BMC Genomics

RESEARCH Open Access

Enhancing RNA-seq bias mitigation with the Gaussian self-benchmarking framework: towards unbiased sequencing data

Qiang Su[1,4*†], Yi Long[2†], Deming Gou[3], Junmin Quan[4*] and Qizhou Lian[1,5,6*]

Abstract

Background RNA sequencing is a vital technique for analyzing RNA behavior in cells, but it often suffers from various biases that distort the data. Traditional methods to address these biases are typically empirical and handle them individually, limiting their effectiveness. Our study introduces the Gaussian Self-Benchmarking (GSB) framework, a novel approach that leverages the natural distribution patterns of guanine (G) and cytosine (C) content in RNA to mitigate multiple biases simultaneously. This method is grounded in a theoretical model, organizing k-mers based on their GC content and applying a Gaussian model for alignment to ensure empirical sequencing data closely match their theoretical distribution.

Results The GSB framework demonstrated superior performance in mitigating sequencing biases compared to existing methods. Testing with synthetic RNA constructs and real human samples showed that the GSB approach not only addresses individual biases more effectively but also manages co-existing biases jointly. The framework's reliance on accurately pre-determined parameters like mean and standard deviation of GC content distribution allows for a more precise representation of RNA samples. This results in improved accuracy and reliability of RNA sequencing data, enhancing our understanding of RNA behavior in health and disease.

Conclusions The GSB framework presents a significant advancement in RNA sequencing analysis by providing a well-validated, multi-bias mitigation strategy. It functions independently from previously identified dataset flaws and sets a new standard for unbiased RNA sequencing results. This development enhances the reliability of RNA studies, broadening the potential for scientific breakthroughs in medicine and biology, particularly in genetic disease research and the development of targeted treatments.

Keywords RNA sequencing, Bias mitigation, Gaussian self-benchmarking (GSB), GC content

†Qiang Su and Yi Long contributed equally to this work.

*Correspondence:
Qiang Su
su@chemie.uni-siegen.de
Junmin Quan
quanjm@pku.edu.cn
Qizhou Lian
qz.lian@siat.ac.cn

Full list of author information is available at the end of the article

Background

RNA sequencing (RNA-seq) is a vital technique for in-depth exploration of the transcriptome, opening windows into gene expression, disease patterns, and cell-type-specific signatures [1, 2]. It utilizes both short-read and long-read base-calling technologies. The advantage of long-read technologies lies in their ability to sequence complete transcripts via single-molecule sequencing, offering a detailed view of entire RNA molecules or their cDNA [3]. Nonetheless, these methods face challenges with low throughput and higher error rates, making them less effective for low to medium expression levels [4–6]. Short-read technologies, by contrast, use a parallel approach to generate numerous short snippets [7, 8], effectively capturing a broad spectrum of transcripts, including those at low abundance, and yielding high-quality data amenable to complex analyses [9]. Nevertheless, the effectiveness of short-read RNA-seq hinges critically on a meticulously organized workflow. This process encompasses several critical steps, including the fragmentation of RNA, synthesis of cDNA, ligation of adaptors, amplification of the library, sequencing, and base calling. This leads to the scattering of sequence information from complete transcripts across several short reads, introducing biases that can adversely impact the accuracy of transcript reconstruction and quantification [10–12]. To ensure precise analysis, it's crucial to address and mitigate these biases in RNA-seq datasets through systematic and comprehensive bias modeling.

In the realm of short-read RNA-seq datasets, the presence of complex co-existing biases is a common challenge. These observed biases encompass GC bias (a correlation between read coverage and GC content), fragmentation or degradation bias (resulting from RNA body positional survivorship), library preparation bias (due to hexamer-associated binding preference and PCR amplification), mapping bias (stemming from specific characteristics of RNA molecules), and experimental bias (caused by factors such as sequencing depth or batch effects) [13–19]. The traditional strategy for mitigating these biases entails calculating bias-specific weighting parameters for each identified influence by consulting empirical sequencing data. This involves the use of statistical models tailored to single out factors like GC content, the location within the transcript, or the likelihood of hexamer binding [16, 20–23]. While these models can indeed diminish bias-specific variability in the distribution of sequencing reads across genomes or transcripts, achieving truly unbiased coverage is still an unachieved objective. This suggests a fundamental flaw in the approach: modeling for a single bias at a time is insufficient to address the challenge of concurrent biases, highlighting the complexity of dealing with multiple biases simultaneously. The intricate dynamics between these biases complicates the task of achieving thorough or even precise mitigation of biases. Therefore, a deeper understanding of the multifaceted mechanisms that lead to phenotypic biases in RNA-seq data is essential. Moreover, there's an inherent flaw in the reliance on empirical sequencing data to set specific bias-related weighting factors in traditional methods. Given that this empirical data is already biased, using it as a foundation for adjustment factors could compromise the effectiveness of bias mitigation measures. This situation indicates a clear need for a novel model that can simultaneously address multiple biases in a cohesive manner and function independently from the biases ingrained in empirical data.

