Supplementary Materials

Graph-theoretic computing procedures of random c matrix

Step 1. Identify the common regions:
Considering a set of m transcripts, we divide the transcript space into $2^m - m - 1$ sections based on the partition of reads and then identify the pre-common regions.

Step 2. Retain regions longer than read length:
The pre-common regions identified in Step 1, whose lengths do not exceed the read length, are removed. This ensures that only regions capable of generating reads of the specified length are retained.

Step 3. Merge adjacent exons as one vertex and add edges to link adjacent vertexes:

After alternative splicing, introns are excised, leaving only exons. Adjacent exons are then merged into a single vertex, with a dotted edge added to connect adjacent vertexes. Additionally, a sliding window approach is employed to illustrate the merging process and the addition of dotted edges, providing a clear interpretation of the vertexes and edges in this context.

Consider a diagram illustrating this step, focusing on a single region as an example. Assume exons e_1, e_2, e_3, and e_4 have lengths of 200 bp, 200 bp, 300 bp, and 100 bp, respectively, while the read length is 250bp.A sliding window is applied presented in Fig.**??**, starting from the first exon in this region .

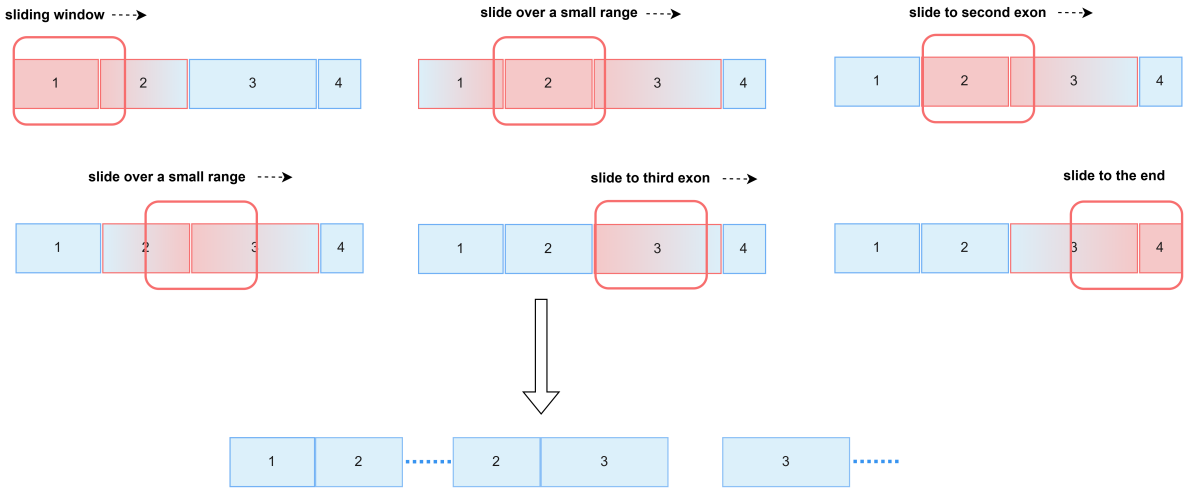
1. Since exon e_1 is shorter than the read length, it cannot independently generate a read. Therefore, e_1 and e_2 are merged into a single vertex, indicating that reads must span both e_1 and e_2.

2. The sliding window is then moved over a small range, and it is found that the reads can span e_1, e_2, and e_3. A dotted edge is added behind the first vertex to reflect this span.

3. The window is then moved to the second exon, e_2. Since e_2 is also shorter than the read length, e_2 and e_3 are merged into a second vertex, indicating that reads must span e_2 and e_3. However, the sliding window reveals that reads cannot span e_2, e_3, and e_4, so no dotted edge is added behind the second vertex.

4. Next, the window is slid to the third exon, e_3, which is longer than the read length and can generate reads on its own. Therefore, e_3 and e_4 are not merged. However, since the reads can span both e_3 and e_4, a dotted edge is added behind the third vertex.

5. Finally, the window is moved to the last exon, e_4. As e_4 is shorter, no further vertexes are generated, and no additional dotted edges are added.



