



Universiteit  
Leiden  
The Netherlands

## Statistical method for modeling sequencing data from different technologies in longitudinal studies with application to Huntington disease

Fuady, A.M.; Roon-Mom, W.M.C. van; Kielbasa, S.M.; Uh, H.W.; Houwing-Duistermaat, J.J.

### Citation

Fuady, A. M., Roon-Mom, W. M. C. van, Kielbasa, S. M., Uh, H. W., & Houwing-Duistermaat, J. J. (2020). Statistical method for modeling sequencing data from different technologies in longitudinal studies with application to Huntington disease. *Biometrical Journal*, 63(4), 745-760. doi:10.1002/bimj.201900235

Version: Publisher's Version  
License: [Creative Commons CC BY-NC 4.0 license](#)  
Downloaded from: <https://hdl.handle.net/1887/3184673>

**Note:** To cite this publication please use the final published version (if applicable).

## RESEARCH PAPER

# Statistical method for modeling sequencing data from different technologies in longitudinal studies with application to Huntington disease

Angga M. Fuady<sup>1,3</sup>  | Willeke M. C. van Roon-Mom<sup>2</sup> | Szymon M. Kiełbasa<sup>1</sup> | Hae-Won Uh<sup>3</sup> | Jeanine J. Houwing-Duistermaat<sup>3,4</sup>

<sup>1</sup> Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

<sup>2</sup> Department Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

<sup>3</sup> Department of Biostatistics and Research Support, Div. Julius Centrum, University Medical Center Utrecht, Utrecht, the Netherlands

<sup>4</sup> Department of Statistics and Alan Turing Institute, University of Leeds, Leeds, United Kingdom

## Correspondence

Angga M. Fuady, Department of Biomedical Data Sciences, Leiden University Medical Center, Postzone S-5-P, PO Box 9600, 2300 RC Leiden, the Netherlands. Email: [A.M.Fuady@lumc.nl](mailto:A.M.Fuady@lumc.nl)

## Funding information

Indonesian Endowment Fund for Education (LPDP); H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 721815; FP7 Health, Grant/Award Numbers: 305121, 305280



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## Abstract

Advancement of gene expression measurements in longitudinal studies enables the identification of genes associated with disease severity over time. However, problems arise when the technology used to measure gene expression differs between time points. Observed differences between the results obtained at different time points can be caused by technical differences. Modeling the two measurements jointly over time might provide insight into the causes of these different results. Our work is motivated by a study of gene expression data of blood samples from Huntington disease patients, which were obtained using two different sequencing technologies. At time point 1, DeepSAGE technology was used to measure the gene expression, with a subsample also measured using RNA-Seq technology. At time point 2, all samples were measured using RNA-Seq technology. Significant associations between gene expression measured by DeepSAGE and disease severity using data from the first time point could not be replicated by the RNA-Seq data from the second time point. We modeled the relationship between the two sequencing technologies using the data from the overlapping samples. We used linear mixed models with either DeepSAGE or RNA-Seq measurements as the dependent variable and disease severity as the independent variable. In conclusion, (1) for one out of 14 genes, the initial significant result could be replicated with both technologies using data from both time points; (2) statistical efficiency is lost due to disagreement between the two technologies, measurement error when predicting gene expressions, and the need to include additional parameters to account for possible differences.

## KEYWORDS

DeepSAGE, linear mixed model, measurement error, quality control, RNA-Seq

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

## 1 | INTRODUCTION

A longitudinal study design comprises multiple observations from each patient over time. This design offers efficiency gain by using multiple observations instead of one per patient. When recruiting more patients is not feasible, this might be the only way to obtain sufficient statistical power to assess the relationship between response and covariables. Rapid advances in sequencing technology may bring a researcher into a situation where the type of technology to measure the response variable has changed over time. For example, in recent years, sequence-based approaches to quantify gene expression levels became available and affordable for large studies. Compared to hybridization-based approaches, the newer technologies enable us to detect novel transcripts, have lower background noise, provide a broad dynamic range, and have high technical reproducibility (’t Hoen et al., 2008; Wang, Gerstein, & Snyder, 2009). Gene expression levels can be quantified by sequencing and counting mRNA fragments (RNA-Seq) or tags (DeepSAGE) (Ozsolak & Milos, 2011; Zhernakova et al., 2013). Although the two techniques provide information on the number of transcripts in each sample, in RNA-Seq, the reads will typically be aligned across the entire transcript, while with DeepSAGE all tags will be aligned with the 3' ends of the transcript.

Our motivating example is a study of the association between gene expression measured from blood samples and a disease severity indicator for Huntington disease (HD). HD is an autosomal dominant neurodegenerative disorder characterized by progressive motor symptoms (Jones & Hughes, 2011; Kent, 2004; Mastrokolias et al., 2015; van der Burg, Björkqvist, & Brundin, 2009; Walker, 2007). In a cross-sectional study, Mastrokolias et al. (2015) assessed the association between gene expression and the motor score representing disease severity, while adjusting for age, gender, and relative cell counts represented by the proportion of hemoglobin tags to the total tags per samples. They used linear regression with DeepSAGE as a dependent variable. They identified 167 genes for which gene expression was associated with the motor score. Of these genes, 20 genes were confirmed by RT-qPCR. To further replicate these findings, 3 years later, follow-up samples were obtained and measured with RNA-Seq technology. In addition, a subset of the samples at the first time point was measured by RNA-Seq, that is, for these samples, measurements of both techniques are available. Unfortunately, the analysis of the RNA-Seq measurements at the second time point did not confirm the first findings. Reasons for this lack of replication might be the difference in technologies used or false positiveness or negativeness of the first or the second findings. Hence, a joint longitudinal model for both measurements might provide a better understanding of the associations between gene expressions measured with the two platforms and the motor score (disease indicator) over time. For the genes identified by Mastrokolias et al. (2015), the association between DeepSAGE-measured gene expression and the severity of HD using the data at both time points was modeled. Moreover, we performed all gene analysis using the RNA-Seq measurement as an outcome at both time points. To deal with unobserved measurements, we used measurement error models.

Much research in the field of measurement error model has been performed. However, most work has been done on the error-in-covariates problem (Buonaccorsi, 2010; Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Gustafson, 2004). In general, measurement error in the response variable might increase the variability of the fitted value; hence the statistical power to detect the true effects is decreased (Abrevaya & Hausman, 2004; Carroll et al., 2006). In linear regression, the additional variability induced by the measurement error will be absorbed by the residual variance. Measurement error in response variables, therefore, is often ignored. However, incorporating measurement error variance in the model might yield an efficiency gain.

When a calibration set is available for a subset of the samples, in which both the response variable subject to measurement error and the true response variable are measured, estimates of the true relationship can be obtained. Buonaccorsi (1991) proposed a moment estimator to combine these two sources of information. They assume that the observed and the true response are related via a simple linear regression model. The method of moments was used for estimation and provided unbiased parameter estimates. Alternatively, one can consider the maximum likelihood estimation. For nonlinear relationships, the pseudo maximum likelihood approach was used to estimate the parameter. Here the estimated error variance was obtained from a calibration set and plugged into the final likelihood function (Buonaccorsi, 1996; Buonaccorsi & Tosteson, 1993). Keogh, Carroll, Tooze, Kirkpatrick, and Freedman (2016) considered both the maximum likelihood method and the method of moments to estimate the relationship between food intake and an intervention. Here the measurement error of the response variable, namely food intake based on the questionnaire, might depend on the intervention, and no valid statistical inference can be performed by using only the questionnaire data. Also, the authors have biomarker information, representing the outcome variable of interest in a small subset. This subset provides information about the relationship between the true response and the questionnaire, which can be used to adjust the responses with

**TABLE 1** Characteristics of the study participants at both time points

Characteristics	Baseline ( <i>N</i> = 124)	Follow-up ( <i>N</i> = 73)
Age, mean (SD)	47.98 (12.09)	52.53 (11.28)
Gender, <i>n</i> (%)		
Female	67 (54.03)	40 (54.79)
Male	57 (45.97)	33 (45.21)
Motor score, mean (SD)	20.36 (24.92)	32.50 (30.81)
Carrier status, <i>n</i> (%)		
Symptomatic	64 (51.61)	55 (75.34)
Presymptomatic	29 (23.39)	18 (24.66)
Control	31 (25)	0 (0)

regard to the questionnaire in the larger set. For this situation, they showed that the gain in efficiency by using the method of moments and maximum likelihood method was similar.

