Supplementary Material for: A Phylogenetic Framework to Simulate Synthetic Inter-species RNA-Seq Data

Paul Bastide[1,*], Charlotte Soneson[2,3], David B. Stern[4,5], Olivier Lespinet[6], and Mélina Gallopin[6,*]

[1] *IMAG, Université de Montpellier, CNRS, Montpellier, France*
[2] *Friedrich Miescher Institute for Biomedical Research, 4058, Basel, Switzerland*
[3] *SIB Swiss Institute of Bioinformatics, 4058, Basel, Switzerland*
[4] *Department of Integrative Biology, 430 Lincoln Drive, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.*
[5] *National Biodefense Analysis and Countermeasures Center (NBACC), Operated by Battelle National Biodefense Institute (BNBI) for the U.S. Department of Homeland Security Science and Technology Directorate, Fort Detrick, MD 21702, USA*
[6] *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.*

Contents

A Review of Methods Used to Compare Level of Expression Across Species.

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.

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.

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).

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: $\mathrm{CPM}_{gi} = \frac{y_{gi}}{M_i/10^6}$. The Reads (or fragments) per kilobase per million mapped reads (RPKM) score incorporates an extra gene-specific normalization as follow: $\mathrm{RPKM}_{gi} = \frac{y_{gi}}{M_i/10^6 \times \ell_{gi}/10^3}$ (Mortazavi *et al.*, 2008). Another way to include the same gene-specific normalisation is to use the Transcripts per million (TPM) score: $\mathrm{TPM}_{gi} = \frac{y_{gi}/\ell_{gi}}{\sum_g y_{gi}/\ell_{gi}/10^6}$ (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, 2018) 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.

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.

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 (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 Section Phylogenetic Comparative Methods) 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, 2018; 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).

B Clade Specific Differential Expression Analysis

Using the linear model, we defined the design matrix to distinguish between 4 groups: sighted species (reference group); blind *Procambarus* species (*pallidus*, *horsti* and *lucifugus*); blind *Cambarus* species (*setosus*, *cryptodytes* and *hamulatus*); and blind *Orconectes* species (*australis* and *incomptus*). Using the same procedure as above, we tested with limma cor for the coefficients associated to each of the three genus of blind species. We report below genes with an adjusted p-value below 0.05.

Only one gene was found differentially expressed in all three groups: OG0002505, with Uniprot top hit XYLA ARATH, name "Xylose isomerase". In addition, two more genes were common to groups *Procambarus* and *Cambarus* only: OG0000233 (RTBS DROME, "Probable RNA-directed DNA polymerase from transposon BS") and OG0000606 (LIN1 NYCCO, "LINE-1 reverse transcriptase homolog"); and two other genes were common to groups *Orconectes* and *Cambarus* only: OG0001105 (PIPA DROME, "1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase") and OG0001281 (OPSD PROCL, "Rhodopsin"). Among these 5 genes, only one (OG0000606) was not found differentially expressed in the first analysis where all the blind species where merged into one single group (see Table 1).

It is not surprising for de-novo assembled transcriptomes of invertebrate species to have a high percentage of genes that do not match any protein in reference databases (Stern and Crandall, 2018), and function identification often comes from distantly related species, which can make it difficult to interpret. In the differentially expressed genes presented above, there were only 5 orthogroups that did not have a significant hit in the Uniprot/Swiss-Prot database. This small percentage could be partly explained by the fact that only highly expressed genes found across all the species were considered in this analysis.

Table S1: Differentially expressed genes across blind *Procambarus* and all sighted Crayfish species found by the limma cor method on $\log_2$ transformed TPM values.

Orthogroup	adj. p-value	Uniprot top hit	Protein name
OG0002505	1.6e-04	XYLA ARATH	Xylose isomerase
OG0000383	1.3e-02	POL3 DROME	Retrovirus-related Pol polyprotein from transposon 17.6
OG0001902	1.3e-02	EF1A BOMMO	Elongation factor 1-alpha
OG0000013	1.6e-02	YI31B YEAST	Transposon Ty3-I Gag-Pol polyprotein
OG0009115	1.6e-02	TMED7 RAT	Transmembrane emp24 domain-containing protein 7
OG0008389	2.2e-02	U389 DROPS	UPF0389 protein GA21628
OG0000233	2.4e-02	RTBS DROME	Probable RNA-directed DNA polymerase from transposon BS
OG0001073	2.4e-02	POL3 DROME	Retrovirus-related Pol polyprotein from transposon 17.6
OG0003669	2.4e-02	GP107 MOUSE	Protein GPR107
OG0000625	4.2e-02	TF29 SCHPO	Transposon Tf2-9 polyprotein
OG0008424	4.3e-02	RL28 SPOFR	60S ribosomal protein L28
OG0008907	4.5e-02	RL44 OCHTR	60S ribosomal protein L44
OG0000606	4.7e-02	LIN1 NYCCO	LINE-1 reverse transcriptase homolog
OG0000165	5.0e-02	RTJK DROFU	RNA-directed DNA polymerase from mobile element jockey

Table S2: Differentially expressed genes across blind *Orconectes* and all sighted Crayfish species found by the limma cor method on $\log_2$ transformed TPM values.

