Benchmark of Differential Gene Expression Analysis Methods for Inter-species RNA-Seq Data using a Phylogenetic Simulation Framework

Paul Bastide[1,*], Charlotte Soneson[2,3], Olivier Lespinet[4], and Mélina Gallopin[4,*]

[1] *IMAG, Université de Montpellier, CNRS, 34000, Montpellier, France*
[2] *Friedrich Miescher Institute for Biomedical Research, 4058, Basel, Switzerland*
[3] *SIB Swiss Institute of Bioinformatics, 4058, Basel, Switzerland*
[4] *Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS, 91198, Gif-sur-Yvette, France*
[*] *Corresponding authors: paul.bastide@umontpellier.fr; melina.gallopin@universite-paris-saclay.fr.*

Abstract

Inter-species RNA-Seq datasets are increasingly common, and have the potential to answer new questions on gene expression patterns across the evolution. Single species differential expression analysis is a now well studied problem, that benefits from sound statistical methods. Extensive reviews on biological or synthetic datasets have provided the community with a clear picture on the relative performances of the available tools in various settings. Such benchmarks are still missing in the inter-species gene expression context. In this work, we take a first step in this direction by developing and implementing a new simulation framework. This tool builds on both the RNA-Seq and the Phylogenetic Comparative Methods literatures to generate realistic count datasets, while taking into account the phylogenetic relationships between the samples. We illustrate the features of this new framework through a targeted simulation study, that reveals some of the strengths and weaknesses of both the classical and phylogenetic approaches for inter-species differential expression analysis. The tool has been integrated in the R package compcodeR freely available on Bioconductor.

Keywords— RNA-Seq, Differential Gene Expression, Phylogenetic Comparative Methods, compcodeR, Orthologous Genes, Comparative Transcriptomic

Contents

1 Introduction

The study and analysis of gene expression differences across species is a long standing problem (King and Wilson, 1975). The development of microarray technologies led to the gathering of the first large scale and across species gene expression datasets, that allowed for the formulation and study of various hypotheses regarding the link between gene expression and evolution (Enard, 2002; Gilad et al., 2006; Khaitovich et al., 2004; Whitehead and Crawford, 2006). RNA-Sequencing technologies have changed the way to measure gene expression (Wang et al., 2009), making comparisons across several species easier, even for species with no reference genome available (Perry et al., 2012; Romero et al., 2012). As a compilation of a large number of datasets and atlases for gene expression of healthy wild-type individuals, the well maintained Bgee database (Bastian et al., 2021) is an important resource to ease the comparison of expression patterns across animal species.

Since changes in expression may underlie complex phenotypes, across species gene expression datasets can be used to test a wide range of evolutionary scenarios (Dunn et al., 2013; Romero et al., 2012). Tested hypotheses include for instance expression divergence (Gu, 2004); strength of expression conservation (Gu et al., 2019); coevolution of gene expression (Cope et al., 2020); test of the orthologous conjecture (Rogozin et al., 2014; Dunn et al., 2018); detection of the "phylogenetic signal" (Musser and Wagner, 2015); equality of within-species variance (Catalán et al., 2019); constant stabilizing selection, loss through drift, parallel or divergent selection (Stern and Crandall, 2018a,b); or detection of duplication-specific effects in expression evolution (Fukushima and Pollock, 2020).

In this review, we focus on the detection of change in gene expression levels across species, in a specific lineage or between different groups of species. This problem can be formalized as an inter-species differential expression analysis, and has been studied in various groups of organisms (Stern and Crandall, 2018b; Cáceres et al., 2003; Zheng-Bradley et al., 2010; Blake et al., 2018; Chen et al., 2019; Blake et al., 2020; Alam et al., 2020). For instance, difference in gene expression levels was found between mammalian lineages and birds (Brawand et al., 2011), across non-model primates species (Perry et al., 2012), between *Drosophila* species (Torres-Oliva et al., 2016) or *Heliconius* butterflies (Catalán et al., 2019). Note that the biological interpretation of changes in the level of expression of a gene across species is not easy (Romero et al., 2012). Shifts in gene expression across species could be molecular signatures of ecological adaptation, or associated with a directional selection scenario, or a relaxation of evolutionary constraints.

From a bioinformatic point of view, the comparison of RNA-Seq samples between multiple species requires, first, the detection of orthologous relationships between genes (Tatusov, 1997; Tekaia, 2016), second, the consideration of differences in genome mappability (Zhu et al., 2014) and, third, the adaptation of alignment and quantification pipelines (LoVerso and Cui, 2015; Chung et al., 2021). Multi-species alignments techniques have also been developed (Bradley et al., 2009; Brawand et al., 2011). In this review, we name orthologous genes (OG), or simply genes, the set of genes having orthologous relationship across species. Once the orthologous gene expression matrix has been created, the level of expression can be transformed into a discrete variable to detect presence versus absence of gene expression (Bastian et al., 2021). Other approaches perform separate differential expression for each species (Dunn et al., 2013; Kristiansson et al., 2013) or focus on pairwise comparisons only (Zhou et al., 2019; Chung et al., 2021). Direct comparisons of expression between species can be complicated by batch effects (Gilad and Mizrahi-Man, 2015), or potential confounding factors (Roux et al., 2015; Cope et al., 2020). Comparative gene expression studies should be carefully designed (Dunn et al., 2013; Romero et al., 2012; Chung et al., 2021). In this work, we focus our attention on genes having a one-to-one relationship across several species

(more than two species). We consider the level of expression of genes as a quantitative trait evolving across several species, and we detect genes with a shift in the level of expression across species as performed in e.g. Brawand et al. (2011); Perry et al. (2012); Torres-Oliva et al. (2016); Stern and Crandall (2018a). Since no statistical method is clearly established to perform this detection across multiple species, we present in the next section a review of all strategies used in practice.

There are several well-established tools to simulate RNA-Seq count data in the classical, intra-species case (Dillies et al., 2013; Soneson and Delorenzi, 2013; Soneson, 2014), which allowed for the benchmark of many differential expression analysis models (Anders and Huber, 2010; Robinson and Oshlack, 2010; Law et al., 2014). Although some methodological questions remain open (Van den Berge et al., 2019), these extensive simulation studies helped setting good practices in terms of model choice or normalization methods in various intra-species RNA-Seq settings. To our knowledge, there exists no extension of these frameworks to the inter-species setting. Simulation of gene expression across species has been performed using linear models and Gaussian variables (Rohlfs et al., 2014; Rohlfs and Nielsen, 2015; Gu et al., 2019), but without taking into account the specificity of RNA-Seq count data and without focusing on the detection of shifts across species. In this review, we propose a framework to simulate RNA-Seq data across species. We use this framework to compare different strategies to detect genes with a expression level shift across multiple species, and draw recommendations for inter-species gene expression comparison.

The paper is structured as follows: first, we describe the main methods used to perform differential analysis or shift detection across multiple species. We then explain our simulation method to generate synthetic inter-species RNA-Seq data using a Poisson log-normal model. Our simulation tool is integrated in the Bioconductor package compcodeR. Finally, a targeted simulation study, that draws its parameters from a recent inter-species RNA-Seq study (Stern and Crandall, 2018a), allows us to compare the current statistical methods, and propose some recommendations.

2 Review of methods used to compare level of expression across species.

2.1 Setting and Notation

For the remainder of this work, y_{gi} denotes the measured level of expression for gene g, $1 \leq g \leq p$, and sample i, $1 \leq i \leq n$. We assume that the species are partitioned into two groups S_1 and S_2, that depends on the biological question at hand. Each sample is associated to a species, and each species belongs to one of the two groups of interest.

Our goal is to detect genes with a shift in expression level across groups. To perform this test, we need to properly model the level of expression of gene g in sample i, taking into account the specificities of inter-species RNA-Seq data, that are multifold. Indeed, RNA-Seq data are counts, usually measured on a low number of samples. In addition, several technical biases affect the measured level of expression y_{gi}, either gene-specific (such as heterogeneity of gene length and GC content across genes and samples), or sample-specific (such as heterogeneity in library size across samples). Finally, since the level of expression of a gene g is measured across several species, the phylogenetic relationships between species induce some correlations in the data. While, ideally, all these specificities should be taken into account in the statistical analysis, to our knowledge there exist no model that includes all these constraints in its hypotheses. Below, we present an overview of the three main strategies adopted to model inter-species RNA-Seq data, that each make different simplifying assumptions.

We denote by m_i the sample specific normalization factor for sample i. Several approaches exist to compute this factor (Dillies et al., 2013), such as the Relative Log Expression (RLE) (Anders and Huber, 2010) method or the Trimmed Mean of M-values (TMM) (Robinson and Oshlack, 2010) method. We further denote by ℓ_{gi} the length of the gene g in sample i, which need to be taken into account as a gene and sample specific normalisation factor.

All the methods described below rely on a (generalized) linear model. The design (or model) matrix $\mathbf{X}$ of the experiment defines the form of this model. For differential analysis, it contains at least a grouping information, specifying which biological replicate belongs to S_1 or S_2. It can include some covariates that might influence the gene expression, such as information about environmental or experimental conditions. The matrix $\mathbf{X}$ has n rows, and as many columns as the number of coefficients in the model.

2.2 Strategy 1: Generalized Linear Model on Raw Count Data

The first option to perform differential expression analysis across species is to use a generalized linear model based on the negative binomial distribution (Anders and Huber, 2010; Robinson and Oshlack, 2010), implemented in several R packages such as DESeq2 or edgeR. In DESeq2 (Love et al., 2014), the random variable modeling the raw level of expression Y_{gi} of gene g in sample i is a negative binomial with expectation $\mu_{gi} = c_{gi}q_{gi}$ and dispersion α_g: $Y_{gi} \sim NB(\mu_{gi}, \alpha_g)$. The coefficient c_{gi} is a sample and gene specific normalization factor that depends on the sample specific normalization factor m_i and on the gene length ℓ_{gi}. The parameter q_{gi} is linked to the true level of expression of sample i, and includes the model design through the relationship $\log_2(q_{gi}) = \mathbf{X}_{i.}\boldsymbol{\theta}_g$, where $\mathbf{X}_{i.}$ denotes the i^{th} line of the design matrix $\mathbf{X}$, and the vector of coefficients $\boldsymbol{\theta}_g$ contains the information on the $\log_2$ fold changes between the two groups of species for gene g.