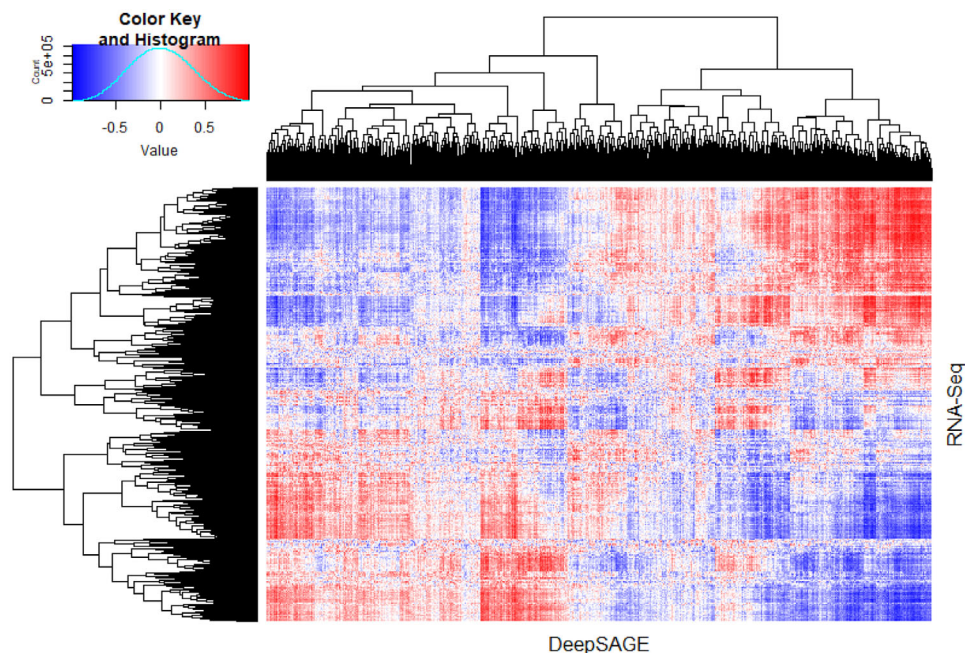
In this paper, we propose a pseudo-likelihood approach which is a combination of the method of moments and the maximum likelihood approach. Like Keogh et al. (2016), we use a linear model to assess the relationship between the two gene expression measurements. Prior to applying the linear model, we select genes for which their expressions are similar by both technologies using a mixture model (De Veaux, 1989; McLachlan & Peel, 2000). Here, two latent classes represent samples with correlated and uncorrelated measurements of the two platforms. The obtained estimates from the linear model are plugged into a linear mixed model (McCulloch and Searle, 2000; Verbeke & Molenberghs, 2000), fitted by maximizing the log-likelihood function over the remaining parameters. The combination of methods of moments and maximum likelihood enables us to process a large number of genes (methods of moments) and modeling the relationship over time (maximum likelihood estimation).

The remainder of the paper is organized as follows. In Section 2, the motivating dataset is described in detail. Notations and statistical methods are formulated in Section 3. In Section 4, a small simulation study is performed. The results of the data application are presented in Section 5. We offer some conclusion and discussion in Section 6.

## 2 | DATA

Gene expressions of HD patients from two different sequencing technologies are available, namely DeepSAGE and RNA-Seq, at two time points. At the first time point, 124 samples were measured by DeepSAGE technology. Additionally, 22 samples from the first time point were measured by RNA-Seq technology. At the second time point, 73 samples were measured by RNA-Seq. For each sample at each time point, the motor section of the Unified Huntington's Disease Rate Scale was used to obtain a severity indicator of HD, namely the motor score. For both platforms at both time points, 19,711 genes were identified. For the 22 overlapping samples from the first time point, 16,798 genes were available. The characteristics of the study participants for both time points are shown in Table 1. Note that the controls are only measured once. At the second time point, there are more symptomatic cases compared to presymptomatic cases.

Prior to performing the analysis, preprocessing steps are needed: removing samples with low detected genes, removing genes having a low- or high-abundance count, and normalizing the count data. From the 22 overlapping samples, we removed three samples because of a small number of detected genes. The corresponding RNA-Seq samples at time point 2 for these genes were also removed. In the DeepSAGE dataset, these three deleted samples have less than 5,000 reads detected, while other samples have a minimum of 5 million reads. Note that in contrast to this paper, these samples were included in the previous study. Thus, we ended up analyzing 121 samples at time point 1 and 70 samples at time point 2. Following the approach of the initial analysis (Mastrokolias et al., 2015), the top three overabundant genes were also removed, that is, *HBA1*, *HBA2*, and *HBB*. Genes having less than one count per million reads in at least three samples in RNA-Seq measurement were removed, while in DeepSAGE, genes with a minimum of 10 counts per million in at least three samples were selected. After quality control, we have the expressions of 5,079 genes. For these genes, Figure 1 shows the heatmap correlation plot of DeepSAGE and RNA-Seq using the data from the overlapping samples. The two measurements appeared to be highly correlated for most of the genes.



**FIGURE 1** Heatmap correlation of DeepSAGE and RNA-Seq measurements based on data from the overlapping samples at time point 1. Plot is based on 5,079 genes selected after the preprocessing step

DeepSAGE produces one read per gene transcript, whereas RNA-Seq generates multiple reads per gene transcript proportionally to transcript length. To create a comparable scale, we divided the reads in RNA-Seq by transcript length. The transcript length is defined as the sum of gene exon lengths. For each sample in each technology, the total number of reads is different. The trimmed mean of M-values (TMM) provides normalization for the total number of reads, which is implemented in the R package edgeR (Robinson, McCarthy, & Smyth, 2010; Robinson & Oshlack, 2010). TMM with no length correction is the count per million, while TMM with a length correction is the reads per kilobase of transcript per million mapped reads (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008). To estimate the overdispersion in the count data, the function `fitdistr` in the R package MASS (Venables & Ripley, 2002) was used. For all analyses, we used  $\log_2$  transformation of normalized counts. The estimated dispersion parameter for DeepSAGE data was ranging from 0.492 to 6.437, with an average of 3.160.

### 3 | METHODS

Let  $I$  be the number of subjects. Let  $X_{ij}$  and  $Z_{ij}$  be the log-transformed counts measured by DeepSAGE and RNA-Seq. The first subscript denotes the subject  $i = 1, \dots, I$ , while the second denotes the time point  $j = 1, 2$ . Further, let  $ms_{ij}$  be the motor score of subject  $i$  at time point  $j$ .

To model the relationship between gene expression and the motor score in a similar way as the previous paper (Mastrokoulas et al., 2015), we consider the following model:

$$\begin{aligned} X_{i1} &= \beta_0 + \beta_2 ms_{i1} + u_i + \epsilon_{i1}, \\ X_{i2} &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) ms_{i2} + u_i + \epsilon_{i2}, \end{aligned} \quad (1)$$

where  $u_i \sim \mathcal{N}(0, \sigma_u^2)$  is the subject-specific random effect and  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$  the random error variable. Here,  $\beta_2$  and  $\beta_2 + \beta_3$  represent the effect of the motor score on gene expression at time point 1 and at time point 2, respectively. Note, we assume that given the information on the time point 2, the information available for time point 1 is independent of the gene expression on the time point 2. For notational simplicity, if possible, we drop the subject index  $i$ .

We first introduce three subsamples of indices based on the availability of sequencing datasets for subject  $i$ . The first subsample,  $S_1$ , consists of the set of indices of the subjects measured by DeepSAGE at the first time point, that is,  $X_{i1}$  is

observed, and by RNA-Seq at the second time point, that is,  $Z_{i2}$  is observed. The second subsample,  $S_2$ , denotes the set of indices of subjects measured by DeepSAGE at the first time point only, that is, only  $X_{i1}$  is observed. The third subsample,  $S_3$ , indicates the set of indices of subjects measured by RNA-Seq at the second time point only, that is, only  $Z_{i2}$  is observed. Finally, we define  $S_1^s \subset S_1$  as the set of indices for which we also have RNA-Seq at the first time point. The sample size for  $S_1$ ,  $S_1^s$ ,  $S_2$ , and  $S_3$  are  $n_1 = 53$ ,  $n_1^s = 19$ ,  $n_2 = 68$ , and  $n_3 = 17$ , respectively.

