A Multivariate Negative-Binomial Model with Random Effects for Differential Gene-Expression Analysis of Correlated mRNA Sequencing Data

Peer-reviewed author version

Kazakiewicz, D; CLAESEN, Jurgen; GORCZAK, Katarzyna; Plewczynski, D & BURZYKOWSKI, Tomasz (2019) A Multivariate Negative-Binomial Model with Random Effects for Differential Gene-Expression Analysis of Correlated mRNA Sequencing Data. In: JOURNAL OF COMPUTATIONAL BIOLOGY, 26 (12) , p. 1339 -1348.

DOI: 10.1089/cmb.2019.0168
Handle: http://hdl.handle.net/1942/30436

A multivariate negative-binomial model for differential gene-expression analysis of correlated RNA-Seq data

Denis Kazakiewicz (dzainis.kazakiewicz@uhasselt.be)[1,2], Jürgen Claesen, (jurgen.claesen@uhasselt.be)[1], Katarzyna Górczak (katarzyna.gorczak@uhasselt.be)[1,3], Dariusz Plewczyński (dariuszplewczynski@cent.uw.edu.pl)[2,4], and Tomasz Burzykowski (tomasz.burzykowski@uhasselt.be)[1,2*]

[1]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt, Belgium

[2]Center for Innovative Research, Medical University of Białystok, Białystok 15-089, Poland

[3]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań 60-637, Poland

[4]Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland

*To whom correspondence should be addressed.

May 6, 2019

Abstract

Some experimental designs, such as matched-pair or longitudinal studies, yield mRNA sequencing (mRNA-Seq) counts that are correlated across samples. Most of the approaches for the analysis of correlated mRNA-Seq data are restricted to a specific design and/or balanced data only (with the same number of observations in each group).

We propose a model which is applicable to the analysis of correlated mRNA-Seq data of different types: paired, clustered, longitudinal or others. Any combination of explanatory variables as well as unbalanced data can be processed within this framework. The model assumes that exon counts of a particular gene of an individual sample jointly follow a multivariate negative-binomial distribution. Additional correlation between exon counts obtained for, e.g., individual samples within the same pair or cluster is taken into account by including into the model a cluster-level normal random effect. An interesting feature of the model is that it provides explicit expressions for marginal correlation between exon counts at different levels. The performance of the model is evaluated by using a simulation study and an analysis of two real-life datasets: a paired RNA-Seq experiment for 24 patients with clear-cell renal cell carcinoma and longitudinal RNA-Seq experiment for 29 patients with Lyme disease.

Python–code to apply the model is available at `https://sourceforge.net/projects/dgeee/`

1 Introduction

mRNA sequencing (mRNA-Seq) is a powerful and versatile high-throughput technique to study gene- and transcript-expression. The output of a mRNA-Seq experiment is typically a set of overdispersed counts. A large number of methods have been specifically developed to conduct differential gene-expression based on mRNA-Seq counts. Soneson et al. Soneson and Delorenzi (2013), Rapaport et al. Rapaport et al. (2013), Seyednasrollah et al. citeSeyednasrollah2015, and Conesa et al. Conesa et al. (2016) give a detailed overview and comparison of these methods.

Some experimental designs, such as matched-pair experiments or longitudinal studies, may yield RNA-Seq counts that are correlated across samples. A number of methods have been proposed to analyze data from such experiments. For instance, Pham and Jimenez Pham and Jimenez (2012), Hardcastle and Kelly Hardcastle and Kelly (2013), and Chung et al. Chung et al. (2013) have developed methods to analyze matched-pair data. For longitudinal studies and/or other types of clustered experiments several methods have been introduced Spies and Ciaudo (2015), including `PLNSeq` Zhang et al. (2015) and DESeq2 Love et al. (2014).

In this article, we propose a hierarchical model for differential gene expression analysis of correlated RNA-Seq data based upon exon counts. Exons are basic units in transcription. Within a gene expression of individual exons vary due to, among the other things, differences in exon lengths and alternative isoform regulation. Conventional methods for differential gene expression

operate on a summarized gene-level and disregard exon-expression variability. On the other hand, the methods assume that expression of a single exon has to necessarily lead to differential expression of the gene that contains that exon Anders et al. (2012). We propose to acknowledge the variation in exon expression when making an inference about gene expression by using a multivariate distribution for the exon-expression levels.

In particular, the model we propose includes two types of random effects which account, respectively, for the correlation between different samples (*cluster random effect*) and for the correlation within a sample (*individual random effect*). The cluster random effects are assumed to be normally distributed, whereas the individual random effects follow a gamma distribution. Consequently, conditionally on the cluster random effect, counts from exons of the same gene in a particular sample follow a multivariate negative-binomial (MVNB) distribution. Essentially, the proposed model falls in the framework developed by Molenberghs et al. Molenberghs et al. (2010). An important advantage of the model is that it can be applied to data coming from various designs that may yield correlated RNA-Seq counts, including matched pairs, clustered sampling, and longitudinal studies. Moreover, it allows computing the conditional and marginal correlation coefficients, which offer insight into the correlation structure of the data.

2 Methodology

We consider per-gene analysis. Thus, in what follows, we drop the index indicating the gene.

Assume that a gene consists of J exons. For a particular sample, the exon counts may be correlated. To account for the correlation, we propose that the counts follow a MVNB distribution Fabio et al. (2012).

In particular, denote by $\boldsymbol{y}_s = (y_{s1}, \ldots, y_{sJ})'$ the vector of exon counts for a particular gene in sample s. Let n_j denote the length of exon j and L_s be the effective library size Robinson and Oshlack (2010) of sample s. Additionally, let $\boldsymbol{x}_s = (1, x_{s1}, \ldots, x_{sp})'$ be the vector of covariates describing the sample. Then the joint probability mass function for $\boldsymbol{y}_s$ is assumed to be given by

$$P(\boldsymbol{y}_s) = \frac{\Gamma(\phi + \sum_{j=1}^{J} y_{sj})}{\Gamma(\phi) \prod_{j=1}^{J} (y_{sj}!)} Q_s^{-\phi} \prod_{j=1}^{J} \left(\frac{\mu_{sj}}{\phi Q_s} \right)^{y_{sj}}, \tag{1}$$

where

$$\mu_{sj} = n_j L_s \exp(\boldsymbol{x}_s' \boldsymbol{\beta})$$

is the expected value of the count for exon j, $\boldsymbol{\beta}$ is the $(p+1)$-dimensional vector of (unknown) coefficients (including the intercept) corresponding to the covariates in $\boldsymbol{x}_s$, $Q_s = 1 + \sum_j \mu_{sj}/\phi$, and ϕ is the overdispersion parameter so that

$$\mathrm{Var}(y_{sj}) = \mu_{sj}(1 + \mu_{sj}/\phi). \tag{2}$$

Note that the use of MVNB distribution implies that

$$\text{Corr}(y_{sj},y_{sk}) = \frac{\mu_{sj}\mu_{sk}}{\sqrt{(\phi+\mu_{sj})(\phi+\mu_{sk})}}. \tag{3}$$

Fabio et al. Fabio et al. (2012) demonstrated that the MVNB distribution, defined by (1), can be obtained as a distribution of a set of independent Poisson-distributed random variables with mean values depending on a gamma-distributed random effect:

$$y_{sj}|\gamma_s \sim \text{Poisson}(\mu_{sj}\gamma_s), \tag{4}$$

$$\gamma_s \sim \text{Gamma}(\phi, 1/\phi). \tag{5}$$

Note that (5) implies that the expected value of γ_s is equal to 1 and the variance is equal to $1/\phi$.

Assume now that exon counts $\boldsymbol{y}_s$ are collected for a set of samples that may be correlated. Thus, we observe N clusters, each with N_c samples. We extend our notation and identify a sample by index c for cluster ($c = 1, \ldots, N$) and s for the sample within the cluster ($s = 1, \ldots, N_c$). To account for the correlation between the exon counts obtained for samples from the same cluster, we include in our model a cluster-specific random effect b_c that is normally distributed with mean zero and variance σ^2. In particular, using the hierarchical representation (4)–(5) of the MVNB (1), we define the following

hierarchical model:

$$\begin{aligned} y_{csj}|\gamma_{cs}, b_c &\sim \text{Poisson}\{n_j L_{cs}\gamma_{cs}\exp(\boldsymbol{x}'_{cs}\boldsymbol{\beta} + b_c)\}, && (6)\\ \gamma_{cs} &\sim \text{Gamma}(\phi, 1/\phi), && (7)\\ b_c &\sim \text{Normal}(0, \sigma^2). && (8) \end{aligned}$$

Note that, alternately, we can describe the model as resulting from the assumption that, conditionally on b_c, exon counts $\boldsymbol{y}_{cs}$ within a sample are correlated as in (3) and distributed according to the MVNB (1) with overdispersion ϕ and with mean values

$$\mu_{csj|b} = n_j L_{cs}\exp(\boldsymbol{x}'_{cs}\boldsymbol{\beta} + b_c) \equiv K_{csj}\exp(b_c). \quad (9)$$

For the hierarchical model (6)–(8) it is possible (2010, Molenberghs et al. (2010)) to derive marginal moments. The marginal mean and variance of the exon count y_{csj} are given by, respectively,

$$\begin{aligned} \text{E}(y_{csj}) &= K_{csj}e^{\sigma^2/2}, && (10)\\ \text{Var}(y_{csj}) &= K_{csj}e^{\sigma^2/2} + K_{csj}^2 e^{2\sigma^2}\left(1/\phi + 1 - e^{-\sigma^2}\right), && (11) \end{aligned}$$

with K_{csj} defined in (9). The marginal covariance between counts for two exons, j and k, observed for two samples from the same cluster c, is given by

$$\text{Cov}(y_{csj}, y_{ctk}) = \begin{cases} K_{csj}K_{csk}e^{2\sigma^2}\left(1/\phi + 1 - e^{-\sigma^2}\right) & \text{if} \quad s = t \quad \& \quad j \neq k, \\ K_{csj}K_{ctk}e^{2\sigma^2}\left(1 - e^{-\sigma^2}\right) & \text{if} \quad s \neq t. \end{cases} \quad (12)$$

Thus, the model implies that counts of exons from different samples ($s \neq t$) that are part of the same cluster are (positively) correlated. The correlation is weaker than the correlation between counts of exons obtained for the same sample ($s = t$).

Note that correlation coefficients are functions of the marginal means defined by (10), which depend on the exon length n_j, library size L_{cs}, and sample-specific covariates. Thus, even for the same sample, the correlation will differ for different pairs of exon counts, unless the exons are of the same length. On the other hand, for a fixed pair of exons, the correlation will differ for different samples, unless the library sizes and sample-specific covariates are exactly the same.

The marginal likelihood for model (6)–(8) for exon-count data observed for N clusters with N_c samples each is given by

$$L(\boldsymbol{\beta}, \phi, \sigma^2) = \prod_{c=1}^{N} \int_{-\infty}^{\infty} \prod_{s=1}^{N_c} P(\boldsymbol{y}_{cs}|b_c) f(b_c) db_c, \tag{13}$$

where $P(\boldsymbol{y}_{cs}|b_c)$ is the probability mass function (1) defined by using mean values $\mu_{csj|b}$ specified in (9), while $f(b_c)$ is the density of the mean-zero normal distribution with variance σ^2. Note that, for brevity, we have not indicated the dependence of the functions involved in the right-hand-side part of (13) on the parameters.

The estimates of the parameters $\boldsymbol{\beta}$, ϕ, and σ are obtained by maximizing the marginal likelihood function (13).

The integral involved in (13) is computed by using the adaptive Gaussian-

Hermite quadrature (AGHQ). In our study we used 10 quadrature points. Variance-covariance matrix of the estimated parameters is obtained from the inverse of the negative Hessian of the logarithm of the marginal likelihood.

The model (6)-(8) is implemented in python. Our open source software is available under `https://sourceforge.net/projects/dgeee/`

3 Data

To investigate the performance of our model, we conducted simulation studies. We also applied the model to two real-life RNA-Seq datasets.