This method properly models counts and is appropriate to analyse data with low sample size thanks to dispersion shrinkage (Anders and Huber, 2010; Robinson and Oshlack, 2010). Sample specific and gene specific technical biases are taken into account directly into the parametrization of the model. Unfortunately, to our knowledge, this model is not flexible enough to account for the correlation induced by the phylogenetic tree. For this reason, this model is usually used to perform pairwise comparison between species (Torres-Oliva et al., 2016).

2.3 Normalization and Transformations

As we will see below, instead of using a generalized linear model on raw count data, it is possible to use a simple linear model on normalized data. The normalization step is essential to transform count measurements into continuous values, and to unlock the use of linear models. The normalization should be designed to temper the sample and gene specific technical biases, as well as to render the data homoscedastic (i.e. with homogeneous variance across samples).

Three main normalization scores are used in the literature. They all rely on the normalized library size M_i for sample i, defined as: $M_i = \sum_g y_{gi} m_i$, with m_i the scaling normalization factor described above. The Count Per Million (CPM) score incorporates sample-specific normalization only: $\text{CPM}_{gi} = \frac{y_{gi}}{M_i/10^6}$. The Reads (or fragments) per kilobase per million mapped reads (RPKM) score incorporates an extra gene-specific normalization as follow: $\text{RPKM}_{gi} = \frac{y_{gi}}{M_i/10^6 \times \ell_{gi}/10^3}$ (Mortazavi et al., 2008). Another way to include the same gene-specific normalisation is to use the Transcripts per million (TPM) score: $\text{TPM}_{gi} = \frac{y_{gi}/\ell_{gi}}{\sum_g y_{gi}/\ell_{gi}/10^6}$ (Wagner et al., 2012). Compared to the RPKM, the TPM scores summed over all genes are

equal to a constant (10^6), which is a property that can be desirable in some settings (Musser and Wagner, 2015).

In addition to the normalization, an extra transformation is often needed to make the data behave closer to a homoscedastic Gaussian. Two transformations are widely used: the $\log_2$ transformation (Law et al., 2014) and the square root transformation (Musser and Wagner, 2015).

For inter-species differential expression analysis, the choice of the right normalization and transformation to perform is not clearly established. Some studies use the $\log_2$-transformed RPKM (Mortazavi et al., 2008; Brawand et al., 2011; Catalán et al., 2019) or CPM (Blake et al., 2018) scores. Other studies advocates for the use of the $\log_{10}$ (Chen et al., 2019) or square-root (Musser and Wagner, 2015; Stern and Crandall, 2018a) transformed TPM.

In the remainder of this work, $\tilde{y}_{gi}$ denotes the normalized and transformed level of expression for gene g and sample i.

2.4 Strategy 2: Linear Model on Normalized Data

Assuming the data has been normalized and transformed properly, it can be modelled, for each gene g, using a simple linear regression:

$$\widetilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g, \tag{1}$$

where $\widetilde{\mathbf{Y}}_g$ is the vector of the n normalized measurements for gene g, $\mathbf{E}_g$ is a vector of Gaussian independent and identically distributed residuals, and, as previously, $\mathbf{X}$ is the design matrix and $\boldsymbol{\theta}_g$ the associated vector of coefficients. This model is implemented in the popular R package limma (Smyth, 2004; Smyth et al., 2005), that uses an empirical Bayes moderated statistic to test whether the coefficient of $\boldsymbol{\theta}_g$ associated with the group segregation is significantly different from zero. This method is appropriate to analyze datasets with low sample size, but a large number of genes that are pooled in a hierarchical model to get a better estimation of the variance.

It can be applied directly to RNA-Seq data, normalized using the previous methods. If the data presents mean-variance trends, which is typically the case in classical intra-species RNA-Seq data due to the presence of a high number of highly variable small counts, this can be taken into account through a weighting method (voom), or through the direct inclusion of the trend in the hierarchical empirical Bayes model (the trend method) (Law et al., 2014).

This method does not take the phylogenetic correlations into account, and has been used to performed pairwise comparisons (Blake et al., 2018, 2020; Torres-Oliva et al., 2016). This model is flexible and can be extended to a linear mixed model that accounts for the correlation between replicates of the same species (Breschi et al., 2016), using the duplicateCorrelation function from limma. However, correlations between species, encoded by the phylogenetic tree, cannot be directly taken into account using this approach.

2.5 Phylogenetic Comparative Methods

The methods described above are tailored for RNA-Seq data, but they are not designed to deal with the correlations introduced in the measurements by the phylogenetic relationships between the samples in an inter-species analysis. In this section, we briefly introduce Phylogenetic Comparative Methods, that have precisely been developed to deal with these correlations, before demonstrating some of their uses in the RNA-Seq literature.

Phylogenetic Comparative Methods. Phylogenetic relationships are known to induce correlations between observed quantitative traits on several species (Felsenstein, 1985). The field

of Phylogenetic Comparative Methods (PCMs) specializes in the comparative study of such phylogenetically related traits, and has been flowering over the last decades (see e.g. Harmon (2019) for a recent review). Conditionally on a phylogenetic tree that links a set of species, PCMs model the evolution of a quantitative trait as a stochastic process running along the branches of the tree (see Fig. 1). This generative model induces a multivariate Gaussian structure of the observed vector of traits across species, with a correlation structure that depends on the tree and on the chosen process. The values of the trait are only observed at the tips of the tree. The values at the root or at the internal nodes are unobserved and are modeled using latent variables.

Brownian Motion on a Tree. The most commonly used process is the Brownian Motion (BM) (Felsenstein, 1985). Under this model, for a given continuous trait $\mathbf{Z}^{'}$ measured at the tips of the tree, the covariance between traits $Z_i^{'}$ and $Z_j^{'}$ is simply proportional to the time of shared evolution between species i and j, i.e. the time t_{ij} between the root of the tree and the most recent common ancestor of i and j: $\mathbb{C}\text{ov}[Z_i^{'}; Z_j^{'}] = \sigma^2_{\text{BM}} t_{ij}$, where σ^2_{BM} is the variance of the BM process. The expectation of each trait is equal to μ, the ancestral value of the process at the root.

Ornstein-Uhlenbeck on a Tree. To model stabilizing selection, the Ornstein-Uhlenbeck (OU) process is often used (Hansen and Martins, 1996; Hansen, 1997). Compared to the BM, it has an equilibrium value β, that represents the "optimal value" of the trait in a given environment. The trait is attracted to this optimum with a speed that is controlled by the selection strength α, or better the phylogenetic half-life $t_{1/2} = \log(2)/\alpha$ (Hansen, 1997): when $t_{1/2}$ is large compared to the total height of the tree t ($t_{1/2} \gg t$), the trait needs a relatively long time to approach its optimum, and the selection strength is weak, while when $t_{1/2}$ is small compared to t ($t_{1/2} \ll t$), the selection strength is considered as strong. This process induces a different correlation structure than the Brownian motion, with stronger selection strength inducing weaker inter-species correlations (Hansen, 1997; Ho and Ané, 2013). Specifically, conditionally on a fixed root, $\mathbb{C}\text{ov}[Z_i^{'}; Z_j^{'}] = \gamma^2(1 - e^{-2\alpha t_{ij}})e^{-\alpha(t_i + t_j - 2t_{ij})}$, with $\gamma^2 = \sigma^2_{\text{OU}}/(2\alpha)$ the stationary variance of the process, and $t_i = t_{ii}$ the time between the root and node i (Ho and Ané, 2013).

Within-Species Variation. The traditional PCM framework assumes that only one measurement is available for each species, and that there is no measurement error, i.e. that all the observed variation can be explained by the evolution process on the tree. However, ignoring measurement error can lead to severe biases (Silvestro et al., 2015; Cooper et al., 2016). In addition, in an inter-species RNA-Seq differential analysis, it is usual to have access to replicated measurements, i.e. to measurements for several individuals of the same species. There is a vast literature on the subject of within-species variation (Grafen, 1989, 1992; Lynch, 1991; Housworth et al., 2004; Ives et al., 2007; Hadfield and Nakagawa, 2010; Goolsby et al., 2017). One simple way to look at the problem in a univariate setting is to assume that all the individuals from a same species are placed on the tree as tips linked to a same species node with a branch of length zero (Felsenstein, 2008) and to add a uniform Gaussian individual variance s^2 to all the tip samples traits (see Figures 1 and 2). In such a framework, the total variance of a sample trait Z_i attached to a latent tip with trait $Z^{'}_{\text{sp}(i)}$ is given by $\mathbb{V}\text{ar}[Z_i] = \mathbb{V}\text{ar}[Z^{'}_{\text{sp}(i)}] + s^2$, where $\mathbb{V}\text{ar}[Z^{'}_{\text{sp}(i)}]$ is determined by the chosen stochastic process to model the latent trait (BM or OU). Similarly, the covariance between two sample traits Z_i and Z_j attached, respectively, to latent tip traits $Z^{'}_{\text{sp}(i)}$ and $Z^{'}_{\text{sp}(j)}$ is given by $\mathbb{C}\text{ov}[Z_i; Z_j] = \mathbb{C}\text{ov}[Z^{'}_{\text{sp}(i)}; Z^{'}_{\text{sp}(j)}]$.