Supplementary figure 1. Sliding window approach. The window sequentially slides across the exons within a region, merging them one by one, and the corresponding weights are computed in the end.

Step 4. Identify the exclusive regions and assign weights to vertexes and edges: After removing repetitive vertexes and edges in all sections, the exclusive regions are identified. Subsequently, a sliding window approach is used to interpret the assignment of weights. The first vertex, which indicates that reads can span exons e_8 and e_9, corresponds to 151 possible starting positions for the reads. Therefore, a weight of 151 is assigned to the first vertex. The dotted edge behind the first vertex, which indicates that reads can span exons e_8, e_9, and e_{10}, corresponds to 49 possible starting positions for the reads, so a weight of 49 is assigned to the dotted edge. Similarly, the second and third vertexes are assigned weights of 200 and 51, respectively, while the dotted edge behind e_{10} is assigned a weight of 100.

Step 5. Compute the random matrix: The total value in each section is computed by summing the weighted vertexes and edges, which represents the numerator of the random matrix elements. The denominator is given by $L(T_j) - L(r) + 1$, where L denotes the length of the respective regions. The elements of the random matrix are then calculated by division.

Fundamental properties of the random matrix

The procedures of random matrix calculation reveal several fundamental properties :

(1) With an increment in read length, the values in the diagonal part of the c matrix are monotonically increaseing.

(2) By the definition of the c-matrix, for each column of the c matrix, the sum of all elements is equal to 1.

(3) The arrangement of non-zero elements in each row is unique to other rows. Specifically, if the first row has only c_{11} as non-zero, then for the m^{th} row, c_{m1} must either be zero or there must be some certain $n \neq 1$ such that both c_{mn} and c_{m1} are non-zero.

(4) With an increment in read length, the values of the c matrix undergo continuous variations without abrupt changes.

The illustration of non-uniqueness of MLE

First, assume that $\boldsymbol{\omega}$ is one of the solutions. Let $\mathbf{h} \in \ker(\mathbf{c})$, where $\mathbf{h} = (h_1, h_2, \dots)^T \neq 0$. The following conditions hold:

$$\begin{cases} \mathbf{ch} = \mathbf{0}, \\ (1,1,\dots,1)^T \times \mathbf{ch} = \sum_{j=1}^{2^m-1} h_j = (1,1,\dots,1)^T \times \mathbf{0} = 0. \end{cases}$$

Next, without loss of generality, assume $\varepsilon > 0$ is sufficiently small such that $\boldsymbol{\omega} + \varepsilon\mathbf{h}$ has non-negative components. In this case, $\boldsymbol{\omega} + \varepsilon\mathbf{h}$ is also a solution to the convex optimization problem.

Since the sum of the components of $\boldsymbol{\omega} + \varepsilon\mathbf{h}$ is equal to 1, we conclude that $\boldsymbol{\omega} + \varepsilon\mathbf{h}$ remains within the feasible domain. Therefore, we have:

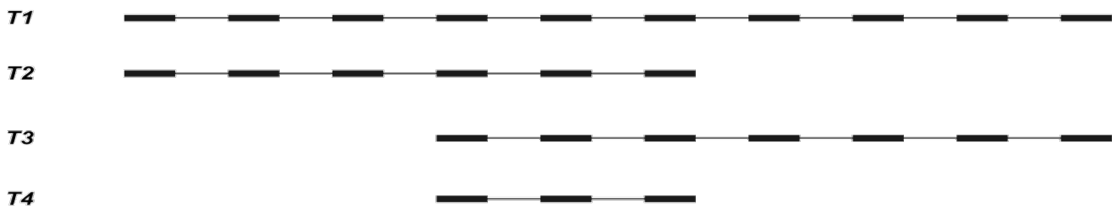
$$\mathbf{c}(\boldsymbol{\omega} + \varepsilon\mathbf{h}) = \mathbf{c}\boldsymbol{\omega} = \mathbf{p}.$$

As a result, we obtain:

$$F(\boldsymbol{\omega}) = F(\boldsymbol{\omega} + \varepsilon\mathbf{h}).$$

This implies that $\boldsymbol{\omega} + \varepsilon\mathbf{h}$ is also a solution. Next, we provide an example to show the multiple solutions.