We considered settings of a matched-pair design and of a clustered experiment. For each of these settings, we generated 10,000 datasets with exon-counts for a gene that consists out of three exons. Data were generated by using the hierarchical model (6)–(8).

We also generated data from a conditional-independence model where, conditionally on the normally-distributed random effect, exon counts were independent and followed negative binomial distributions with mean values given by (9) and overdispersion parameter ϕ.

3.1 A simulated matched-pair RNA-Seq experiment

We simulated datasets for a hypothetical matched-pair experiment. We assumed that that the samples within each pair originated from two different biological conditions, "control" and "experimental," say. Thus, $\boldsymbol{x}_{cs} = (1, x_{cs1})'$,

where x_{cs1} is the binary indicator of the "experimental" condition, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. We assumed that $\beta_0 = 0.5$.

We considered different scenarios by varying the effect of the "experimental" condition (β_1), the number of matched pairs (N), the value of the overdispersion parameter (ϕ), and the standard deviation of the cluster random effect (σ). The scenarios are listed in Table 1. We assumed $\beta_1 = 0$, 0.12, or 1 and combined it with $N = 12$, 24, or 36 to study the Type-I error probability and power in function of the sample size. The value of ϕ was set to be equal to 54.6, 7.4, and 2.7 to investigate the effect of the correlation between exon counts on the estimation and model-based inference. In particular, the values implied the within-sample correlation of exon counts equal to 0.51, 0.87, and 0.95, respectively. The value of σ was set to 0.2 or 0.4 to evaluate the effect of increasing marginal between-sample correlation. We additionally considered the performance of the model in the case when, conditionally on the cluster random effect, there was no within-sample correlation.

In particular, scenarios (1)–(3) correspond to the situation of no difference in gene-expression between the two biological conditions. These scenarios allow, in particular, investigation of the Type-I error probability for testing $\beta_1 = 0$. On the other hand, scenarios (4)–(7) specify the case when there is differential expression, with a smaller ($\exp \beta_1 = \exp(0.12) = 1.12$) and a larger ($\exp \beta_1 = \exp(1) = 2.73$) fold-change. They allow investigation of the power for testing $\beta_1 = 0$. Scenarios (8)–(15) allow evaluation of the Type-I error probability and power for stronger within-cluster and/or within-sample correlations.

Finally, scenarios (16)–(17) focus on the performance of model (6)–(8), which assumes a positive within-sample correlation, when, in fact, data are generated from a simpler model independent exon-counts within the same sample.

3.2 A simulated longitudinal RNA-Seq experiment

We simulated also datasets for a hypothetical longitudinal study. We assumed that samples from the same individual were obtained at three different time points. Thus, $\boldsymbol{x}_{cs} = (1, x_{cs1}, x_{cs2})'$, where x_{cs1} and x_{cs2} are the binary indicators of the second and third measurement occasion, respectively, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. We assumed that $\beta_0 = 0.5$.

Table 2 presents the considered combinations of the effect of the measurement occasion (β_1 and β_2), the number of clusters (N), the value of the overdispersion parameter (ϕ), and the standard deviation of the cluster random effects (σ).

We were particularly interested in the Type-I error probability and power for the Wald test for the following two null hypotheses:

$$\mathrm{H}_0^1: \quad \beta_1 = \beta_2 \equiv 0, \tag{14}$$

$$\mathrm{H}_0^2: \quad \beta_1 = \beta_2. \tag{15}$$

Rejection of H_0^1 indicates that gene-expression changes with time. Rejection of H_0^2 implies that a change occurs between the second and third measurement

Table 1. Parameters of the simulations (10,000 replicates each) of a matched pair RNA-Seq experiment for one gene consisting of three exons. (N is the number of pairs. $\sigma\,(\ln\sigma)$, $\phi\,(\ln\phi)$, β_1 refer to model (6)–(8). In all scenarios $\beta_0 = 0.5$. $\rho|b$ and $\rho^*|b$ are the conditional correlations between exon 1 and exon 3 for control and experimental samples, respectively, for a pair with the random effect equal to zero ($b_c = 0$). $\rho_{s,s}$ and $\rho_{s,t}$ are the marginal correlations between exon 1 and exon 3 within the same control sample and for different samples within the same pair, respectively. All correlations were calculated from the true parameter values.)

Scenario	N	$\sigma\,(\ln\sigma)$	$\phi\,(\ln\phi)$	β_1	$\rho\|b$	$\rho^*\|b$	$\rho_{s,s}$	$\rho_{s,t}$
No differential expression ($\beta_1 = 0$)								
(1)	12	0.2 (-1.6)	54.6 (4)	0	0.51	0.51	0.76	0.52
(2)	24	0.2 (-1.6)	54.6 (4)	0	0.51	0.51	0.76	0.52
(3)	36	0.2 (-1.6)	54.6 (4)	0	0.51	0.51	0.76	0.52
Small fold change ($\beta_1 = 0.12$)								
(4)	12	0.2 (-1.6)	54.6 (4)	0.12	0.51	0.54	0.76	0.52
(5)	24	0.2 (-1.6)	54.6 (4)	0.12	0.51	0.54	0.76	0.52
Large fold change ($\beta_1 = 1$)								
(6)	12	0.2 (-1.6)	54.6 (4)	1	0.51	0.73	0.76	0.53
(7)	24	0.2 (-1.6)	54.6 (4)	1	0.51	0.73	0.76	0.53
Increased marginal within-cluster correlation ($\sigma = 0.4$)								
(8)	12	0.4 (-0.9)	54.6 (4)	0	0.51	0.51	0.91	0.81
(9)	12	0.4 (-0.9)	54.6 (4)	0.12	0.51	0.54	0.91	0.81
Increased conditional within-sample correlation ($\phi = 7.4$)								
(10)	12	0.2 (-1.6)	7.4 (2)	0	0.87	0.87	0.90	0.20
(11)	12	0.2 (-1.6)	7.4 (2)	1	0.87	0.95	0.90	0.21
(12)	24	0.2 (-1.6)	7.4 (2)	1	0.87	0.95	0.90	0.21
(13)	36	0.2 (-1.6)	7.4 (2)	1	0.87	0.95	0.90	0.21
Increased within-cluster ($\sigma = 0.4$) and -sample ($\phi = 7.4$) correlations								
(14)	12	0.4 (-0.9)	7.4 (2)	1	0.87	0.95	0.95	0.50
Increased conditional within-sample correlation ($\phi = 2.7$)								
(15)	12	0.2 (-1.6)	2.7 (1)	1	0.95	0.98	0.96	0.09
Conditional-independence model ($\rho|b = 0$)								
(16)	12	0.2 (-1.6)	54.6 (4)	1	0	0		
(17)	24	0.2 (-1.6)	54.6 (4)	1	0	0		

occasion.

In simulated scenarios (1)–(2), presented in Table 2, both null hypotheses H_0^1 and H_0^2 were true. Thus, the scenarios allowed estimation of the Type-I

error probability. In scenarios (3)–(4), only H_0^2 was true. These scenarios were suitable for evaluating the Type-I error probability for H_0^2 and power for H_0^1. The settings of scenarios (5)–(7) allowed us to investigate the influence of the sample size and of the magnitude of fixed effects on the power. Finally, in scenarios (8)–(11) we analyzed the influence of the magnitude of the marginal and conditional correlations on the power. In all scenarios we also observed the performance of the model in terms of the bias and precision of the parameter estimates.

3.3 Renal-cell carcinoma matched-pair experiment

Metastases in clear cell renal cell carcinoma (RCC) are associated with poor treatment outcomes Capitanio and Montorsi (2016). Recent drug discovery strategies in metastatic RCC have been directed to specific targets in a few biological pathways Capitanio and Montorsi (2016). It is, therefore, important to identify genes associated with differences in metastatic status.

The dataset contained the outcomes of a pre-processed RNA-Seq experiment for 24 RCC patients. Twelve patients with a metastatic disease were matched with twelve non-metastatic patients based on the SSIGN–score Frank et al. (2002). It must be noted, that absence of metastases at the time of the study did not rule out the risk of development of metastases at later stages of the disease.

Pre-processed data were obtained from the Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN.

Table 2. Parameters of the simulations (10,000 replicates each) of a longitudinal RNA-Seq experiment with one gene consisting of three exons. (N is the number of clusters. $\sigma\,(\ln\sigma)$, $\phi\,(\ln\phi)$, β_1, β_2 refer to model (6)–(8). In all scenarios $\beta_0 = 0.5$. $\rho|b$ and $\rho^*|b$ are the conditional correlations between exon 1 and exon 3 for an individual with the random effect equal to zero ($b_c = 0$) at the first and third time point, respectively. $\rho_{s,s}$ is the marginal correlation between exon 1 and exon 3 within the same sample at the first time point. $\rho_{s,t}$ is the marginal correlation between exon 1 from a sample collected at the first point and exon 3 from a sample collected at the third time point. All correlations were calculated from the true parameter values.)

Scenario	N	$\sigma\,(\ln\sigma)$	$\phi\,(\ln\phi)$	β_1	β_2	$\rho\|b$	$\rho^*\|b$	$\rho_{s,s}$	$\rho_{s,t}$
			H_0^1: $\beta_1 = \beta_2 = 0$						
(1)	8	0.2 (-1.6)	54.6 (4)	0	0	0.51	0.51	0.76	0.52
(2)	16	0.2 (-1.6)	54.6 (4)	0	0	0.51	0.51	0.76	0.52
			H_0^2: $\beta_1 = \beta_2 \neq 0$						
(3)	8	0.2 (-1.6)	54.6 (4)	0.15	0.15	0.51	0.55	0.76	0.52
(4)	16	0.2 (-1.6)	54.6 (4)	0.15	0.15	0.51	0.55	0.76	0.52
			$\beta_1 \neq \beta_2$						
(5)	8	0.2 (-1.6)	54.6 (4)	0.15	0.25	0.51	0.57	0.76	0.52
(6)	16	0.2 (-1.6)	54.6 (4)	0.15	0.25	0.51	0.57	0.76	0.52
(7)	8	0.2 (-1.6)	54.6 (4)	1	1.5	0.51	0.81	0.76	0.54
			Increased conditional within-sample correlation ($\phi = 7.4$)						
(8)	8	0.2 (-1.6)	7.4 (2)	0	0	0.87	0.87	0.90	0.20
(9)	8	0.2 (-1.6)	7.4 (2)	0.15	0.25	0.87	0.90	0.90	0.20
			Increased marginal within-cluster correlation ($\sigma = 0.4$)						
(10)	8	0.4 (-0.9)	54.6 (4)	0	0	0.51	0.51	0.91	0.81
(11)	8	0.4 (-0.9)	54.6 (4)	0.15	0.25	0.51	0.57	0.91	0.82

For each patient, expression of 22,334 genes was quantified, yielding measurements for the total of 234,575 exons. There were 24,158 exons with zero counts across all the samples. The number of exons varied between 1 and 468 (mean 10.5, median 7) per gene.

3.4 Tick-born Lyme disease longitudinal study

Lyme disease is a tick-born infection. Some patients report lingering or recurring symptoms lasting months to years after antibiotic treatment Bouquet et al. (2016). Despite growing knowledge on immune response to an acute Lyme disease, pathogenetic molecular mechanisms behind persistent post-lyme symptoms are not well understood Bouquet et al. (2016); Strle et al. (2014). Identifying the genes associated with the dynamics of Lyme disease might bring better understanding of these mechanisms.

Bouquet et al. Bouquet et al. (2016) conducted a study in which RNA-Seq data were collected from 29 patients with tick-born Lyme disease and from 13 healthy controls. We downloaded these freely available data from `https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP049605`

In our analysis, we only use the samples from the patients with Lyme disease. From each patient, blood samples were taken three times: once before the antibiotic treatment was started, once immediately after completion of the treatment, and six months after treatment completion.

