A comparison of metabolic labeling and statistical methods to study genome-wide dynamics of RNA turnover
Supplementary File 5

Using the kinetic equations describing RNA dynamics in the pulseR model for the nucleotide conversion protocols, we investigated key aspects of optimal design using the variance-matrix of the estimator, as described in Uvarovskii *et al.* [1]. Briefly, we computed the Fisher information matrix (FIM) $\mathcal{I}$ using the nucleotide conversion models as described in the manuscript (Methods). To validate the expressions for the elements of the FIM, we used the Yamwi interface to the Maxima software [3]. The supporting file is provided with the source code at https://github.com/dieterich-lab/ComparisonOfMetabolicLabeling.

We define the score as the gradient of $\log \mathcal{L}(\boldsymbol{\theta})$

$$\mathcal{S}(\boldsymbol{\theta}) = \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{1}$$

Under certain regularity conditions, the FIM is described as

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right] \tag{2}$$

See Uvarovskii *et al.* [1] for details. Using the fact that the variance of the score function is

$$\mathrm{var}_{\boldsymbol{\theta}}(\mathcal{S}(\boldsymbol{\theta})) = \mathcal{I}(\boldsymbol{\theta}) \tag{3}$$

and substituting the expression for the logarithm of the likelihood function (Methods), we have

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{k}{m(m+k)} \frac{\partial m}{\partial \theta_i} \frac{\partial m}{\partial \theta_j} \tag{4}$$

Because of the reciprocity of estimator-variance and Fisher information, minimizing the variance corresponds to maximizing the information. The diagonal term of the FIM $\mathcal{I}(\boldsymbol{\theta})_{\delta\delta}$ with respect to the decay rate δ was used as a lower bound to estimate the inverse FIM, and thus the variance of the estimator $\mathrm{var}(\hat{\delta})$. By substituting the expressions for the RNA amounts, we have

$$(\mathcal{I}_{\mathrm{unlabeled}}(\boldsymbol{\theta}))_{\delta\delta} = \frac{k\mu_3^2 t^2 e^{-2\delta t}}{(\mu_1 + \mu_3 e^{-\delta t})(k + \mu_1 + \mu_3 e^{-\delta t})} \tag{5}$$

$$(\mathcal{I}_{\mathrm{labeled}}(\boldsymbol{\theta}))_{\delta\delta} = \frac{k\mu_3^2 t^2 e^{-2\delta t}}{(\mu_2 + \mu_3 (1 - e^{-\delta t}))(k + \mu_2 + \mu_3 (1 - e^{-\delta t}))} \tag{6}$$

and $\mathcal{I} = \mathcal{I}_{\mathrm{unlabeled}} + \mathcal{I}_{\mathrm{labeled}}$. To explore how the variance of decay rates depend on the labeling time, and how the choice of time points affect the confidence intervals, we used the diagonal term $\mathcal{I}(\boldsymbol{\theta})_{\delta\delta}$, abbreviated as $\mathcal{I}_{\delta\delta}$. This term can be interpreted as an information gain assuming all other parameters were known [1]. Even at >80% conversion efficiency, at short labeling times only a fraction of reads contains more than one conversion, *i.e.* the sequencing data consists of unlabeled molecules from genes with slower synthesis. In the limit of very large unlabeled background, these reads do not contribute to the FIM, and only provide information on the nuisance parameter for the background.

In the limit of high unlabeled background level, and assuming no overdispersion, we have $\mu_1 \gg \mu_3$, and the unlabeled reads do not contribute to the FIM term

$$\lim_{\mu_1/\mu_3 \to \infty} (\mathcal{I}_{\mathrm{unlabeled}})_{\delta\delta} = \lim_{\mu_1/\mu_3 \to \infty} \frac{\mu_3^2 t^2 e^{-2\delta t}}{\mu_1 + \mu_3 e^{-\delta t}} = 0 \tag{7}$$

If the sequencing depth is fixed, and in the limit of little background labeled fraction, $\mu_3 \to 0$ as $\mu_1/\mu_3 \to \infty$, and $(\mathcal{I}_{\text{labeled}})_{\delta\delta} \to 0$ also. In such cases, the estimator has a high variance, because $\text{var}(\hat{\delta}) = (\mathcal{I}^{-1}(\boldsymbol{\theta}))_{\delta\delta} \geq 1/\mathcal{I}(\boldsymbol{\theta})_{\delta\delta}$, and $\mathcal{I}(\boldsymbol{\theta})_{\delta\delta} \to 0$. Similar observations were made in Uvarovskii *et al.* [1]. In practice, however, the presence of the background error μ_2 which has to be taken into account in the models presented here renders the estimation of fast turnover genes unreliable, and introduces limitations on the estimates.

