Supplementary Information

Supplementary Methods

We use a block-coordinate descent approach to iteratively update individual parameters of the model, including the transcript abundance vector $\boldsymbol{\beta}$, platform-specific library sizes s_k, and the effect of confounding biases, represented by vector $\boldsymbol{\gamma}$.

1. Inferring $\boldsymbol{\beta}$

We start by showing how the maximum-likelihood estimate (MLE) of $\boldsymbol{\beta}$ can be obtained in the absence of a log-normal prior on $\boldsymbol{\beta}$. We will then show the inference of the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}$ with a log-normal prior. As described in the Methods section, in the absence of a prior, for each platform k we have:

$$\forall k \in \{1, \dots, K\} \,\Big|\, \boldsymbol{\lambda}_k = \left(\lambda_{k,1}, \dots, \lambda_{k,|U_k|}\right)^{\top} = s_k \boldsymbol{P}_k \boldsymbol{\beta}$$

$$\forall u \in U_k \big|\, n_{k,u} \sim \mathrm{Pois}\left(\lambda_{k,u}\right)$$

Let's define the matrix $\boldsymbol{P}$ as:

$$\boldsymbol{P} = \begin{bmatrix} s_1 \boldsymbol{P}_1 \\ \vdots \\ s_K \boldsymbol{P}_K \end{bmatrix}$$

Then we can rewrite the model as:

$$\boldsymbol{\lambda} = \boldsymbol{P}\boldsymbol{\beta}$$

$$\forall u \in \{U_1, \dots, U_U\} |\, n_u \sim \mathrm{Pois}(\lambda_u)$$

The MLE can be obtained by minimizing the negative log-likelihood function:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{u \in U} (\lambda_u - n_u \log \lambda_u)$$

We use a sequential coordinate-wise descent (SCD) approach to iteratively solve each element β_t of $\boldsymbol{\beta}$ ($t \in T$). Consider the current estimate β_t^i, where i denotes the last iteration of the optimization algorithm. Let's assume that in the next iteration (i+1), the updated estimate for β_t will be different from the current estimate by $\delta^{[i+1]}$; i.e., the next estimate will be $\beta_t^{[i+1]} = \beta_t^{[i]} + \delta^{[i+1]}$. If the vector of the current predicted OU abundances is $\boldsymbol{\lambda}^{[i]}$, then the next set of predicted fragment abundances is given by:

$$\lambda_u^{[i+1]} = \lambda_u^{[i]} + \delta^{[i+1]} p_{u,t}$$

Therefore:

$$\beta_t^{[i+1]} = \beta_t^{[i]} + \delta^{[i+1]}$$

$$\delta^{[i+1]} = \arg\min_{\delta} \sum_{u \in U} \left[\lambda_u^{[i]} + \delta p_{u,t} - n_u \log\left(\lambda_u^{[i]} + \delta p_{u,t}\right)\right]$$

To solve for $\delta^{[i+1]}$, we take the derivative of the negative log-likelihood function with respect to δ and set it to zero:

$$\frac{d}{d\delta} \sum_{u \in U} \left[\lambda_u^i + \delta p_{u,t} - n_u \log\left(\lambda_u^i + \delta p_{u,t}\right)\right] = 0$$

$$\sum_{u \in U} p_{u,t} - \frac{n_u p_{u,t}}{\lambda_u^i + \delta p_{u,t}} = 0$$

To find the root of this function, we use Newton's method to update δ in each iteration:

$$f(\delta) = \sum_{u \in U} \left[p_{u,t} - \frac{n_u p_{u,t}}{\lambda_u^i + \delta p_{u,t}}\right] = \sum_{u \in U} p_{u,t} \left(1 - \frac{n_u p_{u,t}}{\lambda_u^i + \delta p_{u,t}}\right)$$

$$\delta^{[i+1]} = \delta^{[i]} - \frac{f\left(\delta^{[i]}\right)}{f'\left(\delta^{[i]}\right)}$$

$$\delta^{[i+1]} = \delta^{[i]} - \frac{\sum_{u \in U} p_{u,t} \left(1 - \frac{n_u}{\lambda_u^i + \delta^{[i]} p_{u,t}}\right)}{\sum_{u \in U} n_u \left(\frac{p_{u,t}}{\lambda_u^i + \delta^{[i]} p_{u,t}}\right)^2}$$