Eighty-seven libraries were sequenced as 100-bp paired-end runs on a HiSeq 2500 (Illumina). Three of them were discarded because of insufficient read counts and transcriptome coverage. We analyzed only single-end reads in order to limit the number of unaligned reads. The preprocessing steps included removal of the 5' and 3' adapters, and trimming low-quality ends from the reads using `cutadapt` Martin (2011). Processed reads shorter than 50bp were discarded.

Single-end processed reads were mapped to the human genome (GRCh38) with Bowtie Langmead et al. (2009) allowing for up to two mismatches and reporting the best mapping location for each alignment. The read counting was performed by using the R function `summarizeOverlaps` Lawrence et al. (2013) according to the exon annotation (GRCh38.82). As a result, 195,785 exons with non-zero sum of counts across all samples were included in the analysis. Finally, we used 19,808 protein coding genes from the annotation file to group exons into genes.

4 Results

4.1 A simulated matched-pair RNA-Seq experiment

Scenarios (1)–(3) (see Table 1) correspond to the situation when there is no differential expression between the two biological conditions. For those cases, Table 3 presents the true values of the parameters, their mean estimated values, relative bias, mean model-based standard error estimate (SE_{model}), empirical standard error of the estimates (SE_{emp}), the estimated coverage (the percentage of cases when the CI does include the true value of the parameter) of the 95% confidence interval (CI), and the percentage of cases when the CI does not include the value of 0. The latter reflects power if the alternative hypothesis is correct (the true value of the parameter is not equal to zero) or the complement of the Type-I error probability if the null hypothesis is correct (the true value of the parameter is equal to zero). Note that, given 10,000 simulations, the standard error of the coverage is equal to

about $\sqrt{0.05 \times 0.95}/100 = 0.002$.

Table 3. A simulated matched-pair RNA-Seq experiment for scenarios (1)–(3) (no differential expression; see Table 1). SE_{model} and SE_{emp} are the mean model-based and empirical standard error estimates, respectively.

	True value	Mean estimate	Relative bias	SE_{model}	SE_{emp}	95% CI coverage
			N=12 (Scenario 1)			
$\ln \sigma$	-1.609	-1.803	0.120	21.201	0.699	0.991
$\ln \phi$	4.000	4.260	0.065	0.617	0.632	0.989
β_0	0.500	0.499	-0.002	0.074	0.072	0.942
β_1	0.000	0.001	NA	0.062	0.060	0.935
			N=24 (Scenario 2)			
$\ln \sigma$	-1.609	-1.674	0.040	0.245	0.237	0.979
$\ln \phi$	4.000	4.108	0.027	0.367	0.371	0.954
β_0	0.500	0.499	-0.001	0.052	0.051	0.948
β_1	0.000	0.000	NA	0.043	0.043	0.943
			N=36 (Scenario 3)			
$\ln \sigma$	-1.609	-1.650	0.025	0.164	0.165	0.969
$\ln \phi$	4.000	4.074	0.018	0.292	0.296	0.952
β_0	0.500	0.499	-0.001	0.042	0.041	0.949
β_1	0.000	0.000	NA	0.035	0.035	0.943

The results presented in Table 3 indicate that, under the null hypothesis $\beta_1 = 0$, both β_0 and β_1 are estimated with a negligible bias. The model-based standard errors slightly overestimate, on average, the empirical standard errors. For $N = 12$ pairs, the estimated coverage of the 95% CI for β_0 is (statistically significantly) slightly below 95%. However, the under-coverage essentially disappears with increasing N. A similar trend is observed for β_1, though for $N = 36$ the coverage is still below 95%. This indicates a slight inflation of the Type-I error probability. Note that we did not investigate the effect of increasing N beyond 36, as $N = 36$ seems to be already a

considerable number of pairs for a matched-pair RNA-Seq experiment.

For the parameters linked to the random effects, i.e., $\ln \sigma$ and $\ln \phi$, the relative bias is equal, respectively, to 12% and 6.5% for $N = 12$. The bias decreases with increasing N.

For $N = 12$, the mean model-based standard error for $\ln \sigma$ is much larger than the empirical standard error. This is largely due to 107 simulations which yielded an extremely large (> 20) model-based standard error. In those cases one can conclude problems with estimation of $\ln \sigma$. The problem disappears with increasing N. It is worth noting that for $\ln \sigma$ and $\ln \phi$ the mean model-based standard error underestimates the empirical standard error. Consequently, the coverage of the 95% CI of both parameters is too high, i.e., above 95%. However, the coverage gets closer to 95% with increasing N. In fact, for $N = 24$ and $N = 36$, it is not statistically significantly different from 95% for $\ln \phi$.

Table 4 shows the results for scenarios (4)–(7) (see Table 1), i.e., for the situation of differential expression ($\beta_1 \neq 0$) between the two biological conditions. Similarly to the case of no differential expression, β_0 and β_1 are estimated with almost no bias. The coverage of the 95% CI of β_1 is lower than 95%. It is worth noting that the power for testing $\beta_1 = 0$ when, in fact, $\beta_1 = 0.12$ is equal to about 0.5 for $N = 12$ and increases to 0.8 for $N = 24$. For $\beta_1 = 1$ the power is essentially equal to 1 even for $N = 12$.

For $\ln \sigma$ and $\ln \phi$, the relative bias is equal to about, respectively, 10% and 6% for $N = 12$. The bias decreases to about, respectively, 4% and 3% for $N = 24$.

Table 4. A simulated matched-pair RNA-Seq experiment – scenarios (4)–(7) (differential expression; see Table 1). SE_{model} and SE_{emp} are the mean model-based and empirical standard error estimates, respectively.

	True value	Mean estimate	Relative bias	SE_{model}	SE_{emp}	95% CI coverage	Power
			N=12 (Scenario 4)				
$\ln\sigma$	-1.609	-1.803	0.120	3.626	0.704	0.993	
$\ln\phi$	4.000	4.243	0.061	0.579	0.592	0.983	
β_0	0.500	0.500	0.000	0.074	0.072	0.941	
β_1	*0.120*	0.120	0.001	0.062	0.060	0.938	0.495
			N=24 (Scenario 5)				
$\ln\sigma$	-1.609	-1.673	0.040	0.363	0.262	0.979	
$\ln\phi$	4.000	4.115	0.029	0.364	0.370	0.954	
β_0	0.500	0.499	-0.002	0.052	0.051	0.946	
β_1	*0.120*	0.120	0.000	0.043	0.043	0.939	0.794
			N=12 (Scenario 6)				
$\ln\sigma$	-1.609	-1.776	0.103	1.909	0.571	0.991	
$\ln\phi$	4.000	4.227	0.057	0.538	0.554	0.958	
β_0	0.500	0.499	-0.002	0.074	0.072	0.948	
β_1	*1.000*	1.001	0.001	0.060	0.059	0.938	1.000
			N=24 (Scenario 7)				
$\ln\sigma$	-1.609	-1.668	0.036	0.207	0.210	0.979	
$\ln\phi$	4.000	4.106	0.027	0.345	0.348	0.948	
β_0	0.500	0.501	0.002	0.052	0.051	0.947	
β_1	*1.000*	0.999	-0.001	0.042	0.042	0.943	1.000

Similarly to the case of no differential expression (see Table 3), the mean model-based standard error for $\ln\sigma$ for $N = 12$ is much larger than the empirical standard error. Increasing N to 24 effectively removes the problem. It also improves the estimation of the standard error of $\ln\sigma$ and $\ln\phi$ and the 95% CI coverage, especially for $\ln\phi$.

The results for scenarios (8)–(15) are presented in Table S4. They lead to conclusions very similar to those presented above. The parameters of primary

interest, β_0 and β_1, are estimated with a negligible bias. The estimates of the parameters of secondary interest, $\ln \sigma$ and $\ln \phi$, have a larger, but reasonably small, bias. For the latter parameters, increasing the number of observations reduces the bias and improves the coverage of the 95% CI.

Table 5 presents results for scenarios (16)–(17), i.e., for the model in which, conditionally on cluster random-effects, exon counts were not correlated. As compared to scenarios (6) and (7) in Table 4, there is essentially no difference in the bias nor precision of estimation of β_0 and β_1. Thus, estimates of the parameters provided from our model seem to be robust to this type of misspecification of the variance-covariance structure.

It is worth noting that, throughout the simulations, the library size was assumed to be the same. Thus, for all scenarios, except of (16) and (17), we could estimate the marginal correlation matrix corresponding to the marginal variances (11) and covariances (12) for each simulated dataset by using the estimated values of the parameters. The estimated matrices were then averaged and compared to the true correlation matrix resulting from true parameter values. In particular, to summarize the relative bias, we calculated the average of the relative differences between the upper triangular elements of the estimated and true correlation matrices. Note that the bias was reported only if the differenced were of the same sign. The obtained results (complete results are available in the supplementary file S3.xlsx) indicated a satisfactory performance of the estimates of the marginal correlations. In general, the correlation coefficients were slightly underestimated. For instance, for scenarios (1)–(3), the average relative bias was equal to -6.8%, -3.2%, and

−2.1% for 12, 24, and 36 pairs, respectively. For scenarios (4)–(5), increasing the sample size from $N = 12$ to $N = 24$ reduced the bias from −5.7% to −2.5%. Only when the within-sample correlation was extremely high (as compared to other scenarios) and, at the same time, the between-sample correlation was extremely low, as in scenario (15) (see Table 1), there was a considerable positive bias of about 35% .

Table 5. A matched pair RNA-Seq experiment – scenarios (16)–(17) (a conditional-independence model; see Table 1). SE_{model} and SE_{emp} are the mean model-based and empirical standard error estimates, respectively.

	True value	Mean estimate	Relative bias	SE_{model}	SE_{emp}	95% CI coverage
			N=12 (Scenario 16)			
log σ	-1.609	-1.723	0.071	0.373	0.319	0.974
log ϕ	4.000	5.025	0.256	0.738	0.757	0.737
β_0	0.500	0.499	-0.002	0.068	0.067	0.936
β_1	1.000	1.000	0.000	0.044	0.044	0.932
			N=24 (Scenario 17)			
log σ	-1.609	-1.662	0.033	0.177	0.178	0.957
log ϕ	4.000	4.858	0.215	0.398	0.399	0.411
β_0	0.500	0.500	0.000	0.047	0.047	0.945
β_1	1.000	1.000	0.000	0.031	0.031	0.940

4.2 A simulated longitudinal RNA-Seq study

For scenarios (1) and (2) (see Table 2), the estimated Type-I-error probability of the joint Wald test for hypotheses H_0^1 and H_0^2 (see equations (14) and (15)) was equal to 6.9% and 6.1%, respectively. Doubling the number of clusters from 8 to 16 in scenario (2) reduced the probability to 5.9% and 5.1%, respectively. Note that, given that 10,000 simulations were conducted,

the former estimate is statistically significantly different from 5%, suggesting a slight inflation of the Type-I-error probability for testing H_0^1.

In scenario (3), only H_0^2 was true. The estimated Type-I-error probability was equal to 6.0%. The power of testing H_0^1 was equal to 51%. Doubling the number of clusters from 8 to 16 in scenario (4) reduced the Type-I-error probability to 5.4% and increased the power to 82%.

Overall, the power of testing the hypotheses increased as the number of clusters increased. For example, for scenario (5) with $N = 8$ clusters, the estimated power for testing H_0^1 and H_0^2 was equal to 0.83 and 0.28, respectively. Doubling the number of clusters in scenario (6) increased the power to 0.99 and 0.48, respectively. Complete results related to the power of the Wald test are available in the supplementary file S3.xlsx.

Table 6 presents the simulation results for scenarios (1) and (2). Similarly to the matched-pair case (see Table 3), under the null hypothesis, the fixed effects (β_0, β_1, β_2) are estimated with a negligible bias. The coverage of the 95% CI is higher than the nominal level, but decreases when the increasing number of clusters. The same is true for the 95% CI and bias of $\log \sigma$ and $\log \phi$.

Similar observations can be made for scenarios (3)–(11). The detailed results are presented in Supplementary Table S5. In particular, they confirm the adequate performance of the model-based estimates of β_0, β_1, and β_2.