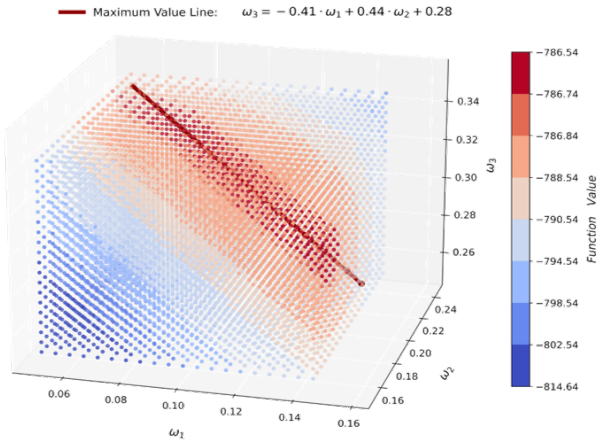
a



b

T_1	0.0	0.0	0.0	0.0
T_2	0.0	0.0	0.0	0.0
T_3	0.0	0.0	0.0	0.0
T_4	0.0	0.0	0.0	0.0
$T_1 \cap T_2$	0.35	0.67	0.0	0.0
$T_1 \cap T_3$	0.47	0.0	0.73	0.0
$T_1 \cap T_4$	0.0	0.0	0.0	0.0
$T_2 \cap T_3$	0.0	0.0	0.0	0.0
$T_2 \cap T_4$	0.0	0.0	0.0	0.0
$T_3 \cap T_4$	0.0	0.0	0.0	0.0
$T_1 \cap T_2 \cap T_3$	0.0	0.0	0.0	0.0
$T_1 \cap T_2 \cap T_4$	0.0	0.0	0.0	0.0
$T_1 \cap T_3 \cap T_4$	0.0	0.0	0.0	0.0
$T_2 \cap T_3 \cap T_4$	0.0	0.0	0.0	0.0
$T_1 \cap T_2 \cap T_3 \cap T_4$	0.18	0.33	0.27	1.00

c



Supplementary figure 2. Example for the multiple solutions. a illustrates the composition of the transcripts. b shows the random matrix of transcripts computed for a read length of 300 bps and exons of 200 bps. c presents the graph of the objective function, highlighting a maximum value of -786.54, with all maxima constrained along the red line.

From the above analysis, it follows that as long as $\boldsymbol{\omega} + \varepsilon\mathbf{h}$ satisfies that the components are non-negative, it must be a solution to the optimization problem. The range of values for ε is given as follows.

For any $\mathbf{h} \in \ker(\mathbf{c})$, we have:

$$\omega_j + \varepsilon h_j \geq 0, \quad j = 1, 2, \ldots, m.$$

To ensure the above condition holds, it is sufficient to satisfy $\varepsilon \leq -\frac{\omega_j}{h_j}$, where $h_j \leq 0$. Therefore, the range of values for ε is:

$$\varepsilon \leq \min\left\{-\frac{\omega_j}{h_j} \mid h_j < 0\right\}.$$

Thus, all solutions to the original convex optimization problem lie within a ball centered at $\boldsymbol{\omega}$ with radius ε. When the upper bound of ε is small, the difference between different solutions is minimal. However, when the upper bound is large, the difference between solutions becomes substantial, leading to instability in the solution when solved using numerical iterative methods. In other words, the solution is highly sensitive to the initial value of the iteration.

Exceptional cases of transcripts complexity

When using data from four organ cells to demonstrate the relationship between complexity ($\kappa(\mathbf{c}^T\mathbf{c})$) and the correlations of different tools, more than 85% of the genes can be explained. However, some undesirable results also arise.

Possible reasons for poor correlation at low condition numbers:

(1) The construction of the transcripts does not conform to our hypothesis, and there is overlap between exon, such as the gene ENSG00000287238.