Orthogroup	adj. p-value	Uniprot top hit	Protein name
OG0004279	1.5e-06	RL22 CAEEL	NA
OG0002505	1.5e-06	XYLA ARATH	Xylose isomerase
OG0007419	2.4e-03	RS14 PROCL	40S ribosomal protein S14
OG0007471	5.7e-03	AN32A DROME	Acidic leucine-rich nuclear phosphoprotein 32 family member A
OG0001081	6.4e-03	MDC1 MACMU	Mediator of DNA damage checkpoint protein 1
OG0001105	6.6e-03	PIPA DROME	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase
OG0002942	8.1e-03	K0100 HUMAN	Protein KIAA0100
OG0000083	1.0e-02	NA	NA
OG0000885	1.2e-02	NA	NA
OG0005282	1.2e-02	ALF DROME	Fructose-bisphosphate aldolase
OG0007052	1.3e-02	PGK DROME	Phosphoglycerate kinase
OG0000698	2.3e-02	LIAT1 HUMAN	Protein LIAT1
OG0005096	2.3e-02	E2AK4 RAT	eIF-2-alpha kinase GCN2
OG0009037	3.1e-02	RS3A BIPLU	40S ribosomal protein S3a
OG0000069	3.3e-02	DNJC5 DROME	DnaJ homolog subfamily C member 5 homolog
OG0001175	3.3e-02	EAA3 RABIT	Excitatory amino acid transporter 3
OG0002503	3.4e-02	TPRGL MOUSE	Tumor protein p63-regulated gene 1-like protein
OG0005913	3.4e-02	EIF3B HUMAN	Eukaryotic translation initiation factor 3 subunit B
OG0002556	3.8e-02	NFKB2 XENLA	Nuclear factor NF-kappa-B p100 subunit
OG0003284	3.8e-02	WDR48 CHICK	WD repeat-containing protein 48
OG0004914	3.8e-02	CDN1B HUMAN	Cyclin-dependent kinase inhibitor 1B
OG0001281	4.3e-02	OPSD PROCL	Rhodopsin
OG0001546	4.4e-02	ZBT49 HUMAN	Zinc finger and BTB domain-containing protein 49
OG0001058	4.8e-02	RTJK DROME	RNA-directed DNA polymerase from mobile element jockey

Table S3: Differentially expressed genes across blind *Cambarus* and all sighted Crayfish species found by the limma cor method on $\log_2$ transformed TPM values.

Orthogroup	adj. p-value	Uniprot top hit	Protein name
OG0002505	2.7e-07	XYLA ARATH	Xylose isomerase
OG0008934	4.9e-05	RS30 ORYLA	40S ribosomal protein S30
OG0000233	6.0e-03	RTBS DROME	Probable RNA-directed DNA polymerase from transposon BS
OG0002370	6.0e-03	ARRH LOCMI	Arrestin homolog
OG0000346	1.8e-02	JERKY HUMAN	Jerky protein homolog
OG0009062	1.8e-02	RL10 BOMMA	60S ribosomal protein L10
OG0009021	2.1e-02	NDKA PONAB	Nucleoside diphosphate kinase A
OG0001281	2.6e-02	OPSD PROCL	Rhodopsin
OG0001008	3.0e-02	RTJK DROME	RNA-directed DNA polymerase from mobile element jockey
OG0007035	3.8e-02	RL19 DROME	60S ribosomal protein L19
OG0000606	3.9e-02	LIN1 NYCCO	LINE-1 reverse transcriptase homolog
OG0001279	3.9e-02	ARRH HELVI	Arrestin homolog
OG0001750	3.9e-02	RTN4 MOUSE	Reticulon-4
OG0005269	3.9e-02	TCPA DROME	T-complex protein 1 subunit alpha
OG0009626	3.9e-02	RL18 TIMBA	60S ribosomal protein L18
OG0009633	3.9e-02	RL6 CHILA	60S ribosomal protein L6
OG0009642	3.9e-02	NACA DROME	Nascent polypeptide-associated complex subunit alpha
OG0000081	4.1e-02	NA	NA
OG0001678	4.1e-02	GNAQ HOMAM	Guanine nucleotide-binding protein G(q) subunit alpha
OG0004085	4.1e-02	RS11 RAT	40S ribosomal protein S11
OG0007915	4.1e-02	H5 CHICK	General transcription factor IIH subunit 5
OG0000025	4.2e-02	DPOL HHBV	Protein P
OG0007885	4.2e-02	NA	NA
OG0001105	4.6e-02	PIPA DROME	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase
OG0000017	4.8e-02	NA	NA
OG0003210	4.8e-02	ELOA1 DROME	Transcription elongation factor B polypeptide 3
OG0003970	4.8e-02	GTD2A HUMAN	General transcription factor II-I repeat domain-containing protein 2A
OG0005931	4.8e-02	7B2 HUMAN	Cytochrome c oxidase subunit 7B2, mitochondrial
OG0006265	4.8e-02	AT1B1 ARTSF	Sodium/potassium-transporting ATPase subunit beta
OG0009074	4.8e-02	PFD1 DANRE	Prefoldin subunit 1

Finally, for comparison, Table S4 presents the list of genes found when applying the limma cor method on the $\log_2$ transformed RPKM values, instead of TPM, for the global analysis. This table is to be compared with Table 1 from the main text.

Table S4: Differentially expressed genes across blind and sighted Crayfish species found by the limma cor method on $\log_2$ transformed RPKM values (adjusted p-values below 0.05).

Orthogroup	adj. p-value	Uniprot top hit	Protein name
OG0002505	9.2e-08	XYLA ARATH	Xylose isomerase
OG0000233	9.5e-03	RTBS DROME	Probable RNA-directed DNA polymerase from transposon BS
OG0000083	4.9e-02	NA	NA
OG0000383	4.9e-02	POL3 DROME	Retrovirus-related Pol polyprotein from transposon 17.6

C Supplementary Simulation Figures

In this Supplementary Section, we present the same results as in the Result section of the main text, but showing for each plot all the same three scores (MCC, FDR and TPR) for a uniform presentation.

