Briefings in Bioinformatics, 2026, 27, bbag134
https://doi.org/10.1093/bib/bbag134
Published: 27 March 2026
Problem Solving Protocol

OXFORD

Quantifying transcript complexity via the condition number of gene-specific random matrix

Bo Zhang[1,2,‡], Yaohui Guo[1,3,‡], Guoping Liu[1,*], Meng Zou[1,*]

[1]School of Mathematics and Statistics, Huazhong University of Science and Technology, Luoyu Road 1037, Hongshan District, Wuhan, Hubei 430074, China
[2]Department of Statistics and Data Science, Tsinghua University, Shuangqing Road 30, Haidian District, Beijing 100084, China
[3]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Longxiang Boulevard 2001, Longgang District, Shenzhen, Guangdong 518172, China

*Corresponding authors. Meng Zou, School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan 430074, China. E-mail: zoumeng@hust.edu.cn; Guoping Liu, School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan 430074, China. E-mail: liuguoping@hust.edu.cn
‡Bo Zhang and Yaohui Guo contribute equally to this work.

Abstract

Accurate transcript quantification remains a central challenge, as expression levels estimated by different computational tools (e.g. Cufflinks, StringTie, featureCounts, RSEM) often exhibit substantial discrepancies. The observed variability stems from intrinsic transcriptional architectures and is termed transcript complexity. Here we present a theoretical framework to quantify the transcript complexity via the Condition Number (CN) of a gene-specific random matrix, which is based on two key factors: the repertoire of transcripts generated by alternative splicing and the length distribution of reads by RNA-seq. The CN defines a theoretical bound for quantification error and strongly correlates with inter-tool concordance in real data. The CN decreases with increasing read length, thereby explaining the advantages of long-read sequencing. Moreover, hybrid-seq, integrating short- and long-read, is mathematically guaranteed to achieve error rates no worse than either approach alone, with an optimal mixing ratio yielding further improvements. Notably, this optimal ratio can be determined through grid search. These findings establish the CN as a principled standard for assessing transcript complexity, elucidating a fundamental source of quantification uncertainty and guiding sequencing strategies.

Keywords transcript complexity, condition number, random matrix, hybrid-seq, quantification error

Introduction

Alternative splicing substantially contributes to the high complexity of transcriptomes by generating large amounts of transcripts from the surprisingly small number of genes [1–3]. High-throughput RNA sequencing (RNA-seq) enables comprehensive analysis of transcriptomes, which is instrumental in accurate transcript identification and quantification that ultimately derive biological processes and complex diseases [4, 5].

Accurate transcript identification and quantification are crucial for studying gene regulatory networks, understanding disease mechanisms, developing therapeutic strategies, and advancing personalized medicine. A number of computational tools, including RSEM [6], eXpress [7], Cufflinks [8], featureCounts [9], and StringTie [10], have been developed to quantify transcript expression from short-read RNA-seq data. However, short-read RNA-seq is inherently limited by its relatively short read lengths (50–300 bp), which is much shorter than the average size of human mRNAs (~3 kb) [11].

Long-read RNA-seq offers significant advantages over short-read RNA-seq, particularly in the accurate identification of full-length transcripts [12, 13]. Tools such as StringTie2 [14], FLAMES [15], IsoQuant [16], and Bambu [17], have been developed to identify the transcripts, providing reliable gene expression estimates [18]. The Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium systematically benchmarked these methods in ~427 million long reads from PacBio and Oxford Nanopore technologies (ONT) protocols [19]. However, long-read RNA-seq data still faces challenges such as sequencing errors, read length bias, and the detection of low-abundance transcripts.

Although a large number of tools have been developed for transcript quantification from short and long reads, the quantification of some genes among these tools varies substantially, which stems from intrinsic transcriptional architectures and is termed transcript complexity. For example, two genes, A and B, each have two transcripts—A1 and A2 for gene A, and B1 and B2 for gene B. If A1 and A2 share over 99% sequence identity, accurately quantifying the

Received: November 11, 2025. **Revised:** January 31, 2026. **Accepted:** March 2, 2026

expression levels of A1 and A2 becomes extremely challenging. In contrast, B1 and B2 differ in 99% of their sequences, obtaining precise expression estimates is considerably easier. Therefore, relatively straightforward computational methods can yield accurate results for gene B. However, different tools are likely to produce widely divergent estimates for gene A, placing high demands on the sophistication of bioinformatic algorithms. In conclusion, gene A has high transcript complexity, while gene B exhibits low transcript complexity.

In this study, we propose the CN of a gene-specific random matrix as a theoretical metric to quantify transcript complexity. The matrix is constructed based on two key factors: the repertoire of transcript isoforms from alternative splicing and the reads length distribution. A rank-deficient matrix indicates multiple possible quantification solutions, signifying high complexity, whereas a full-rank matrix yields an upper bound for the theoretical quantification error. Besides that, the metric could strongly reflect the level of inter-tool (Cufflinks, StringTie, featureCounts, RSEM) consistency in real data. We further show that the CN generally decreases with longer read lengths. While it decreases monotonically for most genes, it exhibits a transient increase for a minority. This indicates that longer reads, while usually improving robustness, do not guarantee more reliable quantification for all genes. Theoretically, the error of a hybrid sequencing strategy (short- and long-read) is bounded by that of either method alone [20–22], and can be further minimized by an optimal mixing ratio derived from grid search. This rigorous framework provides a data-independent metric for transcript complexity, elucidating a fundamental source of quantification uncertainty and offering a principled basis for evaluating and developing computational methods.

Methods

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ denote the set of transcripts under consideration, where each transcript is formed by a specific combination of n exons $e_1, e_2, \ldots, e_n$. The expression level of the k-th transcript is denoted by ω_k, normalized such that

$$\sum_{k=1}^{m} \omega_k = 1.$$

For each read (or fragment) r, define its compatibility set $C(r) \subseteq \mathcal{T}$ as the set of transcripts with which r is compatible, i.e. transcripts from which r could have been generated given its sequence and alignment constraints.

For any non-empty subset $K \subseteq \{1, 2, \ldots, m\}$, we define an equivalence class

$$A_K = \{\, r : C(r) = \{T_k : k \in K\} \,\},$$

consisting of all reads whose compatibility set is exactly the transcript subset indexed by K. Under this definition, all reads can be partitioned into at most $2^m - 1$ mutually disjoint non-empty equivalence classes $\{A_K\}$, corresponding to all possible non-empty compatibility sets.

In particular, the first m equivalence classes correspond to reads uniquely compatible with individual transcripts:

$$A_{\{i\}} = \{\, r : C(r) = \{T_i\} \,\}, \qquad i = 1, 2, \ldots, m.$$

Subsequent classes correspond to reads that are ambiguous among multiple transcripts, such as

$$A_{\{i,j\}} = \{\, r : C(r) = \{T_i, T_j\} \,\}, \qquad 1 \le i < j \le m,$$

followed by higher-order compatibility classes involving three or more transcripts, up to the class compatible with all m transcripts.