Our challenge here is that  $X_{i2}$  is unobserved and  $Z_{i1}$  is only observed for a small subset  $S_1^s$ . We propose therefore to obtain an estimate for  $X_{i2}$ . We use the data from the subsample  $S_1^s$  to relate the DeepSAGE to the RNA-Seq measurements. We assume that for each gene there might be two groups of samples: the first group represents samples for which the gene expressions of the two technologies are highly correlated, and the second group represents samples for which there seems no relationship between the gene expressions measured by the two technologies. Specifically, let  $\alpha$  be the proportion of the samples for which gene expression is well-measured by both technologies, and  $\rho$  be the correlation coefficient between  $X_1$  and  $Z_1$ . For  $i \in S_1^s$ , the joint distribution of  $X_1$  and  $Z_1$  is given as follows:

$$m(\mathbf{x}_1, \mathbf{z}_1) = \alpha g(\mathbf{x}_1, \mathbf{z}_1) + (1 - \alpha)f(\mathbf{x}_1)f(\mathbf{z}_1), \tag{2}$$

where  $g(\mathbf{x}_1, \mathbf{z}_1)$  is the bivariate normal distribution  $\mathcal{N}_{BVN}(\mu_{x_1}, \mu_{z_1}, \sigma_{x_1}^2, \sigma_{z_1}^2, \rho)$ , and  $f(\mathbf{x}_1) \sim \mathcal{N}(\mu_{x_1}, \sigma_{x_1}^2)$  and  $f(\mathbf{z}_1) \sim \mathcal{N}(\mu_{z_1}, \sigma_{z_1}^2)$  are two marginal distributions. Define  $\Theta = (\mu_{x_1}, \mu_{z_1}, \sigma_{x_1}^2, \sigma_{z_1}^2, \rho, \alpha)$ . The corresponding likelihood is given by

$$\begin{aligned} \mathcal{L}(\Theta|\mathbf{x}_1, \mathbf{z}_1) &= \prod_{i \in S_1^s} [\alpha g(\mathbf{x}_1, \mathbf{z}_1) + (1 - \alpha)f(\mathbf{x}_1)f(\mathbf{z}_1)] \\ &= \prod_{i \in S_1^s} [\alpha(1/2)\pi^{-1}|\Sigma_1|^{-1/2} \exp\{-(1/2)(\mathbf{h}_i - \boldsymbol{\mu})^T \Sigma_1^{-1}(\mathbf{h}_i - \boldsymbol{\mu})\} \\ &\quad + (1 - \alpha)(1/2)\pi^{-1}|\Sigma_2|^{-1/2} \exp\{-(1/2)(\mathbf{h}_i - \boldsymbol{\mu})^T \Sigma_2^{-1}(\mathbf{h}_i - \boldsymbol{\mu})\}], \end{aligned}$$

where  $\mathbf{h}_i = (\mathbf{x}_1, \mathbf{z}_1)^T$ ,  $\boldsymbol{\mu} = (\mu_{x_1}, \mu_{z_1})^T$ ,  $\Sigma_1 = \begin{bmatrix} \sigma_{x_1}^2 & \rho\sigma_{x_1}\sigma_{z_1} \\ \rho\sigma_{x_1}\sigma_{z_1} & \sigma_{z_1}^2 \end{bmatrix}$ , and  $\Sigma_2 = \begin{bmatrix} \sigma_{x_1}^2 & 0 \\ 0 & \sigma_{z_1}^2 \end{bmatrix}$ . For each gene, the data are fitted by the model.

We are interested in genes for which the measurements from both technologies are highly correlated and select genes with  $\hat{\rho} \geq .8$  and  $\hat{\alpha} \geq .8$ . For these genes, we assume that two variables representing DeepSAGE and RNA-Seq follow a bivariate normal distribution with mean  $\boldsymbol{\mu} = (\mu_{x_1}, \mu_{z_1})$  and variance  $\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \rho\sigma_{x_1}\sigma_{z_1} \\ \rho\sigma_{x_1}\sigma_{z_1} & \sigma_{z_1}^2 \end{bmatrix}$ . Thus, we assume that  $\alpha = 1$ . Hence for the selected genes we use all samples. Finally, with regard to the correlation over time, we assume that  $\rho_t = \text{corr}(X_1, X_2) = \text{corr}(Z_1, Z_2)$  and the correlation between two different technologies at two different time points is modeled by  $\text{corr}(X_1, Z_2) = \text{corr}(X_2, Z_1) = \rho \times \rho_t$ . Thus the covariance matrix for  $(X_1, Z_1, X_2, Z_2)$  can be written as

$$\Sigma_{(X_1, Z_1, X_2, Z_2)} = \begin{bmatrix} \sigma_{x_1}^2 & \rho\sigma_{x_1}\sigma_{z_1} & \rho_t\sigma_{x_1}^2 & \rho_t\rho\sigma_{x_1}\sigma_{z_2} \\ \rho\sigma_{x_1}\sigma_{z_1} & \sigma_{z_1}^2 & \rho_t\rho\sigma_{x_2}\sigma_{z_2} & \rho_t\sigma_{z_2}^2 \\ \rho_t\sigma_{x_1}^2 & \rho_t\rho\sigma_{x_1}\sigma_{z_1} & \sigma_{x_2}^2 & \rho\sigma_{x_2}\sigma_{z_2} \\ \rho_t\rho\sigma_{x_1}\sigma_{z_1} & \rho_t\sigma_{z_1}^2 & \rho\sigma_{x_2}\sigma_{z_2} & \sigma_{z_2}^2 \end{bmatrix}. \tag{3}$$

### 3.1 | Modeling DeepSAGE measurements

Here, DeepSAGE measurement is considered as the dependent variable in the model. However,  $X_2$  is not available at the second time point and need to be estimated. The overlapping samples which contain a pair of DeepSAGE and RNA-Seq at time point 1 can be used to estimate the joint distribution of DeepSAGE and RNA-Seq. Then RNA-Seq is used to predict DeepSAGE at time point 2. Note that we assume that the joint distribution of  $X_2$  and  $Z_2$  equals the bivariate normal distribution of  $X_1$  and  $Z_1$  (see model (3)). Specifically, it is assumed that  $\sigma_{z_1}^2 = \sigma_{z_2}^2 = \sigma_z^2$  and  $\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_x^2$ . The conditional distribution of  $X_2$  given  $Z_2$  is normal with mean  $c + \rho \frac{\sigma_x}{\sigma_z} z_2$  and variance  $\sigma_\omega^2 = (1 - \rho^2)\sigma_x^2$ . Estimates for  $\sigma_z$ ,  $\sigma_x$ , and  $\rho$  are obtained from the results of fitting model (2) using data from  $S_1^s$  under the assumption of  $\alpha = 1$ . Further, we

do not need to estimate the constant  $c$ , since it will be absorbed in the intercept of our final model. Thus, we can substitute  $X_2$  by  $\tilde{X}_2 = \rho \frac{\sigma_x}{\sigma_z} Z_2$  in model (1) to assess the relationship between gene expression and motor score as follows:

$$\begin{aligned} X_{i1} &= \beta_0 + \beta_2 \text{ms}_{i1} + \gamma_1 \mathbf{D}_{i1} + u_i + \epsilon_{i1}, \\ \tilde{X}_{i2} &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \text{ms}_{i2} + \gamma_2 \mathbf{D}_{i2} + u_i + \omega_i + \epsilon_{i2}, \end{aligned} \quad (4)$$

where  $\beta_3$  represents the combination of the change of the motor score effect between the two time points and of the difference between the effect of motor score on the observed DeepSAGE measurement and the estimated one,  $\omega$  represents the measurement error and  $\mathbf{D}_{i1}$  and  $\mathbf{D}_{i2}$  might be additional covariates included in the model with effect size  $\gamma_1$  and  $\gamma_2$ , respectively. Let  $\Psi = (\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \sigma_u, \sigma_\epsilon)$ . For  $i \in S_1$ , the corresponding log-likelihood is given by

$$\ell(\Psi|Y_i, \text{ms}) = -(1/2) \ln(|\Sigma|) - (1/2) \{(\mathbf{y}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\mu})\} + \text{const},$$

where  $Y_i = [X_{i1}, \tilde{X}_{i2}]$ ,  $\boldsymbol{\mu} = (\beta_0 + \beta_2 \text{ms}_{i1} + \gamma_1 \mathbf{D}_{i1}, (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \text{ms}_{i2} + \gamma_2 \mathbf{D}_{i2})^\top$ ,  $\Sigma = \begin{bmatrix} \sigma_{x_{i1}}^2 & \rho \sigma_{x_{i1}} \sigma_{\tilde{x}_{i2}} \\ \rho \sigma_{x_{i1}} \sigma_{\tilde{x}_{i2}} & \sigma_{\tilde{x}_{i2}}^2 \end{bmatrix}$  with  $\sigma_{x_{i1}}^2 = \sigma_u^2 + \sigma_\epsilon^2$  and  $\sigma_{\tilde{x}_{i2}}^2 = \sigma_u^2 + \sigma_\epsilon^2 + \sigma_\omega^2$ . For the second and third subsamples, the log-likelihood reduces to univariate normal distributions. Finally, the total log-likelihood for the full model is given by

$$\begin{aligned} \sum_{i \in S_1 \cup S_2 \cup S_3} \ell(\Psi|Y_i, \text{ms}_i) &= \sum_{i \in S_1} \ell(\Psi|Y_i, \text{ms}_i) + \sum_{i \in S_2} \ell(\Psi_1|X_{i1}, \text{ms}_{i1}) \\ &\quad + \sum_{i \in S_3} \ell(\Psi|\tilde{X}_{i2}, \text{ms}_{i2}), \end{aligned} \quad (5)$$

where  $\Psi_1 = (\beta_0, \beta_2, \gamma_1, \sigma_u, \sigma_\epsilon)$ . Maximum likelihood estimates of the unknown parameters  $\Psi$  are obtained using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm of `optim` function in R (R Core Team, 2013). This algorithm is a quasi-Newton method that implements approximation of the Hessian matrix using a specified gradient evaluation. The ratio of the subject-specific effect variance to the total variance at the first time point represents the intraclass correlation between DeepSAGE measurements at the two time points if they would have been observed ( $\text{corr}(X_1, X_2)$ ). The ratio of the measurement error variance to the total variance at the second time point represents the dissimilarity between the two measurements, indicating how well DeepSAGE can be constructed by RNA-Seq.