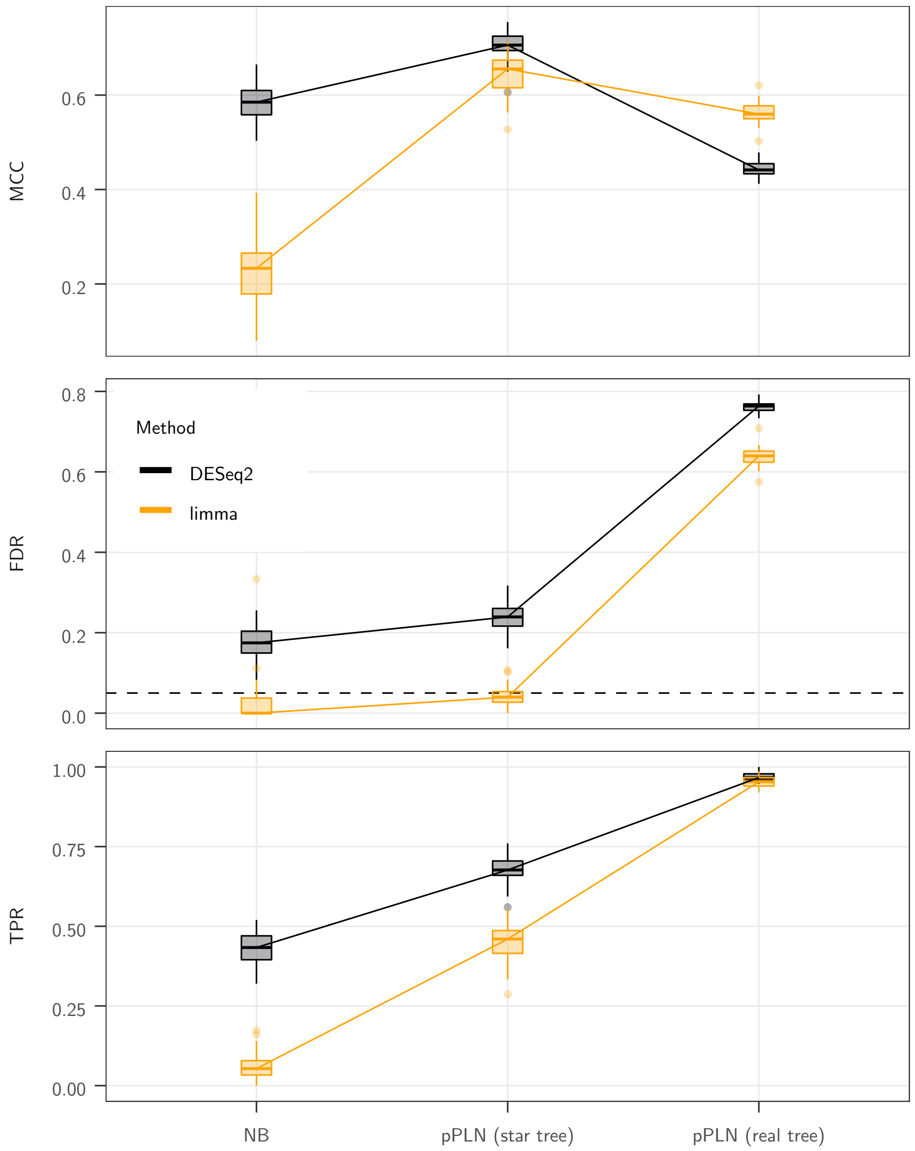

Figure S1: The base scenario (pPLN (real tree), right) had empirical moments drawn from (Stern and Crandall, 2018), with an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree, with the observed "sight" groups (see Fig. 1). It is compared to a pPLN model with the same parameters, but in a case where all samples were 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 were applied to each scenarios.The black dashed line represents the nominal rate of 5% used to call positives. For limma, the counts were normalized using $\log_2$(TPM) values. Boxplots are based on 50 replicates.