For notational convenience, we will subsequently replace the set-indexed notation $\{A_K\}$ by a single-index notation $\{A_i\}_{i=1}^{2^m-1}$, where the mapping $i \leftrightarrow K$ follows a fixed lexicographic order of the non-empty subsets $K \subseteq \{1, \ldots, m\}$. In particular, A_i denotes the equivalence class associated with the i-th subset under this order.

Gene-specific random matrix for reads

Considering r is a random read with fixed length, then the probability of r assigned to the equivalence class A_i could be

$$\begin{aligned} P(r \in A_i, q(r) = s) &= \sum_{j=1}^{m} P(r \in A_i, q(r) = s \mid r \in T_j)\, P(r \in T_j) \\ &= P(q(r) = s) \sum_{j=1}^{m} P(r \in A_i \mid r \in T_j, q(r) = s) \\ &\quad \times P(r \in T_j \mid q(r) = s) \\ &= \frac{P(q(r) = s)}{\sum_{q(T_j) \ge s} \omega_j} \sum_{j=1}^{m} c_{ij}(s)\omega_j, \end{aligned}$$

where $q(x)$ is the length of x and $P(r \in T_j \mid \mathbf{q}(r) = s) = \frac{\omega_j}{\sum_{l(T_j) \ge s} \omega_j}$. $c_{ij}(s) = P(r \in A_i \mid r \in T_j, q(r) = s)$ represents the conditional probability of a read being assigned to A_i given it originates from T_j with length s.

Considering a special case with $(T_k) \ge s$, **for** $\mathbf{k} = 1, 2, \ldots, \mathbf{m}$, then

$$P(r \in A_i, q(r) = s) = \sum_{k=1}^{m} c_{ik}(s)\omega_k,$$

Let $p_i = P(r \in A_i, q(r) = s)$, then the probability of reads assigned to the equivalence class could be

$$\mathbf{p} = \mathbf{c}(\boldsymbol{s})\,\boldsymbol{\omega}$$

Where $\mathbf{c}(\boldsymbol{s}) = (c_{ik})$, $\mathbf{p} = (p_1, p_2, \ldots, p_{2^m-1})^{\mathrm{T}}$, $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)^{\mathrm{T}}$ and $\mathbf{c}(\boldsymbol{s})$ is the gene-specific random matrix. Since the columns of the matrix $\mathbf{c} = (c_{ij})$ represent conditional probabilities, they satisfy the following constraint $\sum_{i=1}^{2^m-1} c_{ij} = 1$.

Transcript quantification via maximum likelihood model

Let random variable X_i represent the number of reads assigned to the equivalence class A_i. Assuming that reads with fixed length are generated at random, $X = (X_1, X_2, \ldots, X_{2^m-1})$ follows a multinomial

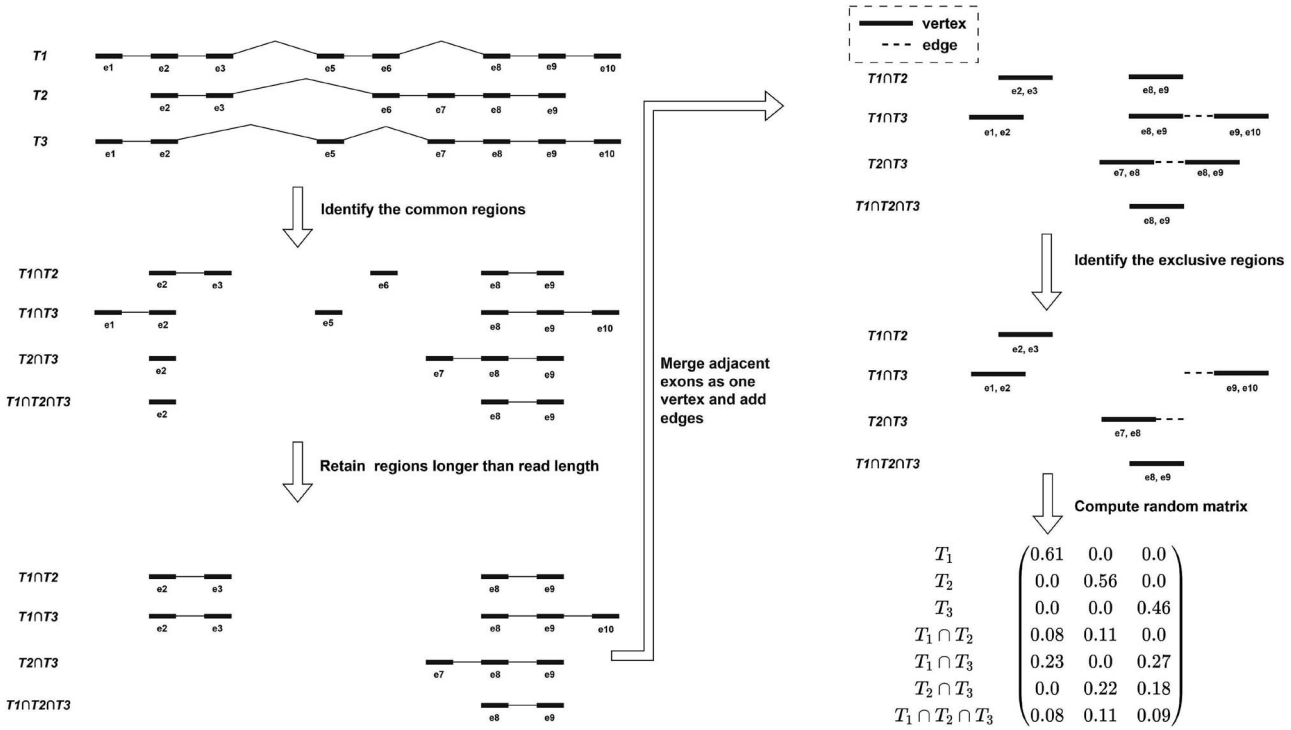

Figure 1 Computational flowchart of random matrix calculation using graph theory. The process of calculating the random matrix consists of five key steps: Step 1, identify common regions across all equivalence classes; step 2, retain regions longer than the read length; step 3, merge adjacent exons into vertices and connect them with edges using a sliding window approach; step 4, remove repetitive regions to identify exclusion regions; step 5, compute the random matrix based on the weights in the graph.

distribution, then the maximum likelihood model could be formulated as

$$\begin{cases} \max & L(\boldsymbol{\omega}) = \prod_i p_i^{x_i} \\ \text{s.t.} & \\ & \mathbf{p} = \mathbf{c}\boldsymbol{\omega} \\ & \sum_{k=1}^{m} \omega_k = 1 \\ & \omega_k \geq 0 \quad k = 1, 2, \ldots, m, \end{cases}$$

where x_i is the observation value of random variables X_i. If the i-th row of the $\mathbf{c}$ contains only zeros, i.e. $c_{ik} = 0$ for all $k = 1, 2, \ldots, m$, then the corresponding $p_i = 0$, indicating that X_i can only take 0. In such cases, p_i does not appear in the objective function $L(\boldsymbol{\omega})$, ensuring that the problem remains well-defined.