Note that since $\delta^{[i+1]}$ is calculated with respect to the current value of $\beta_t^{[i]}$, then $\delta^{[i]}$ must also be calculated relative to $\beta_t^{[i]}$, which means that $\delta^{[i]}$= $\beta_t^{[i]}$–$\beta_t^{[i]}$=0. Therefore:

$$\delta^{[i+1]} = -\frac{\sum_{u\in U} p_{u,t}\left(1-\frac{n_u}{\lambda_u^i}\right)}{\sum_{u\in U} n_u\left(\frac{p_{u,t}}{\lambda_u^i}\right)^2}$$

This provides an iterative procedure where each β_t is updated by adding the value of $\delta^{[i+1]}$ from the equation above, followed by updating each $\lambda_u^{[i]}$ (for $u\in U$) by adding $\delta^{[i+1]}p_{u,t}$ to it.

Now, we will modify the equations above to show how the MAP estimate of $\boldsymbol{\beta}$ can be obtained when a log-normal prior is placed on $\boldsymbol{\beta}$:

$$\log \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$
$$\boldsymbol{\lambda} = \boldsymbol{P\beta}$$
$$n_u \sim \text{Pois}(\lambda_u)$$

In this case, the negative log-likelihood function also includes a regularization term that acts to shrink the logarithm of each β_t:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left[\frac{1}{2\sigma^2}\sum_{t\in T}(\log\beta_t)^2 + \sum_{u\in U}(\lambda_u - n_u\log\lambda_u)\right]$$

Following the same method as above, we can see that for each transcript t, its abundance in iteration i+1 can be updated as:

$$\beta_t^{[i+1]} = \beta_t^{[i]} + \delta^{[i+1]}$$
$$\delta^{[i+1]} = \arg\min_{\delta}\left[\frac{1}{2\sigma^2}\left[\log\left(\beta_t^{[i]}+\delta\right)\right]^2 + \sum_{u\in U}\left[\lambda_u^{[i]} + \delta p_{u,t} - n_u\log\left(\lambda_u^{[i]} + \delta p_{u,t}\right)\right]\right]$$

Again, using Newton's method and following the same method as above, we can see that:

$$\delta^{[i+1]} = -\frac{\frac{\log\beta_t^{[i]}}{\sigma^2\beta_t^{[i]}} + \sum_{u\in U} p_{u,t}\left(1-\frac{n_u}{\lambda_u^i}\right)}{\frac{-\log\beta_t^{[i]}+1}{\sigma^2\left(\beta_t^{[i]}\right)^2} + \sum_{u\in U} n_u\left(\frac{p_{u,t}}{\lambda_u^i}\right)^2}$$

In practice, we have found that in the presence of a log-normal prior, Newton's method occasionally overshoots for some transcripts in some of the early iterations of the optimization algorithm. We detect such overshoot events by examining whether $\delta^{[i+1]}$, as calculated by Newtons's method using the equation above, is outside the range between the value obtained from the MLE estimate and 1 (the latter is equivalent to log $\beta_t^{[i+1]}$=0, i.e., equal to the mean of the prior). When this occurs, we minimize the negative log-likelihood function using the 'optimize' function in R. This procedure, in practice, resolves the overshoot problem in a few iterations, so that in the subsequent iterations Newton's method can be used without any overshoots.

2. The variance of the prior distribution for $\boldsymbol{\beta}$

We use an adaptive prior, which is iteratively updated based on the distribution of all values β_t (for all $t\in T$). In other words, after each iteration i, we update the prior variance σ^2 to be the variance of log(β).