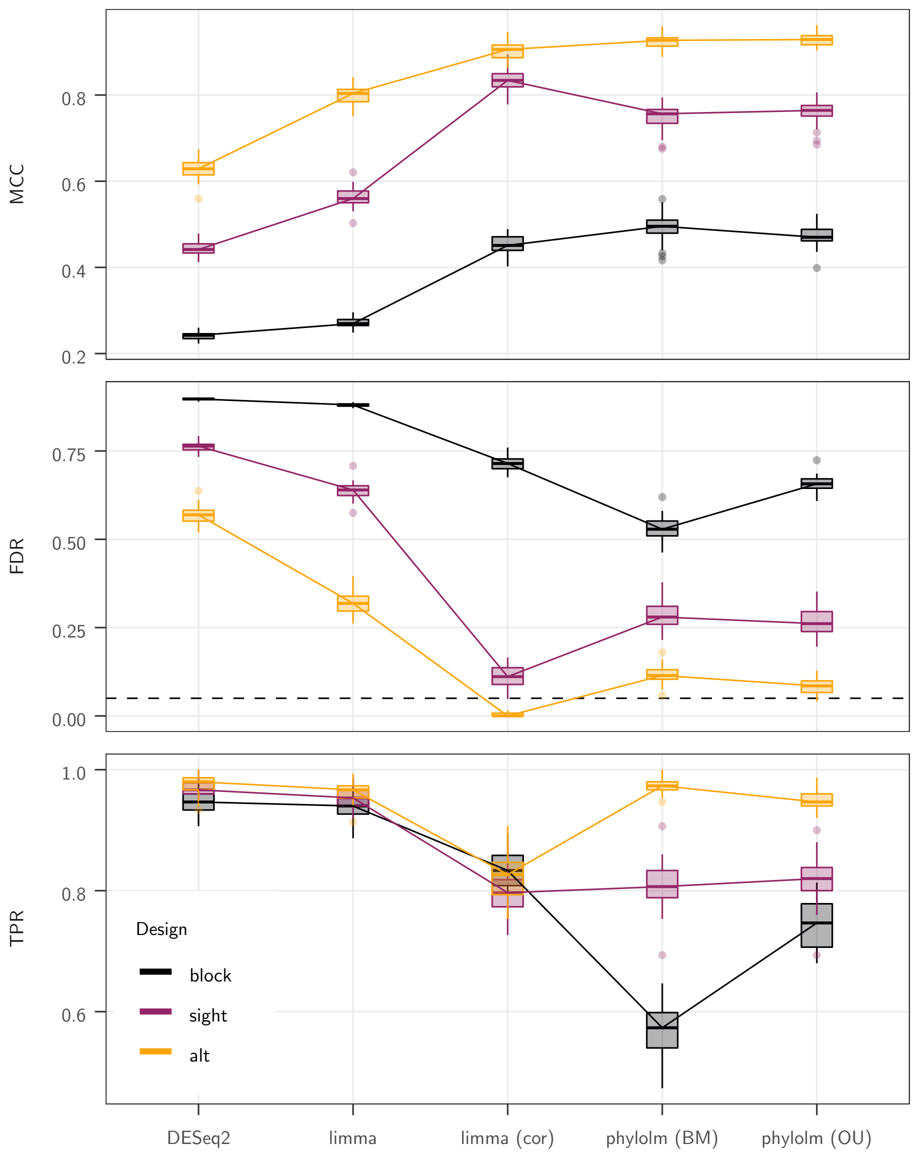

Figure S2: Results in terms of MCC (top), FDR (middle) and TPR (bottom) scores of the five selected statistical methods (x axis) on the pPLN base scenario, that has an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree (Stern and Crandall, 2018), with the observed "sight" groups (dark purple line, see Fig. 1). The "alt" (light orange line) and "block" (black line) groups were 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 were normalized using $\log_2$(TPM) values. Boxplots are based on 50 replicates.

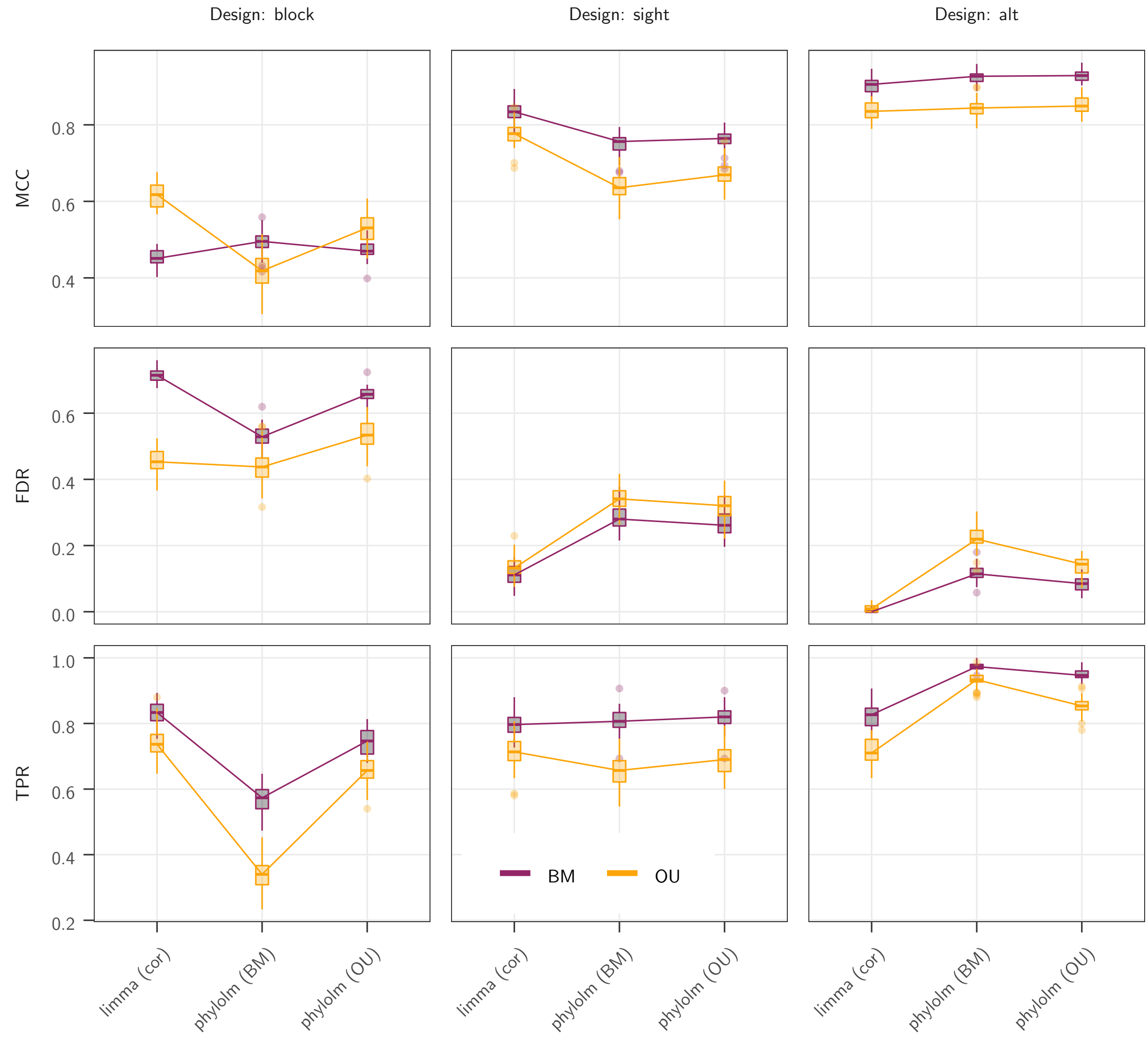

Figure S3: Results in terms of MCC (top), FDR (middle) and TPR (bottom) scores of the three correlation aware statistical methods (x axis) on the pPLN base scenario (effect size of 3, intra-species variation accounting for 20% of the total variance), with a BM (dark purple line) or an OU (light orange line) model of evolution on the maximum likelihood tree (Stern and Crandall, 2018), with the observed "sight" (middle), "block" (left) and "alt" (right) groups. The counts were normalized using $\log_2$(TPM) values. Boxplots are based on 50 replicates. Simulating with an OU weakens the phylogenetic signal (Ho and Ané, 2014b). For the "sight" and "alt" design ($\mathsf{dapaESSn} > 1$), it makes the differential expression detection problem more difficult, while for the block design ($\mathsf{dapaESSn} < 1$), it makes it easier. In the limiting case where the selection strength α goes to infinity, the underlying tree becomes a star tree ans species are no longer correlated (although samples inside a species are). The group design then does not matter in this limiting case, and we expect that the "sight", "alt" and "block" designs converge to the same difficulty ($\mathsf{dapaESSn} = 1$)