We also investigated the bias of the estimates of the marginal correlations, obtained by using the estimates of the model parameters. In most cases,

a negative relative bias was observed. The largest absolute value of the relative bias was equal to 14%. The bias substantially reduced when the number of cluster increased. The detailed results related to the estimation of the marginal correlation structure are available in the supplementary file S3.xlsx.

Table 6. A longitudinal RNA-Seq experiment – scenarios (1)–(2)(a gene without differential expression; see Table 2). SE_{model} and SE_{emp} are the mean model-based and empirical standard error estimates, respectively.

	True value	Mean estimate	Relative bias	SE_{model}	SE_{emp}	95% CI coverage
			N=8 (Scenario 1)			
$\log \sigma$	-1.609	-1.858	0.154	3.886	0.759	0.998
$\log \phi$	4.000	4.271	0.068	0.585	0.535	0.995
β_0	0.500	0.499	-0.002	0.102	0.088	0.961
β_1	0.000	0.000	NA	0.086	0.074	0.964
β_2	0.000	0.000	NA	0.086	0.074	0.964
			N=16 (Scenario 2)			
$\log \sigma$	-1.609	-1.697	0.054	0.327	0.272	0.977
$\log \phi$	4.000	4.122	0.030	0.338	0.325	0.960
β_0	0.500	0.499	-0.001	0.066	0.061	0.956
β_1	0.000	0.000	NA	0.056	0.052	0.955
β_2	0.000	0.000	NA	0.056	0.052	0.958

4.3 Renal-cell carcinoma matched-pair experiment

We applied our model, the `PLNseq` Zhang et al. (2015) and `DESeq2` Love et al. (2014) methods to the renal-cell carcinoma dataset.

To produce gene counts for `PLNseq`, we summed the relevant exon counts. It is worth noting that the current implementation of `PLNseq` requires gene counts to be equal to or larger than 50 in each condition. As a consequence,

21% of 22,334 genes had to be excluded from the analysis. In the remaining set (17,528 genes), 133 genes were found to be statistically significantly (after the correction for multiple testing with the Benjamini-Hochberg (BH) Benjamini and Hochberg (1995) procedure) differentially expressed between the non-metastatic and metatstatic samples.

In contrast to `PLNseq`, our model does not require any minimum value of an exon-count to include the count in the analysis. However, for 6.5% (1448) of the genes we could not obtain estimates due to non-convergence. For information about the `PLNseq` and non-convergence and the convergence criteria of our approach, see Sections A.7 and A.5.2 in Supplementary Materials. Only eight of the remaining 21786 genes were found to be statistically significantly differentially expressed after the BH-correction for multiplicity. None of these eight genes was identified by PLNseq.

We also analyzed the data by using `DESeq2`. Also in that case we produced gene counts by summing the relevant exon counts. We applied a multi-factor design to analyze the paired gene-counts. In particular, the pair number constituted a factor in the design formula Love et al. (2014). The method identified 11 differentially expressed genes, different from those identified by our method.

4.4 Tick-born Lyme disease longitudinal study

Only 8,760 genes were analyzed by the `PLNseq` method Zhang et al. (2015) because the remaining 11,048 (56%) genes had a count smaller than 50 in

any of the conditions. Moreover, as `PLNseq` cannot handle unbalanced data, three patients with only two samples had to also be discarded.

The current implementation of `PLNseq` can only test null hypothesis H_0^1 (see Eq. (14)). For 2,456 genes (28%), the null hypothesis was rejected after the BH-correction for multiplicity.

An advantage of our model is that it can handle unbalanced data. Hence, we could analyze the data for all the Lyme-disease samples.

Our model did not converge for 1,680 (8.5%) genes. Among 18,128 genes for which no convergence issues were noted, H_0^1 was rejected for 4,096 (23%) genes. Among those, 528 were also identified by `PLNseq`. Additionally, for 976 out of the 4,096 genes, H_0^2 was rejected, i.e., we found that there was a statistically significant difference in gene expression between the second and the third measurement occasion. In particular, expression of 552 genes increased at the third occasion as compared to the second occasion, whereas expression of 424 genes decreased. The expression of the 47 significant genes changed more than two-fold, suggesting their active involvement in the long-term reaction to acute Lyme disease.

5 Conclusion

In this article we have presented a model for analyzing differential gene expression in correlated RNA-Seq data based on exon-level counts. The model accounts for the within- and between-sample correlation between the exon counts for a particular gene. The model can be applied to various types

of experiments that might yield correlated RNA-Seq data (matched pairs, longitudinal studies, clustered sampling).

Simulation studies showed that the model is able to correctly estimate differential expression, even when there is no within-sample correlation. Increasing the number of clusters reduces bias and improves precision of the estimation of the random-effect parameters.

The performance of our model has been compared with `PLNSeq` with the help of two real-life experiments. In contrast to `PLNseq`, our method can handle unbalanced data, and a substantially larger number of genes can be tested. Additionally, our model allows testing various hypotheses related to the factors that might influence gene-expression levels.

In the paired experiment, the list of differentially expressed genes, reported by our model, was completely different from the list obtained by the `DESeq2` method (see Section 4.3). This difference may be due to the fact that `DESeq2` is not designed for the analysis of correlated exon counts.

It is possible to use our model for the analysis of differential gene expression in correlated RNA-Seq data with gene counts as input. Mathematically, the model would simplify to a mixed-effects negative-binomial regression with normal cluster-specific random effects. The user could run our model in the same way as if there were just one exon per gene. Alternative solutions could be found in other software like, for instance, Stata (`menbreg` command).

One noticeable drawback of our model is that it assumes that, after adjusting for the exon length and library size, all exons within the same gene have the

same level of expression. However, exon expression can vary due to alternative splicing. In our model alternative spicing is only partially accounted for by using the overdispersion parameter ϕ. An extension of the model that would allow for an explicit adjustment for existence of multiple isoforms of a gene is a topic for further research.

It is worth noting that exon-level analysis of RNA-Seq data is that it disregards exon-exon junctions. In addition, some exons may overlap in an annotation file.

Our method is computationally intensive. It required roughly 5-6 hours per 1000 genes on one core for the real-life datasets described in this work. The study was carried out on either Intel Xeon E5-2670 v3 or Intel Xeon E5-2697 v3 hardware.

The proposed model is implemented in open source Python software, which is available under `https://sourceforge.net/projects/dgeee/`.

Acknowledgements

We thank Wacław Andrzej Sokalski for enabling this collaboration. We also thank Jeanette E. Eckel Passow and Daniel J. Serie for providing renal-cell carcinoma matched-pair experiment data and for helpful discussions.

Funding

This work was supported by the Wrocław Centre for Networking and Supercomputing (grant number 255). DP, TB, and DK were supported by the Medical University of Białystok. DP was supported by the Polish National Science Centre (2014/15/B/ST6/05082) and Foundation for Polish Science (TEAM to DP). DK and KG were supported by BOF bilateral scientific cooperation grant R-6244 and R-7699.

6 References

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017.

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)*, 57:289–300.

Bouquet, J., Soloski, M. J., Swei, A., Cheadle, C., Federman, S., Billaud, J.-N., Rebman, A. W., Kabre, B., Halpert, R., Boorgula, M., Aucott, J. N., and Chiu, C. Y. (2016). Longitudinal Transcriptome Analysis Reveals a Sustained Differential Gene Expression Signature in Patients Treated for Acute Lyme Disease. *mBio*, 7(1):e00100–16.

Capitanio, U. and Montorsi, F. (2016). Renal cancer. *The Lancet*, 387(10021):894–906.

Chung, L. M., Ferguson, J. P., Zheng, W., Qian, F., Bruno, V., Montgomery, R. R., and Zhao, H. (2013). Differential expression analysis for paired RNA-Seq data. *BMC bioinformatics*, 14:110.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13.

Fabio, L. C., Paula, G. A., and de Castro, M. (2012). A Poisson mixed model with nonnormal random effect distribution. *Computational Statistics and Data Analysis*, 56(6):1499–1510.

Fitzmaurice, G., Laird, N., and Ware, J. (2012). *Applied Longitudinal Anal-*

ysis (Wiley Series in Probability and Statistics). Wiley.

Frank, I., Blute, M. L., Cheville, J. C., Lohse, C. M., Weaver, A. L., and Zincke, H. (2002). An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: the SSIGN score. *The Journal of urology*, 168(6):2395–400.

Hardcastle, T. J. and Kelly, K. A. (2013). Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics*, 14(1):135.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):e1003118.

Liu, Q. and Pierce, D. (1994). A note on Gauss—Hermite quadrature. *Biometrika*.

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.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10.

Molenberghs, G., Verbeke, G., Demétrio, C. G. B., and Vieira, A. M. C. (2010). A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects. *Statistical Science*, 25(3):325–347.

Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–773.

Pham, T. V. and Jimenez, C. R. (2012). An accurate paired sample test for count data. *Bioinformatics*, 28(18):i596–i602.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95.

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

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

Spies, D. and Ciaudo, C. (2015). Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Computational and Structural Biotechnology Journal*, 13:469–477.

Strle, K., Stupica, D., Drouin, E. E., Steere, A. C., and Strle, F. (2014). Elevated Levels of IL-23 in a Subset of Patients With Post-Lyme Disease Symptoms Following Erythema Migrans. *Clinical Infectious Diseases*,

58(3):372–380.

Sun, Y., Zhang, J., and Ma, L. (2014). α-catenin. A tumor suppressor beyond adherens junctions. *Cell cycle*, 13(15):2334–9.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*.

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.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560.

A Supplementary materials

Abbreviations

AGHQ – Adaptive Gauss-Hermite quadrature

DE – differential expression

DGE – differential gene expression

EM – expectation maximization

iid – independent with identical distributions

GG – generalized gamma

GHQ – Gauss-Hermite quadrature

GLG – generalized log-gamma GLMM – generalized linear mixed model(s)

LOTUS – Law of the unconscious statistician

ML – maximum likelihood

mRNA – messenger RNA

MVG – multivariate gamma

MVNB – Multivariate Negative Binomial

NB – Negative binomial

OLS – Ordinary Least Squares

pdf – Probability Density Function

pmf – Probability Mass Function

RE – Random Effect(s)

rv – random variable

sd – Standard Deviation

SE – Standard Error

mRNA-Seq – mRNA sequencing

TMM – Trimmed Mean of M-values

A.1 Results

The main results of some simulation scenarios are provided in this section. The output of all simulations, including the tables below can be found in the supplementary file S3.xlsx. Parameter names refer to Model (6)-(8). For each scenario, label from the file file S3.xlsx is followed by scenario number from the main text and then by number of clusters (of 2 for paired experiment or of 3 for clustered experiment).