3. The library size $\boldsymbol{s_k}$

At each iteration i, we update s_k for each platform as:

$$s_k^{[i+1]} = \frac{\sum_{u\in U_k} n_{k,u}}{\sum_{u\in U_k}\sum_{t\in T} p_{k,u,t}\beta_t^{[i]}}$$

4. The sources of bias in long-read RNA-seq data (γ)

As described in the Methods section, we model the submatrix $\boldsymbol{P}_k$ matrix for long-read data as:

$$\boldsymbol{P}_k = \text{diag}[\exp(\boldsymbol{C\gamma})]$$

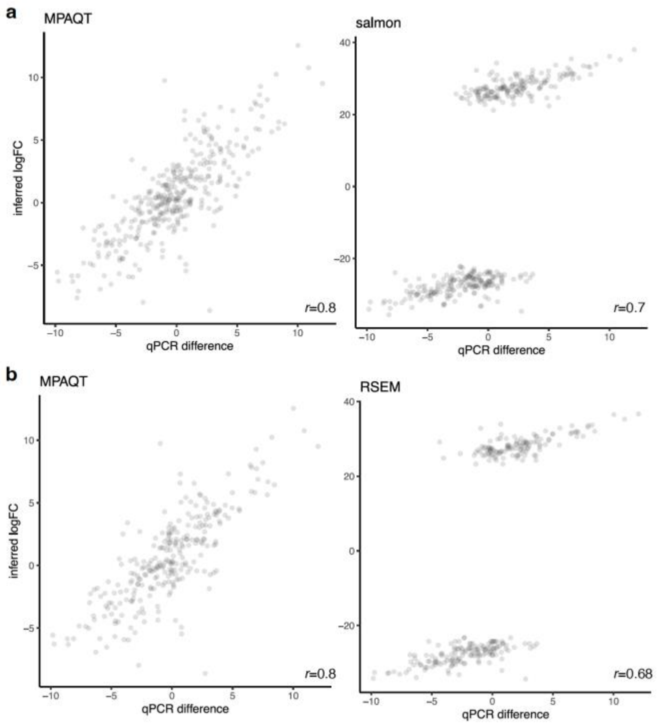
where k is the index of the dataset containing long-read counts, and $\boldsymbol{C}$ is a $|T|\times|D|$ matrix representing the value of variables D, the potential sources of bias, across $|T|$ transcripts. Therefore, we can model the OU counts observed in dataset k as:

$$\lambda_{u\in U_k} = \beta_{t(u)} \sum_{d\in D} \exp\left(c_{u,d}\gamma_d\right)$$

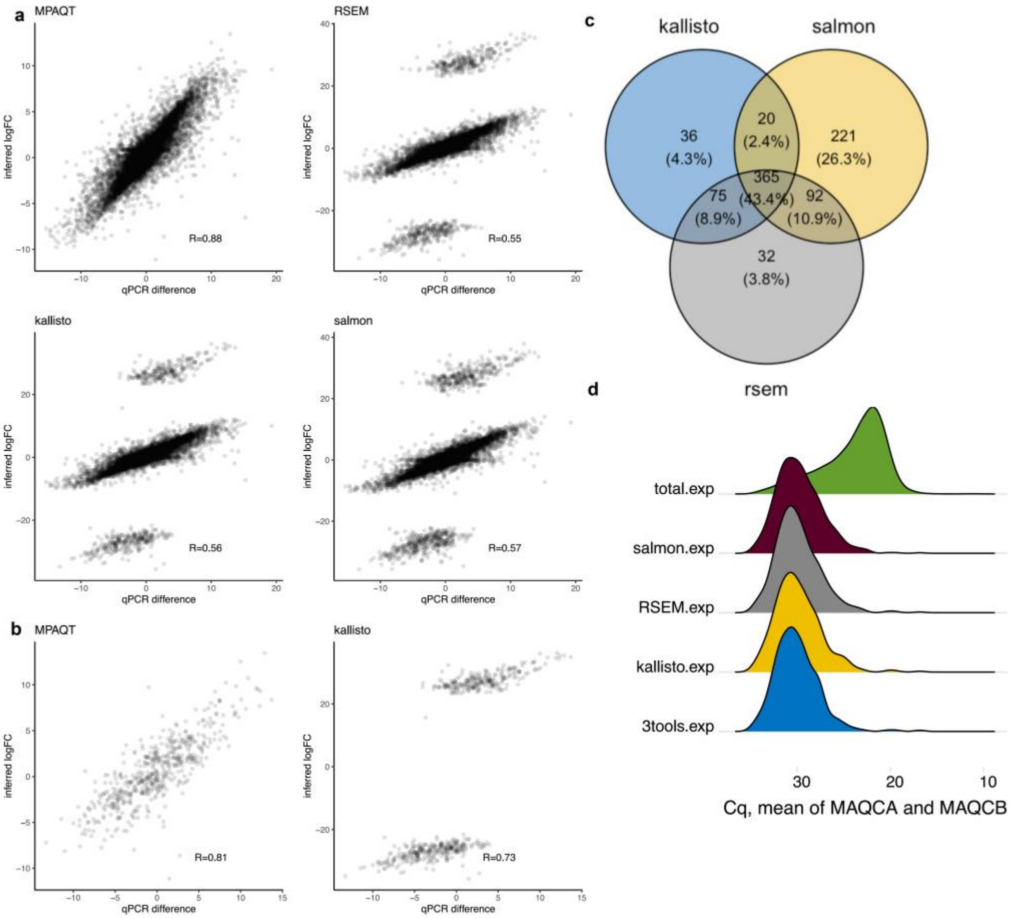
$$n_{u\in U_k} \sim \mathrm{Pois}\left(\lambda_{u\in U_k}\right)$$