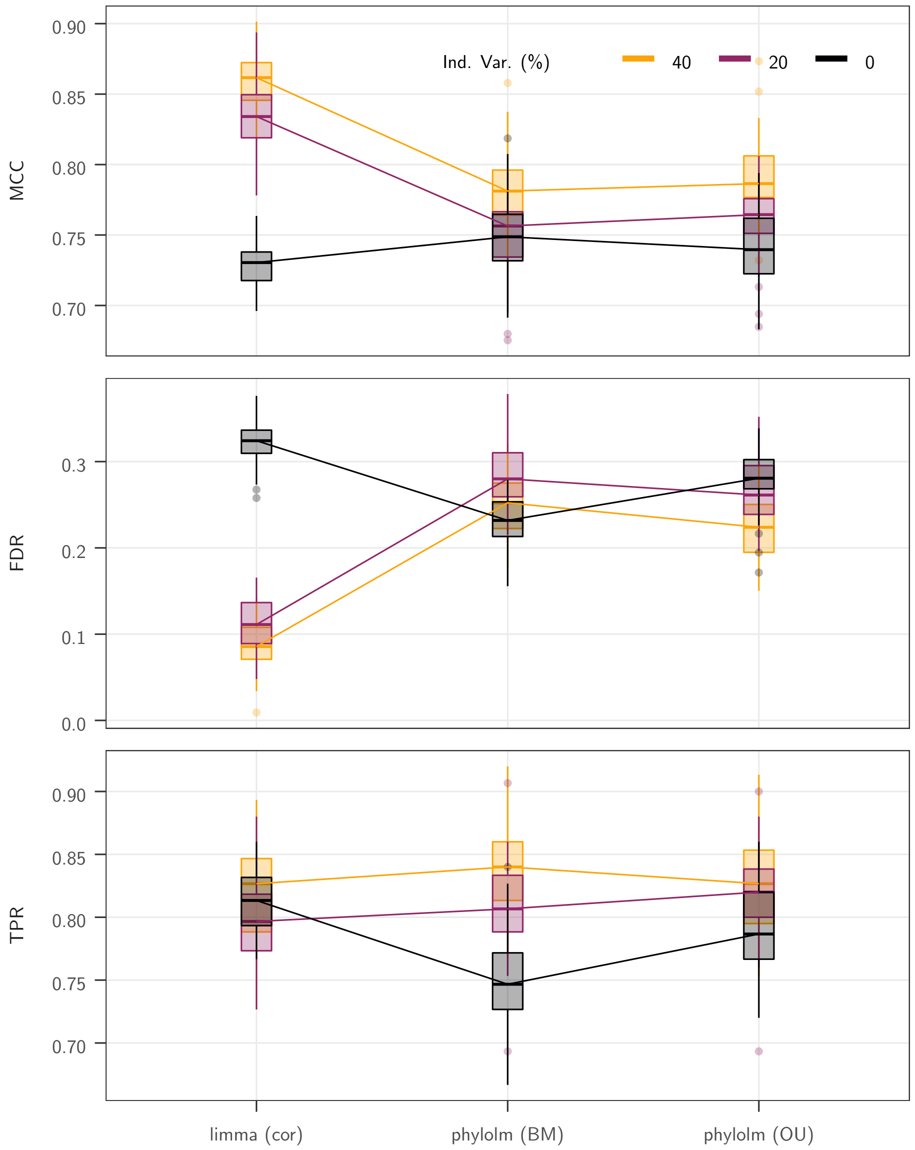

Figure S4: Results in terms of MCC (top), FDR (middle) and TPR (bottom) scores of the three correlation aware statistical methods (x axis) on the pPLN base scenario with an effect size of 3, a BM model of evolution on the maximum likelihood tree (Stern and Crandall, 2018), with the observed "sight" groups, and intra-species variation accounting for 40% (light orange line), 20% (dark purple line), or 0% (black line) of the total variance). The counts were normalized using $\log_2$(TPM) values. Boxplots are based on 50 replicates.