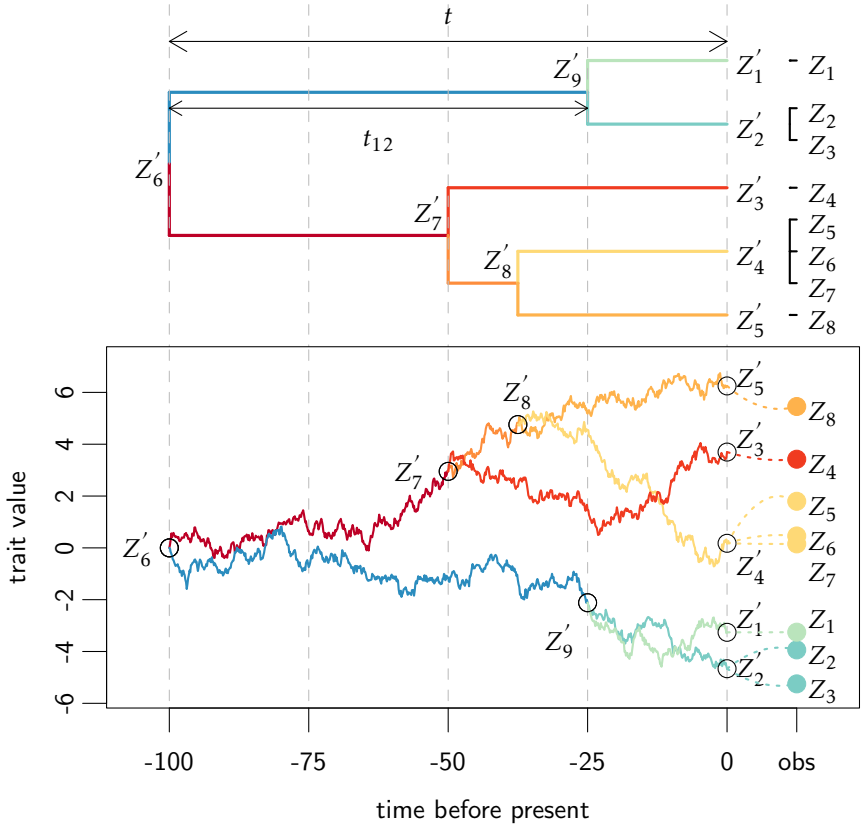

Figure 1: Realization of a Brownian Motion (BM) process (bottom), on a time calibrated ultrametric tree with total height $t = 100$ (top), with replicates and within-species variation. The BM process on the tree controls the distribution of the internal nodes, including ancestral nodes $Z'_6,\dots,Z'_9$, and latent tip traits $Z'_1,\dots,Z'_5$. The ancestral root value of the BM is $\mu = 0$, and its variance is $\sigma^2_{\text{BM}} = 0.1$, so that the latent (unobserved) tip trait variance is $\mathbb{V}\text{ar}[Z'_1] = \cdots = \mathbb{V}\text{ar}[Z'_5] = \sigma^2_{\text{BM}} t = 10$. The covariance of the latent tips trait is proportional to their time of shared evolution, for instance $\mathbb{C}\text{ov}[Z'_1; Z'_2] = \sigma^2_{\text{BM}} t_{12} = 7.5$. Replicated measurements are added on the tree as tips with zero branch lengths (top), with an extra variance of $s^2 = 0.5$. For instance, Z_2 and Z_3 are replicates of the latent tip Z'_2, and their conditional distribution is Gaussian with expectation Z'_2 and variance s^2. The total sample traits variance is hence given by $\mathbb{V}\text{ar}[Z_1] = \cdots = \mathbb{V}\text{ar}[Z_8] = \sigma^2_{\text{BM}} t + s^2 = 10.5$, and the sample traits covariance is given by the tree structure, for instance $\mathbb{C}\text{ov}[Z_1; Z_2] = \mathbb{C}\text{ov}[Z'_1; Z'_2] = \sigma^2_{\text{BM}} t_{12} = 7.5$, and $\mathbb{C}\text{ov}[Z_2; Z_3] = \mathbb{C}\text{ov}[Z'_2; Z'_2] = \sigma^2_{\text{BM}} t = 10$. Note that on this figure, latent internal nodes (internal and external) are numbered from 1 to 9, and observations are numbered from 1 to 8, but these set of indices are distinct. For instance, Z_1 is indeed an observation of Z'_1, but Z_4 is an observation of Z'_3 and is unrelated to Z'_4.

2.6 Strategy 3: Phylogenetic Regression on Normalized Data

One way to include the phylogenetic structure with within-species variation, in statistical analyses is to use a Phylogenetic Mixed Model (PMM (Grafen, 1989, 1992; Lynch, 1991; Housworth et al., 2004)), where the vector $\widetilde{\mathbf{Y}}_g$ of the n normalized and transformed measurement for a given gene g is seen as the sum of a fixed effect, a random phylogenetic effect, and a random independent effect:

$$\widetilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g^{\text{phy}} + \mathbf{E}_g^{\text{iid}}, \tag{2}$$

with $\mathbf{X}$ and $\boldsymbol{\theta}_g$ the design matrix and associated vector of coefficients as in Eq. (1), $\mathbf{E}_g^{\text{phy}}$ a vector of phylogenetically correlated residuals, with correlations given by the chosen process

on the tree (see above) and $\mathbf{E}_g^{\text{iid}}$ independent and identically distributed (iid) residuals, that can capture any non-phylogenetic source of variation of the data, such as within-species variation as described above.

Several methods for gene expression analysis based on models related to the PCM framework have been described in the literature, with different versions of the BM or the OU process, and with or without within-species variation (Khaitovich et al., 2004; Gu, 2004; Gu and Su, 2007; Bedford and Hartl, 2009; Rohlfs et al., 2014; Rohlfs and Nielsen, 2015; Gu et al., 2019), and in particular have been used to detect differences in gene expression across species (Brawand et al., 2011; Rohlfs et al., 2014; Rohlfs and Nielsen, 2015; Stern and Crandall, 2018a; Catalán et al., 2019; Chen et al., 2019).

For differential expression analysis, the *phylogenetic ANOVA* framework (Garland et al., 1993; Grafen, 1989; Rohlfs and Nielsen, 2015; Bastide et al., 2018) is particularly relevant, and can just be seen as the phylogenetic regression above, with the design matrix $\mathbf{X}$ encoding groups of species. This framework is for instance implemented in the popular and computationally efficient R package phylolm (Ho and Ané, 2014a).

3 Probabilistic Models and Data Simulation

Building on existing RNA-Seq methods (Robles et al., 2012; Soneson and Delorenzi, 2013; Soneson, 2014), we developed a new inter-species simulation framework that can generate realistic count datasets, and takes into account, first, the gene expression correlations induced by the phylogeny and, second, the different lengths a given gene can have in different species.

3.1 Realistic Simulations using the Negative Binomial Distribution

We briefly recall here the simulation framework detailed in (Soneson and Delorenzi, 2013), and implemented in compcodeR (Soneson, 2014).

Negative Binomial Distribution. Let Y_{gi} be the random variable representing the count for gene g $(1 \leq g \leq p)$ in sample i $(1 \leq i \leq n)$, with true expression level λ_{gi} and sampling depth M_i. Following Robinson and Oshlack (2010), we model each count independently by a Negative Binomial (NB) distribution with expectation μ_{gi} and dispersion α_g, such that $Y_{gi} \sim \mathsf{NB}\left(\mu_{gi}, \alpha_g\right)$ with:

$$\mu_{gi} = \frac{\lambda_{gi}}{\sum_{h=1}^{p} \lambda_{hi}} M_i. \tag{3}$$

Differential Expression. To model differential expression, we assume that the samples are partitioned into two groups S_1 and S_2. For each gene g, the dispersion parameter α_g is the same for all samples, while the expression level λ_{gi} can only take two values: λ_{gS_1} if i is in S_1 and λ_{gS_2} if i is in S_2. Given λ_{gS_1}, we take λ_{gS_2} as:

$$\lambda_{gS_2} = \begin{cases} \lambda_{gS_1} & \text{if } g \text{ is not differentially expressed;} \\ \lambda_{gS_1} \times (e + X_g^e) & \text{if } g \text{ is up-regulated in } S_2; \\ \lambda_{gS_1} \times (e + X_g^e)^{-1} & \text{if } g \text{ is down-regulated in } S_2; \end{cases}$$

with e the minimal differential effect size, and X_g^e random variables independent identically distributed according to an exponential distribution with parameter 1. The values of the parameters are set to match the empirical counts expectation and dispersion of a real datasets.

3.2 Realistic Simulations using the Poisson Log-Normal Distribution

The Poisson Log-Normal (PLN) distribution has been advocated as an alternative to the NB distribution for the analysis of RNA-Seq data. Being more flexible, it is particularly well suited in the presence of correlations (Gallopin et al., 2013; Zhang et al., 2015; Choi et al., 2017), which proves essential for inter-specific datasets, as demonstrated in the next section. We show here how the parameters of a PLN model can be chosen to match first and second order moments of the NB model described above, making it possible to simulate realistic datasets under this more flexible framework.

The PLN Distribution. Under the PLN model, for each gene g and sample i, we assume that the observed count random variable Y_{gi} follows a Poisson distribution, with log parameter a Gaussian latent variable Z_{gi}, such that:

$$\begin{aligned} Z_{gi} &\sim \mathcal{N}\left(m_{gi}, \sigma_g^2\right) \\ Y_{gi} \,\middle|\, Z_{gi} &\sim \mathcal{P}\left(\exp(Z_{gi})\right). \end{aligned} \tag{4}$$

This model is similar in spirit to the NB distribution, that can be seen as Gamma-Poisson mixture (see e.g. Holmes and Huber, 2019, Chap. 4). Note that in both models the coefficient of variation of the mixing distribution is constant across samples for a given gene (Chen et al., 2014).