To illustrate how the FIM diagonal term for the decay rate δ can depend on the labeling time, we calculated $\mathcal{I}_{\delta\delta}$ for a range of time points normalized by a gene's characteristic time of degradation $\tau = 1/\delta$ (Fig. 1). The optimal labeling time was obtained by maximizing $\mathcal{I}_{\delta\delta}$ using medians of the estimated abundances μ_1, μ_2, and μ_3, as well as the overdispersion parameter k from the model fitted to the full set of points.

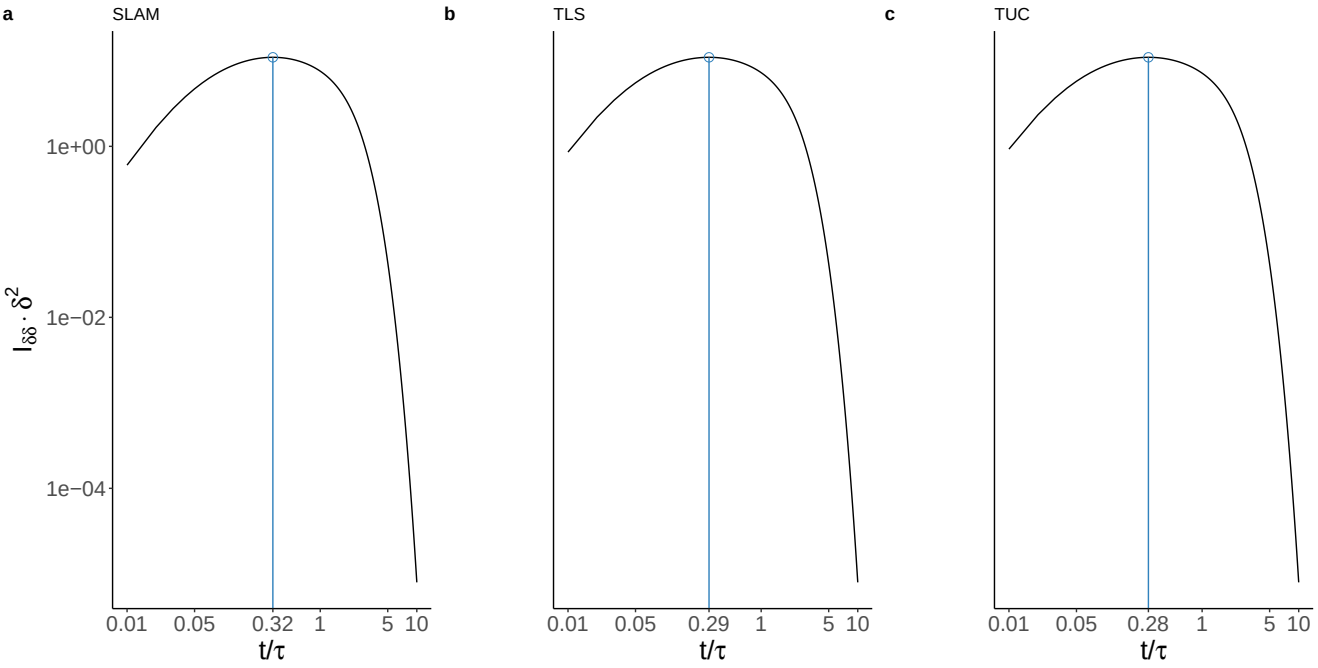

Figure 1: Diagonal term of the normalised Fisher information matrix (FIM) $\mathcal{I}_{\delta\delta}\delta^2$ as a function of pulse time normalized by a gene's characteristic time of degradation $\tau = 1/\delta$, so it corresponds to the lower boundary of the relative variance, for the **a** SLAM-seq, the **b** TLS-seq, and the **c** TUC-seq protocols, using the pulseR kinetic models (Methods). We used parameters values from the model fitted to all time points (0, 1, 2, 4, and 8 h samples).