(2) The exon at the end of a transcript is too long (e.g., gene ENSG00000116260), and when we deal with this kind of data, when the headers are the same, we take the longest length as the length of this exon, so it will lead to the independent part to be longer, and the condition number to be lower, whereas in practice, READ only exists in the independent place.

(3) The presence of two genes overlapping together results in the related transcripts also overlapping together for example the gene ENSG00000233024.

(4) The incomplete transcripts provided by the human genome data led to the calculation of certain transcripts that were missing, which resulted in small condition number calculations, such as gene ENSG00000063244.

(5) From the read data the calculations should be good because the exon that is unique to a transcript is not sampled, but the calculations are not well correlated and it is the software itself that is having problems with the calculations (e.g. gene ENSG00000100804)

(6) There will be errors when constructing the model through the data of transcripts and exons, for example, the tails of two exons differ by a few tens of units of length, at this time our algorithm will think its two exons, while in the actual situation the algorithm will be categorized as one exon, for example, the genes ENSG00000125872, ENSG00000138617.

(7) The transcripts are very similar to each other (e.g., gene ENSG00000090674), but since the dimensionality is not very high (2 or 3 dimensions), the final number of conditions calculated is not very large.

Possible reasons for poor correlation at high condition numbers:

(1) When the number of conditions is higher, the methods used by the algorithms for this gene may be similar, resulting in higher correlation of the final results, e.g., genes ENSG00000047621, ENSG00000113739.

(2) The model construction leads to problems (e.g., gene ENSG00000043355, ENSG00000057019), where we assume that the READ is randomly selected, but in reality the READ is supposed to start from a certain segment, so if there are more UNIQUE parts in a certain end, then it also leads to a better correlation.

(3) Originally the transcripts were all very close to each other, but the fact that an exon belonging to a transcript alone was extracted many times led to the final calculation, e.g. gene ENSG00000078140.

(4) Because we are choosing to use the correlation coefficient as a metric for evaluation, but there are times when the correlation coefficient is relatively large (greater than 0.8) observing the data itself reveals that the proportionality in it does not stand out, which can lead to errors.

The proof of convergence of $\kappa(\boldsymbol{c^T c})$

Since the sum of the elements in each column of the matrix $\mathbf{c}$ equals 1, and applying the Cauchy-Schwarz inequality, it follows that each element of $\mathbf{c}^T\mathbf{c}$ is no greater than 1. Furthermore, the diagonal elements of $\mathbf{c}^T\mathbf{c}$ correspond to the squared lengths of the column vectors of $\mathbf{c}$. Based on previously proven properties of random matrices, the diagonal elements of a random matrix are monotonically increasing with the length of the reads.

To analyze the variation in the diagonal elements of $\mathbf{c}^T\mathbf{c}$ (i.e., the lengths of the column vectors of $\mathbf{c}$), we consider the quadratic function $f(x)$, given by:

$$f(x) = x^2 + (1-x)^2,$$

which is monotonically decreasing on the interval $[0, 0.5]$ and monotonically increasing on the interval $[0.5, 1]$.

Therefore, when the length of the reads is relatively long (i.e., when the diagonal elements of $\mathbf{c}$ exceed 0.5), the diagonal elements of $\mathbf{c}^T\mathbf{c}$ tend to increase as the length of the reads increases. Otherwise, they may initially decrease and then increase.

As for the off-diagonal elements of $\mathbf{c}^T\mathbf{c}$, based on the properties of random matrices, each off-diagonal value is relatively small, and overall, they exhibit a continuous decreasing trend as the length of the reads increases.

According to the Gershgorin Circle Theorem, the eigenvalues of $\mathbf{c}^T\mathbf{c}$ lie within disks centered at the diagonal elements, with radii determined by the sum of the absolute values of the off-diagonal elements in the corresponding rows. As the length of the reads increases, the centers of these disks either continuously approach the vertex $(1, 0)$ in the complex plane or first move away and then move towards it. Meanwhile, the overall trend of the disk radii is a decrease.