Graph-theoretic algorithm for random matrix computation

The random matrix $\mathbf{c}(\boldsymbol{s})$ can be computed using graph-theoretic approaches and involved the following steps (Fig. 1):

Step 1. Identify the common regions. Given m transcripts then the transcript space consists of $2^m - m - 1$ equivalence classes for generating reads. Identify the common regions from all the equivalence classes.

Step 2. Retain regions longer than read length. Remove the pre-common regions whose lengths do not exceed the read length, which 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 vertices. 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 vertices. Additionally, a sliding window approach is employed to illustrate the merging process and the addition of dotted edges. To better demonstrate that this step also holds in the general case (with exons of variable length), we provide an additional example, offering a clearer interpretation of the vertices and edges in this context.

A single region is considered to illustrate the step. 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 250 bp. A sliding window approach is applied which is presented in Supplementary Fig. 1.

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 vertices are generated, and no additional dotted edges are added.

Step 4. Identify the exclusive regions and assign weights to vertices and edges. After removing repetitive vertices and edges in all sections, the exclusive regions are identified. This sliding window approach can also be 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 vertices 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 vertices and edges, which represents the numerator of the random matrix elements. The denominator is given by $q\left(T_j\right)-q(r)+1$. The elements of the random matrix are then calculated by division.

Conditions ensuring uniqueness of the maximum likelihood estimation

Proposition 1. If the random matrix is full rank, the solution to transcript quantification is unique; otherwise, it is not.

Proof. As the Hessian matrix of the log-likelihood function

$$H = \mathbf{c}^T \operatorname{diag}\left(\frac{n_1}{p_1^2}, \frac{n_2}{p_2^2}, \ldots, \frac{n_{2^m-1}}{p_{2^m-1}^2}\right)\mathbf{c}$$

Since $\mathbf{c}$ is full rank, then the objective function is strictly convex. Considering that the constrains are linear, the optimization problem is a convex optimization, which implies that the solution to the optimization problem is unique.

On the other hand, if $\mathbf{c}$ is not of full rank, then the optimal solution not be unique (For further details, see the Supplementary Materials, 'The Illustration of Proposition 1'). For example, considering a gene consists of four transcripts and the transcript space is given in Fig. 2a, where each exon is 200 base pairs (bp) and the read length is 300 bp. In this case, the rank of the random matrix (Fig. 2b) is 3 and the corresponding maximum likelihood function value could be obtained at the red line (Fig. 2c). Similar results would be obtained if the read length is <600 bp, as the random matrix would consistently fail to be of full rank.

Results

Quantification error bounds and a CN-based metric for transcript complexity

If $x_i = E\left(X_i\right)$ then the maximum likelihood estimation (MLE) could obtain $\hat{\mathbf{p}} = \frac{1}{N} E\left(X_1, X_2, \ldots, X_{2^m-1}\right)$. However, the observed x_i values fluctuate around their expectations due to random sampling of reads. When x_i deviates slightly from its expectation, the MLE does not admit a closed-form solution. To address this, we examine the variation of $\mathbf{p}$ by introducing small perturbations $\Delta\mathbf{p}$, thereby assessing the magnitude of the quantification error. The quantification error could be formulated by

$$c\left(\hat{\mathbf{w}} + \boldsymbol{\Delta}\mathbf{w}\right) \quad = \hat{\mathbf{p}} + \boldsymbol{\Delta}\mathbf{p},$$

Using the classical formula in numerical linear algebra,

$$\frac{\| \Delta\boldsymbol{\omega} \|}{\| \boldsymbol{\omega} \|} \leq \kappa\left(\mathbf{c}^{\mathrm{T}}\mathbf{c}\right) \frac{\| \mathbf{c}^{\mathrm{T}}\Delta\mathbf{p} \|}{\| \mathbf{c}^{\mathrm{T}}\mathbf{p} \|},$$

where $\kappa(.)$ denotes the CN.

To demonstrate that quantification error can be measured by $\kappa\left(\mathbf{c}^{\mathrm{T}}\mathbf{c}\right)$, we performed random sampling of reads from human genes in Ensembl release 112 at various read lengths. Simulation results indicate that the quantification error closely tracks changes in the corresponding CN across different read lengths. Overall, the observed trends can be categorized into three distinct scenarios (Fig. 3). In scenario 1, both the CN and the quantification error remain consistently low as the read length increases. For example, in ENSG00000068784, the CN stays below 5, with the corresponding error stabilizing ∼0.05 (Fig. 3a). In scenario 2, the CN decreases with increasing read length, and the quantification error exhibits a similar downward trend. For instance, in ENSG00000012660, the CN declines from over 1000 to nearly 1, accompanied by a reduction in error from above 0.20 to ∼0.05 (Fig. 3b). In scenario 3, the CN first increases and then decreases as the read length grows, and the quantification error follows the same pattern. For example, in ENSG00000152558, the CN rises from 4000 to over 30 000 as the read length grows from 100 bp to 500 bp, with the error increasing from 0.17 to 0.25 (Fig. 3c). The third scenario warrants our primary attention, as it indicates that increasing read length leads to higher transcript complexity and suggests that long-read sequencing may exacerbate quantification errors.

CN strongly correlates with inter-tool concordance in synthetic experiments

To directly evaluate whether CN reflects transcript complexity, we conducted simulations with known ground truth using the BEERS simulator [23]. Synthetic reads were generated from human genes in Ensembl release 112 at a fixed read length of 300 bp, with negligible Polymerase Chain Reaction (PCR) and sequencing biases. Transcript quantification was performed using four widely adopted tools—StringTie, featureCounts, Cufflinks, and RSEM. For each gene, the quantification accuracy was assessed via the Pearson correlation coefficient (PCC) between estimated abundance by these tools and known ground-truth. The statistical analysis demonstrates the CN closely reflects the consistent transcript quantification accuracy (Supplementary Fig. 2a). This monotonic trend indicates that transcripts with higher CN are inherently more complex and thus more challenging to quantify accurately, even under idealized simulation conditions, thereby supporting our theoretical analysis.

To further examine the robustness of the CN, we systematically introduced three common sources of bias into the simulations: read mapping errors, GC(G: Guanine, C:Cytosine) bias, and positional coverage bias (Supplementary Fig. 2b–d). For each bias condition, we computed the Kolmogorov–Smirnov (K–S) statistic comparing PCC distributions across CN intervals between biased and unbiased. In all cases, the K-S test yielded P-values > 0.1, indicating no significant