Matching Moments. Using standard moments expressions for the NB (Holmes and Huber, 2019) and PLN (Aitchison and Ho, 1989) distributions, it is straightforward to show that a PLN distribution with parameters m_{gi} and σ_g^2 yields the same first and second order moments as a NB distribution with expectation μ_{gi} and dispersion α_g if and only if:

$$\begin{cases} \sigma_g^2 = \log(1 + \alpha_g) \\ m_{gi} = \log(\mu_{gi}) - \dfrac{1}{2}\log(1 + \alpha_g). \end{cases} \tag{5}$$

These equations allow us to readily use the framework developed in the previous section also in the case of a PLN simulation.

3.3 Taking the Phylogeny into Account with the Phylogenetic Poisson Log-Normal Distribution

In an inter-specific framework, various samples come from various species, which implies a specific correlation between measures, that can be taken into account in a multivariate PLN model, as shown below.

Continuous Trait Evolution Model. The models of trait evolution used in PCMs and presented in the previous section are generative, and can be used to simulate continuous traits at the tips of a tree (with possible replicates) such that their correlation structure is consistent with their phylogeny (see Fig. 1). Using a simple uniform Gaussian individual variance s_g^2 to model within-species variation, the trait variance Σ_g for the vector $\mathbf{Z}_g$ of continuous traits at the tips of the tree generated by such a process can be expressed as:

$$\begin{cases} [\Sigma_g]_{ij} = \mathbb{C}\mathrm{ov}\left[Z_i; Z_j\right] = \sigma_g^2(\mathrm{sp}(i); \mathrm{sp}(j)) & \text{if } \mathrm{sp}(i) \neq \mathrm{sp}(j), \\ [\Sigma_g]_{ii} = \mathbb{V}\mathrm{ar}\left[Z_i\right] = \sigma_g^2(\mathrm{sp}(i); \mathrm{sp}(i)) + s_g^2 & \text{otherwise,} \end{cases}$$

where $\sigma_g^2(\mathsf{sp}(i);\mathsf{sp}(j))$ is the phylogenetic variance between species $\mathsf{sp}(i)$ and $\mathsf{sp}(j)$ of samples i and j (see Fig. 1), with a structure given by the evolution process (BM or OU, see expressions above), and s_g^2 the added intra-species variation.

The Phylogenetic Poisson Log-Normal Distribution. The models described above are well suited for quantitative traits, but need to be adapted for count measures, such as the one produced by a RNA-Seq analysis. To handle such counts, we propose to add a Poisson layer to the trait evolution models described above, defining a "phylogenetic" Poisson Log-Normal (pPLN) distribution. More specifically, for a given gene g, we simulate a vector of n latent traits $\mathbf{Z}_g$ as the result of such a process running on the tree, and then, conditionally on this vector, draw the observed counts Y_{gi} from a Poisson distribution with parameter $\exp(Z_{gi})$:

$$\begin{aligned} \mathbf{Z}_g &\sim \mathcal{N}\big(\mathbf{m}_g, \Sigma_g\big) \\ Y_{gi} \,\big|\, Z_{gi} &\sim \mathcal{P}\big(\exp(Z_{gi})\big). \end{aligned} \tag{6}$$

In other words, the vector of counts $\mathbf{Y}_g$ for each gene is drawn from a multivariate Poisson Log-Normal distribution, with parameters $\mathbf{m}_g$ and Σ_g obtained from the evolutionary models described above, Σ_g being the structured variance matrix of both phylogenetic and independent effects, and $\mathbf{m}_g$ a vector or expectations values at the tips, that can be set independently from the process.

Matching Moments for Realistic Simulations. Assuming that the diagonal coefficients of Σ_g are all equal to a single value σ_g^2, Equation (5) can be used to ensure that the pPLN model above yields the same marginal expectation and variance as a NB model with expectation μ_{gi} and dispersion α_g. At a macro-evolutionary scale, most of the dated phylogenetic trees encountered are ultrametric, i.e. are such that all the tips are at the same distance t from the root. In that case, all the phylogenetic models described above verify this variance homogeneity assumption. For instance, for the simple BM model with an extra layer of independent variation, we have $\sigma_g^2 = \sigma_{\mathrm{BM}}^2 t + s^2$. Note that, although the NB and pPLN models are set to have the same expectations and variance, they differ significantly in their covariances: while in the standard NB model, all the samples are independent from one another, in the proposed pPLN framework the measurements are correlated, with a structure reflecting both the tree and the selected evolutionary process.

3.4 Taking Differential Gene Lengths into Account

Length Normalisation of Counts. Let ℓ_{gi} denote the length of the gene g for sample i. Following Robinson and Oshlack (2010), we take this length into account by changing Equation (3) to:

$$\mu_{gi} = \frac{\lambda_{gi}\ell_{gi}}{\sum_{h=1}^{p}\lambda_{hi}\ell_{gi}} M_i. \tag{7}$$

Note that the same overall sequencing depth M_i is attributed to each sample, but that, because of the weighted average, it is preferentially allocated to longer genes.

Lengths Simulation. The lengths are simulated according to the pPLN model described above, with expectations and dispersions empirically estimated from the dataset at hand.

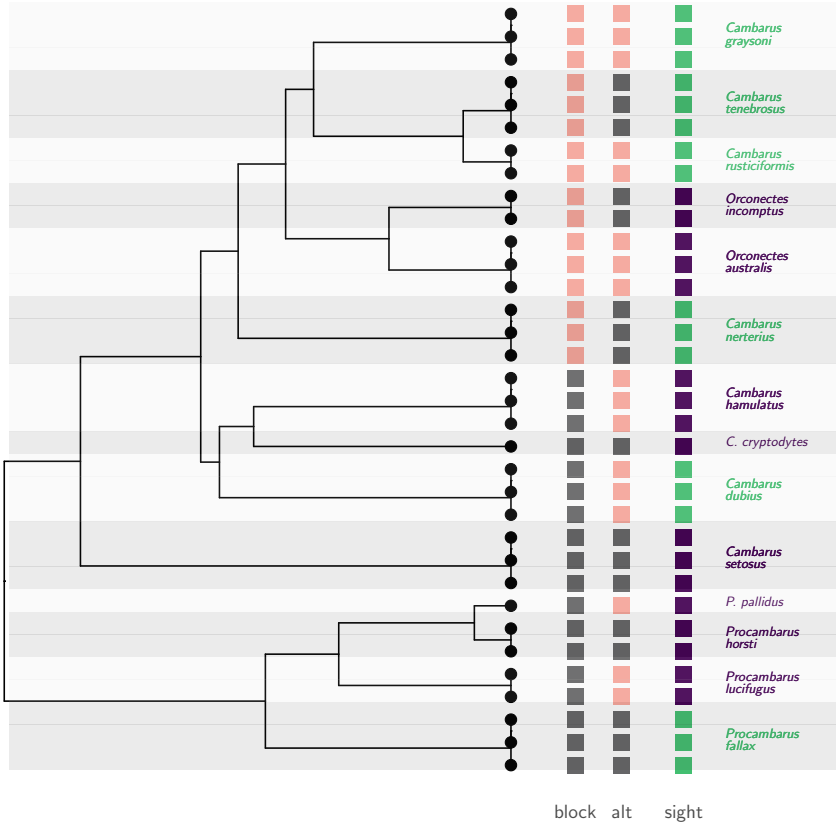

Figure 2: Time-calibrated phylogenetic tree of 8 blind (dark purple) and 6 sighted (light green) crayfish species (Stern et al., 2017). The root was dated to 65 million years before the present (Stern et al., 2017), but the tree was re-scaled to unit height for the analyses. The "sight" design (dark purple and light green squares) matches with the biological vision status of the species studied (Stern and Crandall, 2018a). The "block" and "alt" designs (light pink and gray squares) are artificial extreme scenarios representing, respectively, a situation where the design is almost un-distinguishable from the phylogeny-induced grouping ("block"), and a situation where groups are distributed evenly on the tree to maximize the contrast between sister species ("alt").

4 Simulation Studies

4.1 Material and Methods

Gene Expression Underlying Vision Loss in Cave Animals. We used our new simulation framework to generate realistic synthetic datasets, that were set to mimic the features of a recently published inter-species RNA-Seq dataset (Stern and Crandall, 2018a), while varying the level of evolutionary dependence. In this study (Stern and Crandall, 2018a), the authors analyzed the molecular mechanisms involved in vision loss in the North American family *Cambaridae* of crayfish species. They selected 8 blind and 6 sighted crayfish species, for which a time-calibrated maximum likelihood phylogeny is known (Stern et al., 2017). 3560 orthologous gene expressions were estimated using the method RNA-Seq by Expectation Maximization (RSEM) (Li and Dewey, 2011), with one to three replicates per species (see Fig. 2).

Base Simulation Parameters. Following the methodology described in the previous section, we simulated a "base scenario" dataset using the estimated crayfish tree re-scaled to unit height ($t = 1$), with the observed vision status design ("sight" design, see Fig. 2), and matching the empirical counts and gene lengths expectation and dispersion. The expression level

λ_{gS_1} and the dispersion α_g were estimated from the dataset for each gene g, while for each sample i the simulation sequencing depth M_i was independently drawn from a uniform distribution with bounds $M_{\min}$ and $M_{\max}$ the observed empirical minimal and maximal values of the library size across all samples. We used a BM model of trait evolution, with an independent layer of individual variation s_g^2 representing 20% of the total tip variance σ_g^2 for each gene g: $s_g^2 = 0.2 \times \sigma_g^2$, with $\sigma_g^2 = (\sigma_{\text{BM}}^2)_g t + s_g^2$. We chose a base effect size of 3, with 150 differentially expressed genes out of the 3560 simulated ones. From this base scenario, we varied several parameters in order to study their impacts on the simulated data. Each scenario was replicated 50 times.

Star Tree and NB Simulations. To check that our new pPLN framework produced datasets with properties similar to the well known NB framework, we replaced the crayfish tree with a star-tree, that mimics the NB situation where all species and replicates are independent.

