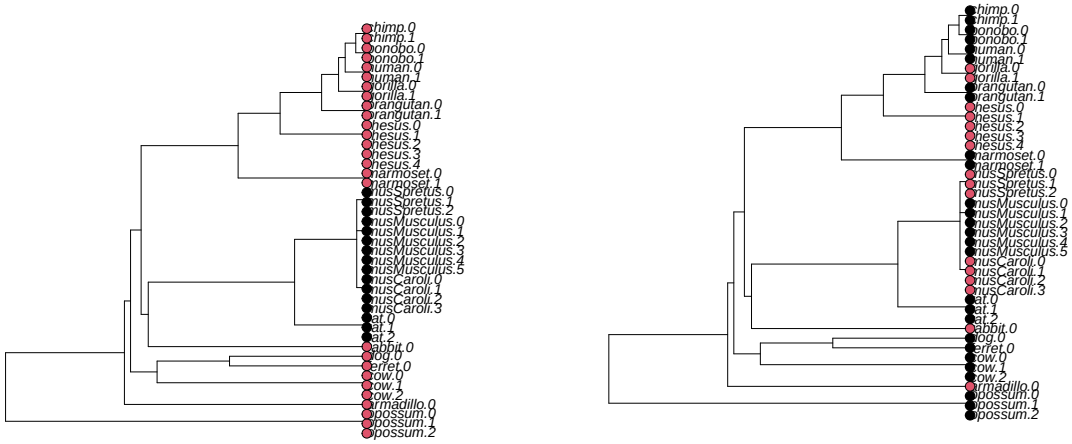
En sélectionnant des espèces de manière aléatoire, nous cassons la structure induite par la phylogénie. Nous nous attendons donc à ce que l'ANOVA réalise de meilleurs résultats que l'ANOVA Phylogénétique en ne prenant pas en compte l'information phylogénétique.

Pour les simulations avec des groupes respectant la structure 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 3a).

Et pour le groupe ne respectant pas la phylogénie, nous avons sélectionnés les espèces en respectant les proportions des groupes définis avant afin de rendre les résultats comparables (voir la figure 3b). Enfin pour que notre analyse soit reproductible nous fixons la graine à 1234.



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

(b) Groupes sélectionnés sans respect de la phylogénie.

FIGURE 3 – Arbre et groupes pour les simulations

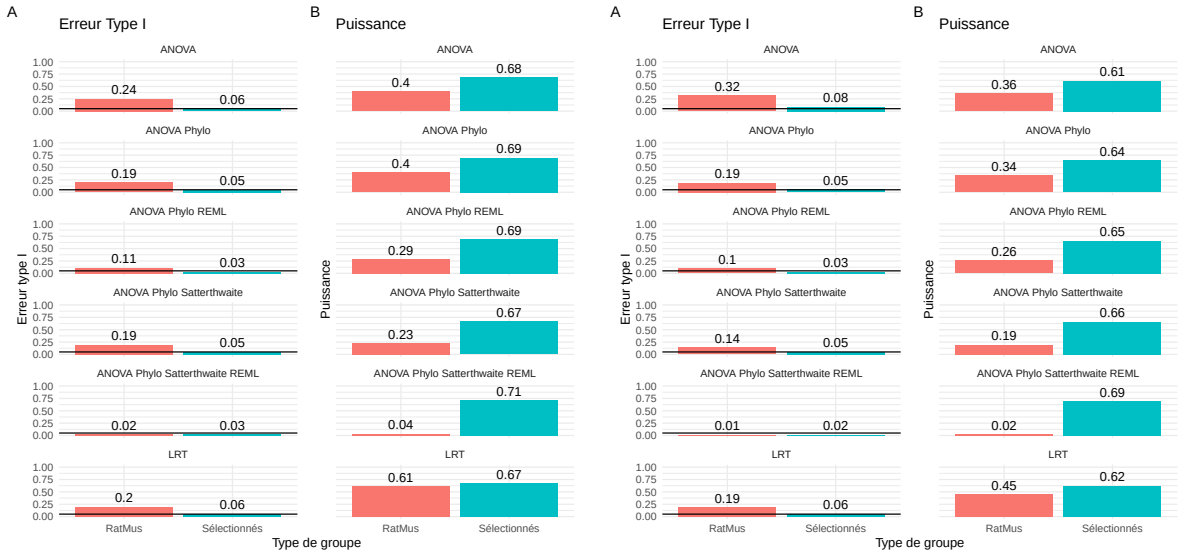
Afin d'avoir un paramètre unique à faire varier, nous re-paramétrisons le modèle, la variance totale v_{tot} suit la relation $v_{tot} = \sigma^2_{phylo} + \sigma^2_{measure} = 1$. Nous allons faire prendre à h, défini comme l'héritabilité, les valeurs $h \in (0.3, 0.5, 0.7, 0.9)$. L'héritabilité est liée à σ^2_{phylo} et $\sigma^2_{phylo} = h \times v_{tot}$. Et alors $\sigma^2_{measure} = (1-h) \times v_{tot}$. Ainsi, $h = 0$ signifie qu'il y a seulement du bruit, et $h = 1$ seulement de l'information phylogénétique.