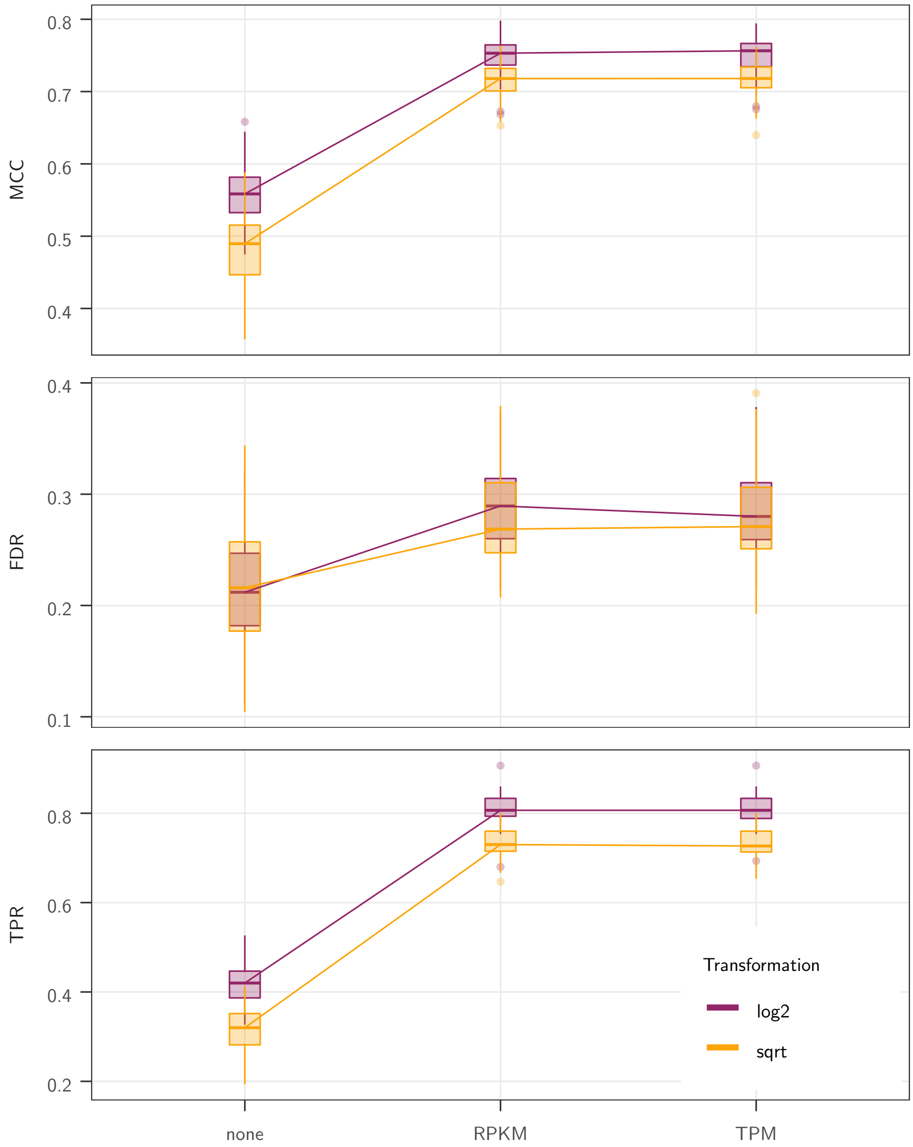

Figure S5: Results in terms of MCC (top), FDR (middle) and TPR (bottom) scores of the phylolm (BM) method on the pPLN base scenario (effect size of 3, BM model of evolution on the maximum likelihood tree (Stern and Crandall, 2018), with the observed "sight" groups, and added intra-species variation accounting for 20% of the total variance). The counts were 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 based on 50 replicates.

D Comparison of NB and pPLN Datasets

In this section, we reproduce and comment some of the results of the countsimQC (Soneson and Robinson, 2018) analysis on the comparison of datasets produced by our simulation schemes. The full analysis is available as an html file in the associated GitHub repository `https://github.com/i2bc/InterspeciesDE/` (see analysis file `R_scripts/06_simulations_data_check.R` and associated result file `2021-12-01_simulations_stern2018/all_with_lengths_3_0.8_1_countsim_report.html`). In this comparison, we considered four different datasets:

- "Original" is the original dataset of Stern and Crandall (2018), that was used as a basis for the calibration of our simulation scheme.
- "pPLN (real tree)" is a dataset produced by our pPLN *base scenario*, that had empirical moments drawn from the original dataet, 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.
- "pPLN (star tree)" is a dataset produced with the same parameters as the previous one, but using a star tree instead of the empirical tree, keeping only one sample per species. All the samples are hence drawn independently in this dataset.
- "NB" is a dataset produced with the classical NB of compcodeR, with parameters set up to match the first two moments of the previous dataset.

Note that we chose to keep only one replicate per species for the star-tree and NB simulation in order to avoid any species structure effect in the data. The first dataset hence had 14 species and 34 samples (with species replicates highly correlated), while the last two only had 14 independent samples. The fact that the "pPLN (real tree)" had more samples than the "pPLN (star tree)" and "NB" datasets makes the failure of DESeq2 and limma to correctly handle the structured dataset even more appreciable (see Fig. 2 and S1).

Dispersion Plot

Figure S6 shows a plot of the dispersion versus the base mean for the four datasets, with the scatter plot replaced with a density plot for better legibility. The dispersion is the final dispersion estimate of DESeq2. Visually, the original dataset has more concentrated base mean values and more dispersed distribution values than all the simulated datasets. The NB and pPLN (star tree) distributions look alike, and seem less diffuse than the pPLN (real tree) one.

Expression Distributions

Figure S7 shows a plot of the distribution over the genes of the sample-wise average log CPM values. As previously, visually, the original dataset has more concentrated mean values than all the simulated datasets, which in turn all look alike.

The Kolmogorov-Smirnov test between the distribution rejected the null hypothesis of equal distribution in pairwise comparisons between any model and the original data, indicating that all models are relatively poor in representing the distibution, NB and pPLN alike (see Table S5). However, we could not reject the null hypothesis when comparing the NB and the pPLN on star tree distribution.

Both comparison tend to show that, with similar parameters, the pPLN (star tree) does mimic the classical NB model, as intended.