In our research, we have developed a Gaussian self-benchmarking (GSB) framework using a Gaussian distribution based on GC content that precisely identifies and corrects for biases within the k-mer counting framework. This methodology leverages the observation that the distribution of guanine (G) and cytosine (C) across natural transcripts inherently follows a Gaussian distribution when k-mer counts are categorized and aggregated by their GC content. In this context, the GC content is not seen as an irregular source of bias, but rather as a foundational element for building a robust, theoretically-derived count-distribution model. This approach enables a self-benchmarking workflow specific to transcripts, relying on a dual-distribution model: a theoretical even distribution for k-mer modeling alongside an empirically observed uneven distribution reflecting the variability of sequencing k-mers along the transcript length. Central to our methodology are the steps of categorizing k-mers, aggregating counts of GC-indexed k-mers, and fitting these count aggregates within a Gaussian distribution keyed to their GC content. The methodology begins with establishing a benchmark for evaluating transcript-specific data, which assumes a uniform distribution of k-mers. This step involves analyzing the collective counts of k-mers grouped by their GC content and projecting these aggregates onto a Gaussian distribution. The success of our approach hinges on accurately establishing key parameters-mean and standard deviation-that embody the unique distribution characteristics of each transcript. Similarly, sequencing data is organized by GC content and subjected to a Gaussian fitting process using these predetermined parameters from modeling data. The Gaussian-distributed counts thus generated act as unbiased indicators of sequencing counts for each GC-content category. The predicting counts obtained for each GC category can be averaged over all the corresponding k-mers within that category, thereby enabling a systematic reduction of bias at targeted positions throughout the transcript. Distinguished by its self-benchmarking capability, our GC-content-based GSB framework surpasses traditional empirical and statistical approaches by offering a theoretical benchmark for adjusting multiple biases simultaneously rather than approximating them individually. Crucially, the foundational parameters of our model are determined independently of any empirical sequencing data, ensuring that the Gaussian function reflects true abundance contributions and excludes all bias-related distortions. To affirm the reliability of our method, we implemented a thorough validation protocol that integrates theoretical constructs with empirical evaluations. This validation process involves conducting experiments with both artificial RNA constructs and genuine human tissue RNA samples, ensuring the robustness and applicability of our approach.

Methods

Data analysis

The software pipeline for processing paired-end raw data employs several key applications: FastQC (version 0.11.8) by Babraham Bioinformatics for quality control checks, Cutadapt (version 2.10) and Trimmomatic (version 0.39) for trimming adapters and low-quality sequences, HISAT2 (version 2.2.1) for aligning reads to the human reference genome GRCh38.p14 (Ensembl), Samtools (version 1.7) for post-alignment processing, and the RSeQC package for quality control of RNA-seq data. The alignment process entails mapping paired-end reads to the aforementioned human reference genome. Additionally, k-mer segments are aligned to transcript references as annotated in Ensembl, ensuring a comprehensive analysis of the genomic data. The data analysis pipeline is implemented using custom scripts, which are provided in the supporting information.

Collection of tissue samples

Human colorectal samples were ethically acquired from the Sun Yat-sen University Cancer Center and Shenzhen University General Hospital. Before the collection of samples, each donor provided informed consent, authorizing the retrieval of biopsies and enabling the conduct of extensive molecular profiling of their transcriptomes.

Cell culture

HEK293T cells were cultured in DMEM high glucose (HyClone, catalog no. SH30022.01) supplemented with 10% FBS (Thermo Fisher, catalog no. 10100147). The cells were maintained under optimal conditions in an incubator set at 37 °C, 5% CO2, and saturated humidity.

Library preparation

The preparation of RNA-seq libraries for spike-ins, HEK293T cells, and colorectal samples follows a standardized protocol, specifically the VAHTS Universal V8 RNA-seq Library Prep Kit. This protocol involves key steps, such as RNA fragmentation, cDNA synthesis via hexamer priming, end repair, and adding an adenine to the 3’ ends of the DNA fragments (dA-tailing), adaptor ligation, PCR amplification of the library, and subsequent sequencing. Notably, for spike-in samples, the protocol incorporates a modification: a tagmentation step is used as an alternative to the end repair, dA-tailing, and adaptor ligation steps, streamlining the preparation process.

RNA isolation

Total RNA was isolated from cells employing the RNAiso Plus kit (TaKaRa Biotechnology, catalog no. 9109), adhering strictly to the provided manufacturer’s protocol. Following extraction, the RNA was dissolved in RNase-free water, a practice maintained across all RNA-centric procedures. The quality of RNA was evaluated using the 2100 Bioanalyzer RNA picochip. Afterwards, the RNA was divided into aliquots of 5 μg each and stored at -80 °C for long-term use in subsequent experiments.

rRNA depletion

To precisely profile non-rRNA molecules in RNA samples, we utilized the Ribo-off rRNA Depletion Kit (Human/Mouse/Rat) (kit no. N406-01, Vazyme), a commercially available solution known for its efficacy in removing ribosomal RNA (rRNA) from the total RNA population. The procedure for rRNA reduction is meticulously designed, comprising several critical phases. Initially, the total RNA sample is combined with specially designed rRNA removal probes, aimed at selectively targeting and binding to rRNA molecules. This mixture is then incubated, a step that allows for the efficient hybridization of the rRNA removal probes to the rRNA entities. After incubation, a removal solution is added to the mix, initiating the breakdown of the rRNA-probe complexes. The RNA that remains, now enriched in non-rRNA molecules, undergoes a purification process to isolate the desired RNA pool suitable for further experiments. Employing the Ribo-off rRNA Depletion Kit is instrumental in diminishing the prevalence of rRNA, thereby facilitating a more exhaustive interrogation of the non-rRNA component. This enhancement in our method markedly increases the sensitivity of ensuing RNA-seq methodologies. The inclusion of an rRNA depletion stage is pivotal in achieving high-caliber data, markedly bolstering the precision and reliability of our molecular investigations.