Here, $t(u)$ represents the transcript that corresponds to OU u (note that there is a one-to-one relationship between OUs in the long-read data and the transcripts). $c_{u,d}$ represents the element (u,d) of matrix $\boldsymbol{C}$. At each iteration i, we update $\boldsymbol{\gamma}$ by maximizing the model likelihood using the 'glm' function in R with a Poisson error distribution and log-link, with the long-read counts as the dependent variable, $\boldsymbol{C}$ as the independent variables, and $\boldsymbol{\beta}^{[i]}$ as the offset.

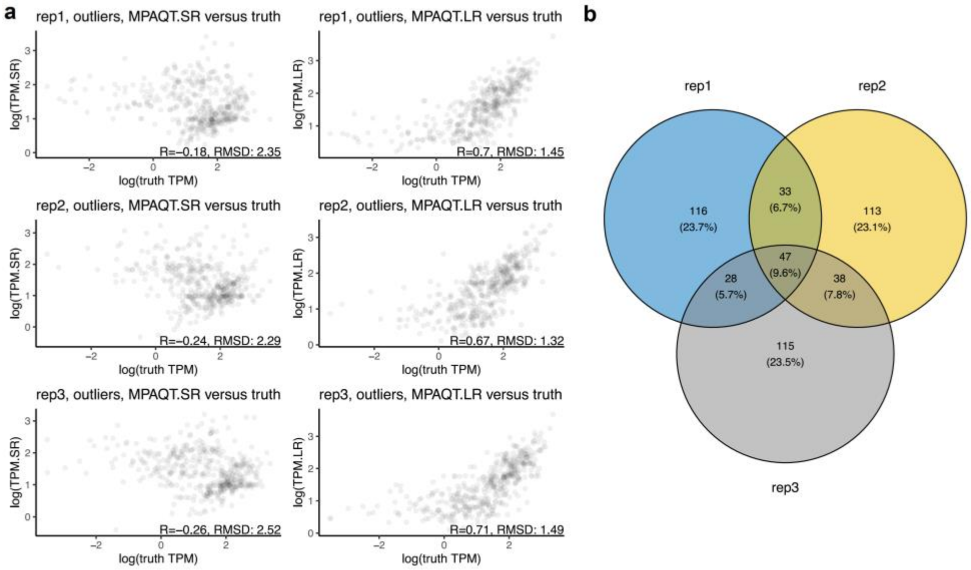
Supplementary Figures



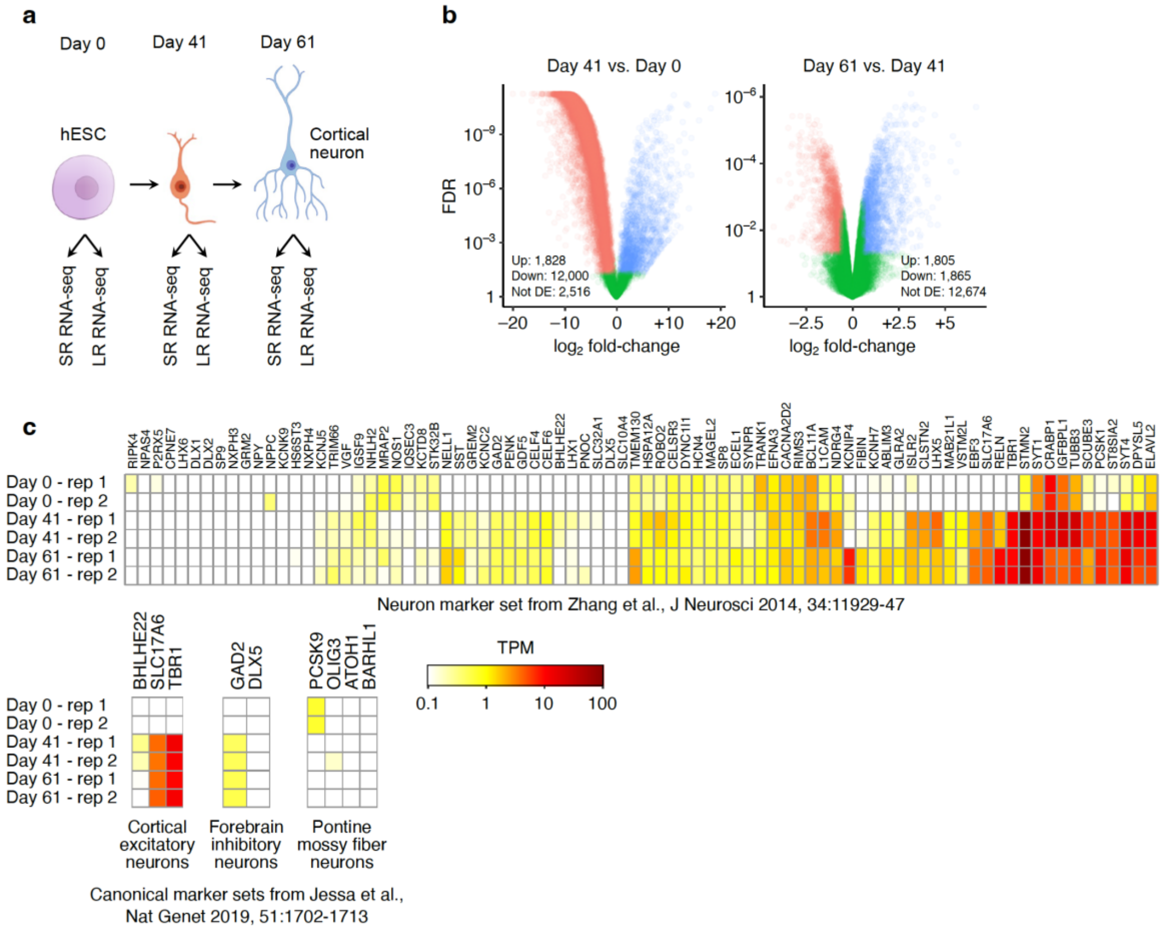
Supplementary Figure 1. MPAQT’s performance on (a) salmon and (b) RSEM’s outliers.