Finally, to assess the statistical significance of the relationship between the motor score and gene expression, the null hypothesis of  $H_0 : \beta_2 = 0$  should be tested. The standard likelihood ratio test, which follows a  $\chi^2$  distribution with one degree of freedom, is used.

### 3.2 | Modeling RNA-Seq measurements

Considering RNA-Seq as a dependent variable allows us to use observed measurements for time point 1 ( $Z_1$ ) and time point 2 ( $Z_2$ ). Based on the overlapping samples, we predict the RNA-Seq from  $X_1$  when  $Z_1$  is missing.

We model the relationship between gene expression represented by RNA-Seq and the motor score as follows:

$$\begin{aligned} Z_{i1} &= \beta_0 + \beta_2 \text{ms}_{i1} + u_i + \epsilon_{i1}, \\ \tilde{Z}_{i1} &= (\beta_0 + \beta_5) + (\beta_2 + \beta_4) \text{ms}_{i1} + u_i + \omega_i + \epsilon_{i1}, \\ Z_{i2} &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \text{ms}_{i2} + u_i + \epsilon_{i2}, \end{aligned} \quad (6)$$

where  $u_i \sim \mathcal{N}(0, \sigma_u^2)$  indicates the subject-specific random effect for RNA-Seq measurement at both time points,  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$  represents the random error variable, and  $\omega_i \sim \mathcal{N}(0, \sigma_\omega^2)$  express the measurement error of predicting RNA-Seq from DeepSAGE. Further,  $\beta_2$  represents the effect of the motor score on the observed measured RNA-Seq,  $\beta_3$  represents the change in the effect of the motor score on the RNA-Seq at the second time point with respect to the first time point, and  $\beta_4$  represents the difference in the effect of the motor score on the predicted RNA-Seq with respect to the observed RNA-Seq.

Since RNA-Seq are partly available at two time points, we predict the unobserved RNA-Seq at the first time point using the overlapping samples of DeepSAGE and RNA-Seq at the same time point. Subsamples  $S_1$  and  $S_2$  are divided into two parts, namely a subset with observed measurement of RNA-seq ( $Z_{i1}$  or  $Z_{i2}$ ) and predicted RNA-seq ( $\hat{Z}_{i1}$ ), respectively. The sample size of the observed and predicted measurements of RNA-Seq in  $S_1$  is 17 and 36, respectively. For  $S_2$ , the sample size for true and predicted measurement of RNA-Seq is 2 and 66, respectively. Finally, the sample size of the true measurement of RNA-Seq for  $S_3$  is 17.

Note that when using RNA-Seq as a dependent variable, the assumption that the joint distribution of  $X_1$  and  $Z_1$  equals the joint distribution of  $X_2$  and  $Z_2$  is not needed because the prediction of RNA-Seq is performed at the same time point as the overlapping samples. Another advantage of using RNA-Seq measurements as a dependent variable is that the difference between the motor score effect on observed RNA-Seq and predicted RNA-Seq could be investigated. Just as for DeepSAGE, we test the null hypothesis of  $\mathcal{H}_0 : \beta_2 = 0$  to assess the relationship between the motor score and the RNA-Seq by using the likelihood ratio test.

## 4 | SIMULATION STUDY

### 4.1 | Simulation setting

Two sets of simulations were conducted to evaluate the robustness of the proposed methods. We focus on the model for DeepSAGE. The first simulation aims to investigate how well our method can predict DeepSAGE from RNA-Seq measurements, assuming a normal distribution for count data. The second simulation aims to study the impact of assuming normal distributions for counts on the size of the likelihood ratio statistic for testing the hypothesis of no association between motor score and gene expression using the model (4). A small overdispersion parameter  $\theta$  yields a dependency between the mean and variance, and a large  $\theta$  will result in skewed distributed data. Hence for both situations, the assumptions of the test statistic may be violated resulting in too liberal or too conservative  $p$ -values.

### 4.2 | Dataset generation

To study the performance of predicting DeepSAGE from RNA-Seq, we simulated a sample of correlated count data of size 100. For the parameter values, we used the corresponding values of one of the genes in our dataset, namely *CYSTMI*. Means  $\mu_1 = 2.15$  and  $\mu_2 = 1.95$  for the log-count data of DeepSAGE and RNA-Seq, respectively, were used, and the dispersion parameter  $\theta$  was assumed to be 3.5. To obtain correlated counts, we first sampled a vector of normally distributed random effects  $v$  with zero mean and variance  $\sigma_v^2 = 0.9$  of length 100. Then vectors of counts for both outcomes were generated using the negative binomial distribution as follows:  $Y_j \sim NB(r = \exp(\mu_j + v), p = \frac{1/\theta}{1+1/\theta})$  for  $j = 1, 2$ , where  $NB(r, p)$  represents the negative binomial distribution with the number of successful trial  $r$  and probability  $p$ ,  $Y_1$  represents the DeepSAGE, and  $Y_2$  represents the RNA-Seq measurements. The count data were then log-transformed and divided into a training (20% of the whole dataset) and a test set (remaining 80%). In the training set, a linear model with DeepSAGE as a dependent variable and RNA-Seq as independent was fitted. The obtained relationship between DeepSAGE and RNA-Seq was used to predict the DeepSAGE measurements from the RNA-Seq measurements in the test set ( $Y_1^{pred}$ ). The agreement between the distributions of the generated  $Y_1$  and the predicted  $Y_1^{pred}$  will be depicted by a QQ plot based on the data in the test set.

The second simulation aims to evaluate the null distribution of the test statistic for the hypothesis of no association between gene expression and motor score. The sample sizes were  $n_1 = 53$ ,  $n_2 = 68$ , and  $n_3 = 17$ , which are the values for our data example. For each sample, two motor score values were generated from a bivariate standard normal distribution with a correlation of .9 between the two time points. The scenario is described in Table 2, which is based on the *NMT2* gene. For the parameter  $\theta$ , we considered five values, namely 0.0001, 0.15, 3.5, 6.5, or 10.

Let  $\sigma_I^2 = \sigma_u^2 + \sigma_\epsilon^2$ , and  $\sigma_{II}^2 = \sigma_u^2 + \sigma_\epsilon^2 + \sigma_\omega^2$ , random effects  $v$  for subjects in each subsample of size  $(n_1, n_2, n_3)$  were generated as follows: for  $i \in S_1$ , Let  $v_i \sim BVN(0, \begin{bmatrix} \sigma_I^2 & .8\sigma_I\sigma_{II} \\ .8\sigma_I\sigma_{II} & \sigma_{II}^2 \end{bmatrix})$ ,  $\mu_{i1} = \exp(\beta_0 + v_{i1})$ , and  $\mu_{i2} = \exp((\beta_0 + \beta_1) + v_{i2})$ . For  $i \in S_2$ , define  $v_i \sim N(0, \sigma_I^2)$  and  $\mu_{i1} = \exp(\beta_0 + v_{i1})$ . For  $i \in S_3$ ,  $v_{i2} \sim N(0, \sigma_{II}^2)$  and  $\mu_{i2} = \exp((\beta_0 + \beta_1) + v_{i2})$  were considered. Finally, multivariate counts with mean  $\mu_i$  and dispersion parameter  $\theta$  were generated by  $NB(\theta^{-1}\mu_i, \frac{\theta^{-1}}{1+\theta^{-1}})$ . For each  $\theta$ , the number of replicates was 10,000. For each replicate, the DeepSAGE model was fitted. Then, the likelihood ratio statistic and corresponding  $p$ -values were obtained.



**TABLE 2** Simulation results: nominal and actual significance levels for the likelihood ratio test for association between log-transformed count data and a covariate. Different values of the dispersion parameter  $\theta$  were considered

$\theta$	Nominal	Significance level
		Actual (95% CI) <sup>a</sup> { $\beta_0 = 2.459, \beta_1 = -0.4078$ } { $\sigma_u = 0.0044, \sigma_\epsilon = 0.3792, \sigma_\omega = 0.4123$ }
0.0001	.0010	.0016 (.0008 , .0024)
	.0050	.0054 (.0040 , .0068)
	.0100	.0108 (.0088 , .0128)
	.0500	.0527 (.0483 , .0571)
0.15	.0010	.0010 (.0004 , .0016)
	.0050	.0045 (.0032 , .0058)
	.0100	.0111 (.0090 , .0132)
	.0500	.0501 (.0458 , .0544)
3.5	.0010	.0006 (.0001 , .0011)
	.0050	.0054 (.0040 , .0068)
	.0100	.0109 (.0089 , .0129)
	.0500	.0497 (.0454 , .0540)
6.5	.0010	.0004 (.0000 , .0008)
	.0050	.0045 (.0032 , .0058)
	.0100	.0097 (.0078 , .0116)
	.0500	.0476 (.0434 , .0518)
10	.0010	.0016 (.0008 , .0024)
	.0050	.0056 (.0041 , .0071)
	.0100	.0105 (.0085 , .0125)
	.0500	.0512 (.0469 , .0555)

<sup>a</sup>Based on the *NMT2* gene.