Tree Group Design. The group design on the tree is known to strongly impact the properties of the data, in particular through its "phylogenetic effective sample size" (Ané, 2008; Bartoszek, 2016). To study its effect in a gene expression context, we replaced the "sight" design with a "block" and "alt" design (see Fig. 2), that were chosen to model two extreme situations. In the "block" design, all the species with a given group are nested within a single clade, so that the differential expression signal is redundant with the phylogenetic signal. At the other end of the spectrum, the "alt" design was chosen so that sister species are in different groups, in order to maximize the contrast between organisms that share a long common history. We expect the "alt" design to produce datasets with a stronger signal.

Differential Analysis Phylogenetic Asymptotic Effective Sample Size. To quantify the intrinsic difficulty of a design compared to another, we propose a new "differential analysis phylogenetic asymptotic effective sample size" (dapaESS). Given a phylogenetic tree $\mathcal{T}$, we first remove all replicates, so that there are no zero-length branches. Then, given a design vector $\mathbf{x}$, we postulate a simple BM model for an hypothetical continuous trait $\mathbf{y}$ at the tips: $\mathbf{y} = \theta_0 \mathbf{1} + \theta_1 \mathbf{x} + \sigma \mathbf{e}^{BM}$, with $\mathbb{V}\text{ar}\left[\mathbf{e}^{BM}\right] = \mathbf{V}^{\text{tree}} = [t_{ij}]_{i,j}$. From standard linear model theory, the variance of the maximum likelihood estimator of the coefficient θ_1 is given by (Ané, 2008): $\mathbb{V}\text{ar}\left[\hat{\theta}_1\right] = \sigma^2 (\mathbf{X}^T \mathbf{V}^{\text{tree}^{-1}} \mathbf{X})^{-1}_{2,2}$, with $\mathbf{X} = (\mathbf{1}\ \mathbf{x})$ the matrix of predictors. We hence define: $\text{dapaESS}(\mathcal{T}, \mathbf{x}) = 1/(\mathbf{X}^T \mathbf{V}^{\text{tree}^{-1}} \mathbf{X})^{-1}_{2,2}$. In the case where all the species are independent (star-tree $\mathcal{T}^*$), we fall back on a standard differential expression analysis, and we get, assuming that there are n species and that the groups are balanced: $\text{dapaESS}(\mathcal{T}^*, \mathbf{x}) = n/4$, which is the standard effective sample size for a balanced two-sample t-test with uniform variance. This gives us a base-line for a "standard" difficulty, and we use in the following the normalized dapaESS: $\text{dapaESSn}(\mathcal{T}, \mathbf{x}) = \text{dapaESS}(\mathcal{T}, \mathbf{x}) / \text{dapaESS}(\mathcal{T}^*, \mathbf{x})$. A value lower than 1 indicates a design that is deemed more difficult than a standard independent design (larger asymptotic variance of the estimator), while a value greater that 1 indicates a problem where the phylogeny actually helps in finding the significant differences. Note that this score can be computed *a priori*, and, as shown below, can be used to asses the quality of the experimental design.

Simulation Process. The simulation process impacts the tree induced correlation between species (Blomberg et al., 2003; Harmon, 2019). To study the impacts of this modeling choice, we replaced the BM process with an OU, with a phylogenetic half-life (Hansen, 1997) $t_{1/2} = \log(2)/\alpha$ fixed equal to 50% of the tree height.

Within-Species Variation Level. We mitigated the effect of the BM model on the tree by varying the level of the independent individual variation representing s_g^2, from 40% to 0% (i.e., all the measurements from a same species are perfectly correlated).

Inference Methods Used. We chose the following statistical inference methods, representing the three main approaches presented above: DESeq2 (Love et al., 2014) assumes a NB distribution on independent counts; limma (Ritchie et al., 2015) applies an Empirical Bayes moderation (without a mean-variance trend correction, unless otherwise specified) on independent normalized counts, possibly assuming that all the samples in a same species are correlated (limma cor (Smyth et al., 2005)); and phylolm (Ho and Ané, 2014a) uses a phylogenetic regression framework based on a BM or OU process, with measurement error. For phylolm, the differential analysis relied on a t statistic computed for each gene independently, conditionally on the estimated maximum likelihood parameters (s_g^2 and α_g for the OU). The raw p-values computed by all methods were adjusted using the BH method (Benjamini and Hochberg, 1995), using the R function p.adjust. Inferred gene expression differences across groups were marked as significant if their associated adjusted p-value was below the threshold of 0.05.

Length Normalisation and Transformation. In DESeq2 (Love et al., 2014), we used the default RLE method (Anders and Huber, 2010) to compute the sample-specific normalization factor m_i. We followed the recommendations of the section "Sample-/gene-dependent normalization factors" from the DESeq2 vignette to compute the coefficients c_{gi} from the coefficients m_i and gene lengths ℓ_{gi} detailed in section 2. For methods requiring a preprocessing normalization of the count data (limma and phylolm), we used the TMM method (Robinson and Oshlack, 2010) implemented in the calcNormFactor function in edgeR, and a TPM length normalization with a $\log_2$ transformation. We studied the effect of these choices by testing combinations of other normalization methods (RPKM length normalization; or a simple CPM, i.e. no length normalization), and an other transformation function (square root, as advocated in Musser and Wagner (2015)).

Scores Used. To assess the performance of the inference methods, based on the list of true (simulated) differentially expressed genes, we computed the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). We used the Matthews correlation coefficient (MCC = $[\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}]\cdot[(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})]^{-1/2}$) as advised in Chicco and Jurman (2020). We also computed the True Positive Rate (TPR = $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$) and the False Discovery Rate (FDR = $\mathrm{FP}/(\mathrm{FP}+\mathrm{TP})$). In addition, we compared the features of the simulated datasets with the empirical one using the countsimQC R package (Soneson and Robinson, 2018).

4.2 Results

PLN and NB Simulation Frameworks Produce Similar Datasets. When parametrized to produce the same moments, the pPLN framework on a star tree produces datasets that are similar in difficulty to the classical NB framework (Fig. 3, first two columns). While limma controls the FDR to the nominal rate, DESeq2 fails to control the FDR in this case with a lot of variance (empirical dispersion range from 0.1 to 5, see also Fig. 5). As it assumes a NB distribution of the counts, DESeq2 suffers from the deviation from this model, as opposed to limma, which performs equally well in both cases. As showed by the countsimQC analysis, the datasets simulated with the pPLN and the NB frameworks have similar features, and are comparable to the original empirical dataset (data not shown, comparison report

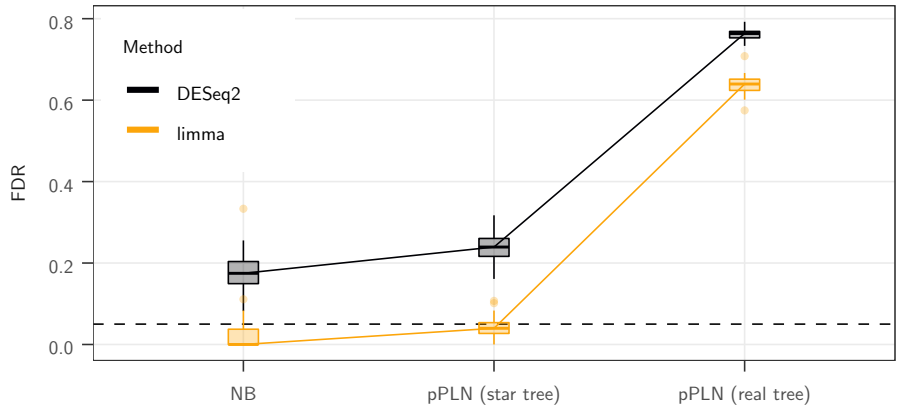

Figure 3: The base scenario (pPLN (real tree), right) has empirical moments drawn from (Stern and Crandall, 2018a), with an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree, with the observed sight groups (see Fig. 2). It is compared to a pPLN model with the same parameters, but in a case where all samples are independent (pPLN (star tree), middle), and to a NB model with the same moments and effect size (NB, left). The DESeq2 (black) and limma (light orange) inference methods are applied to each scenarios, and their FDR is compared. The black dashed line represents the nominal rate of 5% used to call positives. For limma, the counts are normalized using $\log_2$(TPM) values. Boxplots are on 50 replicates.

available on the GitHub repository `github.com/i2bc/InterspeciesDE`). While preserving the univariate moments, the tree included in the framework (Fig. 3, last column), introduces some phylogenetic correlations between the species, and leads to a spectacular loss of power of both methods, with a rate of false discoveries higher than three quarters.

Phylogenetic Data Requires Correlation Modeling. For data simulated according to the base scenario, methods that explicitly model sample correlations (limma cor and phylolm) perform best (Fig. 4, dark purple line). limma cor exhibits the best behavior with the highest MCC, and a TPR reaching about 80%. Its FDR is still above the nominal rate (median around 10%).

Tree group Design Matters. The alt designs produces datasets with the clearest signal, (Fig. 4, light orange line). In this case, limma cor is able to correctly control for the FDR. Although phylolm methods have slightly higher FDR, they achieve a better TPR reaching almost 100%, leading to a better overall MCC score. At the opposite of the spectrum, the block designs produces datasets with a very weak signal, with differentially expressed genes counts M-A values strongly overlapping a very diffuse non-differentially expressed genes distribution (Fig. 5). All methods applied to the block design have FDR higher or equal to about 50% (Fig. 4, black line). The BM phylolm tool has the least bad MCC score (about 0.5), although with the worst TPR (around 50%). The relative difficulties of each design is correctly captured by the normalized dapaESS. While the block design has a lower dapaESS than the independent case (dapaESSn = 0.69), the alt design has a higher one (5.1), and the sight design lies in the middle (1.4).