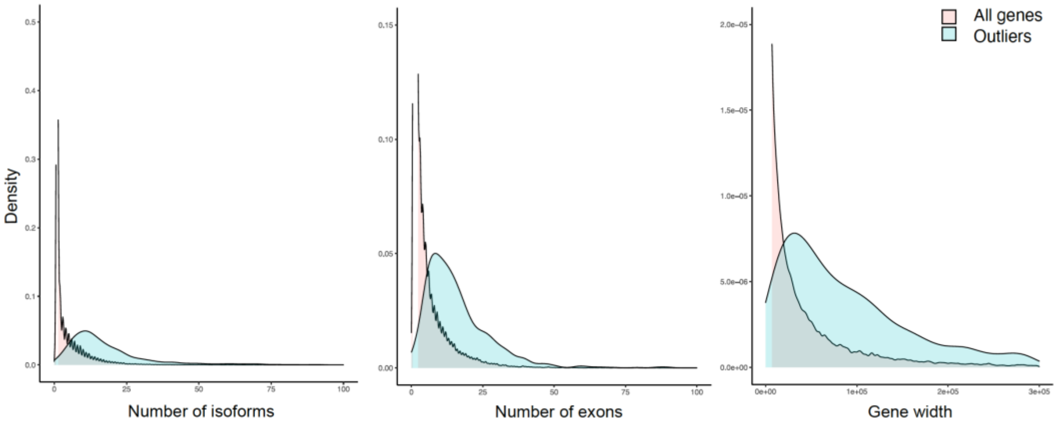
Supplementary Figure 2. Comparison of quantification tools after widening the range of acceptable Cq values. Widening the range of Cq to values between 8-35 does not significantly change the MPAQT performance, while the number of outliers for the other three tools almost doubles. The number of genes remaining after filtering increases from 14,956 to 16,104. **(a-d)** Similar to **Figure 1b-e**, respectively, but for the extended Cq range.



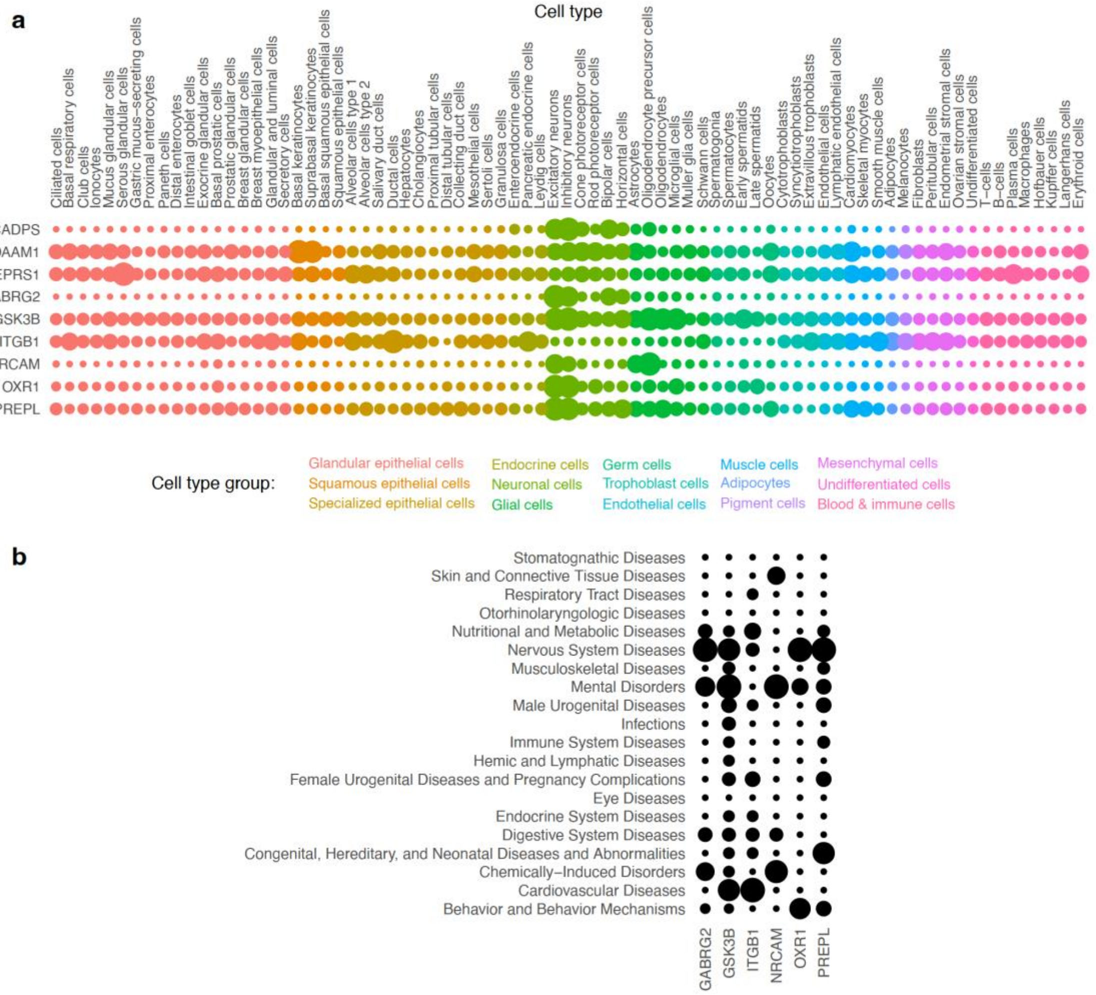
Supplementary Figure 3. Characterization of transcripts differentially quantified by MPAQT upon addition of LR data. Transcripts with substantially different inferences were between SR and LR+SR analyses were extracted (outlier analysis based on Mahalanobis distance >6.36, equivalent to upper-tail P<10–10 for normally distributed data), separately for each of the three randomly generated simulated datasets (with exponential distribution of expression). **(a)** For the subset of outlier genes, quantifications correlate better with ground-truth TPMs in MPAQT analysis of LR+SR data than SR alone across all three replicates. **(b)** Venn diagram of the outliers of the three simulated samples, for genes longer than 250bp.