Spike-in RNA and RNA circularization

To generate the circular RNA spike-ins, a synthesised single-stranded RNA oligonucleotide featuring a 5’ phosphate and a 3’ hydroxyl end was used. The oligo sequence for the spike-ins was as follows: Spike-in: 5’-phosphate-AA AAAAAAGGTAACTGCGNTTANCACNAGCNCCA

NGAGNAACNACANGAATTCTTTATAAAAAAA-OH-3'. For the preparation of the spike-in RNA, 5 μL of the oligonucleotide at a concentration of 10 μM was incorporated into a reaction mixture. This mixture consisted of 1 mM ATP, Rnase inhibitor at a concentration of 2 units/μL, 1 μL of T4 RNA ligase 1 (ssRNA Ligase) from New England Biolabs (Catalog No. M0204S), and 50% PEG8000 diluted to 15%. The ligase reaction was carried out by incubating the mixture at 25 °C for a duration of 1–2 h. The reaction was then halted by heating at 95 °C for 2 min.

Reverse transcription

RNA samples devoid of ribosomal RNA (rRNA) underwent reverse transcription using random hexamers and SuperScript™ IV reverse transcriptase (Invitrogen, no. 18090200), according to the instructions provided by the manufacturer. The reverse transcription was carried out in a 20 μL reaction mixture that included 100 ng of the RNA template, 2.5 μM of random hexamers, a dNTP Mix (at 10 mM of each dNTP), and 200 units (1 μL) of SuperScript™ IV reverse transcriptase. Initially, the mixture was incubated at 25 °C for 10 min, allowing primer annealing. This step was followed by reverse transcription at 42 °C for 50 min. To conclude the reaction, the mixture was heated to 70 °C for 15 min, which inactivated the reverse transcriptase enzyme.

Tagmentation

To perform tagmentation, a mixture was prepared by combining 50 ng of DNA with a 30 μL reaction mixture, which included 1× Insertion Buffer and 2 μL of Tn5-50 adaptor index (10 μM). This mixture was then incubated at 55 °C for 5 min. Following this initial incubation, an additional 30 μL of 2× Tn5 Digestion Mix (from the TransNGS® Tn5 DNA Library Prep Kit for Illumina®, catalog no. KP101) was added to the mixture. The combined mixture was further incubated at 55 °C for an additional 5 min. Through this process, a tagmented DNA library was successfully created. The resulting tagmented library structure is as follows:

5'-AATGATACGGCGACCACCGAGATCTACAC-i5-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-NNNNNN-CTGTCTCTTATACACATCTCCGAGCCCACGAGAC-i7-ATCTCGTATGCCGTCTTCTGCTTG-3'.

PCR amplication

The library underwent PCR amplification employing the 2× HIFI KAPA master mix according to the following formulation: 25 μL of 2× HIFI KAPA master mix was combined with 10 μL of cDNA, 13 μL of H2O, and 1 μL of a primer-either the universal forward or reverse primer, both at a 10 μM concentration. The sequences of the primers, which incorporate phosphorothioate bonds for enhanced stability, are as follows:

Universal forward primer: 5'-AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTT-3'.

Universal reverse primer: 5'-CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTT-3'.

The PCR amplification process was executed in a thermal cycler set to the following program: an initial denaturation step at 95 °C for 5 min; this was followed by 10–15 cycles of 95 °C for 15 s (denaturation) and 60 °C for 30 s (annealing/extension).

After PCR amplification, the library was purified using 1.8× concentration of Ampure XP DNA Beads and resuspended in 20 μL of H2O. The concentration and quality of the purified library were assessed by measuring one microliter of the sample with a Qubit 4 Fluorometer, utilizing the dsDNA HS (High Sensitivity) Assay Kit from Invitrogen.

Sequencing

The PCR libraries, once purified, underwent sequencing utilizing either the Illumina NovaSeq 6000 platform (PE150) or the MGISEQ-2000 platform (PE150), depending on the specific yield requirements set for the data output.

Statistics

To evaluate the effectiveness of the linear and Gaussian function fits, we employed Pearson correlation coefficients and adjusted R-squared values. These metrics are pivotal in discerning the strength and significance of the relationships depicted by the fits. For a comprehensive assessment, we further incorporated parametric statistical methods, particularly the T-test, to deepen our analysis of the data.

Theoretical derivation

A Gaussian-based model has been developed where GC content plays a key role. By quasi-randomly permuting G and C bases within k-mers, we generate distinct Gaussian distribution patterns for each transcript. Starting with a binomial distribution for GC content, we derive a Gaussian approximation, effective for larger n values. This method provides a more flexible analysis of GC patterns. Detailed derivations and methods are relocated to the Supporting Information.

Gaussian self-benchmarking (GSB) framework

The GSB framework processes 50-mers within transcript regions, compiling key metrics such as sequence, modeling count, sequencing counts, normalized counts (e.g., RPKM/TPM), and GC content. These 50-mers are categorized by GC content, and fitted to Gaussian distributions to model transcript-specific data. A fixed-parameter

Gaussian function is applied to sequencing data, revealing biases based on GC content. The GSB method corrects these biases without relying on prior data. Detailed methodology has been relocated to Supporting Information for further reference.