Table S1. Results for the scenario's of a matched pair RNA-Seq experiment
trueVal – true value of the parameter; estim – point estimate of the parameter (mean value of 10,000 point estimates); absBias – absolute bias; RelBias – relative bias; SEmean – Mean value of estimated standard errors; SEmedian – Median of estimated standard errors; SEemp – Empirical standard error; CICover – percentage of confidence intervals which include true value; SECIC – standard error of confidence interval coverage, power – percentage of confidence intervals which do not contain zero

	losig; Scenario (4); 12 pairs									
	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-1.803	-0.193	0.120	3.626	0.291	0.704	0.993	0.001	0.978
log ϕ	4.000	4.243	0.243	0.061	0.579	0.539	0.592	0.983	0.001	0.999
β_0	0.500	0.500	0.000	0.000	0.074	0.073	0.072	0.941	0.002	1.000
β_1	0.120	0.120	0.000	0.001	0.062	0.061	0.060	0.938	0.002	0.495
	LOSIG; Scenario (5); 24 pairs									
	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-1.673	-0.064	0.040	0.363	0.196	0.262	0.979	0.001	0.998
log ϕ	4.000	4.115	0.115	0.029	0.364	0.358	0.370	0.954	0.002	1.000
β_0	0.500	0.499	-0.001	-0.002	0.052	0.051	0.051	0.946	0.002	1.000
β_1	0.120	0.120	0.000	0.000	0.043	0.043	0.043	0.939	0.002	0.794
	rho9; Scenario (11); 12 pairs									
	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-2.914	-1.305	0.811	126.623	0.735	2.346	0.963	0.002	0.557
log ϕ	2.000	2.217	0.217	0.108	0.415	0.437	0.428	0.924	0.003	1.000
β_0	0.500	0.493	-0.007	-0.015	0.128	0.128	0.127	0.941	0.002	0.963
β_1	1.000	0.998	-0.002	-0.002	0.154	0.154	0.156	0.933	0.003	1.000
	RHO9; Scenario (12); 24 pairs									
	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-2.467	-0.858	0.533	120.791	0.495	1.923	0.954	0.002	0.708
log ϕ	2.000	2.101	0.101	0.051	0.285	0.295	0.296	0.926	0.003	1.000
β_0	0.500	0.497	-0.003	-0.007	0.089	0.089	0.088	0.948	0.002	1.000
β_1	1.000	0.999	-0.001	-0.001	0.109	0.109	0.109	0.941	0.002	1.000

rho9 36; Scenario (13); 36 pairs

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-2.167	-0.557	0.346	36.996	0.391	1.543	0.951	0.002	0.806
log ϕ	2.000	2.063	0.063	0.031	0.232	0.238	0.234	0.935	0.002	1.000
β_0	0.500	0.497	-0.003	-0.005	0.072	0.072	0.072	0.946	0.002	1.000
β_1	1.000	0.999	-0.001	-0.001	0.089	0.089	0.088	0.949	0.002	1.000

nullsd4; Scenario (8); 12 pairs

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-0.916	-1.019	-0.103	0.112	0.379	0.240	0.260	0.945	0.002	0.998
log ϕ	4.000	4.257	0.257	0.064	0.597	0.546	0.640	0.983	0.001	0.998
β_0	0.500	0.497	-0.003	-0.005	0.126	0.126	0.123	0.939	0.002	0.967
β_1	0.000	0.000	0.000	NA	0.062	0.062	0.060	0.935	0.002	0.065

sd 4 losig; Scenario (9); 12 pairs

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-0.916	-1.011	-0.095	0.103	0.245	0.240	0.247	0.946	0.002	0.998
log ϕ	4.000	4.260	0.260	0.065	0.593	0.542	0.619	0.982	0.001	0.999
β_0	0.500	0.500	0.000	0.000	0.127	0.126	0.123	0.941	0.002	0.969
β_1	0.120	0.120	0.000	-0.003	0.062	0.061	0.060	0.939	0.002	0.494

sd4 rho9; Scenario (14); 12 pairs

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-0.916	-1.288	-0.372	0.406	12.806	0.340	1.186	0.992	0.001	0.846
log ϕv	2.000	2.192	0.192	0.096	0.443	0.447	0.452	0.928	0.003	0.999
β_0	0.500	0.493	-0.007	-0.014	0.165	0.163	0.162	0.944	0.002	0.839
β_1	1.000	1.001	0.001	0.001	0.158	0.158	0.156	0.935	0.002	1.000

null rho9; Scenario (10); 12 pairs

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-3.070	-1.460	0.907	475.250	0.749	2.627	0.957	0.002	0.547
log ϕ	2.000	2.214	0.214	0.107	0.420	0.443	0.426	0.929	0.003	1.000
β_0	0.500	0.493	-0.007	-0.013	0.128	0.128	0.126	0.944	0.002	0.963
β_1	0.000	0.000	0.000	NA	0.155	0.155	0.156	0.930	0.003	0.070

rho99; Scenario (15); 12 pairs

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-3.510	-1.901	1.181	276.854	2.629	2.651	0.908	0.003	0.186
log ϕi	1.000	1.219	0.219	0.219	0.373	0.380	0.385	0.924	0.003	0.940
β_0	0.500	0.473	-0.027	-0.054	0.198	0.196	0.189	0.950	0.002	0.674
β_1	1.000	1.000	0.000	0.000	0.254	0.254	0.253	0.936	0.002	0.968

Table S2. Results for the scenario's of a longitudinal RNA-Seq experiment
trueVal – true value of the parameter; estim – point estimate of the parameter (mean value of 10,000 point estimates); absBias – absolute bias; RelBias – relative bias; SEmean – Mean value of estimated standard errors; SEmedian – Median of estimated standard errors; SEemp – Empirical standard error; CICover – percentage of confidence intervals which include true value; SECIC – standard error of confidence interval coverage, power – percentage of confidence intervals which do not contain zero

null rho9; Scenario (8); 8 clusters

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-3.015	-1.405	0.873	110.383	0.816	2.471	0.993	0.001	0.539
log ϕ	2.000	2.225	0.225	0.112	0.436	0.450	0.384	0.951	0.002	1.000
β_0	0.500	0.491	-0.009	-0.018	0.176	0.175	0.152	0.967	0.002	0.815
β_1	0.000	-0.003	-0.003	NA	0.217	0.216	0.191	0.962	0.002	0.038
β_2	0.000	-0.002	-0.002	NA	0.217	0.216	0.190	0.964	0.002	0.036

lsrho9; Scenario (9); 8 clusters

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-3.006	-1.396	0.867	265.258	0.812	2.454	0.993	0.001	0.543
log ϕ	2.000	2.229	0.229	0.114	0.434	0.448	0.382	0.951	0.002	1.000
β_0	0.500	0.491	-0.009	-0.018	0.176	0.175	0.151	0.969	0.002	0.815
β_1	0.150	0.150	0.000	0.002	0.216	0.216	0.188	0.963	0.002	0.091
β_2	0.250	0.251	0.001	0.006	0.216	0.216	0.190	0.962	0.002	0.198

null sd04; Scenario (10); 8 clusters

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-0.916	-1.066	-0.150	0.163	0.452	0.332	0.328	0.960	0.002	0.944
log ϕ	4.000	4.283	0.283	0.071	0.585	0.551	0.536	0.990	0.001	0.999
β_0	0.500	0.501	0.001	0.001	0.174	0.172	0.151	0.955	0.002	0.820
β_1	0.000	-0.002	-0.002	NA	0.086	0.086	0.074	0.964	0.002	0.036
β_2	0.000	-0.001	-0.001	NA	0.086	0.086	0.075	0.959	0.002	0.041

lssd04; Scenario (11); 8 clusters

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-0.916	-1.067	-0.151	0.165	0.408	0.331	0.328	0.957	0.002	0.947
log ϕ	4.000	4.281	0.281	0.070	0.565	0.539	0.518	0.983	0.001	1.000
β_0	0.500	0.499	-0.001	-0.002	0.173	0.171	0.149	0.957	0.002	0.825
β_1	0.150	0.150	0.000	-0.002	0.085	0.085	0.073	0.963	0.002	0.418
β_2	0.250	0.251	0.001	0.003	0.085	0.084	0.074	0.963	0.002	0.850

soft; Scenario (7); 8 clusters

	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC	power
log σ	-1.609	-1.826	-0.217	0.135	1.584	0.374	0.621	0.997	0.001	0.985
log ϕ	4.000	4.238	0.238	0.060	0.505	0.497	0.453	0.963	0.002	1.000
β_0	0.500	0.499	-0.001	-0.003	0.102	0.101	0.088	0.963	0.002	0.998
β_1	1.000	1.000	0.000	0.000	0.084	0.083	0.072	0.964	0.002	1.000
β_2	1.500	1.500	0.000	0.000	0.083	0.083	0.072	0.966	0.002	1.000

A.2 Example of a significant gene

Here we consider a gene which is significantly differentially expressed in the RCC paired data (Section 3.3).

CTNNA1 gene has 18 exons and 27 protein coding transcripts. This gene encodes a protein called α-catenin, which part of protein complex in the junctions of neighboring cells. It is generally considered as a tumor-supressing agent, although it's role in metastases is not fully established Sun et al. (2014). As Figure S1 indicates, CTNNA1 is actually over-expressed in metastatic RCC. The Y axis presents the normalized counts (counts divided by length and effective library size). Distribution of each exon is depicted with a boxplot, separately for each group. We can observe that on average express of each exon is higher in the patients with metastases (on the right) in comparison to the same exon in metastases-free patients (on the left). Note that variability of all exons with a group is larger than a variability of any particular exon across conditions.

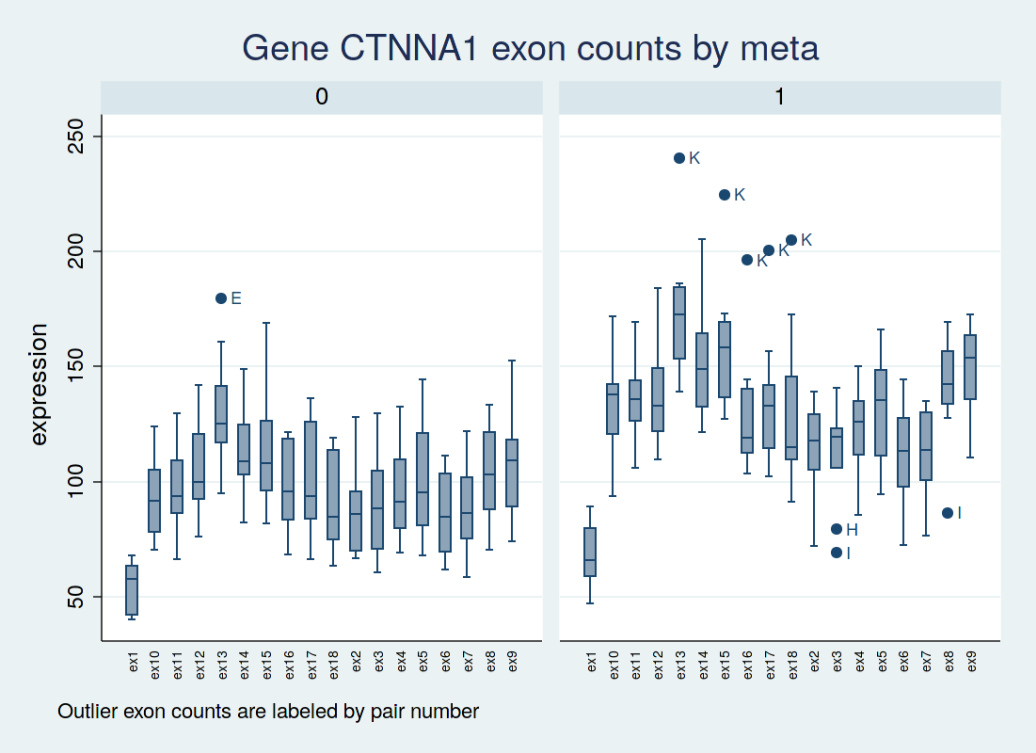

Figure S1. Normalized exon counts of gene CTNNA1. There are two boxblots for each exon. Exons of metastases-free patients are shown on the left, and exons of patients with metastases are shown on the right.On the Y axis there are normalized counts (counts divided by length and effective library size). Effective library size is defined as a product of TMM normalization factor and sample's library size

A.3 Simulation of a paired experiment without true differential expression

In this section we provide detailed results of one representative set of simulations of a paired experiment for "scenario (2)" presented in Tables 1 and 3. We simulated the RNA-Seq data from 24 pairs of patients. The patients in the same pair shared the random effect b_p. Data were generated such that true value of β_1 in (6) was equal to zero. The results of the simulation are provided in the Table S3.

param	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICoverage	SECIC
$\ln \sigma$	-1.609	-1.674	-0.064	0.040	0.245	0.197	0.237	0.979	0.001
$\ln \phi$	4.000	4.108	0.108	0.027	0.367	0.361	0.371	0.954	0.002
β_0	0.500	0.499	-0.001	-0.001	0.052	0.051	0.051	0.948	0.002
β_1	0.000	0.000	0.000	NA	0.043	0.043	0.043	0.943	0.002

Table S3. Summarized analysis of 10,000 replicates of artificial RNA-Seq experiment with known true values for one gene with 3 exons. param – parameter name; trueVal – true value of the parameter (used in data generation); estim – point estimate of the parameter (mean value of 10,000 point estimates, obtained by applying the model to the generated data); absBias – absolute bias in parameter estimation; RelBias – relative bias in parameter estimation; SEmean – Mean value of estimated standard errors of the parameter in the simulation; SEmedian – Median of estimated standard errors of the parameter; SEemp – Empirical standard error (standard deviation of 10,000 point estimates) ; CICover – confidence interval coverage (percentage of confidence intervals which include true value); SECIC – standard error of confidence interval coverage

The main parameter of interest is β_1, because it reflects log-fold-change between conditions (after taking into account correlation between patients within a pair). Point estimates of β_1 are virtually without bias (Table S3). Distribution of β_1 estimates is well-centered around 0 (Figure S3).