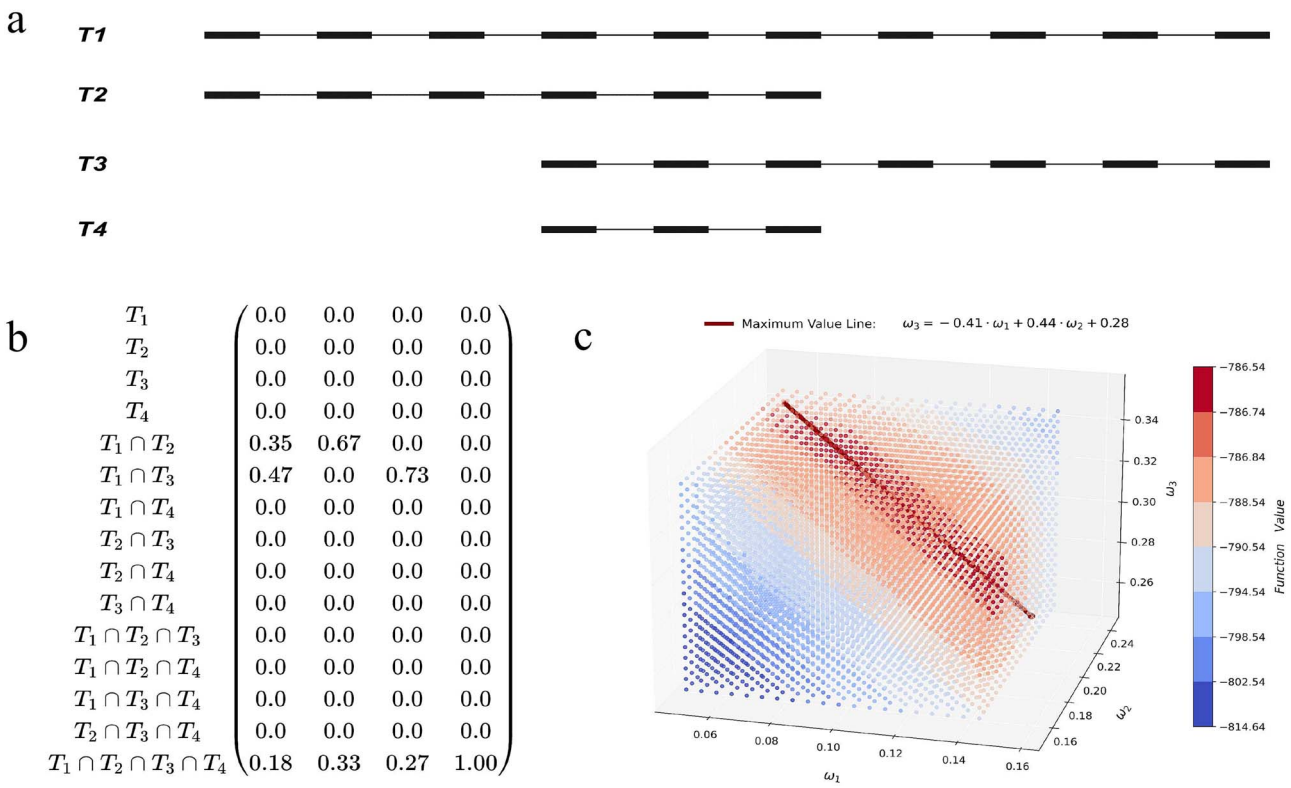

Figure 2 Example of the multiple solutions. (a) Illustrates the composition of the transcripts. (b) Shows the random matrix of transcripts computed for a read length of 300 bp and exons of 200 bp. (c) Presents the graph of the objective function, highlighting a maximum value of −786.54, with all maxima constrained along the red line.

difference in the PCC distributions. These results demonstrate that the relationship between CN and quantification difficulty remains stable in the presence of common experimental biases, supporting the practical utility of CN in real-world RNA-seq applications. In addition, we compared CN with the number of equivalence classes—a widely used proxy for transcript complexity. Our analysis revealed that CN provides a more accurate and mathematically grounded characterization of quantification difficulty. Specifically, we observed that genes with a high number of equivalence classes may still exhibit low CN values (Supplementary Fig. 3), highlighting that structural distinguishability, rather than isoform count alone, drives quantification uncertainty.

CN strongly correlates with inter-tool concordance in real data

Since the true transcript abundances are unknown in real data, we used the inter-tool correlation among four widely adopted quantification tools—StringTie, featureCounts, Cufflinks, and RSEM—as a proxy for quantification error in H1, H9, WTC11, and H1-DE cell lines from the ENCODE database. Considering that read lengths in real datasets are not fixed but follow a distribution roughly between 200 and 400 bp, we approximated the read length as 300 bp to construct the gene-specific random matrix for each gene and calculate its CN. PCC is calculated for each gene between transcript-level expression estimates from every pair of quantification tools (StringTie, featureCounts, Cufflinks, and RSEM). The statistical analysis demonstrates the CN closely reflects the level of inter-tool consistency (Fig. 3d). Genes with CN greater than 250 exhibit poor inter-tool correlation, with a large proportion of correlation coefficients clustering around −0.25, indicating low consistency among quantification results obtained by different tools. In contrast, for genes with CN below 250, the correlation coefficients cluster ∼0.8, reflecting a high degree of consistency across different quantification results. Therefore, we set 250 as the threshold for distinguishing between ‘high’ and ‘low’ CN.

Furthermore, we performed a statistical analysis of the genes from the four cell lines based on their CN and inter-tool correlation levels, and plotted the corresponding histograms (Fig. 3e). Specifically, we quantified the proportions of four gene categories: those with high CN (>250) and low correlations (<0.4); those with low CN (<250) and high correlations (>0.8); those with high correlations where the CN decreases due to reduced transcriptome complexity (e.g. caused by the absence of expression of certain transcripts), which we also classified into the low CN category; and those falling into other special cases. We found that ∼60% of genes with low CN in H1 had a correlation greater than 0.8, while ∼30% of genes with high CN showed poor correlations. A similar pattern was observed in other cell types. Therefore, the quantification of nearly 30% of genes with high CN may be unreliable across different methods, raising concerns about the fairness of their inclusion in benchmark studies. Other special cases include genes characterized by high CN coupled with low correlations, or by low CN coupled with high correlations. Other special cases include genes characterized by high CN coupled with low correlations, or by low CN coupled with high correlations (Supplementary Table 1). The causes of these anomalies may arise from the special structural features of certain genes (e.g. exon overlap, which induces dependence between them). as well as other unexplained factors. For further details, see the Supplementary Materials, ‘Exceptional Cases of Transcript Complexity.’