Results

Theoretical consideration and validation of GC content-based Gaussian distribution

To investigate the pattern of Gaussian distribution relative to GC content, we designed an experiment using synthesized, fixed-length spike-in RNA sequences as our analysis subject for RNA-seq. The basis of this experiment lies in the principle that each position within the RNA structure can harbor one of the four nucleotides, resulting in a plethora of potential RNA sequences-precisely 4^k distinct types for an RNA chain with k nucleotides (as demonstrated in Fig. 1a). These spike-ins were specifically varied by their GC contents, which differ based on the RNA fragment length, providing k+1 distinct categories of GC content for analysis. Central to our focus on elucidating the relationship between RNA GC content and its Gaussian distribution within RNA-seq was the use of circular RNA as our experimental model (as shown in supporting information Fig. S1). This multifaceted approach includes generating circular RNA templates, reverse transcribing using hexamer primers for continuous cDNA synthesis, incorporating adapter-flanked tags through tagmentation, proceeding with PCR amplification, and ultimately sequencing. The structural characteristics of circular RNA highlight the considerable influence of GC content on RNA-seq, offering a rich context for our examination. A key aspect of our template design was the inclusion of a 50-nucleotide RNA core flanked by pairs of 8-mer poly-A tails. This design choice aims at minimizing ligation bias, facilitating the equal formation of circular RNA complexes. Our analysis endeavors to accurately count reads for each RNA sequence, leveraging the identification of repetitive sequences to ensure precise quantification for each equimolar RNA sequence presented. Through this approach, our study not only delves into read count examination but also illuminates the complex biases inherent in RNA-seq data. This comprehensive analysis advances our understanding of how GC-based Gaussian distribution influences RNA-seq outcomes, thereby enriching our insights into the nuances of RNA sequencing data interpretation.

In data processing, entire spike-in sequences are utilized to determine the read count for each RNA sequence (as illustrated in Fig. 1b). Sequencing reads for each unique RNA sequence (represented by 8Ns in a 50-mer) are methodically profiled in the order of A, T, C, and G at N position. Notably, there's a serious variance in the sequencing count, especially when contrasted with the uniform distribution observed in model data. In our experimental setup, we utilize a pool of spike-in RNAs, each present in equimolar concentrations, as a sequencing substrate. This configuration aligns with the assumptions of our theoretical models, which predict a uniform distribution across all spike-in types. To validate the significance of these variations, we performed a Mann-Whitney U Test comparing the sequencing counts of each spike-in against the modeled counts. The results show that the p-value obtained is 0, which is well below the standard significance threshold of 0.05. This disparity in coverage among the spike-in RNAs suggests the presence of biases inherent to RNA-seq. To confirm the complex yet repeatable effects of bias in RNA-seq, two rounds of spike-in-based RNA sequencing were executed, demonstrating consistent replication with strong correlations between the replicates (as shown in Fig. 1c). This consistency lays a solid foundation for comparative analysis of crucial datasets later on. Moreover, by organizing the sequencing counts of spike-in RNAs according to their GC content, we observe a distinct stratification across nine GC-content categories (as illustrated in Fig. 1d). This stratification results in a bell-shaped distribution of sequencing counts when sorted by GC content. The aggregation of counts within each GC category transforms the initially uneven distribution of sequencing counts across spike-in RNAs into a more regular curve. This observation underscores the important influence of the GC permutation mechanism. Consequently, this approach demonstrates that complex, biased sequencing data can be effectively organized through GC content categorization, providing a clearer and more structured representation of the data. It's crucial to recognize that RNAs with extremely high or low GC content are underrepresented not due to random variation, but because of their intrinsic properties.

The joint mitigation of co-existing biases through GSB framework

The GSB framework introduces an innovative methodology for comprehensive bias mitigation. Central to the GSB framework is the implementation of pre-determined core parameters that are meticulously chosen independent of empirical sequencing data. This independence is crucial as it allows for an unbiased benchmarking process, free from the variability and potential biases inherent in empirical data. The procedure starts by developing a theoretical model for spike-in RNA data, based on the assumption that there is consistent coverage across all distinct spike-in RNA templates. Following categorization and aggregation, this idealized data undergoes a mathematical fitting process employing the Gaussian distribution function (as illustrated in Fig. 2a). The fitting