Maximum likelihood standard errors tend to underestimate true standard errors. As a result, the null hypothesis is rejected too often. In order to solve this issue we considered multiplying the ML standard errors by an inflation factor equal to $n/(n-2)$, where n was the number of pairs and 2 was the number of βs (see Section A.5.1 for more details). After the adjustment, the percentage of CIs for β_1, which included the true value, was equal to 0.943 with standard error of 0.002 (Table S3). Thus, it was less than 95%, though slightly.

In this simulation the null hypothesis that $\beta_1 = 0$ was true. Hence, the probability of Type-I error could be calculated as a percentage of CIs which did not include zero. That value is equal to one minus confidence interval

coverage, and it is slightly more that 5%.

The distribution of β_0-estimates is fairly close to normal and there is no substantial bias, despite a small dip near the mean value on the density plot (Figure S2).

Unlike distributions of βs, distributions of $\ln\sigma$ and $\ln\phi$ are skewed (Figures S4 and S5) and their estimates biased by 4% and 3% respectively. It is our conjecture that estimates for association parameters still converge to normal distribution, though more slowly than parameters for the mean.

The model-based mean standard error of $\ln\sigma$ was notably larger than the model-based median standard error, and the model-based standard errors in four of the simulated experiments were extremely large and equal to 21, 91, 95, 111. In all four cases the point estimates of $\ln\sigma$ were small, so σ was estimated as 0.004 or less. More details on the relationship between σ estimates and their SE are given in Section A.4.

It is worth noting that there is a non-negligible correlation (see Table S4) between point estimates of parameters, which capture correlation between $\ln\sigma$ and $\ln\phi$ and, separately, between the estimates of β_0 and β_1.

To sum up, this example of simulation demonstrates that the parameter of interest β_1 is estimated virtually without any bias.

	lnsig	lnphi	beta0	beta1
lnsig	1.00	0.21	-0.03	0.01
lnphi	0.21	1.00	-0.03	-0.01
beta0	-0.03	-0.03	1.00	-0.42
beta1	0.01	-0.01	-0.42	1.00

Table S4. Correlation matrix of the parameter estimates in the simulation of paired experiment without true differential gene expression

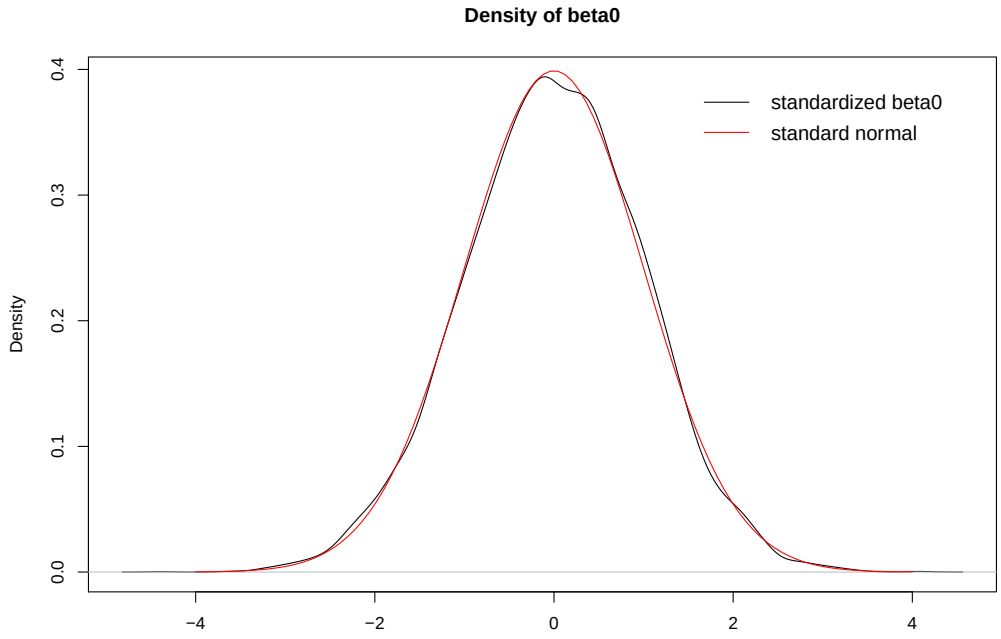

Figure S2. Density plot for the estimate of β_0 in the simulation of paired experiment without true differential gene expression

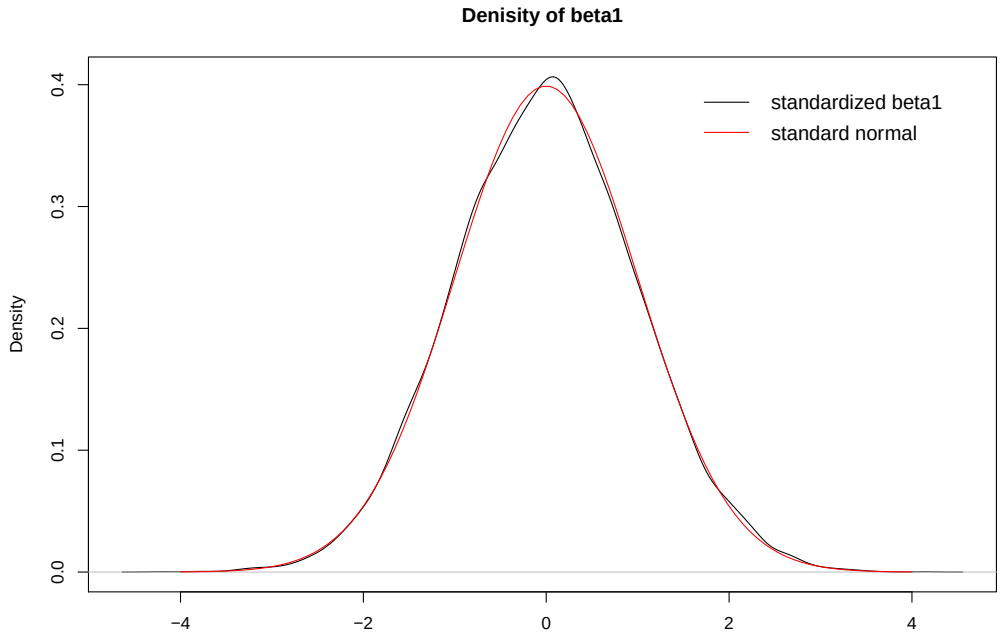

Figure S3. Density plot for the estimate of β_1 in the simulation of paired experiment without true differential gene expression

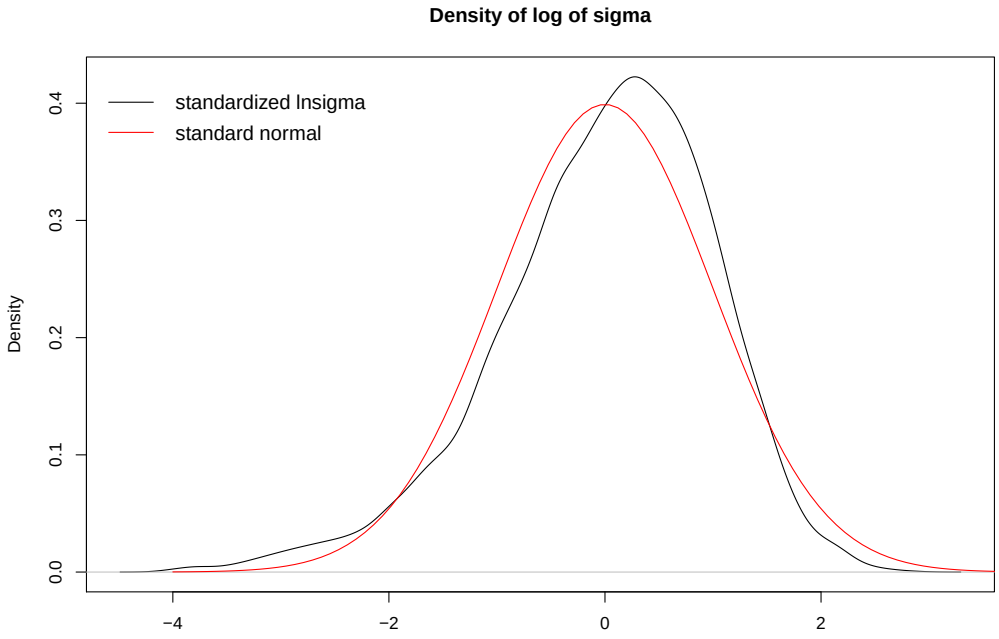

Figure S4. Density plot for the estimate of log of normal random effect σ in the simulation of paired experiment without true differential gene expression

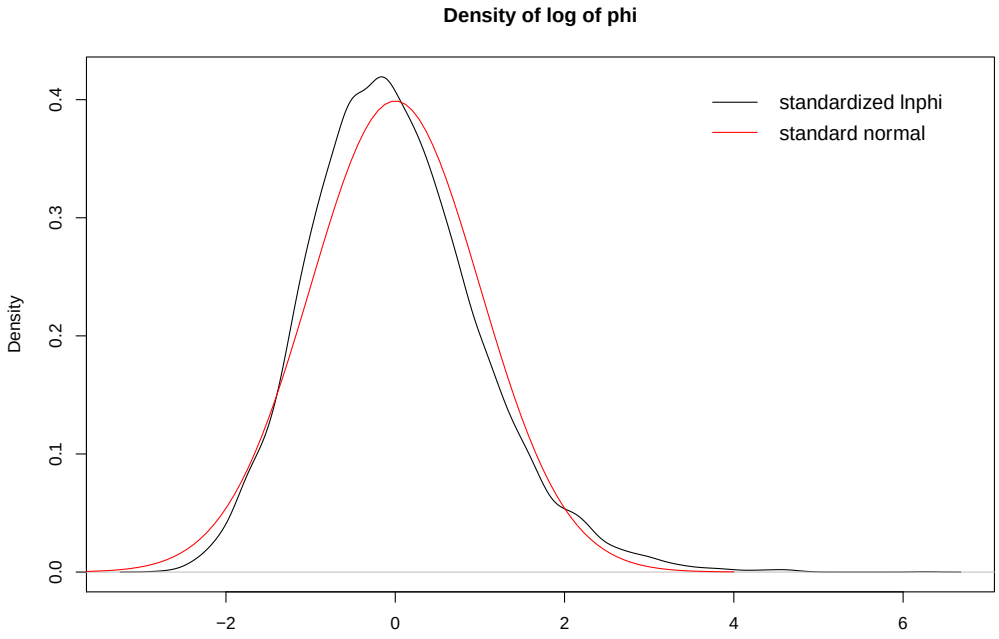

Figure S5. Density plot for the estimate of $\log \phi$ in the simulation of paired experiment without true differential gene expression

A.4 SE of $\ln\sigma$

In some simulation scenarios with a high conditional correlation $\rho|b$ there was a notable fraction of very large SE of $\ln\sigma$.

Here were focus on scenario (10) in Table 1, which is a representative example of such a case. Table S5 provides results of the simulation of an experiment under the null hypothesis $\beta = 0$ not too dissimilar from the simulation considered in Section A.3. There are two differences between simulations summarized in Tables S3 and S5 though: number of pairs was twice smaller in the latter (24 vs 12) and $\ln\phi$ was twice smaller in the latter (4 vs. 2); all other parameters were the same.