At earlier time points (1 h and 2 h), there is no distinct minimum in relative confidence intervals, since genes with $1/0.32 \approx 3$ h and $2/0.32 \approx 6$ h are not sufficiently represented, and the majority of the decay rates were not identifiable (Figs. 2 to 4, 0, 1, 2 h, see also 0, 2, 4 h). At longer pulse times, the estimates for faster genes were worse (Figs. 2 to 4, 0, 4, 8 h), due to information loss (see below, and Fig. 7b). Nevertheless, even despite their large confidence intervals, we know that these estimates are reliable, since we have the results from the BSA model (Figs. 5 and 6). In contrast to the BSA model, however, the labeled and unlabeled 'fractions' contribute differently to the FIM diagonal term (Fig. 7). In the BSA model, the FIM diagonal term for the labeled fraction (eluate) was higher at short labeling times, but the contribution from the unlabeled fraction (supernatant) increased for the majority of genes at longer times. As noted earlier [1], the proportion of RNA amounts from genes with higher decay rates in the supernatant fraction exponentially decreased, resulting in low counts for the fastest genes (tail in Fig. 7a, 8 h). The limiting values (dashed lines) show how the terms are bounded by the presence of overdispersion. An increase in sequencing depth cannot improve these limits [1]. With the nucleotide conversion model, the information gain from the labeled read fraction almost always remained higher, and the proportion of RNA amounts for the unlabeled read fraction also decreased for genes with the fastest decay rates, but this time according to the asymptotic limit of the labeled fread fraction (Fig. 7b). At longer labeling times, the information gain from the unlabeled read fraction increased for slower genes, but fast genes became almost completely unidentifiables (Fig. 7b, 8 h). It can be clearly seen that increased labeling time had an effect similar to increasing sequencing depth for the labeled read fraction.

In summary, the definition of kinetic equations describing the RNA populations in pulseR directly influence parameter optimization, and the required number of time points. The presence of background fractions introduces limitations affecting the confidence intervals, which are calculated using the profile likelihood, taking account of the

presence of the mean background labeled and unlabeled fractions as well as the difference between the maximum and the background labeled fraction, all considered nuisance parameters (Methods). While adjustments in likelihood functions may reduce the impact of nuisance parameters in estimating confidence intervals for the decay rates, sensitivity to potential model misspecifications is also an issue, and must be considered carefully together with experimental design.