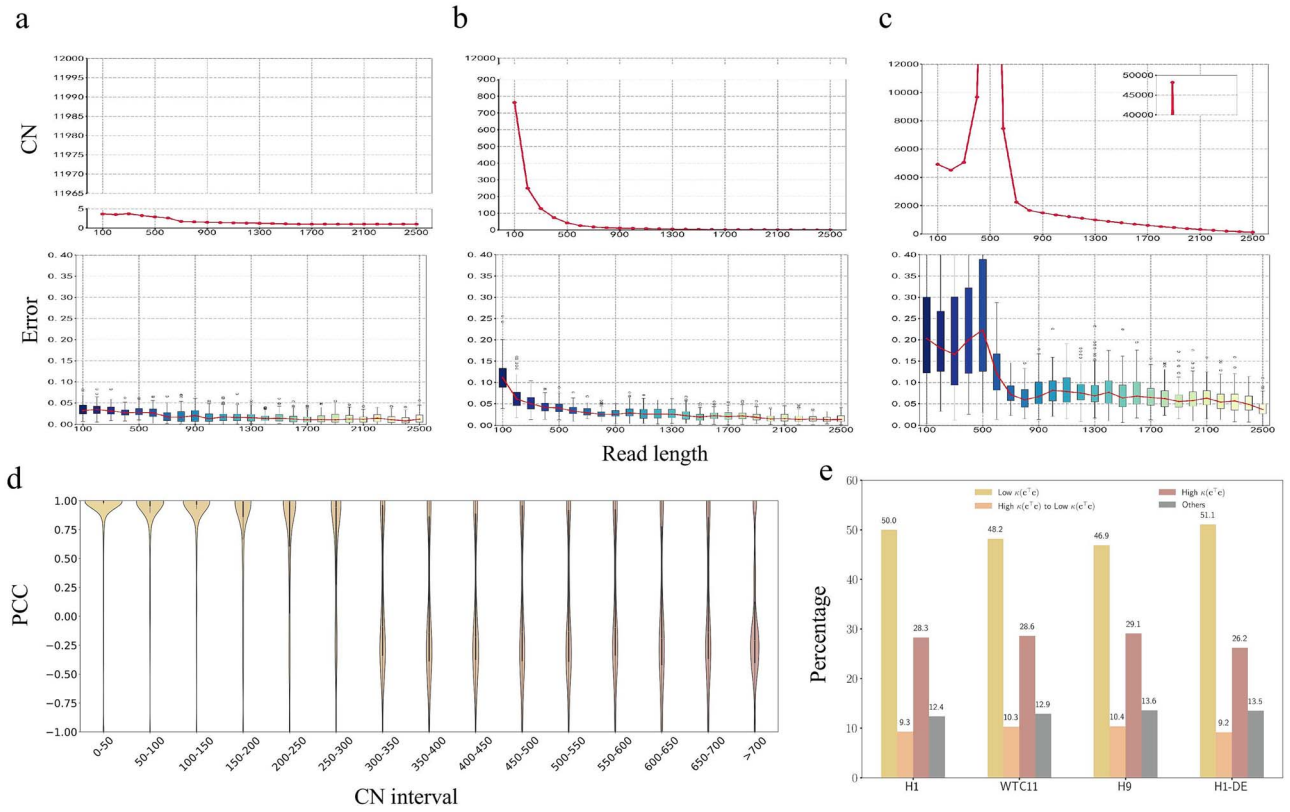

Figure 3 The correlations between CN and quantification error. (a–c) The trends of CN aligns with those of quantification error for all three scenarios. The first two rows denote CN and the corresponding estimation error with the increasing read length, individually for three representative genes (a) gene ID: ENSG00000068784, (b) gene ID: ENSG00000012660, (c) gene ID: ENSG00000152558. (d) The distribution of PCC across CN intervals. Genes are stratified by CN (x-axis), and PCC is calculated for each gene between transcript-level expression estimates from every pair of quantification tools (StringTie, featureCounts, cufflinks, and RSEM). PCC values are aggregated within each CN interval to generate the corresponding distribution. (e) The histogram of CN and correlations across H1 cell line (a similar pattern was observed in other cell lines). High CN means >250 and low CN means <250.

Longer reads may yield lower CN

Proposition 2. For sufficiently long reads, the CN is close to 1.

Proof. Since the sum of the elements in each column of the gene-specific random matrix $\mathbf{c}$ equals 1, and applying the Cauchy-Schwarz inequality, it follows that each element of $\mathbf{c}^{\mathrm{T}}\mathbf{c}$ is no greater than 1. Furthermore, the diagonal elements of $\mathbf{c}^{\mathrm{T}}\mathbf{c}$ correspond to the lengths of the column vectors of $\mathbf{c}$.

Based on the computational procedure of the gene-specific random matrix, it can be shown that the diagonal elements of matrix are monotonically increasing with the length of the reads (For further details, see the Supplementary Materials, 'Fundamental Properties of the Random Matrix').

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 point $(1, 0)$ in the complex plane or first move away and then move toward it. Meanwhile, the overall trend of the disk radii is a decrease.

As the read length increases, the CN of the corresponding random matrix shows a decreasing trend. This observation validates the use of long reads for more accurate estimations. To illustrate this behavior, Fig. 4 depicts the evolution of Gershgorin Circles as read length increases. The ratio of the endpoints serves as an upper bound for the CN. As the endpoints contract toward (1,0) from opposite directions, the CN decreases from 109 to 1.

As the read length increases beyond a certain degree, some diagonal elements of $\mathbf{c}$ approach 1, meaning $c_{ii} = 1$ (Supplementary Fig. 4, also read 'The Proof of Dimensionality Reduction' in the Supplementary Materials for details.). As a result, dimensionality reduction occurs, implying that the reads originating from the i-th transcript can be excluded from the dataset without affecting the estimation results. We provide an example showing that when (for read lengths between 2950 and 3100 bps), the PCC exceeds 0.95. In contrast, for short reads (200–400 bps), the PCC are negative, with values such as −0.36, −0.64, and −0.44 (Supplementary Fig. 2). Furthermore, 14.6%, 18.7%, 23.8%,