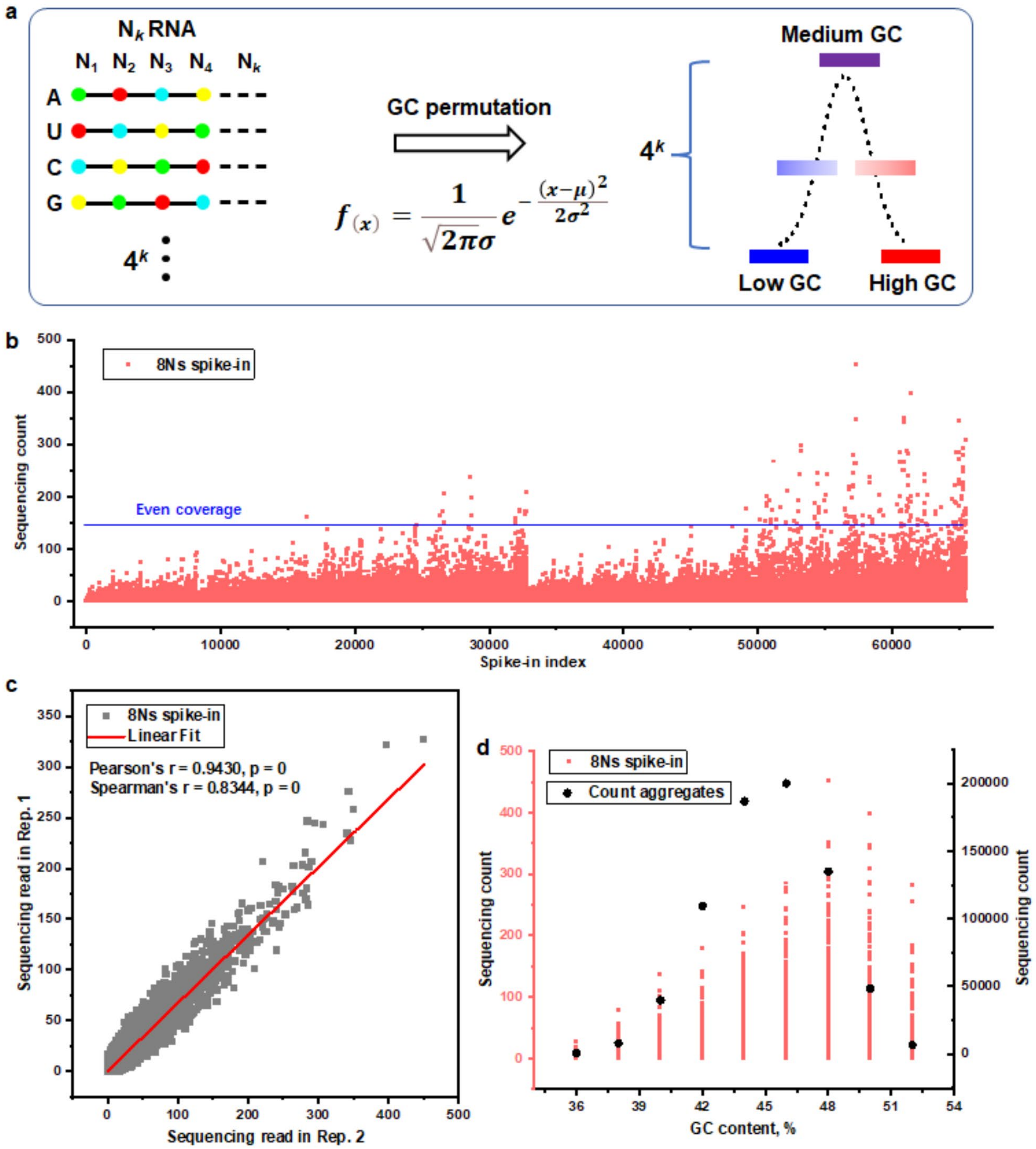

Fig. 1 Analysis of sequencing counts distribution based on GC content within random permutations. **a**) The GSB principle emerges from examining the patterns of guanine (G) and cytosine (C) bases that are shuffled randomly within a k-mer sequence. Utilizing the binomial distribution, this approach underscores how occurrences of k-mers with the same GC content can be systematically explored. It introduces a technique for sorting k-mers according to their GC content. By accumulating the counts of k-mers within each defined GC content category, a Gaussian distribution curve of these counts becomes apparent. **b**) The analysis highlights the variability in sequencing read distribution by examining 65,536 unique spike-in RNA sequences, each characterized by a unique nucleotide composition that follows a 4^8 variation pattern. This approach illustratively maps out the distribution differences across a broad spectrum of sequences. **c**) A linear regression analysis was conducted on biological replicate data. This analysis involved plotting the sequencing read counts for each of the 65,536 unique spike-in RNA templates, measured in duplicates. Both Pearson's correlation coefficient and Spearman's rank correlation were calculated, respectively, between the datasets. **d**) Compiles the sequencing counts for spike-in RNA sequences grouped by identical GC-content value, showcasing how distribution biases emerge across different GC content categories