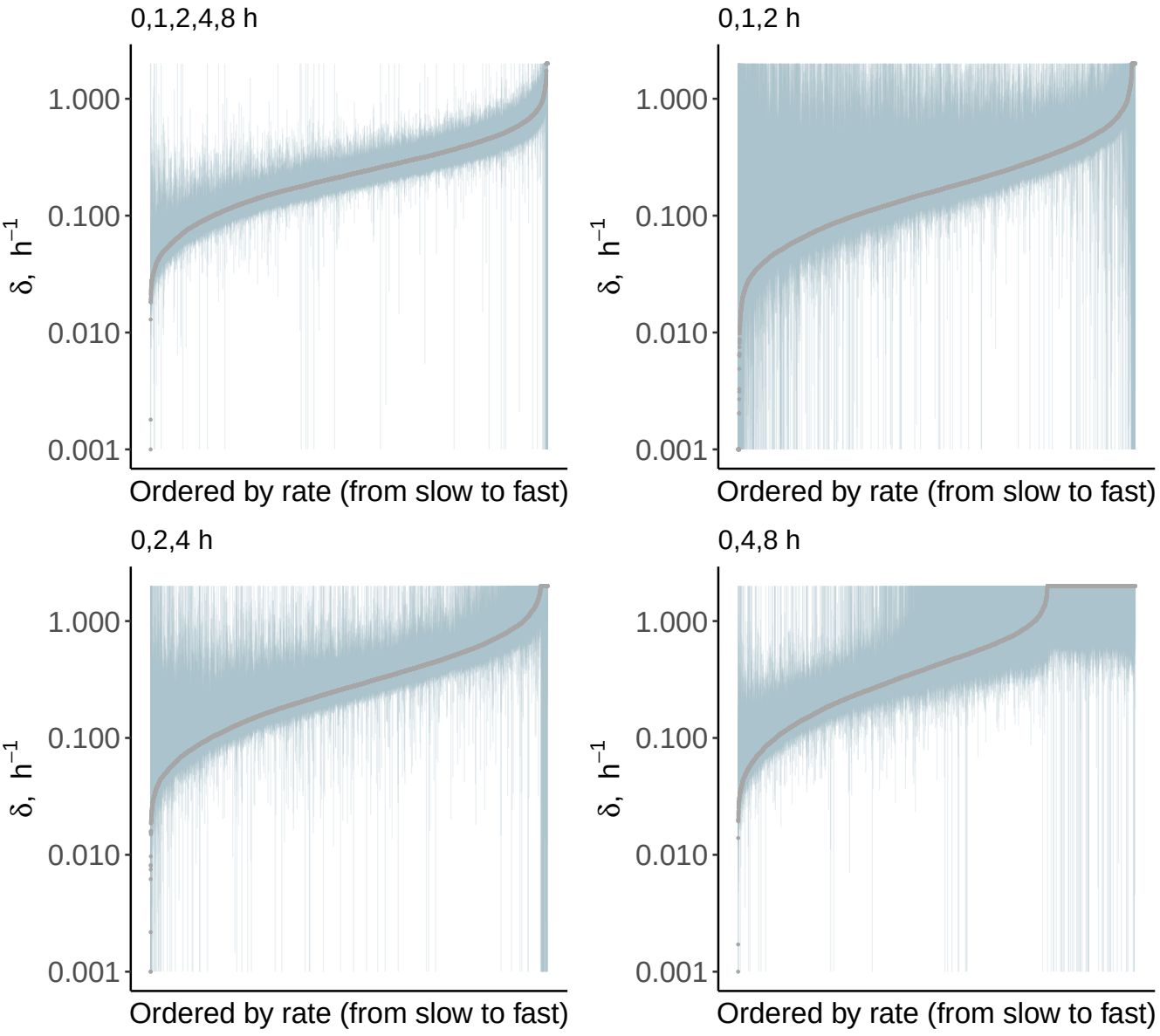

Figure 2: Estimates for the SLAM-seq data using pulseR. Degradation rates and 95% confidence intervals for different labeling time points.