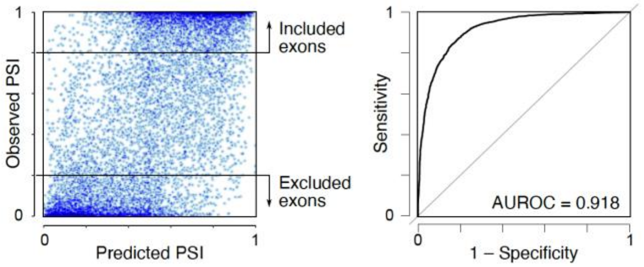
Supplementary Figure 4. Differentiation of hESCs to neurons. (**a**) Schematic of neuronal differentiation. Two replicates were collected for each sample, followed by RNA-seq of each replicate using either short-read or long-read sequencing. (**b**) Volcano plots of upregulated (blue), downregulated (red) and non-DE (green) genes between days 0 and 41 (left) and days 41 and 61 (right). (**c**) Heatmaps showing the TPM of neuronal marker genes from ref[1] (top), and different neuronal subclasses from ref[2] (bottom).



Supplementary Figure 5. Characteristics of differentially quantified transcripts between SR-only and LR+SR inferences at day 61.



Supplementary Figure 6. Cell type and disease associations of top genes whose cassette exons are differentially quantified between SR-only and LR+SR analyses. (**a**) A dot plot representing the cell type-specific expression, based on RNA single-cell data from Human Protein Atlas[3]. Dot size represents TPM of each gene in each cell type, relative to the cell type with maximum expression of that gene. (**b**) Reported gene-disease associations (GDAs); data from DisGenNet[4]. The rows represent disease classes, and the columns are the genes. The size of each dot represents the number of associations with GDA score >0.1, normalized for each gene to the disease class with the largest number of associations.



Supplementary Figure 7. Predicting cassette exon inclusion. The scatterplot of predicted vs. observed PSI (similar to **Figure 6d**) is shown on the left. Exons with observed PSI >0.8 are considered as “positive” observations, and those with PSI <0.2 as “negative” observations, in order to construct the classification ROC curve shown on the right.

References

1. Zhang, Y. et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-11947 (2014).
2. Jessa, S. et al. Stalled developmental programs at the root of pediatric brain tumors. *Nat Genet* **51**, 1702-1713 (2019).
3. Karlsson, M. et al. A single-cell type transcriptomics map of human tissues. *Sci Adv* **7** (2021).
4. Pinero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* **48**, D845-D855 (2020).