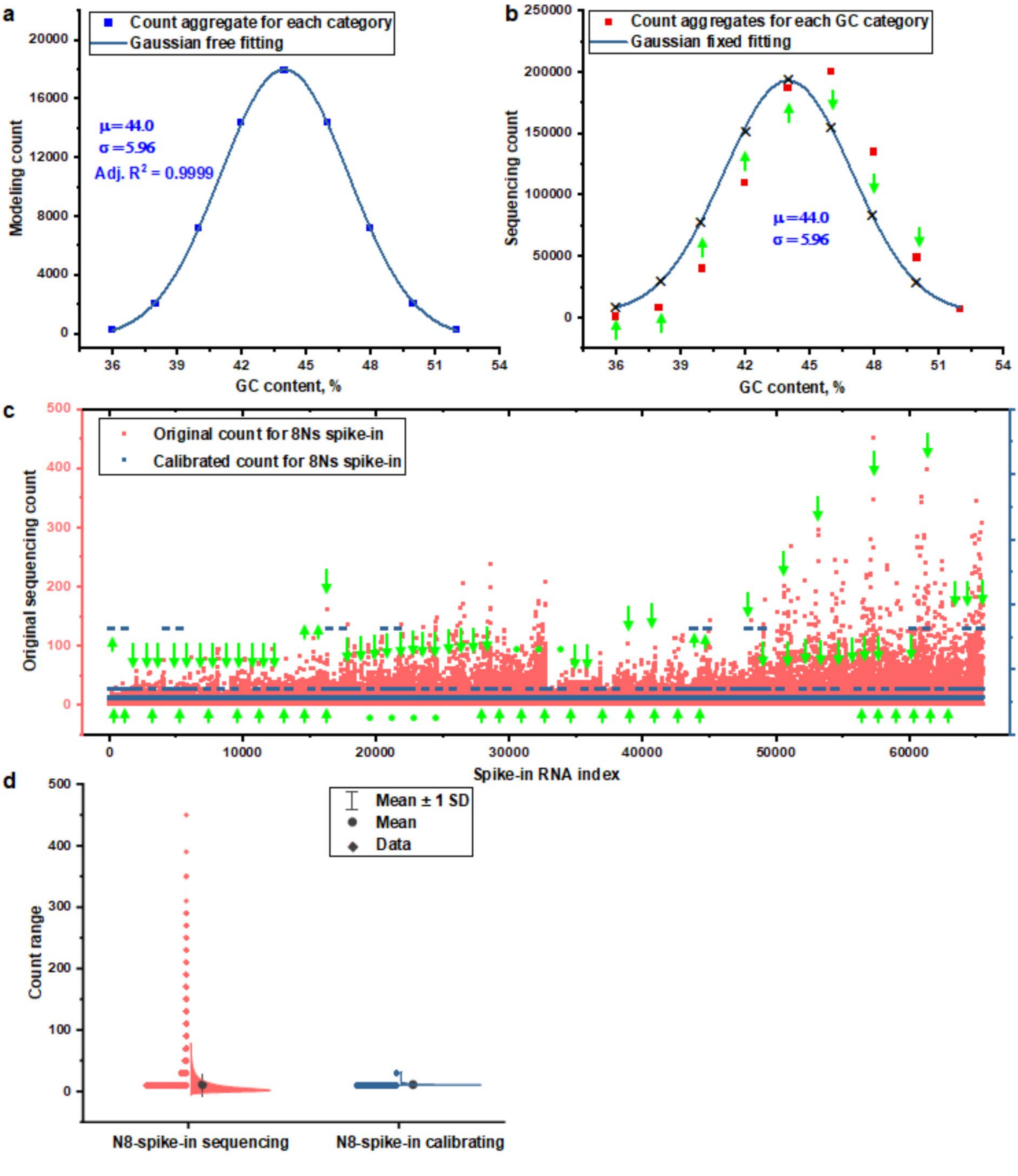

Fig. 2 The GC content-based Gaussian distribution method for self-benchmarking. **a**) Modeling counts categorized by their GC content are analyzed using a Gaussian distribution to derive essential parameters that effectively represent a desired uniform distribution for spike-ins, laying down an initial calibration benchmark. **b**) Sequencing data, sorted by GC content, is then adjusted to align with the Gaussian model guided by these parameters, with discrepancies between the predicting count and original counts at different GC content levels highlighted by arrows. **c**) A finer level of adjustment is carried out for each individual spike-in, where actual sequencing counts are compared against the calibrated benchmark. Necessary adjustments are marked by arrows, striving for higher precision in count metrics at this detailed analysis stage. **d**) The effectiveness of these calibrations is showcased through a box plot comparison of sequencing counts before and after calibration for all spike-ins

process is executed flawlessly, with the predefined mean and standard deviation parameters adeptly capturing the characteristics of the artificial spike-in RNA components. Following the establishment of these theoretical parameters, the predefined Gaussian distribution, characterized by the set mean and standard deviation, is employed to fit the sequencing data categorized and aggregated according to GC content (as illustrated in Fig. 2b). Any discrepancies between the predicted count aggregates from the sequencing data and the actual counts for each category are addressed through a comprehensive calibration process. This procedure systematically adjusts for potential biases, ensuring the accuracy and reliability of our sequencing results. The use of fixed parameters in this step ensures the creation of an unbiased representation of count aggregates for each GC category.

The technique described employs counts that follow a Gaussian distribution, enabling the generation of unbiased estimators for sequencing across various GC-content categories. This is accomplished by averaging the predicted counts for all related k-mers within a category, effectively reducing the complex biases usually associated with individual spike-ins (as illustrated in Fig. 2c). This approach highlights the robust ability of the technique to correct biases, ensuring more accurate and reliable sequencing results. The process is further elaborated on in the supporting information (as detailed in supporting information Fig. S2), ensuring a thorough understanding of the methodology. By accurately calibrating the counts for each RNA spike-in, this approach allows for a direct comparison between the observed sequencing counts and the theoretical expectations for each RNA spike-in template, ensuring precise analytical outcomes. The comparison of original versus calibrated sequencing counts across all spike-ins (as depicted in Fig. 2d) highlights the substantial biases introduced during sequencing and demonstrates the effective mitigation of these complex biases. Importantly, within the GSB framework, GC content is integral to the generation of a robust, theoretically-derived count distribution, rather than just serving as a smoothing parameter in bias-specific modeling. This novel approach, which employs predetermined parameters to systematically address bias, sets a new benchmark in RNA sequencing analysis. It represents a great leap forward in achieving truly unbiased quantification and enhances the accuracy of RNA sequencing data interpretation.

Validation of the GSB framework for bias mitigation in natural transcript sequencing data

To bolster the credibility of the GSB framework, it is imperative to rigorously assess its precision and reliability. This involves utilizing real sequencing data sourced from an authentic transcriptome. Our approach entails segmenting a naturally occurring human transcriptome sequence from its 5' end into subsets of defined lengths, known as k-mers, and subsequently evaluating the GC content for each k-mer (as depicted in Fig. 3a). In this context, each k-mer derived from the natural transcript sequence serves as an individual spike-in RNA template unit in the GSB validation process. The rich and varied GC content found in these natural transcript sequences should significantly aid in the development of the GSB framework, offering a striking contrast to the relatively homogeneous GC content seen in sequences fashioned through synthetic N8-based methods. This natural variation in GC content is crucial for refining the GSB framework's efficacy and precision, providing a deeper insight into the GC-based GSB model for bias mitigation. As a case in point, we delve into a detailed study focusing on the gene USF2, which is characterized by a moderate transcript length and relative abundance. We construct a stack plot that concurrently maps the profile of 50-mer sequencing count against its GC content, leveraging sequencing data derived from HEK293T cells (as illustrated in Fig. 3b). This analysis highlights the substantial variation in GC content and sequencing counts across 50-mers distributed along natural transcripts.