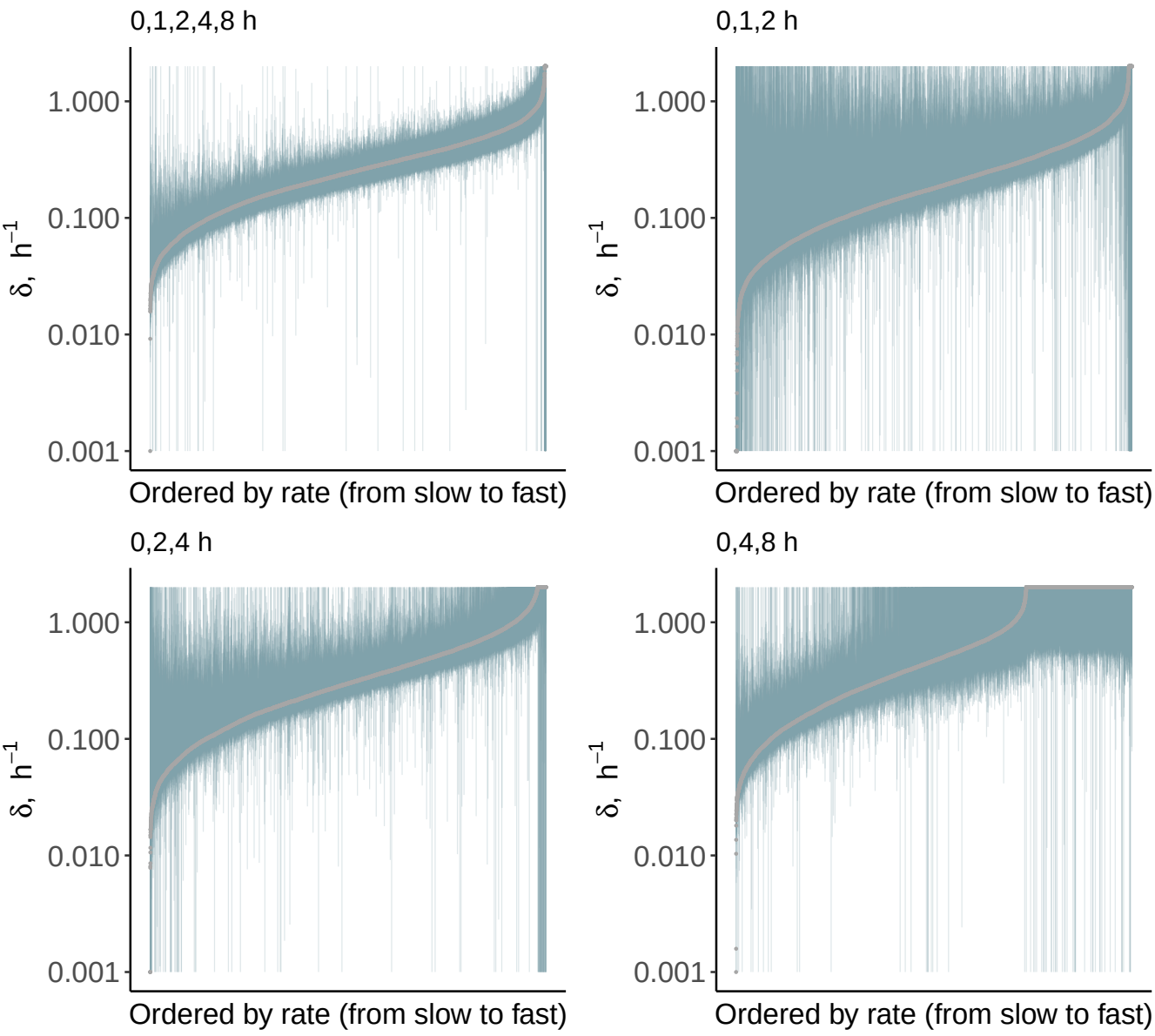

Figure 3: Estimates for the TLS-seq data using pulseR. Degradation rates and 95% confidence intervals for different labeling time points.

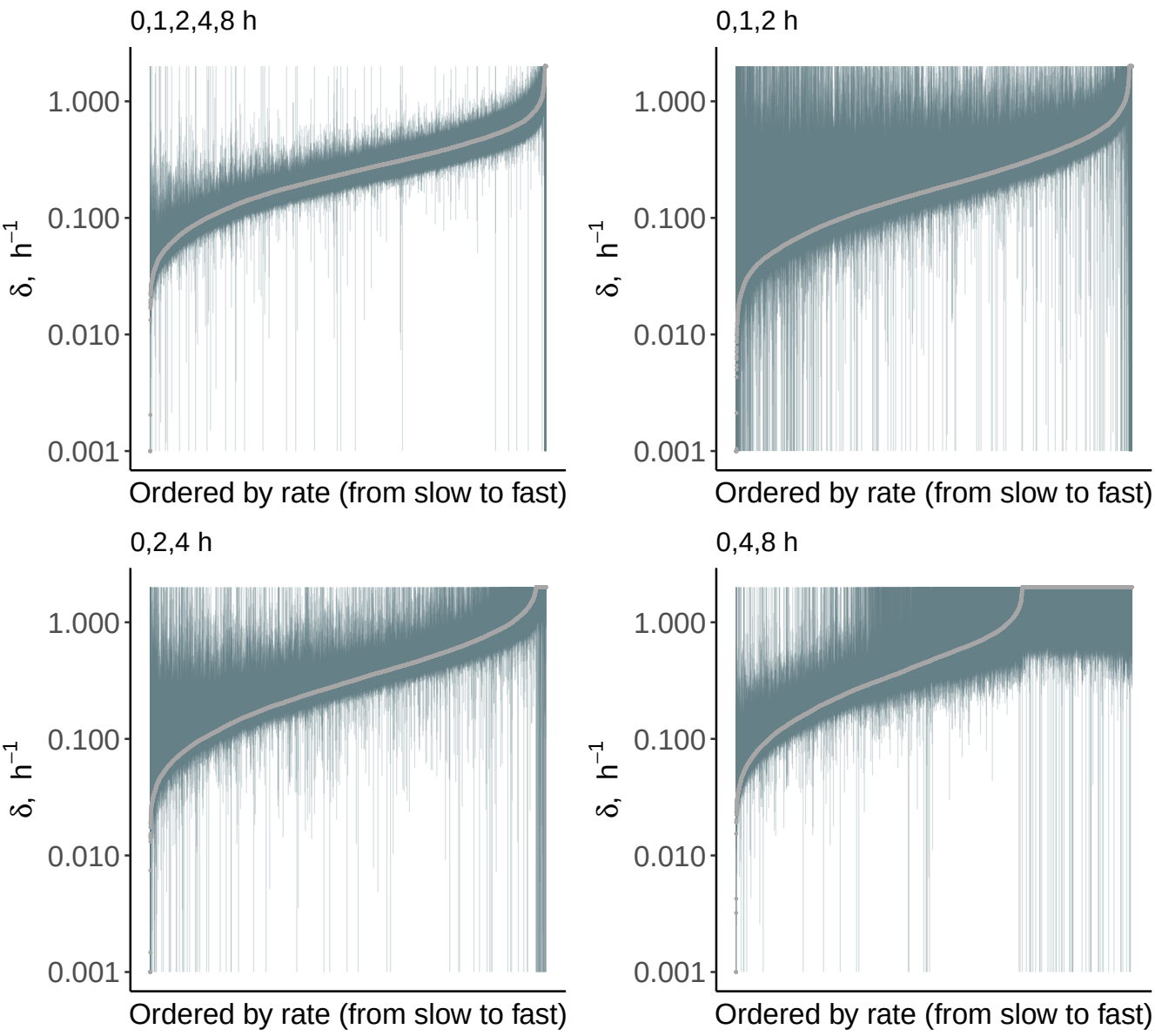

Figure 4: Estimates for the TUC-seq data using pulseR. Degradation rates and 95% confidence intervals for different labeling time points.