Pour les valeurs quantitatives des 2 groupes, nous avons 2 valeurs différentes :

$$\mu_1 = 0, \qquad \mu_2 = snr \times v_{tot} = \frac{taille\ d'effet}{\cancel{v_{tot}}} \times \cancel{v_{tot}} = 1 \tag{13}$$

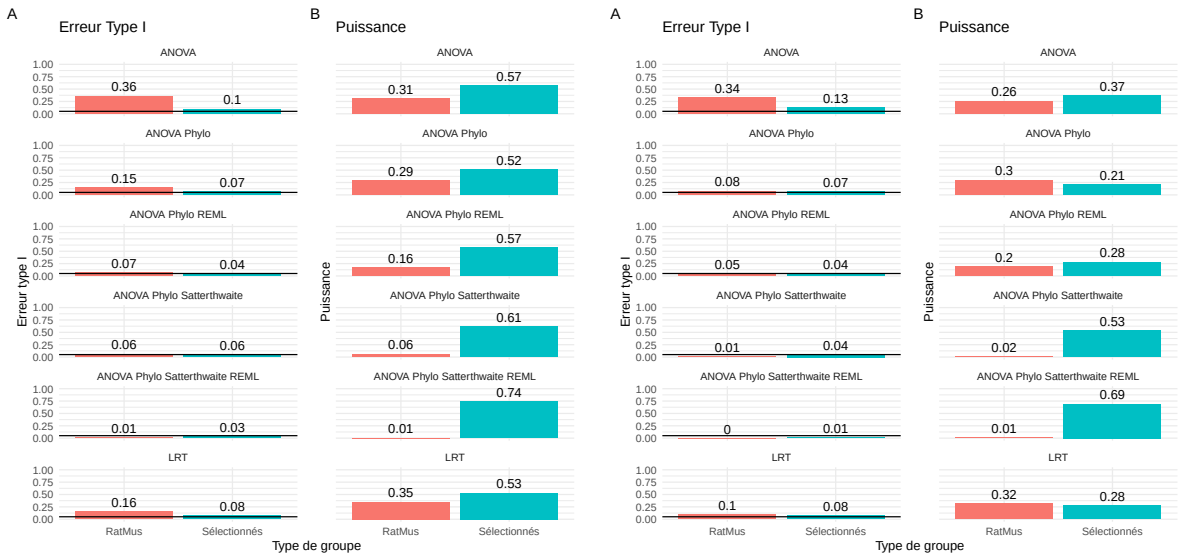
Note : snr signifie *signal noise ratio* et comme indiqué est donc le rapport entre la taille d'effet et la variance totale. Et dans l'équation 13, μ_1 et μ_2 correspondent aux β_1 et β_2 définis dans la sous-section 2.2.

Pour chaque valeur d'héritabilité, nous allons générer 500 jeux de données différents sur lesquels les méthodes sont utilisées avec les valeurs définies dans l'équation 13.



(a) $h = 0.3$

(b) $h = 0.5$



(c) $h = 0.7$

(d) $h = 0.9$

FIGURE 4 – Erreur de type I et puissance pour les simulations en faisant varier l'héritabilité

Sur toutes les sous-figures de la figure 4, les étiquettes A présentent les erreurs de type I commises par les méthodes et les étiquettes B présentent les puissances des mêmes méthodes.