R implementations are available in GitHub (<https://github.com/Fuady/DeepSAGE>).

### 4.3 | Simulation results

For the first set of simulations, Figure 2 shows the QQ plot of the observed and predicted DeepSAGE measurement for sample size  $N = 100$ , that is, the size of the test set is 80 samples. It appears that the distribution of the predicted was almost similar to the distribution of the observed DeepSAGE. Hence we conclude that under a similar setting as the dataset, our method can predict DeepSAGE well from RNA-Seq.

The results of the second set of simulations aimed to evaluate the size of the likelihood ratio statistic are presented in Table 2. The actual significance levels for the nominal levels of .001, .005, .01, and .05 and dispersion parameters  $\theta$  of 0.0001, 0.15, 3.5, 6.5, and 10 are given. For a  $\theta$  of 0.0001 or 6.5, the actual size of the likelihood ratio statistic deviates from the nominal size (for a level of .05, the actual levels are .053 and .048, respectively). For the other  $\theta$ s, the actual and nominal significance levels are almost the same. Thus our proposed method performs well when the overdispersion  $\theta$  is between 0.15 and 6.5, which appears to be the range of values in our data. Hence we conclude that the likelihood ratio statistic for testing  $\beta_2 = 0$  in model (4) has the correct size.

## 5 | RESULTS

### 5.1 | Selection of genes

To select genes from the set of 5,079 for which both technologies provide similar measurements, we apply the mixture model (2) to the gene expressions from the 19 overlapping samples at the first time point. It appears that the gene expressions of 811 genes satisfy the criteria of a high correlation between the two measurements in a large proportion of

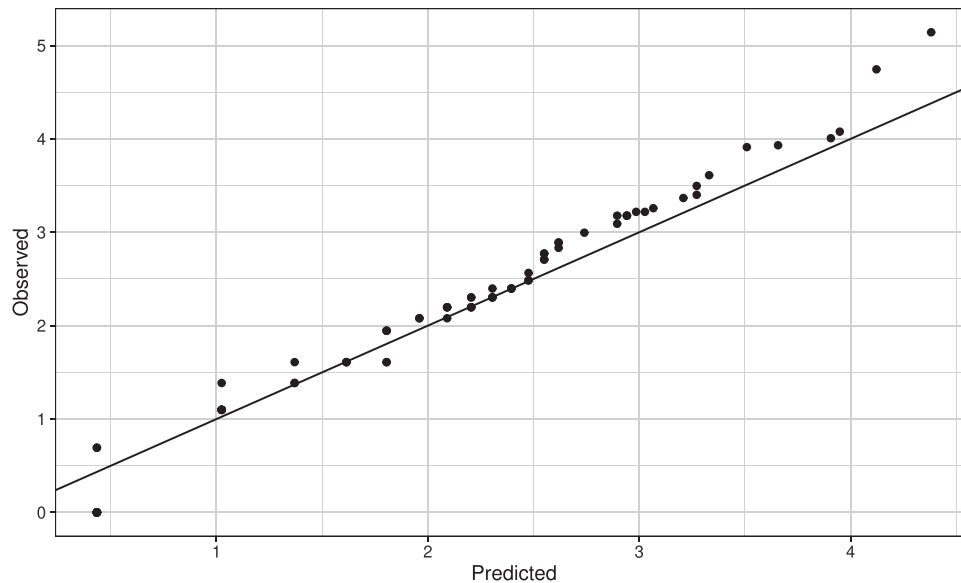


FIGURE 2 QQ plot of  $Y_1^{pred}$  versus  $Y_1$  for the test set of 80 samples

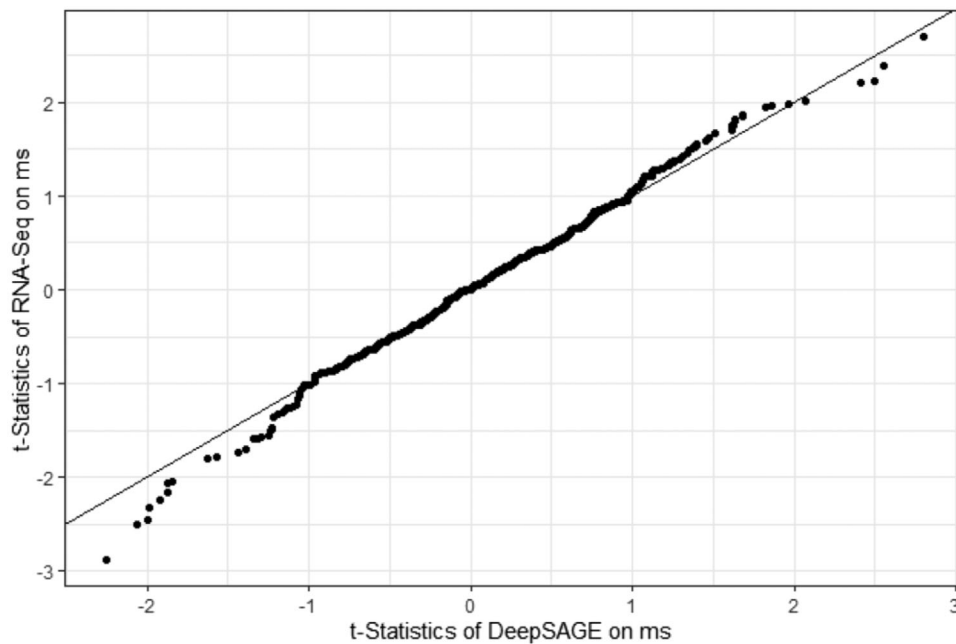


FIGURE 3 The QQ plot of  $t$ -statistics under the null hypothesis of no motor score effect on the DeepSAGE and RNA-Seq measurements in the overlapping samples

the overlapping samples ( $\hat{\alpha} \geq .8$  and  $\hat{\rho} \geq .8$ ). To further investigate the agreement between the two measurements, we studied the distribution of the agreement between the  $t$ -statistics for testing the null hypothesis of no effect of the motor score on the gene expression level measured by the two technologies at the first time point in the overlapping samples. In Figure 3, the QQ plot of the  $t$ -statistics shows that the distributions agree for most of the genes.

## 5.2 | Replication of the previous results

We focus on the genes that have been identified in the previous study by Mastrokolas et al. (2015) using the DeepSAGE data at the first time point. They identified 167 genes for which gene expression was significantly associated with the motor

score after adjusting for age, gender, and relative cell content. The last variable represents the ratio of hemoglobin reads to the total reads per sample. False discovery rate (FDR) was applied to correct for multiple testing. Of these genes, 20 were validated by RT-qPCR. Only 14 of these 167 genes passed our quality control, as described in the previous section. Table 3 shows the estimates of parameters of the mixture model and the linear mixed model for these genes. Also, the following results are presented: the  $p$ -values corresponding to the test for association between gene expression and motor score using only the data of the first time point, the  $p$ -values corresponding to the one degree of freedom test for the null hypothesis of testing  $\beta_2 = 0$  in the mixed model (4) with DeepSAGE measurements as an outcome, and the original  $p$ -values as given in the paper of Mastrokoulas et al. (2015). Note that since we applied a more stringent data cleaning procedure, these latter two  $p$ -values might be different. For the sake of comparisons, we included both the unadjusted and the adjusted  $p$ -values in Table 3. The adjusted  $p$ -values are the FDR-corrected  $p$ -values applying the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) on these 14 genes.

After multiple testing corrections, the gene expressions for all 14 genes are significantly associated with the motor score when using only the data for time point 1 as well as when using the data for both time points. For almost all genes, the  $p$ -values based on the data from two time points are slightly larger than when using only data from the first time point, except for *SIK1*. When considering only data at the first time point, the unadjusted  $p$ -values of our method are similar or lower than the original analysis results for five genes, namely *PTPN4*, *CYSTMI*, *NMT2*, *RASA3*, and *GNPTAB*, reflecting a different quality control compared to the original analysis. The effect sizes of the motor score on gene expression using the data from time point 1 analysis and using all data with the full model are similar. The largest absolute effect size of the motor score on the gene expression at the first time point is for *CYSTMI* with 0.0114.

For eight genes, the intraclass correlation ( $\sigma_u^2 / (\sigma_u^2 + \sigma_\epsilon^2)$ ) is  $\leq .001$ . The largest intraclass correlation is obtained for *CCR2* (.0329). The gene *RHOB* shows the lowest measurement error variation to the total variance ( $\sigma_\omega^2 / (\sigma_\omega^2 + \sigma_u^2 + \sigma_\epsilon^2)$ ) of .1601.

To further study these results, we fitted the linear mixed model (6) with RNA-Seq measurements as the dependent variable. The results are given in Table 4. The correlation between the measurements at time point 1 and 2 was much larger for the RNA-Seq than for the DeepSAGE measurements. After multiple testing corrections, 2 out of 14 genes were statistically significant, namely the genes *RAPGEFL1* and *SIK1*.