OU Makes the Signal Weaker and is Hard to Correct For. When simulating the counts using an OU model of trait evolution for the latent trait instead of a BM, the signal becomes weaker, and all methods achieve lower MCC scores (Fig. 6). The limma cor methods performs the best in this case, even when compared to a phylolm method that explicitly takes the OU

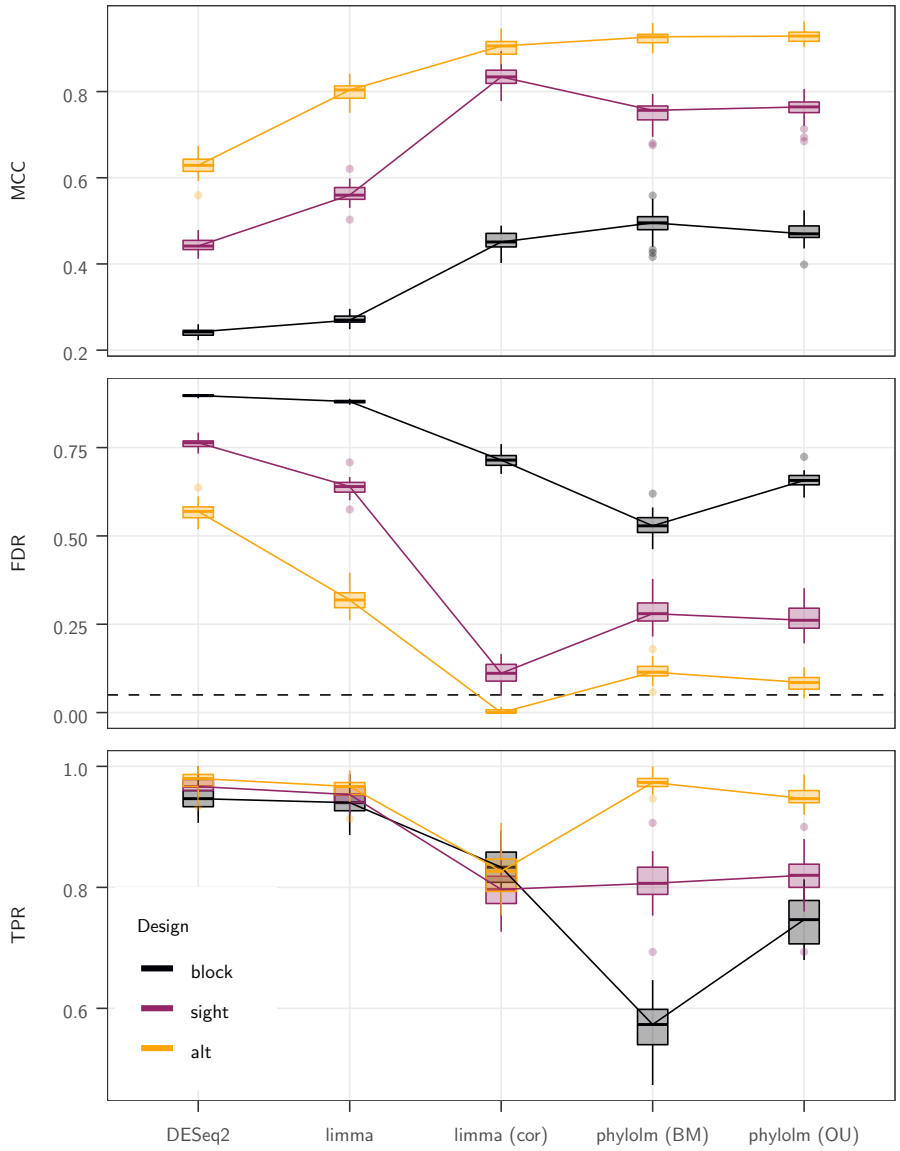

Figure 4: Results in term of MCC (top), FDR (middle) and TPR (bottom) scores of the five selected statistical methods (x axis) on the pPLN base scenario, that has an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree (Stern and Crandall, 2018a), with the observed sight groups (dark purple line, see Fig. 2). The alt (light orange line) and block (black line) groups are also tested, with the same parameters. For the FDR, the black dashed line represents the nominal rate of 5% used to call positives. When required, the counts are normalized using $\log_2$(TPM) values. Boxplots are on 50 replicates.

model into account.

Phylogenetic Methods are Robust to Intra-Species Variations. When reducing the intra-specific variance to 0 (inducing a correlation of 1 between sample values of the same species), the limma cor method loses its advantage compared to the phylolm methods, which performances are less affected by the level of intra-specific noise (Fig. 7).

$\log_2$(TPM) **Normalisation is Slightly Better on Phylogenetic Data.** Taking gene lengths into account, using either TPM or RPKM, significantly improves the power of the methods, in particular in term of TPR (Fig. 8). Although TPM normalization leads to a slightly better

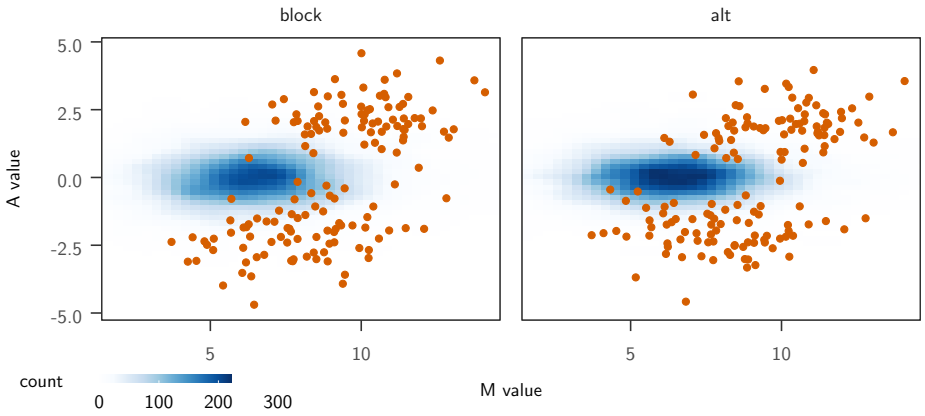

Figure 5: M-A plots (log2 fold change as a function of the mean of normalized counts for each gene) of the datasets produced with base pPLN parameters (effect size of 3, BM model with added intra-species variation accounting for 20% of the total variance), on the maximum likelihood tree (Stern and Crandall, 2018a), with the block (left) and alt (right) designs. The M-A values distribution for the 3410 non-differentially expressed genes is shown as a tile plot, with deeper blues representing high probability values. The M-A values of the 150 differentially expressed genes are shown as red dots.

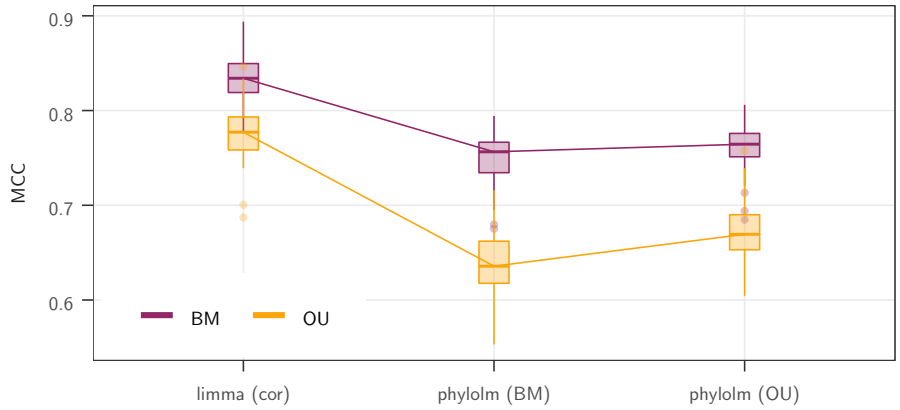

Figure 6: Results in term of MCC scores of the three correlation aware statistical methods (x axis) on the pPLN base scenario (effect size of 3, intra-species variation accounting for 20% of the total variance), with a BM (dark purple line) or an OU (light orange line) model of evolution on the maximum likelihood tree (Stern and Crandall, 2018a), with the observed sight groups (see Fig. 2). The counts are normalized using $\log_2$(TPM) values. Boxplots are on 50 replicates.

MCC median, its performances are largely similar to the RPKM normalization. On this base scenario, the $\log_2$ transformation leads to a consistent gain of about 10% in TPR compared to the square root (going from around 70% to 80%, Fig. 8).

No Small Counts in De Novo Assembled Data. Including a mean-variance trend correction in the limma cor method did not change its performance on the base scenario, producing very similar MCC values (the median MCC on all 50 runs differ by less than 0.002). This is consistent with the fact that the original dataset uses *de novo* assembled data, that naturally exclude any small counts, and hence the need for a mean-variance trend correction (see Discussion).

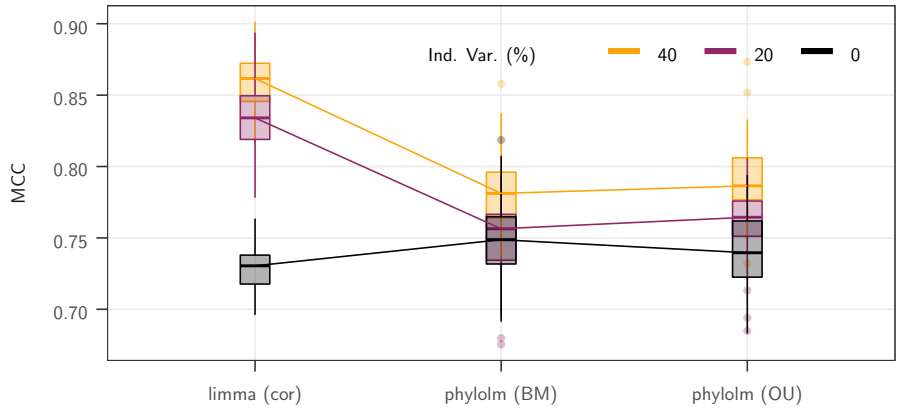

Figure 7: Results in term of MCC scores of the three correlation aware statistical methods (x axis) on the pPLN base scenario with an effect size of 3, a BM model of evolution on the maximum likelihood tree (Stern and Crandall, 2018a), with the observed sight groups, and intra-species variation accounting for 40% (light orange line), 20% (dark purple line), or 0% (black line) of the total variance). The counts are normalized using $\log_2$(TPM) values. Boxplots are on 50 replicates.