Analyse pour les groupes respectant la phylogénie

Analyse des erreurs de type I L'erreur de type I est particulièrement importante à contrôler, en effet elle indique le nombre de faux positifs et l'on veut pouvoir en déterminer le seuil α avec comme seuil classique 0.05.

Nous constatons que dans le cas des groupes respectant la phylogénie, l'ANOVA a une erreur de type I très forte dans toutes les simulations. Pour l'expliquer nous avons deux interprétations principales. Tout d'abord, l'ANOVA suppose des observations indépendantes et identiquement distribuées et les en accord avec la phylogénie ne respectent pas cette hypothèse. De plus, n'ayant pas l'information de la dérive génétique sous-jacente, elle ne peut pas différencier ce qui est dû à la dérive et à de vraies différences entre les groupes. C'est la raison de son fort de taux de faux-positifs pour les groupes qui respectent la structure phylogénétique.

Par exemple, deux clades peuvent être éloignés à cause de leur éloignement temporel. L'oubli de la structure peut suggérer de mettre un saut alors que cet éloignement est seulement dû à la dérive.

Pour les autres méthodes, elles ont toutes tendances à avoir de forts taux de faux-positifs, exceptée l'ANOVA phylogénétique REML avec approximation de Satterthwaite qui respectent le seuil de 5% dans toutes nos conditions.

Nous remarquons qu'en général, plus l'héritabilité augmente et plus les méthodes incluant l'information phylogénétique contrôle l'erreur de première espèce.

Importance de l'erreur de type I Nous insistons particulièrement sur le contrôle de l'erreur de type I, car dans le cadre des analyses de données transcriptomiques cette phase d'analyse statistique permet d'identifier des gènes différentiellement exprimés et pouvant donc potentiellement intervenir dans des réseaux de gènes d'intérêt.

Une fois les gènes identifiés il faut faire des expériences qui sont particulièrement onéreuses et donc on ne souhaite pas faire des expériences "pour rien".

Analyse des puissances Le revers de la médaille se fait sentir sur les puissances. La méthode d'ANOVA phylogénétique REML avec approximation de Sattertwhaite, a les puissances les plus faibles de toutes les méthodes, ce qui fait sens, étant plus conservatrice elle sélectionne moins. Et nous observons donc que les méthodes avec les puissances les plus fortes sont le LRT et l'ANOVA.

REML vs Maximum Likelihood (ML) D'après nos simulations, les méthodes utilisant le REML contrôle toujours mieux l'erreur de première espèce que les méthodes utilisant le maximum de vraisemblance. Les paramètres de variance étant mieux estimés dans ce cas, ce résultat est cohérent avec les résultats classiques sur le REML. Mais à cause de ce meilleur contrôle, les méthodes REML ont donc des puissances plus faible, comme décrit plus haut.

Analyse pour les groupes choisis

Nous analysons ici les groupes sélectionnés pour ne pas respecter la phylogénie. Ils correspondent aux barres de couleurs bleues sur la figure 4

Analyse des erreurs de type I Toutes les erreurs de types sont proches d'être sous la barre des 5%. Les méthodes qui ne sont pas sous les 5% sont l'ANOVA et le LRT [1]. Cela indique peut-être que malgré notre sélection que nous avons souhaité la plus aléatoire possible [2], nous n'avons peut-être pas cassé toute la structure phylogénétique existante. Il faudrait investiguer avec d'autres simulations.

Analyse des puissances Comme l'on pouvait s'y attendre cette fois-ci toutes les puissances sont relativement élevées. Nous remarquons que la méthode la plus puissante est l'ANOVA phylogénétique REML avec approximation de Satterthwaite. Aux vues du doute émis au paragraphe précédent cela pourrait être dû à la persistance d'une structure phylogénétique.

REML vs Maximum Likelihood (ML) Ici aussi les méthodes REML contrôlent mieux l'erreur de type I mais fait intéressant elles obtiennent aussi de meilleures puissances. Cela pourrait être dû au fait que leur estimation de la variance est meilleure.