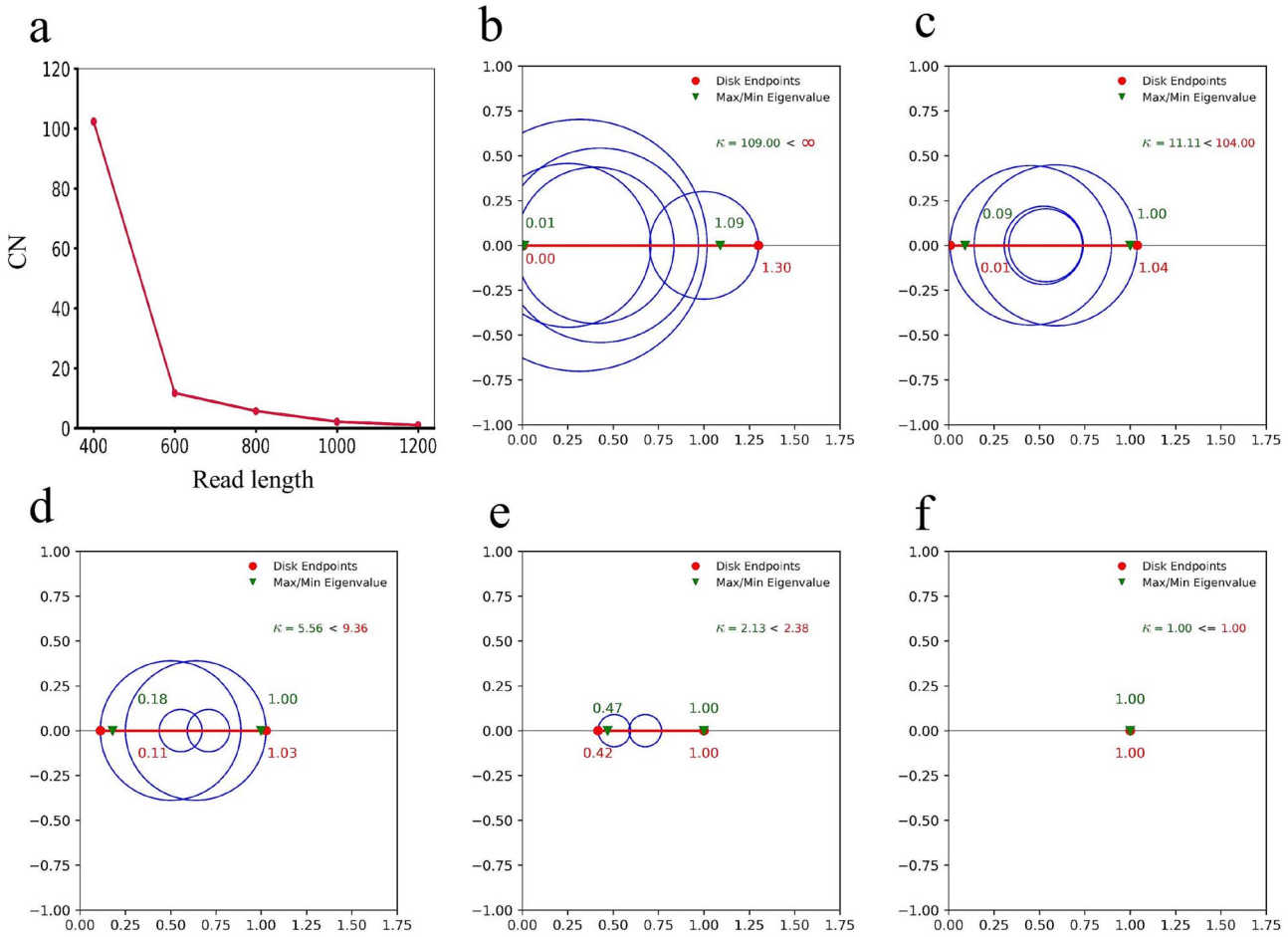

Figure 4 Illustration (b–f): Gershgorin circles showing that CN approaches 1 with increasing read length. The blue circles are Gershgorin circles of a random matrix for ENSG00000136932, the green points are max or min eigenvalues, the red points are disk endpoints.

and 17.8% of genes in H1, WTC11, H9, and H1-DE, respectively, follow a similar trend, indicating that long-read sequencing helps reduce the transcript complexity and makes a more reliable quantification.

Improved quantification with hybrid-seq strategy

To demonstrate the transcript complexity of hybrid-seq, the short-read, and long-read sequencing is used. Assume the ratio of short reads is h_1, then the model could be

$$\begin{cases} \max \quad f(\boldsymbol{\omega}) = \prod_{j=1}^{2^m-1} p_j^{x_{1j}} \prod_{k=1}^{2^m-1} q_k^{x_{2k}} \\ \text{s.t.} \quad \mathbf{p} = \begin{bmatrix} h_1\mathbf{c_1} \\ h_2\mathbf{c_2} \end{bmatrix} \omega \\ \sum_{j=1}^{m} \omega_j = 1 \\ \omega_j \geq 0 \quad \text{for } j = 1, 2, \ldots, m \end{cases}$$

where x_{1j}, x_{2k} are the observed value of X_{1j} and X_{2k} from short-read and long read individually. Similarity, the CN of the hybrid gene-specific random matrix $\begin{bmatrix} h_1\mathbf{c_1} \\ h_2\mathbf{c_2} \end{bmatrix}$ could be $\kappa\left(h_1^2\mathbf{c_1^T c_1} + h_2^2\mathbf{c_2^T c_2}\right)$.

Proposition 3. The CN of hybrid-seq is no greater than the maximum of the CN of the two individual sequencing methods.

Proof. By the Weyl's inequality, $\lambda_i(C)$ denotes the i-th largest eigenvalue of $\mathbf{c}$, we have

$$\begin{aligned} \kappa\left(h_1^2\mathbf{c_1^T c_1} + h_2^2\mathbf{c_2^T c_2}\right) &= \frac{\lambda_1\left(h_1^2 c_1^T c_1 + h_2^2 c_2^T c_2\right)}{\lambda_m\left(h_1^2 c_1^T c_1 + h_2^2 c_2^T c_2\right)} \\ &\leq \frac{\lambda_1\left(h_1^2 c_1^T c_1\right) + \lambda_1\left(h_2^2 c_2^T c_2\right)}{\lambda_m\left(h_1^2 c_1^T c_1\right) + \lambda_m\left(h_2^2 c_2^T c_2\right)} \\ &\leq \max\left\{\frac{\lambda_1\left(h_1^2 c_1^T c_1\right)}{\lambda_m\left(h_1^2 c_1^T c_1\right)}, \frac{\lambda_1\left(h_2^2 c_2^T c_2\right)}{\lambda_m\left(h_2^2 c_2^T c_2\right)}\right\} \\ &= \max\left\{\kappa\left(c_1^T c_1\right), \kappa\left(c_2^T c_2\right)\right\} \end{aligned}$$

By mathematical induction, this conclusion can be extended to multiple sequencing methods with varying read lengths. Therefore,

$$\kappa\left(\sum_{i=1}^{q} h_i \mathbf{c_i^T c_i}\right) \leq \max_i \left\{\mathbf{c_i^T c_i}\right\}$$

Where $\mathbf{c_i}$ is the i-th sequencing method.

Although a general closed-form expression for $\kappa\left(h_1^2\boldsymbol{c_1^T c_1}+h_2^2\boldsymbol{c_2^T c_2}\right)$ is not available, we investigated the entire human genes from ensemble 112. The results indicate that $\kappa\left(h_1^2\boldsymbol{c_1^T c_1}+h_2^2\boldsymbol{c_2^T c_2}\right)$ typically takes values between $\kappa\left(\mathbf{c_1^T c_1}\right)$ and $\kappa\left(\mathbf{c_2^T c_2}\right)$ (Fig. 5a). With the gradual increase of the ratio h_2, certain genes exhibit a marked reduction in their CN, such as ENSG00000160094, which exhibits a dramatic decrease (Fig. 5a). With sequencing depth held constant, the quantification error of this gene decreases initially and then increases as the ratio $h_2$grows, reaching a minimum lower than that observed in both the $h_2 = 0$ and $h_2 = 1$ case (Fig. 5b). This phenomenon is termed as the 'concave' pattern which suggests that hybrid-seq can contribute to more accurate transcript quantification. To identify the optimal $h_2$value that yields the minimal quantification error, a grid search strategy could be adopted via simulation experiments. Beyond the concave pattern, a decline pattern is characterized by a monotonic reduction in quantification error with increasing long-read length. For genes exhibiting this behavior, we recommend employing long-read sequencing technologies directly for quantification, since hybrid strategies do not offer additional improvements. However, for genes exhibiting a continuous increase in quantification error, we recommend employing short-read sequencing technologies for quantification. Subsequently, to evaluate the number of genes for which the hybrid strategy enhances quantification, we conducted simulation experiments on all human genes by mixing short reads of length 300 with long reads ranging from 1000 to 5000, while keeping sequencing depth constant. Statistical analysis demonstrated that the proportion of genes exhibiting a concave pattern rose from 8.5% to 17.7% as the length of the long reads increased (Fig. 5c). For genes whose CN remain consistently low with increasing read length (Fig. 5b), such as ENSG00000068784, we recommend the direct use of short-read sequencing technologies to minimize costs, as neither hybrid nor long-read strategies offer considerable gains in quantification accuracy. Conversely, for genes with consistently high CN, such as ENSG00000158352, the quantification results remain unreliable irrespective of the employed strategy. In addition, the proportion of genes with low CN increases to 93%, compared to 56% with short-read sequencing alone, indicating that ∼37% of genes are more reliably quantified due to the advantages of long-read sequencing.