Efforts were made to classify each 50-mer, together with its associated sequencing count and a modeled count, based on its specific GC content along the USF2 transcript. The process of modeling the count aggregates for each GC-content category employs a Gaussian distribution (as highlighted in Fig. 3c), drawing upon data that depict even coverage. Utilizing 50-mer segments, thirty-nine distinct GC content types were extracted from the USF2 transcript. To accurately describe the distribution of counts across these GC-content categories in a transcript-specific manner, the key parameters of the Gaussian distribution-mean and standard deviation-were ascertained. Consequently, the Gaussian function, defined by these parameters, was used to fit the aggregate sequencing counts for each GC-content category (as demonstrated in Fig. 3d). A notable deviation between the empirical sequencing counts and those predicted by the Gaussian fit was observed across each GC-content category. This discrepancy underscores the intricate biases influencing the distribution of sequencing counts, prompting the initiation of a bias mitigation process. To enhance clarity, a comprehensive, step-by-step explanation of the computational procedure is provided (see Supporting Information, Fig. S3). The modeling process begins by preparing the initial data for each 50-mer along the transcript. This involves constructing a table that includes the 50-mer sequence, an initial modeling count set to 1, sequencing counts from cleaned data, normalized counts using RPKM or TPM, and GC content. Using a simplified dataset of 10 example 50-mers, we group

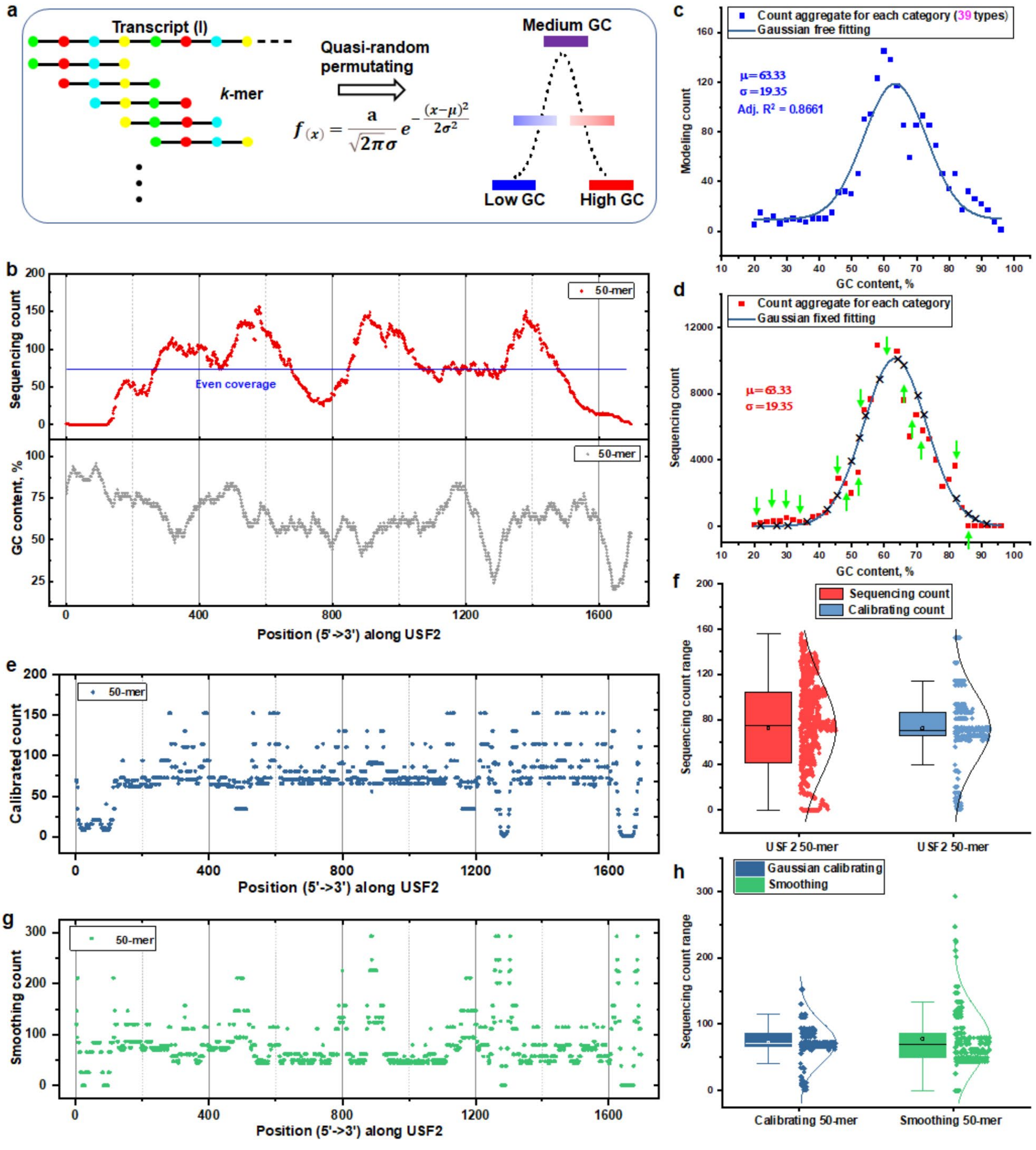

Fig. 3 (See legend on next page.)