4 Application aux données réelles

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

Les données compilées par Chen et al. 2019 sont des données de RNA-seq, c'est-à-dire des données quantifiant l'expression des gènes, par le biais du transcriptome, parmi les différentes espèces du bout de l'arbre.

Le but est alors d'identifier les gènes différentiellement exprimés, au sens de nombre d'ARN par gène différent entre les espèces.

4.1 Modalités des tests

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

Ci-dessous la figure 5 présente les p-values ordonnées des différentes méthodes. Il s'agit d'une visualisation classique pour les données RNA-seq. Il est important de noter que ce graphique présente les p-values *non ajustées*.

1. Ainsi que pour les valeurs d'héritabilité de $h = 0.7$ et $h = 0.9$ l'ANOVA phylogénétique et l'ANOVA phylogénétique avec approximation de Satterthwaite, qui sont légèrement au-dessus. Un point intéressant à remarquer est que leurs contreparties utilisant le REML ne présentent pas ces problèmes.
2. Cela en respectant la contrainte de ne pas séparer les individus d'une même espèce.

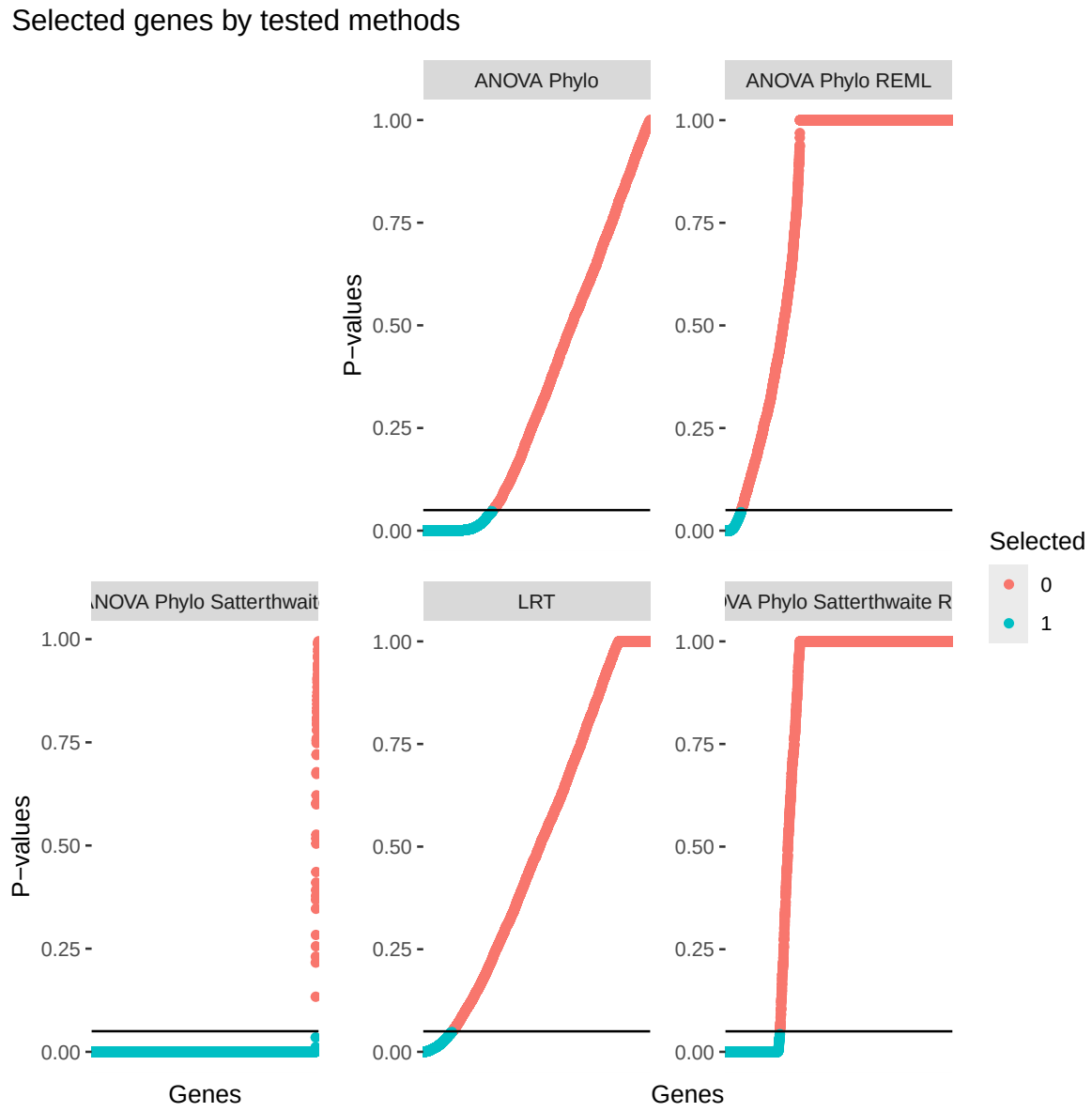


FIGURE 5 – *p-values* ordonnées pour les différents tests

Pour la suite de cette analyse, nous allons appliquer un ajustement des p-values pour les test multiples, nommément la correction de BENJAMINI et HOCHBERG 1995.

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

On peut voir que la méthode de Satterthwaite sans REML a sélectionné énormément de gènes, 5346 comme étant différentiellement exprimés.

Ce résultat n'étant pas biologiquement crédible, nous préférons ne pas l'afficher dans le *UpSet diagram*, figure 6.

4.2 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 pour cette problématique.

Elle détecte ici 209 gènes différentiellement exprimés.

Son principe de fonctionnement suppose que les traits évoluent selon un processus d'Ornstein-Uhlenbeck et le test réalisé est un *Likelihood Ratio test*.