Therefore, CN enables three key applications to enhance RNA-seq workflows. First, computing CN prior to sequencing identifies transcripts that are challenging to resolve with short reads, thereby guiding targeted experimental strategies such as long-read or hybrid sequencing. Second, reporting CN as a confidence metric alongside expression estimates allows researchers to systematically distinguish robust quantification results from those sensitive to methodological variation. Third, CN-stratified evaluation in method benchmarking distinguishes true algorithmic performance from limitations imposed by inherent transcriptome complexity. Collectively, these applications translate CN from a theoretical metric into a practical tool for guiding experimental design, supporting data interpretation, and refining methodological assessment in transcript quantification studies.

Discussions

In this study, we presented a systematic approach for quantifying transcript complexity using the CN of a random matrix for each gene. The CN serves as a reliable upper bound for quantification errors in RNA-seq data. Our findings reveal that genes with low CN exhibit consistent correlations across inter-tools, including Cufflinks, StringTie, and RSEM, whereas genes with high CN pose greater challenges for accurate quantification. These results underscore the importance of accounting for transcript complexity in RNA-seq analyses and suggest that methods for handling more complex genes may require further refinement to address these inherent difficulties.

Our study also uncovered a significant relationship between read length and the CN. For sufficiently long reads, the CN is close to 1. This supports the notion that long-read RNA-seq technologies are particularly advantageous for improving quantification accuracy, especially for genes with high CN in short reads. Long reads capture more information from the transcript, thereby mitigating uncertainties in expression estimates and improving consistency across quantification methods. In contrast, hybrid sequencing approaches, which combine short and long reads, did not directly reduce the CN but resulted in lower quantification errors. Short reads provide high throughput, while long reads help resolve ambiguities in transcript assembly, ultimately reducing quantification errors.

Besides, we find that transcript expression level has limited influence on quantification estimates. Rather, concordance across tools is governed by identifiability—low $\kappa\left(c^\top c\right)$ yields stable estimates, whereas high κdrives discrepancies. For further details, See 'Impact of Different Transcripts Expression Levels on Estimation' in the Supplementary Materials.

We acknowledge that certain real-world factors influencing RNA-seq quantification—including read mapping errors, transcript discovery uncertainty, mRNA degradation bias, and residual adapter reads—were intentionally excluded from the current model to maintain theoretical tractability. While these factors can affect observed read distributions and downstream expression estimates, it should be noted that the CN is a structural metric determined solely by transcript architecture and sequencing read length. Consequently, CN remains invariant to experimental noise, which instead introduces stochastic variation on top of an underlying identifiability structure. In this sense, CN reflects the intrinsic complexity of transcripts independent of technical variability.

Looking ahead, future refinements of this framework could formally incorporate such experimental factors to enhance practical utility. For example, transcript discovery uncertainty could be addressed through model selection approaches such as the Bayesian information criterion to infer an optimal transcript set. Similarly, biases stemming from degradation or adapter contamination could be integrated within a probabilistic likelihood model that simultaneously accounts for structural identifiability and technical artifacts. We consider these as promising directions for advancing CN-guided quantification in real-world RNA-seq studies, while underscoring that the current formulation provides a foundational architecture-aware perspective on quantification complexity.

In conclusion, this study underscores the necessity of systematically characterizing transcriptome complexity and integrating such insights into RNA-seq analytical workflows. By adopting the CN as a measure of intrinsic quantification uncertainty and employing complementary sequencing strategies, we establish a framework that enhances the accuracy and reliability of transcript-level quantification. More importantly, CN provides a mathematically grounded uncertainty metric that improves the interpretability of transcriptomic data for downstream functional analyses. This methodological advance strengthens the analytical foundation of molecular clinical resources, such as the tumor treatment drug database TGDD [24], thereby supporting

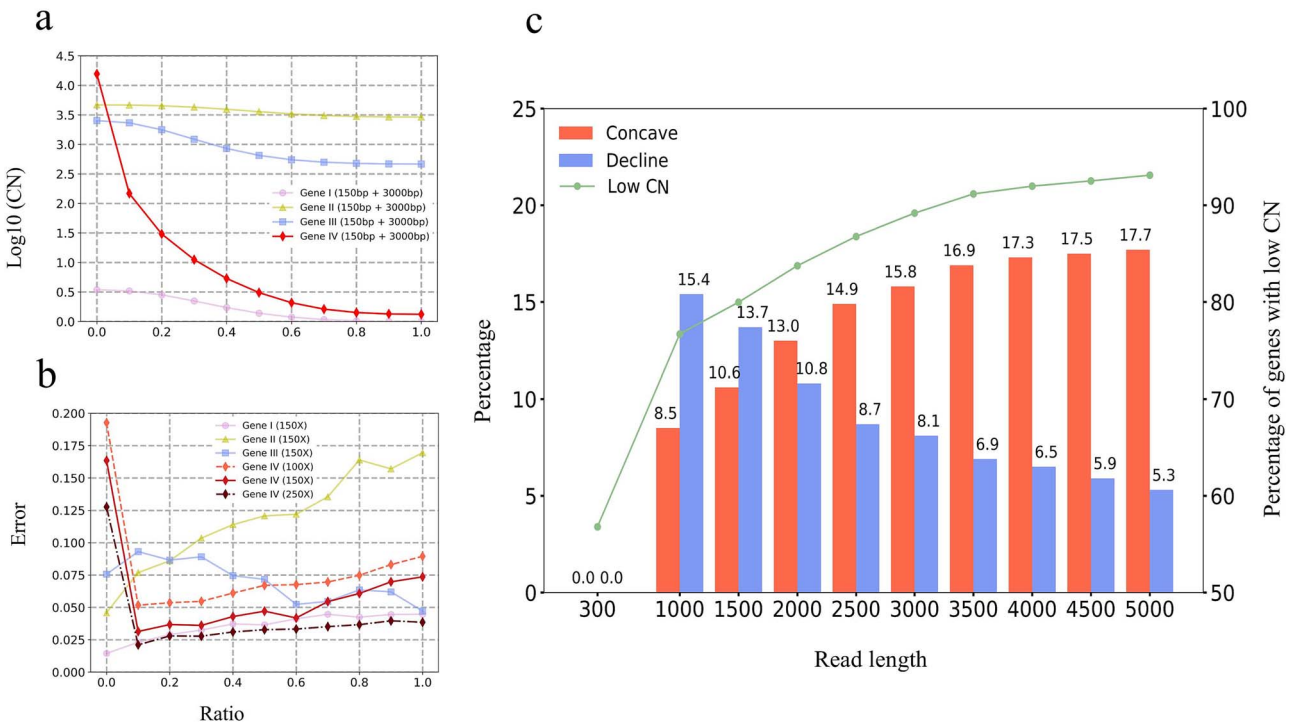

Figure 5 The CN and the corresponding quantification error in hybrid-seq. (a) The CN varies with different ratios of short-to-long reads for four representative genes: Gene I: ENSG00000068784, gene II: ENSG00000158352, gene III: ENSG00000100764, gene IV: ENSG00000160094. (b) The corresponding quantification error varies with different ratios of short-to-long reads for four representative genes with different sequencing depth. (c) The histogram of concave and decline pattern with different long-read length. When the long-read length increases to 5000 bps, the proportion of genes with a concave pattern increases to 17.7%, and the percentage of genes with low CN rises to 93%.

more robust drug-target mapping and informed therapeutic decision-making in precision oncology. Future research will address the limitations outlined above, incorporating additional sources of error and refining our models to better capture the complexities of transcriptome analysis. We anticipate that these efforts will lead to more precise and robust methods for transcript quantification, ultimately enhancing our ability to characterize gene expression across diverse biological contexts.