Consequently, we can conclude that as the length of the reads increases, the condition number $\kappa(\mathbf{c}^T\mathbf{c})$ is likely to decrease, indicating a reduction in systematic error.

The proof of dimensionality reduction

Without loss of generality, we may assume that the first k (k <m) columns of the diagonal array have a value of 1, i.e., $c_{ii} = 1, i = 1, 2 \cdots k$. In addition to this, we will set:

$$\begin{cases} f(\omega_1, \omega_2 \cdots \omega_m) = \prod_{j=1}^{2^m-1} p_j^{n_j} \\ g(\omega_1, \omega_2 \cdots \omega_k) = \prod_{j=1}^{k} p_j^{n_j} = \prod_{j=1}^{k} \omega_j^{n_j} \\ h(\omega_{k+1} \cdots \omega_m) = \prod_{j=k+1}^{2^m-1} p_j^{n_j} \\ f(\omega_1, \omega_2 \cdots \omega_m) = g(\omega_1, \omega_2 \cdots \omega_k) \times h(\omega_{k+1} \cdots \omega_m) \end{cases} \tag{1}$$

With the introduction of the intermediate variable s, the original optimization problem can be equated to two optimization problems (2) (3).

$$\begin{cases} \textbf{max} \;\; g(\omega_1, \omega_2 \cdots \omega_k) = \prod_{j=1}^{k} p_j^{n_j} = \prod_{j=1}^{k} \omega_j^{n_j} \\ \quad s.t. \quad \boldsymbol{p^{(1)}} = \boldsymbol{c^{(1)}\omega^{(1)}} \\ \qquad \sum_{j=1}^{k} \omega_j = s \\ \qquad \omega_j \geq 0 \quad j = 1, 2 \cdots k \end{cases} \tag{2}$$

$$\begin{cases} \textbf{max} \;\; h(\omega_{k+1} \cdots \omega_m) = \prod_{j=k+1}^{2^m-1} p_j^{n_j} \\ \quad s.t. \quad \boldsymbol{p^{(2)}} = \boldsymbol{c^{(2)}\omega^{(2)}} \\ \qquad \sum_{j=k+1}^{2^m-1} \omega_j = 1 - s \\ \qquad \omega_j \geq 0 \quad j = k+1 \cdots 2^m - 1 \end{cases} \tag{3}$$

Where $s \in (0, 1)$ is the introduced to-be-determined intermediate variable; $\boldsymbol{c}$ and $\boldsymbol{\omega}$ is the chunking matrix, we have:

$$\boldsymbol{c} = \begin{bmatrix} \boldsymbol{c^{(1)}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{c^{(2)}} \end{bmatrix}, \boldsymbol{\omega} = \begin{bmatrix} \omega^{(1)} \\ \omega^{(2)} \end{bmatrix} \tag{4}$$

In addition to this there are

$$\textbf{max} \;\; f(\omega_1, \omega_2 \cdots \omega_m) = \textbf{max} \;\; g(\omega_1, \omega_2 \cdots \omega_k) \times \textbf{max} \;\; h(\omega_{k+1} \cdots \omega_m) \tag{5}$$

The following normalizes $\sum_{j=k+1}^{2^m-1} \omega_j = 1 - s$ in problem (3). Set

$$\begin{cases} \bar{\omega}_j = \omega_j/(1-s) \\ \bar{p}_j = p_j/(1-s), j = k+1 \cdots 2^m - 1 \\ \bar{h}(\omega_{k+1} \cdots \omega_m) = \prod_{j=k+1}^{2^m-1} \bar{p_j}^{n_j} \end{cases} \tag{6}$$

then, we have:

$$\textbf{max} \;\; h(\omega_n, \omega_{n+1} \cdots \omega_m) = \bar{h}(\omega_{k+1} \cdots \omega_m)/(1-s)^{n_{k+1}+\cdots+n_{2^m-1}} \tag{7}$$

Since the constant does not change the optimum vertex, the constant can be transferred to problem (2), then problem (3) is transformed into an s-independent optimization problem (8)