When comparing the direction of the estimates of  $\beta_2$  in the model for DeepSAGE with those for RNA-Seq, the majority of estimates have the same direction, except for *KLRG1* and *SIK1*. Note that the expression of *SIK1* is significantly associated with the motor score for both measurements. The  $\beta_2$  parameters in the two models have a slightly different interpretation. In the RNA-Seq model  $\beta_2$ , which is positive, represents the effect of the motor score on the observed RNA-Seq measurements in the overlapping samples, and  $\beta_2 + \beta_4$ , which appears to be negative, represents the effect of the motor score on the predicted RNA-measurements. The  $\beta_2$  in the model for DeepSAGE, which is negative, represents the effect of the motor score on the observed DeepSAGE measurements in the whole sample. Further analyses revealed that the relationship between motor score and DeepSAGE measurements is positive in the overlapping samples and negative in the remaining samples at time point 1. Thus, the two platforms appear to agree for this gene.

Next, we studied the effect of plugging the value for  $\sigma_\omega$  obtained from the model (2) instead of estimating it when fitting the model (4). The results are given in Supplementary Table S1. As expected, the  $\beta_2$  estimates are similar and for most genes the  $p$ -values are slightly larger when plugging in  $\sigma_\omega$  instead of estimating it.

Finally, we checked the effect of including the covariates age, gender, and hemoglobin percentage (HB) in the mixed model (4). The inclusion of additional covariates might be beneficial if the covariates explain a part of the variance; hence inclusion reduces the noise. However, when the covariate is a collider, the results might be biased. To investigate the effect of including these covariates in the model, we have plotted the  $\beta_2$  estimates of the models with and without additional covariates (Supplementary Figure S1) using the data of all genes. It appears that the two estimates are very similar. The average estimated residual variance is smaller when including the covariates (0.554 vs. 0.588); hence there is a small increase in efficiency when including these covariates.

### 5.3 | All gene analysis

We also analyzed all available data without considering the initial results (Mastrokoulas et al., 2015) using gene expression measured by RNA-Seq as a response. Out of 811 genes that satisfy the criteria of  $\hat{\alpha} \geq .8$  and  $\hat{\rho} \geq .8$ , we identified 89 genes for which their expressions were associated with the motor score using the full model. These genes were significant at the .05 level without multiple testing corrections. After multiple testing corrections using FDR, we identified 59 genes with a

TABLE 3 Results of the replication study with DeepSAGE measurements as a dependent variable. Results are sorted based on Mastrokolias et al. (2015) Adjusted  $p$ -value

Gene	Mixture model			Linear model			Full model			Original results							
	$\rho_p^a$	$\rho$	$\alpha$	est	se	$p$ -value <sup>b</sup>	Adj- $p$ -value <sup>b</sup>	est	se	$\frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$ c	$\frac{\sigma_w^2}{\sigma_u^2 + \sigma_w^2 + \sigma_e^2}$ d	$p$ -value <sup>e</sup>	Adj- $p$ -value <sup>e</sup>	$p$ -value <sup>f</sup>	Adj- $p$ -value <sup>f</sup>	est $\theta^g$	
<i>PTPN4</i> <sup>h</sup>	.9440	.9439	.9996	-.0067	.0015	.0000	.0001	-.0067	.0015	.0001	.2130	.0003	.0007	.0000	.0091	.0091	2.5690
<i>CYSTMT</i> <sup>b</sup>	.8510	.8515	.9981	.0114	.0027	.0000	.0002	.0114	.0027	.0130	.2845	.0003	.0007	.0000	.0129	.0129	3.5114
<i>NMT2</i>	.5470	.8496	.8096	-.0083	.0018	.0000	.0001	-.0083	.0016	.0001	.5417	.0000	.0003	.0000	.0129	.0129	2.7230
<i>TLR2</i>	.8424	.9227	.8974	.0078	.0019	.0001	.0003	.0078	.0018	.0003	.3245	.0002	.0007	.0000	.0196	.0196	3.8387
<i>CEP19</i>	.7806	.8527	.8875	.0098	.0026	.0003	.0005	.0098	.0023	.0012	.2694	.0011	.0018	.0001	.0302	.0302	3.2096
<i>CCR2</i>	.8832	.8832	.9999	.0066	.0018	.0003	.0005	.0066	.0016	.0329	.2538	.0009	.0016	.0001	.0324	.0324	3.3441
<i>RAPGEFL1</i>	.8543	.8542	.9994	.0083	.0024	.0006	.0008	.0083	.0023	.0001	.2113	.0042	.0054	.0001	.0324	.0324	3.9099
<i>TLR6</i>	.6130	.8360	.8467	.0074	.0020	.0002	.0005	.0074	.0017	.0000	.3492	.0003	.0007	.0001	.0324	.0324	4.4352
<i>RASA3</i>	.5468	.9989	.9975	-.0057	.0015	.0001	.0004	-.0057	.0013	.0002	.4402	.0003	.0007	.0002	.0338	.0338	3.3716
<i>GNPTAB</i>	.8285	.8286	.9991	-.0055	.0014	.0002	.0004	-.0055	.0013	.0028	.3499	.0006	.0012	.0002	.0340	.0340	3.0003
<i>IFNGR2</i>	.8342	.9135	.8603	.0056	.0016	.0006	.0008	.0056	.0015	.0001	.2779	.0169	.0169	.0003	.0454	.0454	5.0748
<i>RHOB</i>	.8705	.8707	.9997	.0069	.0020	.0007	.0008	.0069	.0019	.0013	.1601	.0112	.0121	.0003	.0458	.0458	4.4057
<i>KLRC1</i>	.8497	.9371	.8893	-.0094	.0029	.0015	.0016	-.0094	.0027	.0272	.2381	.0099	.0116	.0004	.0468	.0468	1.7891
<i>SIK1</i>	.8151	.9051	.8742	-.0062	.0020	.0028	.0028	-.0062	.0019	.0000	.3876	.0013	.0019	.0005	.0498	.0498	2.5768

<sup>a</sup>Correlation between two measurements in the overlapping samples.

<sup>b</sup>Model using DeepSAGE as a dependent variable at time point 1. The  $p$ -value is obtained from testing  $\beta_2 = 0$  using the  $t$ -test. The adjusted  $p$ -value based on 14 genes tested.

<sup>c</sup>The ratio of the subject-specific effect to the total variance at time point 1.

<sup>d</sup>The ratio of the measurement error effect to the total variance at time point 2.

<sup>e</sup>Model using DeepSAGE as a dependent variable at both time points. The  $p$ -value is obtained from testing  $\beta_2 = 0$  using the likelihood ratio test. The adjusted  $p$ -value is based on 14 genes tested.

<sup>f</sup>The  $p$ -value is obtained from Mastrokolias et al. (2015) analysis. The adjusted  $p$ -value is based on 16,657 genes tested.

<sup>g</sup>Estimated dispersion parameter of the DeepSAGE dataset using function `fitdistr` from the R package MASS.

<sup>h</sup>Confirmed by RT-qPCR in Mastrokolias et al. (2015) analysis.

TABLE 4 Results of the replication study with RNA-Seq as a dependent variable using data from both time points

Gene	Full model				p-value <sup>c</sup>	Adj. p-value <sup>c</sup>
	$\beta_2$ est	se	$\frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$ <sup>a</sup>	$\frac{\sigma_\omega^2}{\sigma_u^2 + \sigma_\omega^2 + \sigma_\epsilon^2}$ <sup>b</sup>		
<i>PTPN4</i> <sup>d</sup>	-.0134	.0138	.1895	.1836	.0686	.2619
<i>CYSTM1</i> <sup>d</sup>	.0196	.0211	.1350	.3344	.3530	.6856
<i>NMT2</i>	-.0176	.0099	.4746	.6689	.0748	.2619
<i>TLR2</i>	.0079	.0145	.2469	.3550	.5895	.7177
<i>CEP19</i>	.0205	.0181	.2995	.3057	.2588	.6039
<i>CCR2</i>	.0060	.0130	.2213	.2876	.6430	.7177
<i>RAPGEFL1</i>	.0256	.0150	.4850	.2720	.0000	.0003
<i>TLR6</i>	.0090	.0139	.0000	.3852	.5159	.7177
<i>RASA3</i>	-.0128	.0086	.1969	.5769	.1372	.3842
<i>GNPTAB</i>	-.0066	.0095	.3074	.4071	.4825	.7177
<i>IFNGR2</i>	.0058	.0140	.0849	.2641	.6796	.7177
<i>RHOB</i>	.0127	.0148	.3226	.1810	.3918	.6856
<i>KLRG1</i>	.0070	.0193	.4452	.2754	.7177	.7177
<i>SIK1</i>	.0096	.01s32	.3922	.4639	.0027	.0189

<sup>a</sup>The ratio of the subject-specific effect to the total variance at time point 1.

<sup>b</sup>The ratio of the measurement error effect to the total variance at time point 2.

<sup>c</sup>Model using RNA-Seq as a dependent variable at both time points. The p-value is obtained from testing  $\beta_2 = 0$  using the likelihood ratio test. The adjusted p-value is based on 14 genes tested.