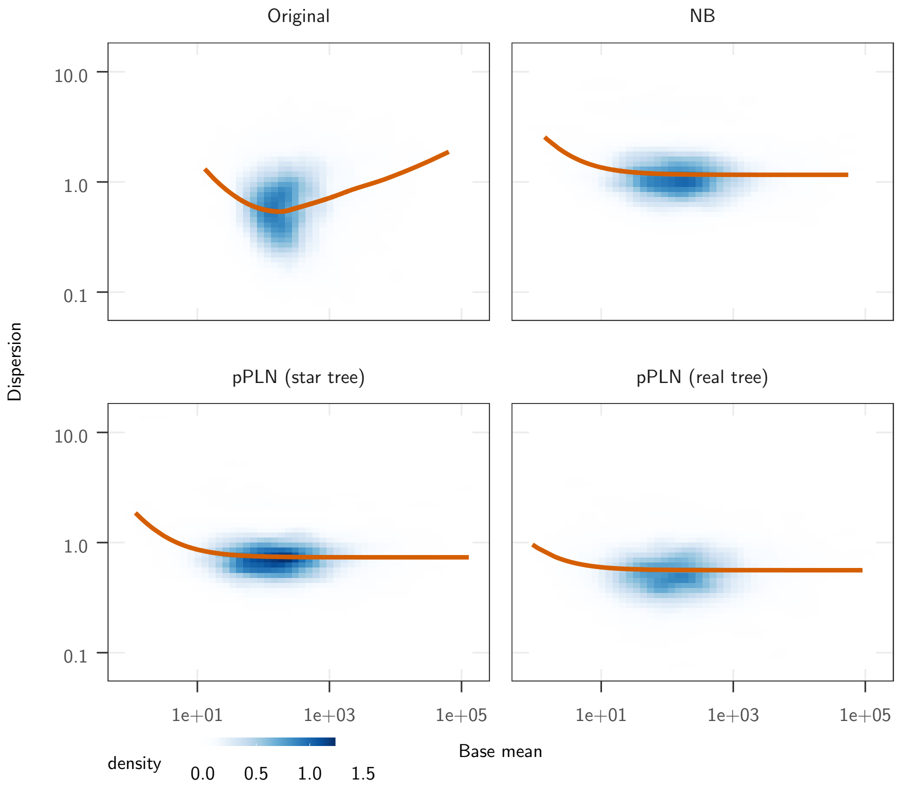

Figure S6: Plot of the average abundance against the dispersion, as calculated by DESeq2 (Love *et al.*, 2014). The final dispersion estimates distribution for the 3560 genes is shown as a tile plot, with deeper blues representing regions with a high density of points. The red curve is the fitted mean-dispersion relationship. Both axis are on the $\log_{10}$ scale. The Original dataset is the one from (Stern and Crandall, 2018). The base scenario (pPLN (real tree), right) had empirical moments drawn from (Stern and Crandall, 2018), 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. It is compared to a pPLN model with the same parameters, but in a case where all samples were independent (pPLN (star tree), middle), and to a NB model with the same moments and effect size (NB, left).

Table S5: Pairwise comparison of the expression distribution of the four samples, using the Kolmogorov–Smirnov test.

dataset1	dataset2	K-S statistic	K-S p-value
pPLN (real tree)	pPLN (star tree)	0.03	0.04
pPLN (real tree)	NB	0.03	0.03
pPLN (real tree)	Original	0.31	0.00
pPLN (star tree)	NB	0.01	0.98
pPLN (star tree)	Original	0.28	0.00
NB	Original	0.28	0.00

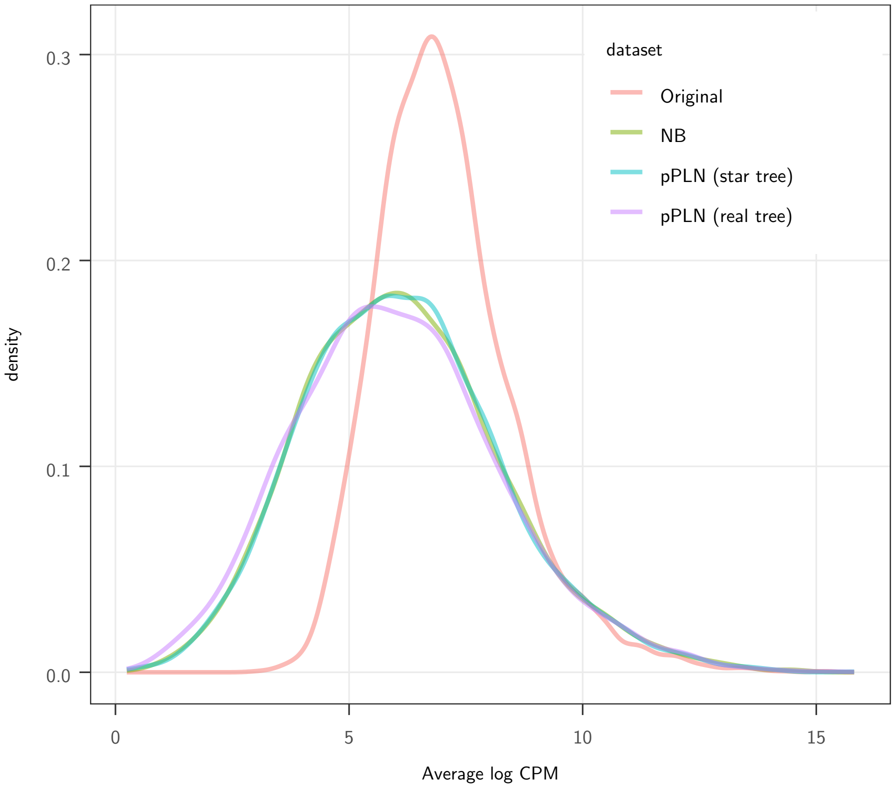

Figure S7: Plot of the distribution of the average abundance values for the genes. The Original dataset is the one from (Stern and Crandall, 2018). The base scenario (pPLN (real tree), right) had empirical moments drawn from (Stern and Crandall, 2018), with an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree, with the observed sight groups. It is compared to a pPLN model with the same parameters, but in a case where all samples were independent (pPLN (star tree), middle), and to a NB model with the same moments and effect size (NB, left).

E Surrogate Variable Analysis

Setting

Surrogate variable analysis (SVA) can be used as a pre-processing step before a differential analysis, to detect for hidden heterogeneity in the data, such as batch effects (Leek and Storey, 2007; Leek, 2014). The phylogenetic structure could be seen as a hidden variable, and a SVA step might help in correcting for the tree-induced structure. To test this hypothesis, we ran a limma sva analysis, that uses package sva (Leek, 2014) to first detect surrogate variables, and the proceed with a classic limma framework, with the added variables in the design. We tested two configurations for the SVA analysis: limma sva (one) was forced to detect only one surrogate variable, while limma sva (auto) used the asymptotic approach to estimate the number of surrogate variables to include (Leek, 2011). In both cases, the control probes were empirically estimated (method="irw"). The general setting was the same as the one used in our base scenario.

Results

The limma sva (one) method performed similarly as the vanilla limma method, while performing significantly worse than the limma cor method (Fig. S9). The limma sva (auto) method had a very high variance, and its best performances matched those of the vanilla limma method.