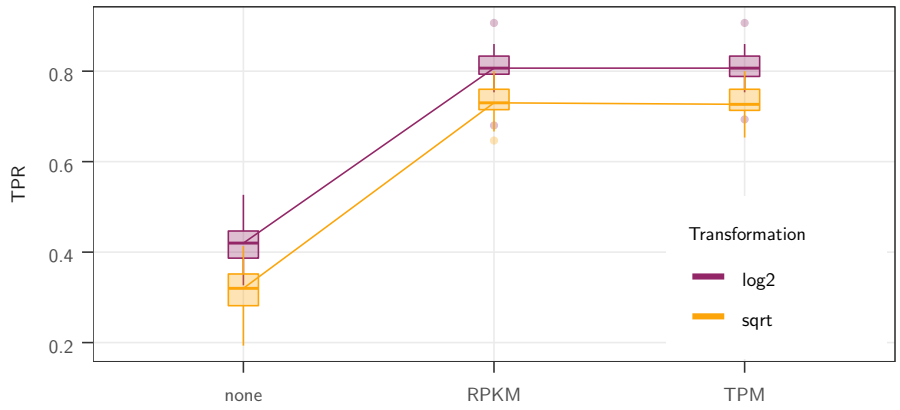

Figure 8: Results in term of TPR score of the phylolm (BM) method on the pPLN base scenario (effect size of 3, BM model of evolution on the maximum likelihood tree (Stern and Crandall, 2018a), with the observed sight groups, and added intra-species variation accounting for 20% of the total variance). The counts are length-normalized (x axis) using CPM (length not taken into account, none), RPKM or TPM, and transformed using the square root (light orange) or the $\log_2$ (dark purple) functions. Boxplots are on 50 replicates.

5 Discussion and Conclusion

5.1 Simulation Study

Our targeted simulation study illustrates some of the specificities of inter-species RNA-Seq differential expression analysis. First, it is essential to take the correlation between replicates within a given species into account. Failure to do so leads to very high rates of false discoveries (Fig. 4), that make the analysis unreliable and hard to exploit. Indeed, the limma method with added correlation seems to outperform other tools, including phylogenetic comparative methods, in many settings. These results tend to indicate that, even if the full tree is not included in the analysis, incorporating these simple correlations between replicates might be sufficient to efficiently analyse inter-species datasets, at least for some simulation designs. The group design on the tree was indeed found to be extremely important (Fig. 4). A balanced design, were the groups are evenly spread over all clades, has a stronger signal (Fig.

5), and allows the analysis to be abstracted from the phylogeny to some extent, as classical tools for differential expression analysis work best in this configuration. On the other hand, when the groups are clustered in the phylogeny, the signal is weaker as it becomes more difficult to distinguish the real group effect from the simple drift that tends to isolate clades from one another. This is in particular the case of designs where one clade or species is tested against out-groups, that is sometimes encountered in the literature (Brawand et al., 2011; Rohlfs and Nielsen, 2015). In this configuration, phylogenetic comparative methods, although imperfect, are essential. Finally, this study confirms the importance of length normalisation for inter-species differential gene expression analysis to achieve acceptable power detection levels (Fig. 8). Although we did not find any significant difference in performance between RPKM and TPM normalizations, the $\log_2$ transformation seemed to have a slight advantage over the square root in this simulation setting.

5.2 Simulation Design

In this work, we proposed a method to simulate RNA-Seq gene expression across multiple species. Similar to intra-species simulation tools (Dillies et al., 2013; Soneson and Delorenzi, 2013; Soneson, 2014), our simulation method can use empirical datasets to set the value of parameters such that the simulated datasets are as close as possible to the real ones, with matching empirical marginal expectation and variance. When applied to independent species, it produces datasets with comparable features (Fig. 2). In our specific simulation studies, we use the dataset from Stern and Crandall (2018a). This dataset was obtained using *de novo* assembled data. In addition, we focused on genes with one-to-one orthologuous relationships across species. As a consequence, this dataset had a low number of zeros and small counts, and a large variance across samples. The simulated datasets had similar characteristics, which could explain the low performance of DESeq2, even when the data was simulated without correlation (Fig. 2), and the fact that the trend procedure did not add any power to the limma method. Inter-species RNA-Seq gene expression datasets are very diverse, with specificities depending on the underlying biological question being studied. This work provides a first step toward realistic simulation of such datasets.

5.3 Simulation Tool

Compared to classical intra-species simulation tools (Dillies et al., 2013; Soneson and Delorenzi, 2013; Soneson, 2014), our simulation framework incorporates the species tree and the gene length, which may vary across species. It makes it possible to model the evolution of gene expression on the tree using two different processes (BM or OU), and it allows for additional independent variation, that can model e.g. inter-specific variation or measurement error. This complex model leads to new effects, that can be difficult to predict. In particular, we showed that the distribution of the groups on the tree had strong effects on the ability of all methods to detect a group expression shift. We proposed a normalized criterion (dapaESS) to assess the difficulty of the group design for the differential gene expression analysis problem. Although it does not take into account the number of replicates or the specific evolution model, we showed that it could well represent the difficulty of an experimental design. The strength of this criterion is that it only depends on the timed species tree and the tips group allocation, and can be computed before any statistical inference or even data collection. It can hence be used as a practical guide on the expected power of the experimental design. In this review, we focused our attention of the detection of shifts of expression between groups spanning across species. However, inter-species datasets are also used to address many other questions, such as equality of within-species variance, expression divergence, or detection of neutral versus directed evolution regimes. Several tools from the

PCM literature have been used to this end, that rely on various models of trait evolution with appropriate parameter constraints. Since our simulation tool is modular, those various processes could be implemented, in order to produce realistic RNA-Seq datasets with the desired structure. Such an extended framework could help researchers to test the statistical properties of these complex inference models.

5.4 Inference Tools

In this study, we focused on a few inference tools, that come either from the RNA-Seq or the PCM literature, limiting ourselves to methods implemented in R and that can do differential analysis. Although a more comprehensive simulation design would be needed to draw stronger conclusions, our results show that simulations under the OU model lead to more difficult datasets, and that even methods that include the OU model in their framework fail to completely correct for this effect. This could be linked with the fact that the estimation of the selection strength in an OU model is a notoriously difficult question, especially on an ultrametric tree (Ho and Ané, 2014b; Cooper et al., 2016). Having to estimate this parameter for thousands of genes is bound to generate some instability, and to deteriorate the performance of those tools. Gu et al. (2019) recently proposed an empirical Bayes approach to deal with this parameter in an RNA-Seq setting. One possible direction could be to adapt this method to a differential analysis problem. More generally, our simulation studies illustrate the need for new statistical tools for inter-species differential analysis, that would combine the strengths of both the classical RNA-Seq literature, that can deal with the specificities of this noisy data, and the PCM literature, that takes into account the phylogeny, an information that can be crucial to correctly interpret inter-species data.

6 Key Points

- Inter-species RNA-Seq datasets have a complex structure, and require a dedicated simulation tool that can generate count data with phylogeny induced correlations and that can take varying gene lengths into account.
- Differential analysis for inter-species RNA-Seq data requires a tool that can take at least within-species sample correlations into account and an adequate length normalisation procedure.
- The experimental design of the group allocation on the phylogeny has a strong impact on the differential expression signal, and is well captured by the dapaESS score, that can be computed *a priori* before any statistical analysis or data collection.

7 Data and Code Availability

The simulation tool is integrated into the compcodeR package, that is freely available on the Bioconductor platform, and documented through a specific vignette (`doi.org/10.18129/B9.bioc.compcodeR`). The data and code used for the simulation study are available on the following GitHub repository: `github.com/i2bc/InterspeciesDE`.

8 Acknowledgments

We thank David Stern for sharing his insights on the dataset, and for giving us access to the raw count data. P.B. and M.G are grateful to Sylvain Merlot for initiating this project, to

Marie-Laure Martin and Guillem Rigaill for useful discussions, and to Claire Ducos, Marie Michel and Sarah Jelassi for their work during their master internship. This work was partly funded by the I2BC and the MI CNRS through the MODELCOG (M.G.) and X-TrEM projects (Sylvain Merlot). We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources.

References

Aitchison, J. and Ho, C. H. (1989). The Multivariate Poisson-log Normal Distribution. *Biometrika*, 76(4):643–653.

Alam, T., Agrawal, S., Severin, J., Young, R. S., Andersson, R., Arner, E., Hasegawa, A., Lizio, M., Ramilowski, J. A., Abugessaisa, I., Ishizu, Y., Noma, S., Tarui, H., Taylor, M. S., Lassmann, T., Itoh, M., Kasukawa, T., Kawaji, H., Marchionni, L., Sheng, G., R.R. Forrest, A., Khachigian, L. M., Hayashizaki, Y., Carninci, P., and de Hoon, M. J. (2020). Comparative transcriptomics of primary cells in vertebrates. *Genome Research*, 30(7):951–961.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Ané, C. (2008). Analysis of comparative data with hierarchical autocorrelation. *The Annals of Applied Statistics*, 2(3):1078–1102.

Bartoszek, K. (2016). Phylogenetic effective sample size. *Journal of Theoretical Biology*, 407:371–386.

Bastian, F. B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S., de Farias, T. M., Moretti, S., Parmentier, G., de Laval, V. R., Rosikiewicz, M., Wollbrett, J., Echchiki, A., Escoriza, A., Gharib, W. H., Gonzales-Porta, M., Jarosz, Y., Laurenczy, B., Moret, P., Person, E., Roelli, P., Sanjeev, K., Seppey, M., and Robinson-Rechavi, M. (2021). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, 49(D1):D831–D847.

Bastide, P., Solís-Lemus, C., Kriebel, R., Sparks, K. W., and Ané, C. (2018). Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic Biology*, 67(5):800–820.