<sup>d</sup>Confirmed by RT-qPCR in Mastrokolias et al. (2015) analysis.

TABLE 5 Results of the top 10 most significant genes when analysing all genes. The motor score effects on the gene expression at time point 1 is represented by  $\beta_2$ 

Gene	Mixture model			Full model				p-value <sup>d</sup>	Adj. p-value <sup>e</sup>
	$\rho_p$ <sup>a</sup>	$\rho$	$\alpha$	$\beta_2$ est	se	$\frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$ <sup>b</sup>	$\frac{\sigma_\omega^2}{\sigma_u^2 + \sigma_\omega^2 + \sigma_\epsilon^2}$ <sup>c</sup>		
<i>ATOH8</i>	.9303	.9311	.9894	.0197	.0243	.9237	.1504	.0000	.0000
<i>PSMD5-AS1</i>	.9097	.9098	.9994	-.0167	.0128	.9397	.1621	.0000	.0000
<i>MYOM2</i>	.9633	.9792	.9328	.0285	.0312	.9005	.0767	.0000	.0000
<i>PAX8-AS1</i>	.9930	.9930	.9988	.0261	.0392	.8850	.0122	.0000	.0000
<i>COL9A3</i>	.9629	.9634	.9936	.0136	.0281	.8927	.0701	.0000	.0000
<i>TNNT1</i>	.9751	.9748	.9930	-.0300	.0249	.8109	.0751	.0000	.0000
<i>PVRL2</i>	.9630	.9814	.9244	-.0117	.0243	.8361	.1140	.0000	.0000
<i>KIAA1671</i>	.8418	.8419	.9983	-.0204	.0180	.8097	.2069	.0000	.0000
<i>TMEM45B</i>	.8930	.8930	.9997	.0028	.0213	.7696	.0607	.0000	.0000
<i>ANKRD55</i>	.8267	.8269	.9988	-.0105	.0109	.7900	.4664	.0000	.0000

<sup>a</sup>Correlation between two measurements in the overlapping samples.

<sup>b</sup>The ratio of the subject-specific effect to the total variance at time point 1.

<sup>c</sup>The ratio of the measurement error effect to the total variance at time point 2.

<sup>d</sup>The p-value is obtained from testing  $\beta_2 = 0$  using the likelihood ratio test.

<sup>e</sup>The adjusted p-value is based on 811 genes tested.

significant association between gene expression and the motor score. The full list of 89 genes can be found in Table S2 in the Supporting Information, including the two significant genes (*SIK1* and *RAPGEFL1*) in the replication study. When using gene expression measured with DeepSAGE as an outcome, the number of significant genes was only six after multiple testing corrections (see the Supporting Information).

The parameter estimates of the 10 most significant genes of these 89 genes are summarized in Table 5. Specifically, parameter estimates of the mixture model (2) and the RNA-Seq full model (6) are presented. Only one gene had a Pearson

correlation coefficient of less than .83, namely *ANKRD55*. The adjusted *p*-value was based on the testing of 811 genes. The largest absolute effect size within these genes was obtained for *TNNT1*. The largest intraclass correlation was obtained for *PSMD5-AS1* with .9397. The lowest intraclass correlation was obtained for *TMEM45B*. The lowest measurement error ratio was obtained for *PAX8-AS1*, namely .0122.

## 6 | CONCLUSION AND DISCUSSION

We presented a new method to model the association between gene expressions measured with two technologies at two-time points and a disease indicator. This method was used to test for an association between the motor score and gene expression in HD patients and controls. At the first time point, gene expression was measured by DeepSAGE technology. A small subset was also measured by RNA-Seq. At the second time point, only RNA-Seq measurements were available. The data at the first time point were analyzed and published by Mastrokoulas et al. (2015). From the 167 genes identified to be related to the motor score by this study only for 14 genes, the measurements were sufficiently similar to enable analysis at both time points. Using DeepSAGE as a response, the association with motor score remained significant for these genes when we included the data at the second time point to the analysis.

We used a mixture model to select genes for which the gene expressions in the overlapping samples were highly correlated between the two measurements. The rationale behind this step is to select genes for which the difference between the two measurements is minimal. For the genes satisfying the requirements, a measurement error model was used to estimate the latent variable from the observed variable. Finally, a linear mixed-effects model was used to assess the relationship between gene expression and the motor score at both time points. We performed two analyses, namely a replication study of the identified genes by Mastrokoulas et al. (2015) using DeepSAGE as the outcome and the analysis of all genes using RNA-Seq measurements as the outcome. When modeling these data, we made several assumptions.

First, we assumed that the log-transformed count variables representing DeepSAGE and RNA-Seq at the two time points follow a multivariate normal distribution. To evaluate the robustness of our methods against the deviation of the normal distribution, we performed two sets of simulations. We studied the effect of assuming a normal distribution for log-transformed count data on predicting DeepSAGE from RNA-Seq measurements and on the size of the likelihood ratio test statistic to assess the relation between gene expression and the motor score. For a sample size of 100 and 20% of the training set, the predicted values appeared to be almost similar to the observed data. The second set of simulations showed that the size of the likelihood ratio test was correct for the range of overdispersion parameters found in our dataset. We used the linear mixed model, since this model was used by the original study, and it is straightforward to fit for various random effects structures. Alternatively, the negative-binomial mixed model might be used to assess the relationship between motor score and gene expressions. However, this model requires numerical integration over the random effects structure and is therefore more challenging to fit.

Second, for the replication study of modeling DeepSAGE, we assumed that the joint distribution of DeepSAGE and RNA-Seq at time point 2 was equal to the bivariate normal distribution of DeepSAGE and RNA-Seq at time point 1. This assumption was needed since we estimated the relationship between the two technologies using data from time point 1 to estimate DeepSAGE from RNA-Seq responses at time point 2. This assumption was not needed when predicting RNA-Seq from DeepSAGE since for RNA-Seq, we had to predict RNA-Seq from DeepSAGE responses at time point 1. Other advantages for using RNA-Seq instead of DeepSAGE measurements as a dependent variable are that the number of observed data points of RNA-seq is larger, and that the effects of the motor score due to different time points and due to predicting the outcome variable can be distinguished.

Third, we included all samples in our analysis while for some samples the relationship between the two measurements might be small. After fitting the mixture model, we chose genes for which the  $\alpha$  is larger than 0.8. Thus for these genes, the fraction of samples with no correlation is smaller than 20%. We chose all samples since selection on motor score (case-control sample) and selection on gene expression may result in biased parameter estimates. Moreover for most samples, we have only one measurement.

In this paper, the aim is to identify genes for which the severity of HD influences gene expression. The underlying hypothesis is that the progress of the disease triggers reactions from several genes. We used total motor score as a surrogate for HD severity, which is commonly applied in the literature. For example, the motor score was used to detect metabolic markers of HD (Mastrokoulas et al., 2016). Here, the metabolites were the responses. In another study, Mina et al. (2016) identified two disease signatures of HD that related to the motor score using gene expression in blood and multiple brain region samples. Here a network analysis was used to identify common modules of gene expression in blood and brain

samples, and then correlation analysis was performed to relate the motor score and the common modules. Also, when studying other traits, gene expressions are often modeled as an outcome. For example, Li et al. (2013) modeled the effect of somatic genetic and epigenetic factors on gene expression levels in breast cancer tumors.

Only for 14 of the original significant 167 genes (Mastrokolias et al., 2015), the measurements of the two technologies were sufficiently correlated to be selected for further analyses. Applying our DeepSAGE model (4) to these data yielded replication of the results. We also analyzed these genes using RNA-Seq as an outcome. It appeared that only for two genes the association is significant (*RAPGEFL1*, *SIK1*). However, for *SIK1*, the direction of the  $\beta_2$  was not the same for the two models. The reason for the opposite direction of the  $\beta_2$  estimate could be explained by the slightly different interpretation between the two models. An additional parameter was included in the model to allow for a difference in the effect of the motor score on observed versus predicted RNA-Seq responses. For both measurements, the relationship in the overlapping samples was positive, while the relationship in the remaining samples was negative. In the model with DeepSAGE as an outcome, these two subsets were not distinguished since, for both subsets, the measurements were observed. For RNA-Seq, the measurements of the remaining samples were predicted. With regard to the other 12 genes, the loss of significance might be caused by the fact that the prediction of RNA-Seq was not sufficiently accurate (see Supplementary Figure S2). Thus, we conclude that for the initially identified genes, only one gene could be replicated when using data from both time points for both technologies, namely *RAPGEFL1*.

Concerning the analysis of all measured 19,711 genes, only 811 genes could be analyzed using both time points. The RNA-Seq model was used to analyze these genes. Out of 811 genes, 59 genes were found significant after multiple testing corrections. The list includes the two significant genes from the replication study. Reasons for more significant findings using RNA-Seq are that for most genes, the variance of RNA-Seq measurements is larger than for DeepSAGE and that for RNA-Seq, we have more observed measurements. The low number of replications using RNA-Seq instead of DeepSAGE might be explained by the fact that the prediction of RNA-Seq from DeepSAGE was less accurate. Indeed the relationship between the measurements in the overlapping samples is more linear in the top 10 significant genes of all genes than in the genes from the replication study (see Supplementary Figures S2 and S3).