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

Toutes les méthodes

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

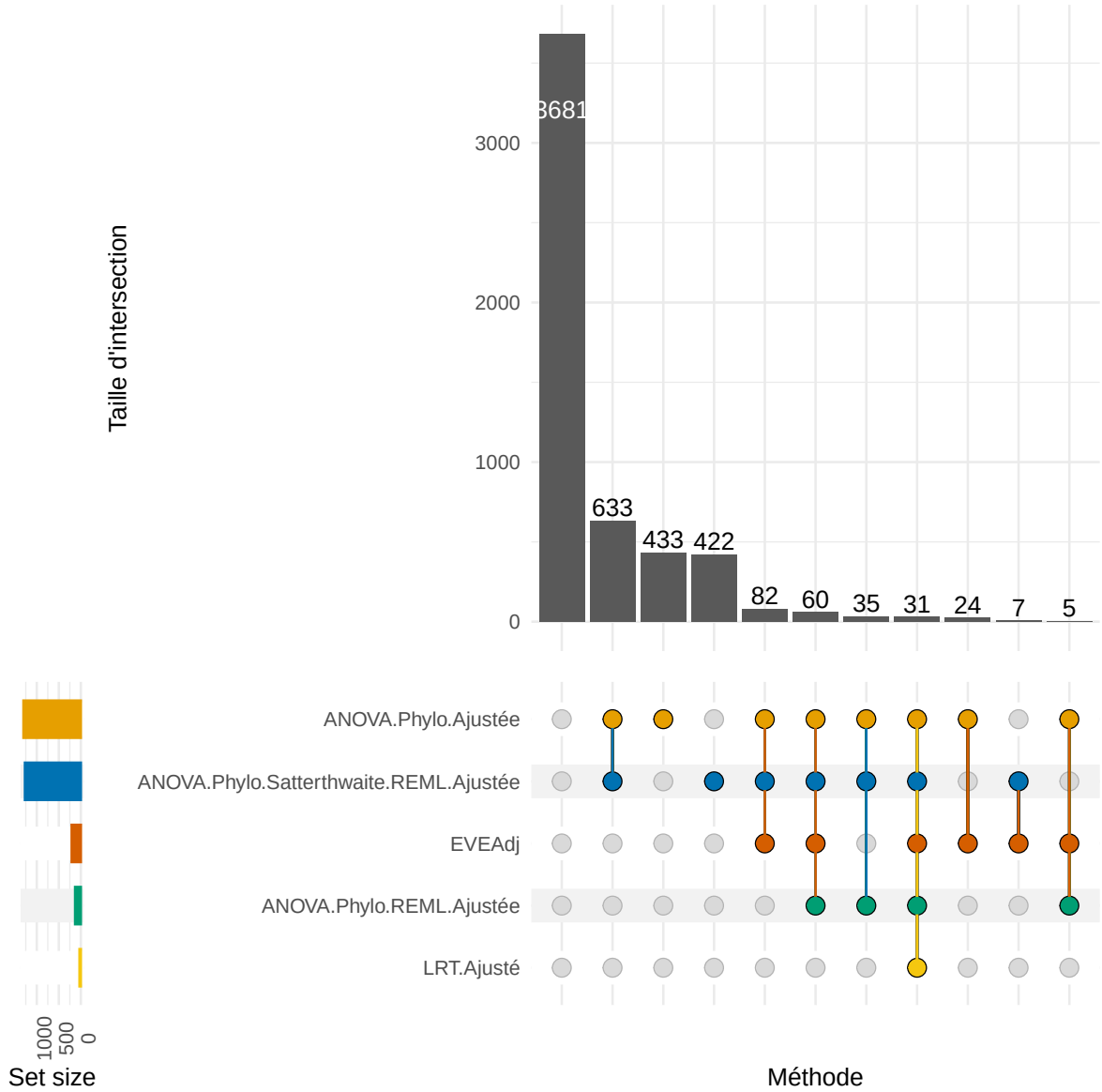


FIGURE 6 – *UpSet diagram* de toutes les méthodes en incluant la méthode EVE

Analyse des résultats Nous pouvons voir que la méthode la plus parcimonieuse est celle utilisant le LRT, qui sélectionne 31 gènes qui sont eux-mêmes **sélectionnés par toutes les méthodes**. Cette unanimité sur ces gènes nous invite à penser qu'ils sont en effet différentiellement exprimés.

La seconde méthode sélectionnant le moins de gènes est l'ANOVA Phylogénétique avec REML. Elle sélectionne 131 gènes. Ces sélections se décompose en plusieurs sous ensembles

TODO Ici nous avons supposé un mouvement brownien comme processus sous-jacent de l'arbre mais ce n'est peut-être pas le meilleur modèle et un OU pourrait être intéressant. Intéressant pour l'ouverture.

5 Conclusions sur le projet

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.

TODO : problèmes qu'on peut avoir eu : Satterthwaite estimation de la Hessienne pas stable, donc utilisation de l'analytique

Références

Bartlett, Maurice Stevenson et Ralph Howard Fowler (jan. 1997). « Properties of Sufficiency and Statistical Tests ». In : *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 160.901, p. 268-282. doi : 10.1098/rspa.1937.0109. url : https://royalsocietypublishing.org/doi/10.1098/rspa.1937.0109 (visité le 17/03/2024).