Table S5 demonstrates that the fixed effects and overdispersion parameter ϕ are, on average, estimated reasonably well, but the estimate for the $\ln\sigma$ is nearly twice smaller than true value and the mean SE of that estimate is extremely large.

param	trueVal	estim	absBias	RelBias	SEmean	SEmedian	SEemp	CICover	SECIC
$\ln\sigma$	-1.609	-3.070	-1.460	0.907	475.250	0.749	2.627	0.957	0.002
$\ln\phi$	2.000	2.214	0.214	0.107	0.420	0.443	0.426	0.929	0.003
β_0	0.500	0.493	-0.007	-0.013	0.128	0.128	0.126	0.944	0.002
β_1	0.000	0.000	0.000	NA	0.155	0.155	0.156	0.930	0.003

Table S5. Summarized analysis of 10,000 replicates of artificial RNA-Seq experiment with known true values for one gene with 3 exons. param – parameter name; trueVal – true value of the parameter (used in data generation); estim – point estimate of the parameter (mean value of 10,000 point estimates, obtained by applying the model to the generated data); absBias – absolute bias in parameter estimation; RelBias – relative bias in parameter estimation; SEmean – Mean value of estimated standard errors of the parameter in the simulation; SEmedian – Median of estimated standard errors of the parameter; SEemp – Empirical standard error (standard deviation of 10,000 point estimates) ; CICover – confidence interval coverage (percentage of confidence intervals which include true value); SECIC – standard error of confidence interval coverage

Maximal SE of $\ln \sigma$ is larger than 10^6 and 23% of $\ln \sigma$ estimates are smaller than -6. As Figure S6 demonstrates that standard error of $\ln \sigma$ gets extremely large if estimate of $\ln \sigma$ is very small This trend holds in other simulations with high conditional correlation.

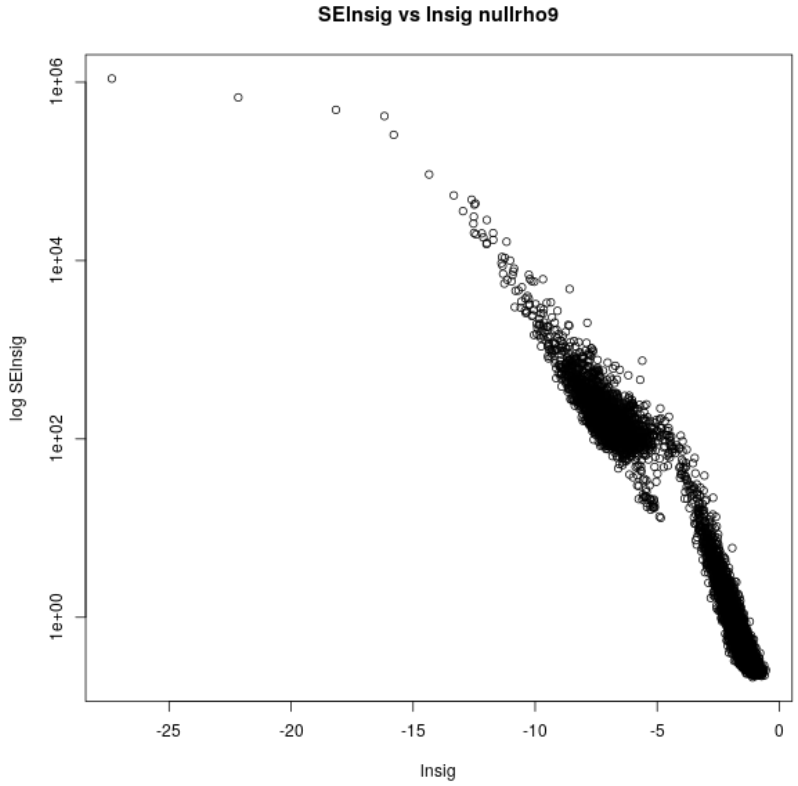

Figure S6. Standard error of $\ln \sigma$ estimate vs $\ln \sigma$ estimate in the simulation of 10,000 replicates of paired RNA-Seq experiment (12 pairs)

We calculated profile likelihood as a function of $\ln \sigma$ for an experiment with SE of $\ln \sigma$ equal to more than 10^6 as follows by fixing possible values of $\ln \sigma$ and computing the ML estimates of all other parameters. The profile likelihood is presented in Figure S7. The result of the full maximization of the likelihood function over all parameters (including $\ln \sigma$) is shown as a red

dot. It is worth noting that the ML estimate of $\ln \sigma$ could vary depending on a software and hardware, but in any case $\ln \sigma$ always stayed in the flat part of the graph (and σ was estimated as approximately equal to zero). Thus, the issue with estimation of $\ln \sigma$ can be explained by the fact that maximum of the likelihood function is located at the boundary of the parameter space for σ. As the likelihood function is flat in the proximity of the $\ln \sigma$ point estimate (red dot on Figure S7), the second partial derivative with respect to $\ln \sigma$ is close to zero, and its inverse is large. That is why the estimate of SE of $\ln \sigma$ is large.

We checked several other experiment-replicates with large values of SE of $\ln \sigma$, and their profile likelihood plots were nearly identical in shape to Figure S7. In each case the point estimate for σ was close to zero. Because of the shape of the likelihood function, the actual point estimate of $\ln \sigma$ in each case is to some extent a random draw between roughly -30 and -6.

The relationship between SE of $\ln \sigma$ and $\ln \sigma$ on Figure S6 may reflect asymptotic behavior of likelihood function at $\sigma = 0$. Starting from $\ln \sigma = -15$ and looking to the right, SEs of $\ln \sigma$ are less and less extreme. Moreover, there is a decreasing linear trend, despite the fact that in the region where $\ln \sigma$ is less than -6, likelihood function is presumably flat, similar to Figure S7.

Of note, the profile likelihood plots for experiments with non-zero σ estimates have a parabolic-like shape and do not contain flat regions on the proximity of ML estimates.

It is interesting to understand why in the simulations, presented in Table S3, there were only four cases of zero estimates of σ (Section A.3), whereas

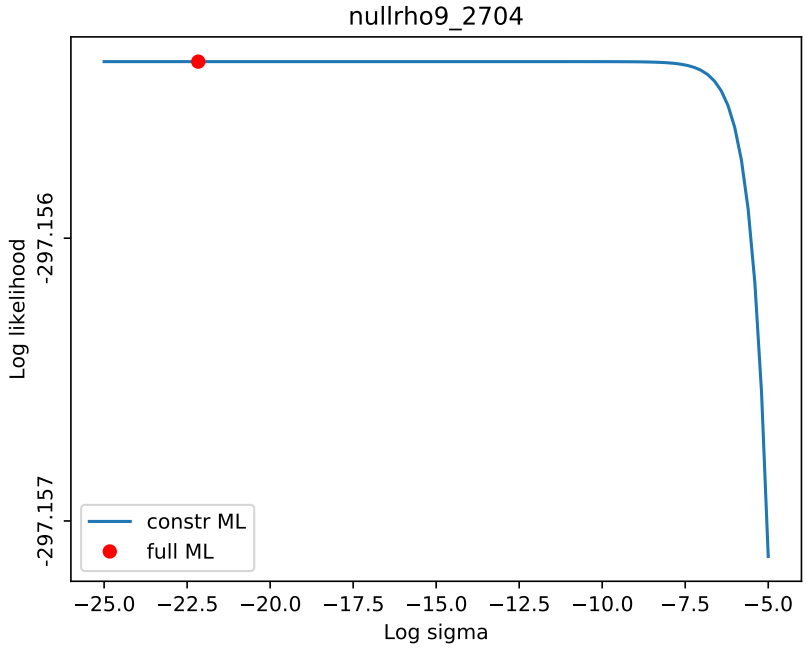

Figure S7. Likelihood as a function of $\ln \sigma$ for an experiment-replicate with large SE of $\ln \sigma$

there were more that 23 hundred such cases in the simulations presented in Table S5. In the supplementary file `S3.xlsx`, these simulations are labeled as `NULL24p` and `nullrho9`, respectively.

According to the model assumptions there are three sources of variation: the gamma random effect, the normal random effect, and the Poison variation conditional on the random effects.

In simulations `NULL24p`, the value of $\ln \phi$ is equal to 4, so the variance of the gamma random effect is approximately equal to 0.02 (see Equation (5)). In simulations `nullrho9`, the value of $\ln \phi$ is twice smaller, so the variance of gamma random effect becomes almost seven times larger and is equal to approximately 0.14. Hence, in the latter case it becomes much more difficult

to discern the normal-random-effect contribution to the overall variability. The problem is exacerbated by the small sample size; there are only 12 pairs in simulations `nullrho9`, twice less than in simulations `NULL24p`.

As shown in supplementary file `S3.xlsx`, increasing the number of pairs or increasing the magnitude of σ notably reduces (but does not eliminate completely) the percentage of zero σ estimates in the simulations with small value of ϕ.

A.5 Implementation

A.5.1 Estimation procedure

The initial values for (13) is obtained through the following procedure. The initial value of $\boldsymbol{\beta}$ is obtained by treating the observed exon counts as independent realizations of Poisson-distributed random variables with mean values $\mu_{csj} = K_{csj}$.

The initial value of ϕ is calculated by assuming independence of the exon counts, obtaining moment-based estimates of ϕ_j from (2) for each exon, and setting ϕ to the average of the so-computed ϕ'_js.

The initial value of σ is obtained by calculating the average exon count for each cluster and then by calculating standard deviation of the logarithms of these average counts.

After defining the initial values, $\boldsymbol{\beta}$ and ϕ are updated by fitting the gamma-Poisson model (4)–(5) with mean values $\mu_{csj} = K_{csj}$.

Finally, by using the so-obtained initial values of the parameters, the (logarithm of the) marginal likelihood (13) is approximated with AGHQ and then maximized. Note that, to improve numerical stability, the likelihood is re-parameterized by using $\ln \phi$ and $\ln \sigma$. Standard errors of the estimated parameters is obtained from the inverse of the negative Hessian - which is computed numerically with `numdifftools` - of the logarithm of the marginal likelihood.

The model (6)-(8) was implemented in set of Python scripts which are available at `https://sourceforge.net/projects/dgeee/`.

Maximum-likelihood estimation was conducted by using the L-BFGS-B algorithm was carried out by L-BFGS-B FORTRAN routines via Python package `scipy.optimize.minimize` Zhu et al. (1997). It is essentially a quasi-Newton method which iteratively approximates Hessian, defines quadratic model of the objective function, and minimizes it Nocedal (1980); Zhu et al. (1997). First-order partial derivatives can be approximated numerically. Otherwise, analytical first-order partial derivatives can be provided as an input to `scipy.optimize.minimize` command. Despite the fact that the software actually executes minimization of negative likelihood, we will refer to it with the conventional notion of maximization of likelihood function.

Overdispersion ϕ and standard deviation of normal pair random effect are bounded to be non-negative, so log transformation were used.

Parameter estimation is carried out in three stages:

1. Calculate initial values $\hat{\boldsymbol{\beta}}$, $\widehat{\ln \phi}$ and $\widehat{\ln \sigma}$ using moments.

2. Run the model without normal RE with $\hat{\boldsymbol{\beta}}$, $\widehat{\ln \phi}$ as a starting values;
   if *converged* **then** *update values of* $\hat{\boldsymbol{\beta}}$, $\widehat{\ln \phi}$.

3. Run the full model starting from $\hat{\boldsymbol{\beta}}$, $\widehat{\ln \phi}$ and $\widehat{\ln \sigma}$.

The very first step, initial value generation, is outlined in Section A.5.1 Next, a simplified model without the normal random effects is applied to the data. Each sample is treated as independent and the model is applied to the data disregarding the clustered nature of the data [function `NREIndivLL` in `stvgen.Stvgen` class]. So, b_c is set to zero for each cluster at this step.