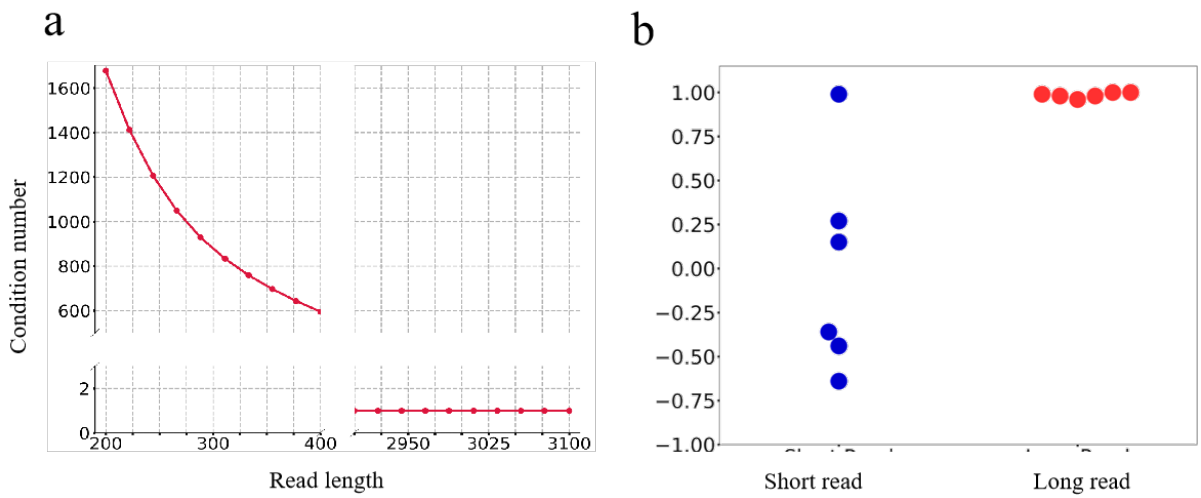
$$\begin{cases} \textbf{max} \;\; \bar{h}(\omega_{k+1} \cdots \omega_m) = \prod_{j=k+1}^{2^m-1} \bar{p_j}^{n_j} \\ \quad s.t. \quad \bar{\boldsymbol{p}}^{(\boldsymbol{2})} = \boldsymbol{c}^{(\boldsymbol{2})} \bar{\boldsymbol{\omega}}^{(\boldsymbol{2})} \\ \qquad \sum_{j=k+1}^{2^m-1} \bar{\omega}_j = 1 \\ \qquad \bar{\omega}_j \geq 0 \quad j = k+1 \cdots 2^m - 1 \end{cases} \tag{8}$$

Noting that $\bar{g}(\omega_1, \omega_2 \cdots \omega_k) = g(\omega_1, \omega_2 \cdots \omega_k) \times (1-s)^{n_{k+1}+\cdots+n_{2^m-1}}$, then the corresponding problem (2) is also transformed into (9)

$$\begin{cases} \textbf{max} \;\; \bar{g}(\omega_1, \omega_2 \cdots \omega_k) = (1-s)^{n_{k+1}+\cdots+n_{2^m-1}} \times \prod_{j=1}^{k} \omega_j^{n_j} \\ \quad s.t. \quad \sum_{j=1}^{n} \omega_j = s \\ \qquad \omega_j \geq 0 \quad j = 1, 2 \cdots k \end{cases} \tag{9}$$

Problem (9) is a simple conditional optimal problem, after calculation it can be seen that $\omega_j = n_j/N$, i.e., the frequency, the assertion is proved.

Then we select a human gene to demonstrate this dimensionality reduction. In the case of short reads the correlations are poor, while the correlations turn good in the case of long reads.