Data cleaning is one of the critical aspects of our analysis. Lowly expressed genes contain more noise and might obscure the real picture of gene expression as well as overestimate the true effects. Discarding these genes might be beneficial for increasing the power of high-throughput experiments (Bourgon, Gentleman, & Huber, 2010; Ignatiadis, Klaus, Zaugg, & Huber, 2016). However, one of the consequences is that we identified a relatively small number of genes compared to the cross-sectional study (Mastrokolias et al., 2015). Besides low expression genes, too much disagreement between the two measurements is also an essential aspect of a limited number of genes identified.

By jointly modeling the two types of measurements, we were able to obtain insight into the relationship between the two technical platforms, the gene expression, and the motor score. We were able to replicate one gene, namely *RAPGEFL1*, using data from both time points and using both technologies. In addition, we replicated 13 genes for DeepSAGE using data from two time points and 58 genes for RNA-Seq using two time points. Note that the models need to account for differences between observed and predicted measurements when adding the second time point data. Therefore, the efficiency for estimation of the parameters is only increased due to the availability of more information to estimate the effect of the other covariates and the variance components.

Using different technologies in the same study brings all kinds of challenges. Each technology produces different reads per gene transcript. DeepSAGE only provides one read, while RNA-Seq produces multiple reads per gene transcript. To make them comparable, we applied several data cleaning and normalization procedures. To be selected for further analyses, both measurements for a gene needed to satisfy the data cleaning criteria. Further, only genes with sufficient correlation between the two measurements were included for the downstream analyses. Thus for each gene the relationship between the two measurements needed to be estimated. Here we defined sufficiently correlated as a correlation larger than .8 in 80% of the samples. By doing so, we might have missed genes whose expression is associated with motor score. On the other hand, lowering the threshold may result in more noise and a higher testing burden. In larger samples, it may be worthwhile to consider to lower the threshold. Finally, it was necessary to account for the different techniques by including additional parameters to the models.

We modeled the relationship between the two measurements per gene, assuming a mixture of bivariate normal distributions. Alternatively, one may consider modeling the relationship between the two datasets by assuming a multivariate normal distribution and using multiple multivariate regression techniques such as partial least squares (PLS) (Wold, 1966). One interesting approach here is two-way orthogonal partial least square (O2PLS), which estimates the joint space and considers data-specific spaces (Bouhaddani et al., 2016; Trygg & Wold, 2003). Such analyses will provide information about the genes highly represented by the two datasets. Another extension was developed by Bouhaddani, Uh, Hayward,

Jongbloed, and Houwing-Duistermaat (2018); they embedded PLS in a probabilistic framework to facilitate statistical inference and unique identification of the parameters. One of the future directions is to compare the performance of the measurement error models and PLS methods, where the former uses additional information about the structure of the data, and the latter estimates this structure from the datasets.

Several extensions of our proposed method are possible. Instead of a two-step model, we might consider one model using the full likelihood approach, which includes all measurements from two technologies. In our proposed method, the overlapping samples were used twice, that is, for the mixture models and the linear mixed effect models. Moreover, the uncertainty in the estimate of the variance of the measurement error was not taken into account. In our analysis, two measurements' joint distribution was assumed to be similar at both time points. Relaxing this assumption can be considered by incorporating a more complex covariance structure, where all variances and covariances vary across measurements at each time point. Finally, method development to model the count data might be considered.

## ACKNOWLEDGMENTS

This work was supported by Indonesian Endowment Fund for Education (LPDP), Ministry of Finance, Indonesia, the European Union's Horizon 2020 grants IMforFUTURE (grant agreement No. 721815), the European Union's Seventh Framework Programme FP7-Health-F5-2012 MIMOmics (grant agreement No. 305280), Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research and Dutch Centre for Biomedical Genetics and the European Community's Seventh Framework Programme (FP7/2007-2013) (grant agreement no. 2012-305121) 'Integrated European -omics research project for diagnosis and therapy in rare neuromuscular and neurodegenerative diseases (NEUROMICS)'.


## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Angga M. Fuady  <https://orcid.org/0000-0002-4012-9779>

## REFERENCES

- Abrevaya, J., & Hausman, J. (2004). Response error in a transformation model with an application to earnings-equation estimation. *Econometrics Journal*, 7(2), 366–388.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.
- Bouhaddani, S. e., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., & Uh, H.-W. (2016). Evaluation of O2PLS in OMICS data integration. *BMC Bioinformatics*, 17(2), S11.
- Bouhaddani, S. E., Uh, H.-W., Hayward, C., Jongbloed, G., & Houwing-Duistermaat, J. (2018). Probabilistic partial least squares model: Identifiability, estimation and application. *Journal of Multivariate Analysis*, 167, 331–346.
- Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 9546–9551.
- Buonaccorsi, J. (1991). Measurement errors, linear calibration and inferences for means. *Computational Statistics & Data Analysis*, 11(3), 239–257.
- Buonaccorsi, J. (1996). Measurement error in the response in the general linear model. *Journal of the American Statistical Association*, 91(434), 633–642.
- Buonaccorsi, J. (2010). *Measurement error: Models, methods, and applications*. Chapman & Hall–CRC Interdisciplinary Statistics Series. Boca Raton, FL: Chapman and Hall/CRC.



- Buonaccorsi, J., & Tosteson, T. (1993). Correcting for nonlinear measurement errors in the dependent variable in the general linear model. *Communications in Statistics—Theory and Methods*, 22(10), 2687–2702.
- Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, FL: Chapman and Hall/CRC.
- De Veaux, R. (1989). Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3), 227–245.
- Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Boca Raton, FL: Chapman and Hall/CRC.
- Ignatiadis, N., Klaus, B., Zaugg, J., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13, 577–580.
- Jones, L., & Hughes, A. (2011). Pathogenic mechanisms in Huntington's disease. *International Review of Neurobiology*, 98, 373–418.
- Kent, A. (2004). Huntington's disease. *Nursing Standard (Royal College of Nursing (Great Britain): 1987)*, 18, 45–51; quiz 52–53.
- Keogh, R., Carroll, R., Toozé, J., Kirkpatrick, S., & Freedman, L. (2016). Statistical issues related to dietary intake as the response variable in intervention trials. *Statistics in Medicine*, 35, 4493–4508.
- Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., ...Freedman, M. L. (2013). Integrative EQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152, 633–641.
- Mastrokolas, A., Ariyurek, Y., Goeman, J., van Duijn, E., Roos, R., van der Mast, R., ...van Roon-Mom, W. (2015). Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *European Journal of Human Genetics: EJHG*, 23, 1349–1356.
- Mastrokolas, A., Pool, R., Mina, E., Hettne, K. M., van Duijn, E., van der Mast, R. C., ... van Roon-Mom, W. (2016). Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics*, 12(8), 137.
- McCulloch, C., & Searle, S. (2000). *Generalized, linear, and mixed models*. New York, NY: John Wiley & Sons.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
- Mina, E., van Roon-Mom, W., Hettne, K., van Zwet, E., Goeman, J., Neri, C., ...Roos, M. (2016). Common disease signatures from gene expression analysis in Huntington's disease human blood and brain. *Orphanet Journal of Rare Diseases*, 11, 97.
- Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5, 621–628.
- Ozsolak, F., & Milos, P. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews. Genetics*, 12, 87–98.
- R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Robinson, M., McCarthy, D., & Smyth, G. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26, 139–140.
- Robinson, M., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biology*, 11, R25.
- 't Hoen, P., Ariyurek, Y., Thygesen, H., Vreugdenhil, E., Vossen, R., de Menezes, R., ...den Dunnen, J. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, 36, e141.
- Trygg, J., & Wold, S. (2003). O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, 17(1), 53–64.
- van der Burg, J., Björkqvist, M., & Brundin, P. (2009). Beyond the brain: Widespread pathology in Huntington's disease. *The Lancet. Neurology*, 8, 765–774.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer-Verlag.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Berlin, Germany: Springer-Verlag.
- Walker, F. (2007). Huntington's disease. *Lancet (London, England)*, 369, 218–228.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10, 57–63.
- Wold, H. (1966). Multivariate analysis. In *Proceedings of an International Symposium* (pp. 391–420). New York, NY: Academic Press.
- Zhernakova, D., de Klerk, E., Westra, H.-J., Mastrokolas, A., Amini, S., Ariyurek, Y., ...Franke, L. (2013). DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genetics*, 9, e1003594.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Fuady AM, van Roon-Mom WC, Kiełbasa SM, Uh Hae-Won, Houwing-Duistermaat JJ. Statistical method for modeling sequencing data from different technologies in longitudinal studies with application to Huntington disease. *Biometrical Journal*. 2021;63:745–760. <https://doi.org/10.1002/bimj.201900235>