In that simplified model, the log-likelihood of the data is the sum of individual logs where each log is given by

$$\ln P(\boldsymbol{y}_{cs}|b) = \ln\Gamma(\phi + \sum_j y_{csj}) - \ln\Gamma(\phi) - \sum_j \ln\Gamma(y_{csj}+1) - \phi \ln Q_{cs} - \sum_j y_{csj} \ln Q_{cs} + \sum_j y_{csj} \ln(\mu_{csj}/\phi) \quad (16)$$

where

$$Q_{cs} = 1 + \frac{\sum_j \mu_{csj}}{\phi}$$

Strarting values (will all the steps listed above) are generated by `stvgen.py` script.

L-BFGS-B maximization of log likelihood is conducted with analytical first derivatives [function `NREIndiv_der` in `stvgen.Stvgen` class].

$$\frac{\partial}{\partial \beta_l} \ln P(\boldsymbol{y}_{cs}|b=0) = x_l \frac{\partial}{\partial b} \ln P(\boldsymbol{y}_{cs}|b)|_{b=0}$$

$$\frac{\partial}{\partial\phi}\ln P(\boldsymbol{y}_{cs}|b=0)=\psi(\phi+\sum_j y_{csj})-\psi(\phi)-\ln Q_{cs}+\frac{(\phi+\sum_j y_{csj})\sum_j \mu_{csj}}{Q_{cs}\phi^2}-\frac{\sum_j y_{csj}}{\phi}$$

where $\psi(\cdot)$ is a digamma function and

$$\frac{\partial}{\partial b_c}\ln P(\boldsymbol{y}_{cs}|b_c)=\sum_j y_{csj}-\frac{(\phi+\sum_j y_{csj})}{Q_{cs}}\sum_j \xi_{csj} \tag{17}$$

where, in turn $\xi_{csj}=\mu_{csj}/\phi$) Thus, at step 2 of Algorithm **??** maximum likelihood point estimates generated [function `SetStartVals` in `stvgen.Stvgen` class].

Next, the full model (6)-(8) is applied to the data. L-BFGS-B algorithm is used for likelihood function maximization of the log of likelihood function (13) [function `RUNminimizer` in `minimizer.Minimizer` class]. First-order partial derivatives of likelihood function are calculated numerically. Total log likelihood for data is summarized from pair log likelihoods by `negLLcalculator.NegLLcalculator` class. At each iteration during the maximization process (16) needs to be evaluated for each pair. Integrals in (13) are approximated by (A.6), which in turn requires calculation of mode of the integrand $\hat{\mu}$ and posterior standard deviation of the integrand $\hat{\sigma}$. Of note, this terminology owes to Bayesian origins of (A.6); Liu and Pierce invented approximation of $\hat{\mu}$ and $\hat{\sigma}$, which is used in (A.6) in this work and in SAS procedure NLMIXED Liu and Pierce (1994). Note on notation: traditional $\hat{\mu}$ and $\hat{\sigma}$ AGHQ notation is not related to μ_{cij} and σ in Model (6)-(8).

Functions, related to adaptive quadrature, are located in `adapt.Adapt` class. As explained in the section A.6, mode of $g(t)$ is calculated by maximizing

$\ln g(t)$. $\ln g(t)$ is maximized inside `PosteriorMuFUN` function with the use of analytical first derivatives; thus,posterior mode $\hat{\mu}$ is obtained. Function `PosteriorSigmaHatFun` calculates posterior standard deviation $\hat{\sigma}$. $\hat{\mu}$ and $\hat{\sigma}$ rely, in turn, upon log of probability of exon counts in individual given normal random effect (see section A.6). $\ln P_{i|c}(\boldsymbol{y}_{ci}|b_c)$ is calculated by function `IndLLgREFUN` in `IndegrandFUN.IndegrandFUN` class. Second derivative of $\ln g(t)$, which is required for calculation of $\hat{\sigma}$, is obtained by the function `Indiv_SecDer`.

After ML point estimates are obtained, Hessian matrix is numerically evaluated at ML point estimates (function `CalcHessian` in `minimizer.Minimizer` class) with python package `numdifftools`. ML estimates of standard errors are calculated as square roots of diagonal of inverse of Hessian. Standard errors reported by the our software are unadjusted. It is important to adjust them, because ML estimates of standard errors underestimate true standard error (see eg section 4.5 in Fitzmaurice et al. (2012)). In this work we use factor $\frac{N}{N-p}$ to inflate standard errors, where n is the number of clusters and p is the number of fixed effects (in other words, number of betas). Inflation of ML-based SE of point estimates is performed in *R*, *outside* the python scripts.

On the contrary, inside calculated variance-covariance matrix, which is used in composite Wald test, SE is inflated by the factor $\frac{N}{N-p}$. Calculation of Wald test statistic is performed inside `CalcHessian` function after line `CalcHessian function`.

There is a subtle technical detail in numerical evaluation of the Hessian:

(16) needs to be approximated many times as a function of parameters and then as a function of parameters with tiny increments (like in mathematical definition of derivative as a limit). The downside of an evaluation with adaptive Gaussian quadrature is that it is computationally demanding. So, exact approximation of (16) at each function call at Hessian calculation be inefficient. We applied the following workaround. The integrand in (16) changes very little (we've checked that) if we calculate (16) as a function of a parameter x or a function of $x + \Delta x$. Therefore, it is plausible to apply the same posterior mode $\hat{\mu}$ to approximate the integrand in both cases. So, before `CalcHessian` function is invoked $\hat{\mu}$ are calculated at ML point estimates and stored for each pair (function `MuHatExtractor` in `minimizer.Minimizer` class).

A.5.2 Nonconvergence criteria

Nonconvergence per se can occur due to failure in maximizing likelihood function. In the case where ML estimates are successfully obtained, standard errors might not be calculated when either Hessian is non-ivertible or there are negative values in the inverse of diagonal of Hessian. If standard error for parameter of interest (e.g. β_1 in paired data example) was not calculated, then this gene considered as not converged despite point estimates were calculated successfully. It is also possible to set an arbitrary cut-off in the number of exons. If number of exons in a gene exceeds cut-off value then calculations of point estimates is skipped all together and this gene is considered as not converged. This cut-off can be set to arbitrary value in

`init` function of `minimizer.Minimizer` class.

A.6 AGHQ quadrature

We have used AGHQ approximation, proposed by Liu and Pierce Liu and Pierce (1994), for the evaluation of the integral (13).

$$\int_{-\infty}^{\infty} g(b)db \approx \sqrt{2}\hat{\sigma} \sum_{k=1}^{d} w_k e^{x_k^2} g(\hat{\mu} + \sqrt{2}\hat{\sigma}x_k) \tag{18}$$

where $\hat{\mu}$ is the mode of $g(t)$, x_k and w_k are standard Gauss-Hermite evaluation points and weights, $\hat{\sigma} = 1/\sqrt{\hat{j}}$ and

$$\hat{j} = -\frac{\partial^2}{\partial t^2} \ln g(b)\Big|_{b=\hat{\mu}}$$

Note, that we've preserved original notation from the work of Liu and Pierce Liu and Pierce (1994) in ; $\hat{\mu}$ and $\hat{\sigma}$ are not related to μ and σ in (6)-(8).

The requirements for (A.6) to work are that $g(b) > 0$ and that ratio of $g(b)$ to some Gaussian curve be a moderately smooth function. Product of likelihood function and a Gaussian density (as in (13)) meets these requirements Liu and Pierce (1994). Mathematical machinery behind (A.6) is based on the first order Laplacian approximation Tierney and Kadane (1986); Rabe-Hesketh et al. (2002).

Posterior mode of $g(b)$:

We find the mode of $g(b)$ in (A.6) by maximizing $\ln g(b)$. As follows from

(13),

$$g(b_c) = f(b_c) \prod_{s=1}^{n_c} P_{s|c}(\boldsymbol{y}_{cs}|b_c)$$

and

$$\ln g(b_c) = \ln f(b_c) + \sum_{s=1}^{n_c} \ln P_{s|c}(\boldsymbol{y}_{cs}|b_c) \tag{19}$$

Partial derivative of (19) with respect to b_c is the sum of the partials for the following individual terms.

$$\frac{\partial \ln f(b_c)}{\partial b_c} = \frac{\frac{\partial f(b_c)}{\partial b_c}}{f(b_c)} = \frac{\frac{-b}{\sigma^2} f(b_c)}{f(b_c)} = \frac{-b_c}{\sigma^2} \tag{20}$$

and $\frac{\partial}{\partial b_c} \ln P_{c|s}(\boldsymbol{y}_{cs}|b_c)$ is given in (17).

Second partial derivatives are required to calculate $\hat{\sigma}$ in (A.6).

$$\frac{\partial^2}{\partial b_c^2} \ln P_{s|c}(\boldsymbol{y}_{cs}|b_c) = -\frac{(\phi + \sum_j y_j) \sum_j \xi_j}{Q^2} \tag{21}$$

Avoiding numerical underflow :

$P_{c|s}(\boldsymbol{y}_{cs}|b_c)$ is a probability that exon counts of all the exons in the particular gene is equal to some particular numbers. This probability alone is numerically an extremely small number, but in this case the issue is even more serious, because integrand in (13) contains product of such probabilities $P_{c|s}$. Obvious overcome an issue like this is to work on log scale. Technically it is possible to calculate in such a case $\ln g(b_p)$, however the exponent of this this number $\exp \ln g(b_p)$ would be less than a smallest positive floating point number in the computer system. The output of this calculation would be

zero To avoid this underflow we can divide and multiply the integral by some large number M.

$$\begin{aligned} \int g(b)db &= M \int \frac{g(b)}{M} db \\ \ln \int g(b)db &= \ln g(\hat{\mu}) + \ln \int \exp(\ln g(b) - \ln g(\hat{\mu}))db \end{aligned} \tag{22}$$

$\hat{\mu}$ and $\hat{\sigma}$ in the integrand in (22) remain the same as in (A.6) due to

$$\frac{\partial}{\partial t} \ln kg(t) = \frac{1}{\cancel{k} g(t)} \cancel{k} \frac{\partial}{\partial t} g(t) = \frac{\partial}{\partial t} \ln g(t)$$

Note, that the integral in (22) is approximated by weighted sum of some array, where each array member is a result of applying some function of Gauss-Hermite locations. Therefore, it seems natural to set $\ln M$ to the maximum of that array. That standard technique is known as log-sum-exp formula (equation 16.1.9 in Press et al. (2007)).

$$\log \left(\sum_i \exp(z_i) \right) = z_{\max} + \log \left(\sum_i \exp(z_i - z_{\max}) \right)$$

A.7 Existing methods: PLNSeq

In this section, we briefly discuss a `PLNseq` method which we have applied to the data for the purposes of comparison with our method.

Zhang et al. Zhang et al. (2015) proposed a model (termed `PLNseq`) for the analysis of matched-pair RNA-Seq data, which can also be applied in a longitudinal setting.

Let X_{ctg} denote read-count for cluster c, condition t, and gene g. Denote by Z_{ctg} the corresponding gene-expression level. It is assumed that, conditionally on Z_{ctg}, counts X_{ctg} for different conditions $t = 1, \ldots, T$ are independent and follow a Poisson distribution:

$$X_{ctg}|Z_{ctg} = z_{ctg} \sim \text{Poisson}(k_{ct} z_{ctg}),$$

where k_{ct} is the library-size-related normalizing factor. On the other hand, the vector of the logarithms of the gene-expression levels,

$$\widetilde{\boldsymbol{Z}}_{cg} = (\ln Z_{c,1g}, \ldots, \ln Z_{cTg})'$$

follows a multivariate normal distribution:

$$\boldsymbol{Z}_{cg} \sim MVN(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}_g = (\mu_{1g}, \ldots, \mu_{Tg})'$ and $\boldsymbol{\Sigma}$ is a $T \times T$ variance-covariance matrix with condition-specific variances σ^2_{tg} and correlation coefficients $\rho_{t_1,t_2,g} = \text{Corr}\{\ln(Z_{t_1,g}), \ln(Z_{t_2,g})\}$.

Note that, in practice, the `PLNseq` method assumes that $\sigma^2_{tg} \equiv \sigma^2_g$ and that $\rho_{t_1,t_2,g} \equiv \rho_{t_1,t_2}$.