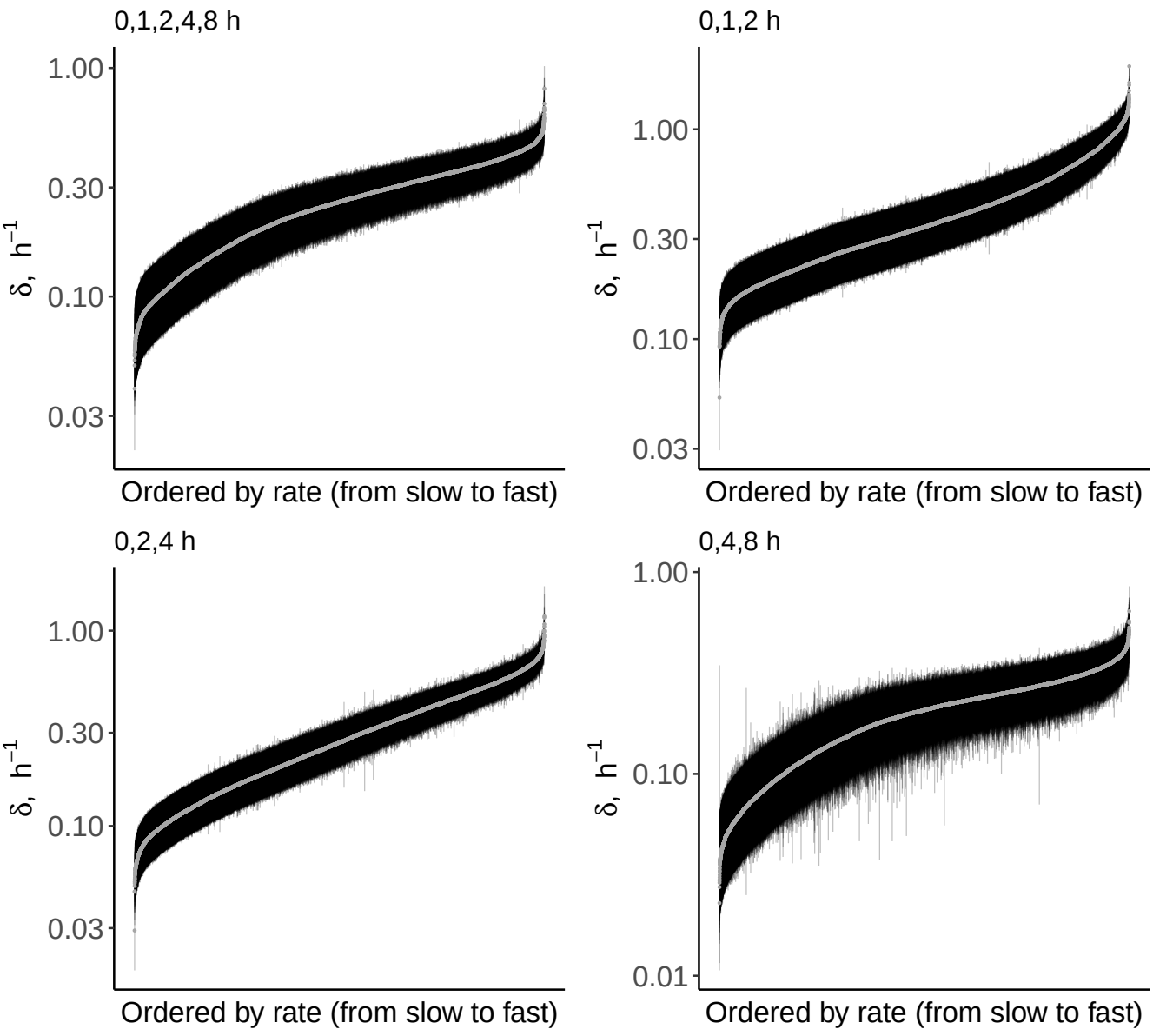

Figure 5: Estimates for the ERCC data using pulseR. Degradation rates and 95% confidence intervals for different labeling time points.