Supplementary figure 3. Correlations heatmap comparison in case of dimensionality reduction. a shows the condition number of short reads (200-400 bps) and long reads (2950-3100 bps). b represents the correlations between four computational tools. Gene ID:ENSG00000063241

Impact of Different Transcripts Expression Levels on Estimation

The expression levels of transcripts have a significant impact on $\Delta \boldsymbol{p}$. This can be explained as follows. By the Central Limit Theorem, when the number of samples N is sufficiently large, the random vector $\boldsymbol{n}$ can be reasonably approximated as following a multivariate normal distribution. For simplicity, we assume that the components of $\boldsymbol{n}$ are approximately independent (though this is not strictly true). Under this assumption, we have:

$$\frac{1}{N}(n_1, n_2, n_3, \ldots, n_{2^m-1}) \sim \text{Multinomial}(p_1, p_2, p_3, \ldots, p_{2^m-1}), \tag{10}$$

with expectations and covariances given by:

$$E(n_j) = N \cdot p_j, \tag{11}$$

$$\text{Cov}(\boldsymbol{n}) = \begin{bmatrix} N \cdot p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & N \cdot p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N \cdot p_{2^m-1}(1-p_{2^m-1}) \end{bmatrix}. \tag{12}$$

By the 3σ-rule of the normal distribution, the probability that $\boldsymbol{n}$ falls within the range $E(\boldsymbol{n}) \pm 3\sqrt{\text{Var}(\boldsymbol{n})}$ is 0.9974. Since $\boldsymbol{p} = \boldsymbol{c\omega}$, the variation $\Delta \boldsymbol{p}$ is affected by the expression levels $\boldsymbol{\omega}$. After each random sampling, $\boldsymbol{n}$ takes a specific value, thereby determining the maximum likelihood function. Due to the complexity of deriving the MLE through theoretical analysis, when the sample size N is sufficiently large, it is reasonable to approximate the MLE by the frequency of $\boldsymbol{n}$, i.e., $\frac{n_j}{N}$. Consequently, the distribution of $\hat{\boldsymbol{p}}$ can be simulated by introducing normal noise, which then allows for the determination of the distribution of $\hat{\boldsymbol{\omega}}$ through the relationship $\hat{\boldsymbol{p}} = \boldsymbol{c}\hat{\boldsymbol{\omega}}$.

To analyze the impact of $\kappa(\boldsymbol{c^T c})$, we conduct a series of simulations by introducing normal noise using human genome data. The simulation error is computed as follows:

$$\begin{aligned} \frac{\boldsymbol{p} + \boldsymbol{\Delta p}}{\|\boldsymbol{p} + \boldsymbol{\Delta p}\|_1} &= \boldsymbol{c}(\boldsymbol{\omega} + \boldsymbol{\Delta \omega}), \\ \boldsymbol{p} &= \boldsymbol{c\Delta \omega}, \end{aligned} \tag{13}$$

leading to:

$$\boldsymbol{\Delta \omega} = \frac{1 - \|\boldsymbol{p} + \boldsymbol{\Delta p}\|_1}{\|\boldsymbol{p} + \boldsymbol{\Delta p}\|_1}\boldsymbol{\omega} + \frac{(\boldsymbol{c}^T\boldsymbol{c})^{-1}\boldsymbol{c}^T\boldsymbol{\Delta p}}{\|\boldsymbol{p} + \boldsymbol{\Delta p}\|_1}. \tag{14}$$

We then select data from two different organelles to illustrate the impact. Our observations indicate that when $\kappa(\boldsymbol{c^T c})$ is low, the effect of $\Delta \boldsymbol{p}$ on the results is minimal, as the correlation coefficients remain high. However, in cases where $\kappa(\boldsymbol{c^T c})$ is large, controlling for systematic error becomes more critical than focusing solely on $\Delta \boldsymbol{p}$ (expression levels).

Organcell	**Transcript**	**StringTie**	**Cufflinks**	**FeatureCount**	**RSEM**
H1-DE	1	6.110411	0.25136593	26.054724	4.13
	2	0.967163	0.00000000	6.810161	0.81
	3	4.931979	0.23184203	19.643231	3.88
WTC11	1	1.951116	0.01853847	8.130225	1.10
	2	0.000000	0.00000000	0.000441	0.00034
	3	7.487917	0.06493518	34.593193	5.52

Table 1: Comparison of different expression levels for gene ENSG00000092978 across two organcells, evaluated using different tools.