Bastide, Paul et Julien Clavel (déc. 2022). « 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*.

Benjamini, Yoav et Yosef Hochberg (1995). « Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, p. 289-300. issn : 0035-9246. JSTOR : 2346101. url : https://www.jstor.org/stable/2346101 (visité le 17/03/2024).

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).

Harville, David A. (1er juin 1977). « Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems ». In : *Journal of the American Statistical Association* 72.358, p. 320-338. issn : 0162-1459. doi : 10.1080/01621459.1977.10480998. url : https://www.tandfonline.com/doi/abs/10.1080/01621459.1977.10480998 (visité le 17/03/2024).

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).

Patterson, H. D. et R. Thompson (1er déc. 1971). « Recovery of Inter-Block Information When Block Sizes Are Unequal ». In : *Biometrika* 58.3, p. 545-554. issn : 0006-3444. doi : 10.1093/biomet/58.3.545. url : https://doi.org/10.1093/biomet/58.3.545 (visité le 17/03/2024).

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

$$\begin{pmatrix} & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \\ Y_1 & 10 & a_{12} & a_{13} & a_{14} & a_{15} \\ Y_2 & a_{21} & 10 & a_{23} & a_{24} & a_{25} \\ Y_3 & a_{31} & a_{32} & 10 & a_{34} & a_{35} \\ Y_4 & a_{41} & a_{42} & a_{43} & 10 & a_{45} \\ Y_5 & a_{51} & a_{52} & a_{53} & a_{54} & 10 \end{pmatrix}$$