The high variance of the limma sva (auto) can be linked with the high variance of the number of surrogate variable selected, that goes from 1 to 26, with a median of 10 (Fig. S8). Interestingly, this number almost did not vary for the block design, wich is also the design for which the performance of limma sva (auto) relative to limma is the worse.

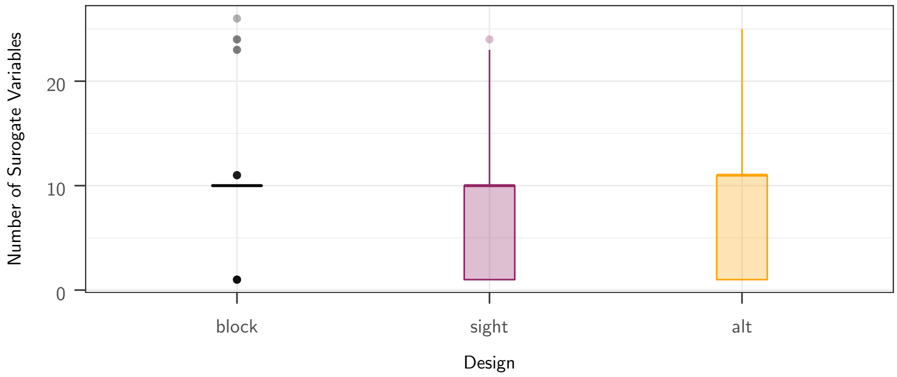

Figure S8: Number of surrogate variables selected in limma sva (auto), for the three designs (x axis) on the pPLN base scenario, that has an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree (Stern and Crandall, 2018), with counts normalized using $\log_2$(TPM) values. Boxplots are based on 50 replicates.

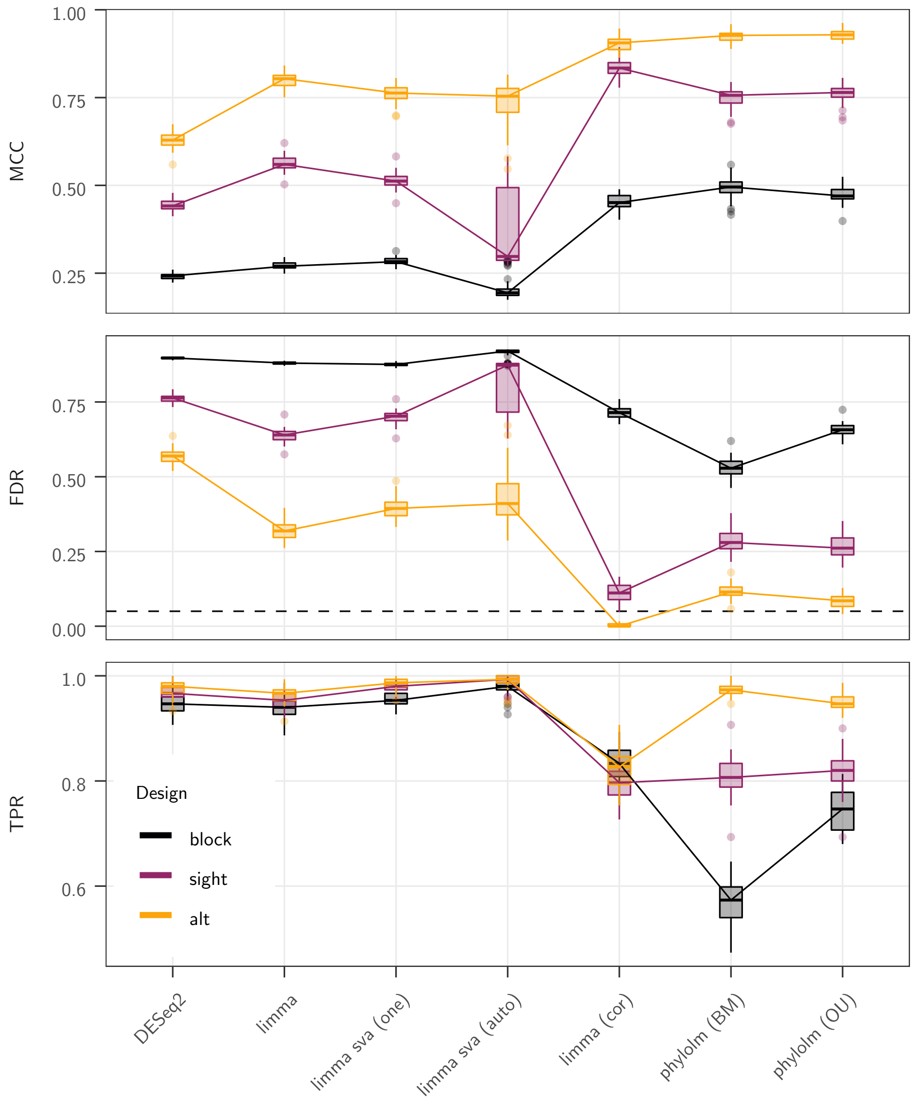

Figure S9: Results in terms of MCC (top), FDR (middle) and TPR (bottom) scores of the five selected statistical methods (x axis) on the pPLN base scenario, that has an effect size of 3, a BM model of evolution with added intra-species variation accounting for 20% of the total variance, on the maximum likelihood tree (Stern and Crandall, 2018), with the observed “sight” groups (dark purple line, see Fig. 1). The “alt” (light orange line) and “block” (black line) groups were 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 were normalized using $\log_2$(TPM) values. Boxplots are based on 50 replicates.

References

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

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.

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.

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.

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.

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.

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.

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.

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. 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.

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.

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.

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

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.

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.

Leek, J. T. 2011. Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data. *Biometrics*, 67(2): 344–352.

Leek, J. T. 2014. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21): e161.

Leek, J. T. and Storey, J. D. 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, 3(9): e161.

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.

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.

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

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.

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.

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

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

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).

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.