Bedford, T. and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences*, 106(4):1133–1138.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Blake, L. E., Roux, J., Hernando-Herraez, I., Banovich, N. E., Perez, R. G., Hsiao, C. J., Eres, I., Cuevas, C., Marques-Bonet, T., and Gilad, Y. (2020). A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Research*, 30(2):250–262.

Blake, L. E., Thomas, S. M., Blischak, J. D., Hsiao, C. J., Chavarria, C., Myrthil, M., Gilad, Y., and Pavlovic, B. J. (2018). A comparative study of endoderm differentiation in humans and chimpanzees. *Genome Biology*, 19(1):162.

Blomberg, S. P., Garland, T., and Ives, A. R. (2003). Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile. *Evolution*, 57(4):717–745.

Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., and Pachter, L. (2009). Fast statistical alignment. *PLoS Computational Biology*, 5(5):e1000392.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Breschi, A., Djebali, S., Gillis, J., Pervouchine, D. D., Dobin, A., Davis, C. A., Gingeras, T. R., and Guigó, R. (2016). Gene-specific patterns of expression variation across organs and species. *Genome Biology*, 17(1):151.

Cáceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., Lockhart, D. J., Preuss, T. M., and Barlow, C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22):13030–13035.

Catalán, A., Briscoe, A. D., and Höhna, S. (2019). Drift and Directional Selection Are the Evolutionary Forces Driving Gene Expression Divergence in Eye and Brain Tissue of Heliconius Butterflies. *Genetics*, 213(2):581–594.

Chen, J., Swofford, R., Johnson, J., Cummings, B. B., Rogel, N., Lindblad-Toh, K., Haerty, W., Di Palma, F., and Regev, A. (2019). A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research*, 29(1):53–63.

Chen, Y., Lun, A. T. L., and Smyth, G. K. (2014). Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. In Datta, S. and Nettleton, D., editors, *Statistical Analysis of Next Generation Sequencing Data*, pages 51–74. Springer International Publishing, Cham.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13.

Choi, Y., Coram, M., Peng, J., and Tang, H. (2017). A poisson log-normal model for constructing gene covariation network using RNA-seq data. *Journal of Computational Biology*, 24(7):721–731.

Chung, M., Bruno, V. M., Rasko, D. A., Cuomo, C. A., Muñoz, J. F., Livny, J., Shetty, A. C., Mahurkar, A., and Dunning Hotopp, J. C. (2021). Best practices on the differential expression analysis of multi-species RNA-seq. *Genome Biology*, 22(1):121.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., and Freckleton, R. P. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, 118(1):64–77.

Cope, A. L., O'Meara, B. C., and Gilchrist, M. A. (2020). Gene expression of functionally-related genes coevolves across fungal species: detecting coevolution of gene expression using phylogenetic comparative methods. *BMC Genomics*, 21(1):370.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., and Jaffrezic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683.

Dunn, C. W., Luo, X., and Wu, Z. (2013). Phylogenetic Analysis of Gene Expression. *Integrative and Comparative Biology*, 53(5):847–856.

Dunn, C. W., Zapata, F., Munro, C., Siebert, S., and Hejnol, A. (2018). Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences*, 115(3):E409–E417.

Enard, W. (2002). Intra- and Interspecific Variation in Primate Gene Expression Patterns. *Science*, 296(5566):340–343.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15.

Felsenstein, J. (2008). Comparative Methods with Sampling Error and Within-Species Variation: Contrasts Revisited and Revised. *The American Naturalist*, 171(6):713–725.

Fukushima, K. and Pollock, D. D. (2020). Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. *Nature Communications*, 11(1):4459.

Gallopin, M., Rau, A., and Jaffrézic, F. (2013). A Hierarchical Poisson Log-Normal Model for Network Inference from RNA Sequencing Data. *PLoS ONE*, 8(10).

Garland, T., Dickerman, A. W., Janis, C. M., and Jones, J. A. (1993). Phylogenetic analysis of covariance by computer simulation. *Systematic Biology*, 42(3):265–292.

Gilad, Y. and Mizrahi-Man, O. (2015). A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research*, 4(May):121.

Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P., and White, K. P. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, 440(7081):242–245.

Goolsby, E. W., Bruggeman, J., and Ané, C. (2017). Rphylopars : fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, 8(1):22–27.

Grafen, A. (1989). The Phylogenetic Regression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 326(1233):119–157.

Grafen, A. (1992). The uniqueness of the phylogenetic regression. *Journal of Theoretical Biology*, 156(4):405–423.

Gu, X. (2004). Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics*, 167(1):531–542.

Gu, X., Ruan, H., and Yang, J. (2019). Estimating the strength of expression conservation from high throughput RNA-seq data. *Bioinformatics*, 35(23):5030–5038.

Gu, X. and Su, Z. (2007). Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences*, 104(8):2779–2784.

Hadfield, J. D. and Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, 23(3):494–508.

Hansen, T. F. (1997). Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution*, 51(5):1341.

Hansen, T. F. and Martins, E. P. (1996). Translating Between Microevolutionary Process and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data. *Evolution*, 50(4):1404.

Harmon, L. J. (2019). *Phylogenetic Comparative Methods: Learning From Trees*. Center for Open Science, version 1. edition.

Ho, L. S. T. and Ané, C. (2013). Asymptotic theory with hierarchical autocorrelation: Ornstein?Uhlenbeck tree models. *The Annals of Statistics*, 41(2):957–981.

Ho, L. S. T. and Ané, C. (2014a). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic Biology*, 63(3):397–408.

Ho, L. S. T. and Ané, C. (2014b). Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*, 5(11):1133–1146.

Holmes, S. and Huber, W. (2019). *Modern Statistics for Modern Biology*. Cambridge University Press, Cambridge.

Housworth, E. a., Martins, E. P., and Lynch, M. (2004). The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96.

Ives, A. R., Midford, P. E., Garland, T., and Oakley, T. (2007). Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods. *Systematic Biology*, 56(2):252–270.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Pääbo, S. (2004). A Neutral Model of Transcriptome Evolution. *PLoS Biology*, 2(5):e132.

King, M. and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Kristiansson, E., Österlund, T., Gunnarsson, L., Arne, G., Larsson, D. G. J., and Nerman, O. (2013). A novel method for cross-species gene expression analysis. *BMC Bioinformatics*, 14(1).

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

LoVerso, P. R. and Cui, F. (2015). A computational pipeline for cross-species analysis of RNA-seq data using r and bioconductor. *Bioinformatics and Biology Insights*, 9:BBI.S30884.

Lynch, M. (1991). Methods for the Analysis of Comparative Data in Evolutionary Biology. *Evolution*, 45(5):1065–1080.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628.

Musser, J. M. and Wagner, G. P. (2015). Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called "species signal". *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(7):588–604.

Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., Michelini, K., Zehr, S., Yoder, A. D., Stephens, M., Pritchard, J. K., and Gilad, Y. (2012). Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*, 22(4):602–610.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., and Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing. *BMC Genomics*, 13(1):484.

Rogozin, I. B., Managadze, D., Shabalina, S. A., and Koonin, E. V. (2014). Gene Family Level Comparative Analysis of Gene Expression in Mammals Validates the Ortholog Conjecture. *Genome Biology and Evolution*, 6(4):754–762.

Rohlfs, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation. *Molecular Biology and Evolution*, 31(1):201–211.

Rohlfs, R. V. and Nielsen, R. (2015). Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Systematic Biology*, 64(5):695–708.

Romero, I. G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516.

Roux, J., Rosikiewicz, M., and Robinson-Rechavi, M. (2015). What to compare and how: Comparative transcriptomics for Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(4):372–382.

Silvestro, D., Kostikova, A., Litsios, G., Pearman, P. B., and Salamin, N. (2015). Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution*, 6(3):340–346.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Smyth, G. K., Michaud, J., and Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075.

Soneson, C. (2014). compcodeR–an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*, 30(17):2517–2518.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91.

Soneson, C. and Robinson, M. D. (2018). Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics*, 34(4):691–692.

Stern, D. B., Breinholt, J., Pedraza-Lara, C., LÃ³pez-MejÃa, M., Owen, C. L., Bracken-Grissom, H., Fetzner, J. W., and Crandall, K. A. (2017). Phylogenetic evidence from freshwater crayfishes that cave adaptation is not an evolutionary dead-end. *Evolution*, 71(10):2522–2532.

Stern, D. B. and Crandall, K. A. (2018a). The evolution of gene expression underlying vision loss in cave animals. *Molecular Biology and Evolution*, 35(8):2005–2014.

Stern, D. B. and Crandall, K. A. (2018b). Phototransduction gene expression and evolution in cave and surface crayfishes. *Integrative and Comparative Biology*, 58(3):398–410.

Tatusov, R. L. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637.

Tekaia, F. (2016). Inferring orthologs: Open questions and perspectives. *Genomics Insights*, 9:GEI.S37925.

Torres-Oliva, M., Almudi, I., McGregor, A. P., and Posnien, N. (2016). A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. *BMC Genomics*, 17(1).

Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., Patro, R., and Robinson, M. D. (2019). RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annual Review of Biomedical Data Science*, 2(1):139–173.

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Whitehead, A. and Crawford, D. L. (2006). Variation within and among species in gene expression: raw material for evolution. *Molecular Ecology*, 15(5):1197–1211.

Zhang, H., Xu, J., Jiang, N., Hu, X., and Luo, Z. (2015). PLNseq: a multivariate poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. *Statistics in Medicine*, 34(9):1577–1589.

Zheng-Bradley, X., Rung, J., Parkinson, H., and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11(12):R124.

Zhou, Y., Zhu, J., Tong, T., Wang, J., Lin, B., and Zhang, J. (2019). A statistical normalization method and differential expression analysis for RNA-seq data between different species. *BMC Bioinformatics*, 20(1):163.

Zhu, Y., Li, M., Sousa, A. M., and Šestan, N. (2014). XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. *BMC Genomics*, 15(1):343.