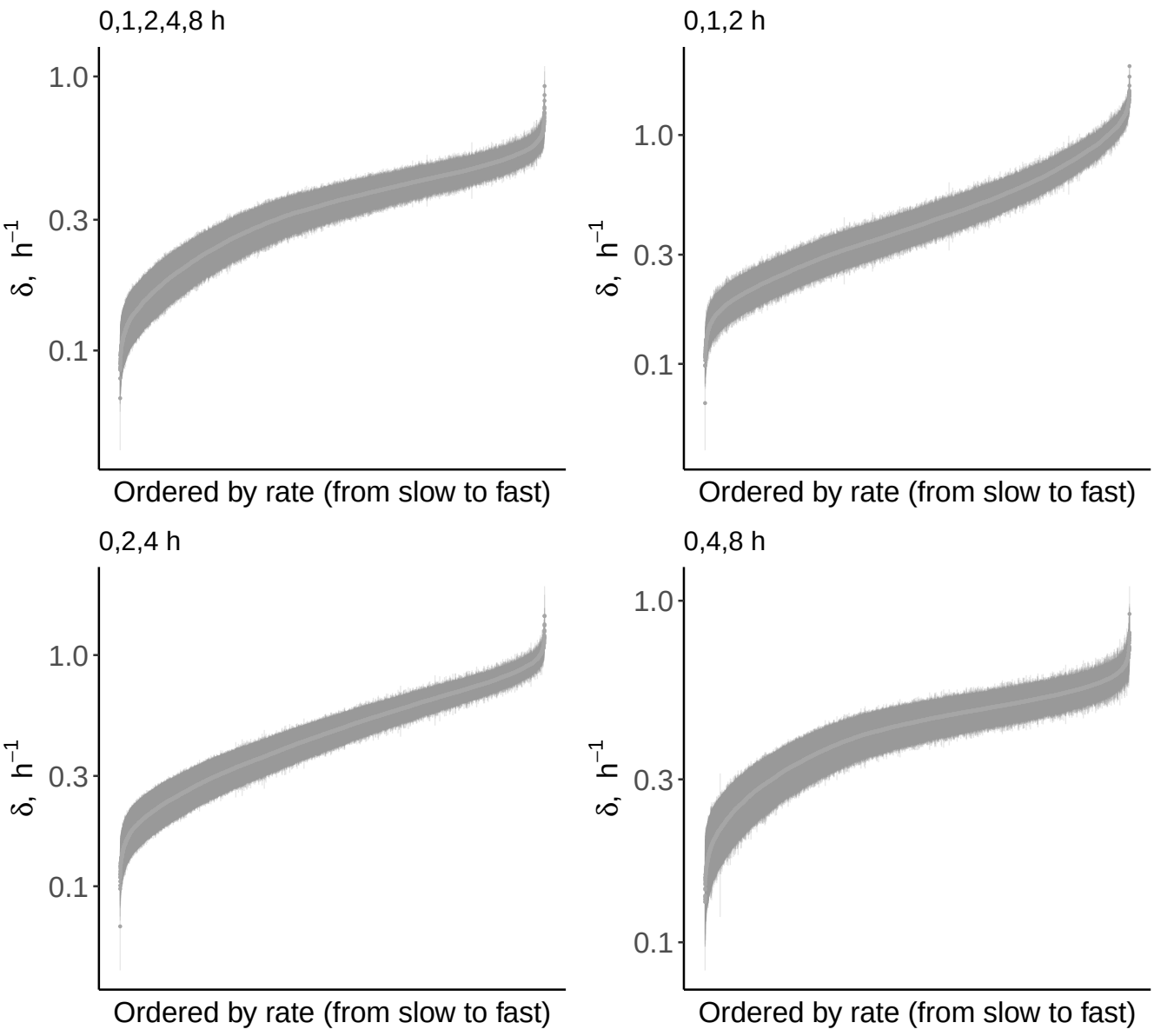

Figure 6: Estimates for the BSA (without ERCC) data using pulseR. Degradation rates and 95% confidence intervals for different labeling time points.