Key Points

- We present a theoretical framework to quantify the transcript complexity via the Condition Number (CN) of a gene-specific random matrix, which is based on two key factors: the repertoire of transcripts generated by alternative splicing and the length distribution of reads by RNA-seq.
- The CN provides a theoretical bound for quantification error and strongly correlates with inter-tool concordance in real data like H1, H9, WTC11, and H1-DE.
- The CN decreases with increasing read length, and it is close to 1 for sufficiently long reads, thereby explaining the advantages of long-read sequencing.
- Theoretically, the error of a hybrid sequencing strategy (short- and long-read) is bounded by that of either method alone, and can be further minimized by an optimal mixing ratio derived from grid search.

Acknowledgments

The authors are supported by the National Natural Science Foundation of China (NSFC) under Grants No. 11422108 and 12001215. The authors are supported by Interdisciplinary Research Program of HUST 2024JCYJ004. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

Author contributions

MZ and GL design and support the project. MZ, BZ, and YG complement the theoretical and experiment study. MZ, GL, BZ, and YG finish the paper.

Conflict of interests

None of conflicts declared.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Data availability

The human cell line gene data used in this study were obtained from the Encode Database. Specifically, for the cell line H1, the short-read data can be accessed at https://www.encodeproject.org/experiments/ENCSR588EJX/ with accession numbers ENCFF644AQW

and ENCFF766OAK, and the long-read data at https://www.encodeproject.org/experiments/ENCSR339KRF/ with accession number ENCFF153SIE. For the cell line WTC11, short-read data is available under accession numbers ENCFF644AQW and ENCFF766OAK, and the long-read data is accessible under accession number ENCFF153SIE. For the cell line H9, short-read data can be accessed using accession number ENCFF778LAE, and the long-read data with accession number ENCFF688QGB. For the cell line H1-DE, short-read data is accessible under accession numbers ENCFF451LZF and ENCFF783SRR, while the long-read data is available under accession number ENCFF884GOA. All the computation code used for data analysis is available at https://github.com/HUSTzoulab/TCCN, where all tools, parameters, and workflows used for gene processing are also provided.

References

1. Modrek B, Lee C. A genomic view of alternative splicing. *Nat Genet* 2002;**30**:13–9. https://doi.org/10.1038/ng0102-13
2. Marasco LE, Kornblihtt AR. The physiology of alternative splicing. *Nat Rev Mol Cell Biol* 2023;**24**:242–54. https://doi.org/10.1038/s41580-022-00545-z
3. Stamm S, Ben-Ari S, Rafalska I *et al.* Function of alternative splicing. *Gene* 2005;**344**:1–20. https://doi.org/10.1016/j.gene.2004.10.022
4. Costa V, Angelini C, De Feis I *et al.* Uncovering the complexity of transcriptomes with RNA-Seq. *Biomed Res Int* 2010;**2010**:853916.
5. Marguerat S, Bahler J. RNA-seq: from technology to biology. *Cell Mol Life Sci* 2010;**67**:569–79. https://doi.org/10.1007/s00018-009-0180-6
6. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform* 2011;**12**:1–16. https://doi.org/10.1186/1471-2105-12-323
7. Forster SC, Finkel AM, Gould JA *et al.* RNA-eXpress annotates novel transcript features in RNA-seq data. *Bioinformatics* 2013;**29**:810–2. https://doi.org/10.1093/bioinformatics/btt034
8. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat Biotechnol* 2010;**28**:511–5. https://doi.org/10.1038/nbt.1621
9. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656
10. Pertea M, Pertea GM, Antonescu CM *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5. https://doi.org/10.1038/nbt.3122
11. Piovesan A, Antonaros F, Vitale L *et al.* Human protein-coding genes and gene feature statistics in 2019. *BMC Res Notes* 2019;**12**:315. https://doi.org/10.1186/s13104-019-4343-8.
12. Amarasinghe SL, Su S, Dong X *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**:30. https://doi.org/10.1186/s13059-020-1935-5
13. Hoang NV, Furtado A, Mason PJ *et al.* A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics* 2017;**18**:395. https://doi.org/10.1186/s12864-017-3757-8
14. Kovaka S, Zimin AV, Pertea GM *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* 2019;**20**:278. https://doi.org/10.1186/s13059-019-1910-1
15. Tian L, Jabbari JS, Thijssen R *et al.* Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* 2021;**22**:310. https://doi.org/10.1186/s13059-021-02525-6
16. Prjibelski AD, Mikheenko A, Joglekar A *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* 2023;**41**:915–8. https://doi.org/10.1038/s41587-022-01565-y
17. Chen Y, Sim A, Wan YK *et al.* Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* 2023;**20**:1187–95. https://doi.org/10.1038/s41592-023-01908-w
18. Ament IH, DeBruyne N, Wang F *et al.* Long-read RNA sequencing: a transformative technology for exploring transcriptome complexity in human diseases. *Mol Ther* 2025;**33**:883–94. https://doi.org/10.1016/j.ymthe.2024.11.025
19. Pardo-Palacios FJ, Wang D, Reese F *et al.* Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* 2024;**21**:1349–63. https://doi.org/10.1038/s41592-024-02298-3
20. Au KF, Sebastiano V, Afshar PT *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci* 2013;**110**:E4821–30. https://doi.org/10.1073/pnas.1320101110
21. Rhoads A, Au KF. PacBio sequencing and its applications. *Genom Proteomics Bioinform* 2015;**13**:278–89. https://doi.org/10.1016/j.gpb.2015.08.002
22. Weirather JL, Afshar PT, Clark TA *et al.* Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res* 2015;**43**:e116–6. https://doi.org/10.1093/nar/gkv562
23. Brooks TG, Lahens NF, Mrčela A *et al.* BEERS2: RNA-Seq simulation through high fidelity in silico modeling. *Brief Bioinform* 2024;**25**:3. https://doi.org/10.1093/bib/bbae164
24. Li W, Xiong Z, He K *et al.* TGDD: a genomics-based database for tumor treatment drugs. *iCell* 2024;**1**:1. https://doi.org/10.71373/UYBT8671

Briefings in Bioinformatics, 2026, **27**, bbag134
https://doi.org/10.1093/bib/bbag134
Problem Solving Protocol