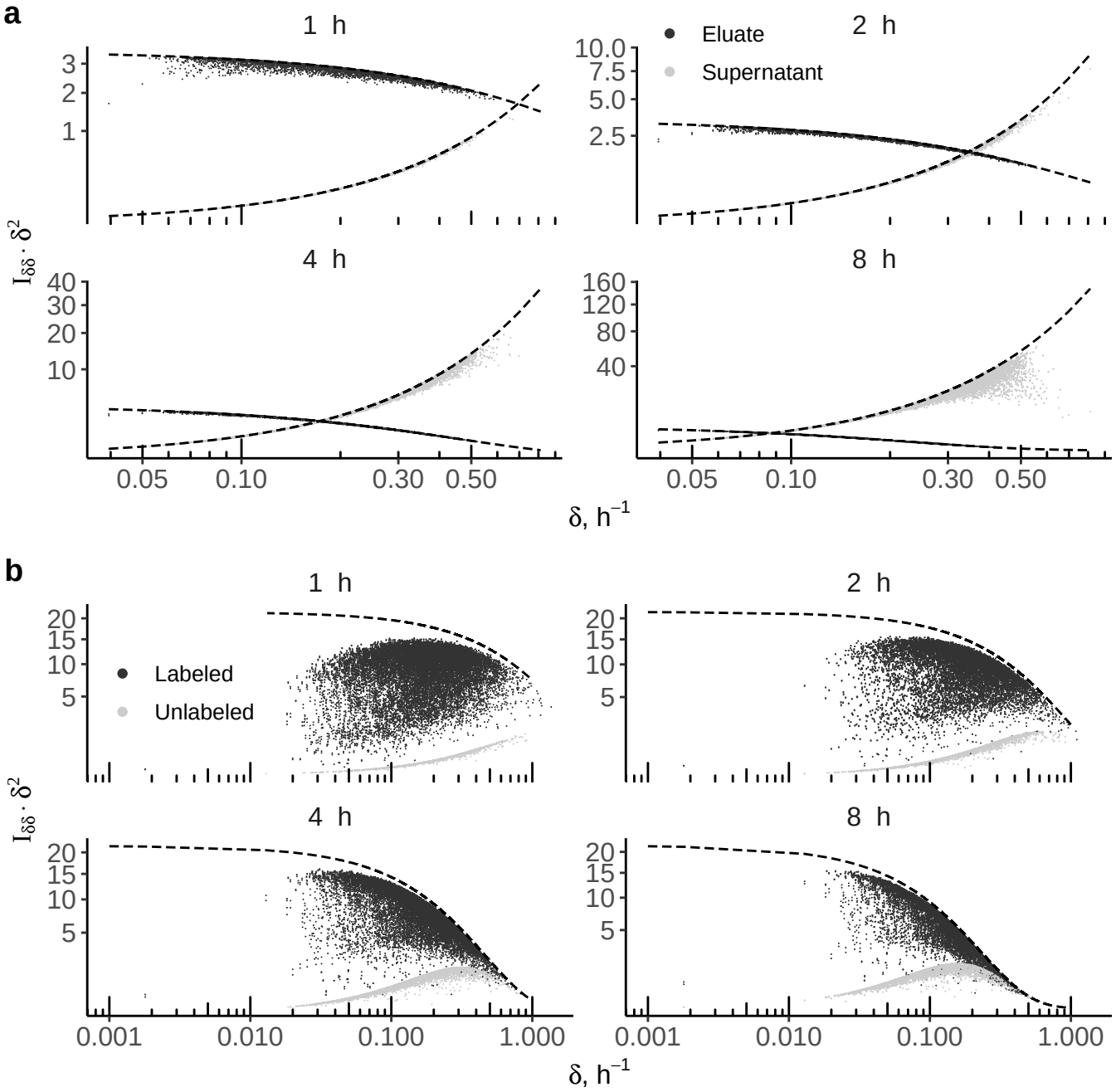

Figure 7: Diagonal term of the FIM multiplied by $\hat{\delta}^2$ to illustrate the contribution of the eluate (labeled) and supernatant (unlabeled) fractions to decay rates δ for different time points. The term was computed at estimated parameters for the **a** ERRC model, and for the **b** nucleotide conversion model (SLAM-seq). The dashed lines are the limiting values for the diagonal term $\mathcal{I}_{\delta\delta}$, bounded by the presence of overdispersion. An increase in sequencing depth cannot improve these limits. For the nucleotide conversion model, the limiting value for the diagonal term $(\mathcal{I}_{\text{labeled}})_{\delta\delta}$ of the ERCC model is shown. We used SLAM-seq as example for the nucleotide conversion protocols, but similar results were obtained with the TLS-seq and TUC-seq data. Scatter plots were downscaled by randomly sampling points for representation.

References

[1] Uvarovskii, A., Naarmann-de Vries, I. S. & Dieterich, C. On the optimal design of metabolic rna labeling experiments. *PLOS Computational Biology* **15**, 1–22, https://doi.org/10.1371/journal.pcbi.1007252 (2019).

[2] Uvarovskii, A. & Dieterich, C. pulseR: Versatile computational analysis of RNA turnover from metabolic labeling experiments. *Bioinformatics* **33**, 3305–3307, https://doi.org/10.1093/bioinformatics/btx368 (2017).

[3] Maxima. Maxima, a computer algebra system version 5.34.1. http://maxima.sourceforge.net/ (2014).