these sequences by identical GC content and construct a Gaussian distribution for each group based on their GC content. Key parameters such as mean and standard deviation are derived from the GC content distribution within the modeling data, independent of empirical sequencing data. Sequencing counts are then grouped by GC content and plotted to visualize their relationship. A Gaussian distribution with fixed parameters is fitted to this plot to address complex biases, resulting in simulated counts that serve as unbiased abundance estimates. Finally, discrepancies in GC-assigned counts are averaged across all 50-mers to correct for sequencing biases, ensuring accurate contribution to overall abundance calculations. In an approach focusing on individual 50-mer segments, the process entails averaging the fitted aggregate counts for each GC-content category specifically

(See figure on previous page.)

Fig. 3 Overview of a GC content-based self-benchmarking approach for adjusting natural transcript bias in sequencing data analysis. **a**) Illustration of the development of a pseudo-Gaussian distribution specifically designed for the counts of 50-mer sequences categorized by GC content, incorporating the innate sequence complexity in natural transcript. **b**) Display of both sequencing count and GC-content metrics for 50-mers derived from the 5′-start position of the USF2 human transcript (ENST00000229239), demonstrating an even distribution of sequence coverage in accordance with the theoretical model. **c**) Compilation of the counts for GC-indexed 50-mers, aligned with the model predicting uniform 50-mer coverage, and subsequent application of a Gaussian distribution fit across 39 distinct GC-content levels. Parameters such as mean, standard deviation, and coefficient of determination (R square) are presented, reflecting the accuracy of the fitting process. **d**) Adaptation of the actual sequencing data's GC-content categorized 50-mer counts to a predefined Gaussian distribution function with established parameters from part (c), with necessary calibration adjustments per GC-content category marked by directional arrows. **e**) Visualization of sequencing counts for each 50-mer along the USF2 transcript following the GC content-based calibration adjustment highlighted in part (d). **f**) Comparative box-plot analysis demonstrating the original versus calibrated sequencing counts for individual 50-mers throughout the transcript, emphasizing the effectiveness of the GC content-based self-benchmarking calibration technique. **g**) Introduction to an alternative calibration method using a cubic polynomial smoothing function aimed at refining adjustments at the single 50-mer level. **h**) Comparative assessment through box-plot analysis of the calibrated sequencing counts for individual 50-mers across the entire transcript, juxtaposing the efficiency of the GC-based Gaussian self-benchmarking technique against the polynomial smoothing method, showcasing their respective impacts on data normalization and bias reduction

for each k-mer segment along the transcript (as depicted in Fig. 3e). This step has a centralizing effect on the sequencing counts, markedly reducing variability (as evidenced in Fig. 3f). Although the calibrated sequencing counts for individual segments don't match the uniform distribution predicted by theoretical models exactly, their correlation with a Gaussian distribution at the GC-content category level is markedly improved. Normal Q-Q plots (as depicted in supporting information Fig. S4) demonstrate that the original and calibrated sequencing count datasets possess identical means. Nonetheless, the notable reduction in standard deviation within the dataset processed through the GSB framework underscores its effectiveness in diminishing variability across sequencing counts. This enhancement points to the GSB framework's utility in achieving more consistent and reliable sequencing data.

In the enhancement of the GSB framework comparison, we've introduced a polynomial regression-based smoothing technique that directly correlates GC content with sequencing counts for USF2 transcript. The outcomes (as illustrated in supporting information Fig. S5) showcase predicted counts across various GC-content categories. Notably, these predictions diverge from counts calibrated using the traditional GSB method. Subsequently, we derived the average count for each 50-mer based on these predicted counts (as displayed in Fig. 3g). This analysis reveals numerous outliers with markedly altered counts for various 50-mers. By examining the distribution of these individual counts (refer to Fig. 3h), it is evident that while the smoothing technique tends to recenter the predicted individual counts to a certain degree, the presence of outliers substantially skews the overall count distribution, occasionally extending it beyond the scale of the initial sequencing counts (as depicted in supporting information Fig. S6). Such distortion may introduce new biases or exacerbate existing ones.

Refining bias mitigation with optimized k-mer length in the GSB framework

The GSB framework utilizes GC content-the proportion of guanine (G) and cytosine (C) bases within k-mers-as a critical component in its bias mitigation model. The efficacy of calibration within the GSB model is mainly influenced by the length of these k-mers, which subsequently dictates the GC content. To better understand this relationship, an in-depth analysis was conducted to examine how varying k-mer lengths, particularly the use of extended k-mers, impact the outcomes of GSB modeling. In the specific case of the EMP1-211 isoform, the analysis of increased k-mer lengths revealed substantial changes in the GC-categorized count distribution (as depicted in supporting information Fig. S7). Notably, an increase in k-mer length led to a contraction in the range of GC content, alongside a reduction in the number of read counts, ultimately resulting in greater diversity among GC content categories. The employment of longer k-mers appears to enhance the model's sensitivity in detecting subtle differences in GC content across different transcript segments. Initially, the model demonstrated strong performance, closely aligning with a GC-dependent Gaussian distribution. However, the incorporation of longer k-mers into the analysis introduced a marked shift in the distribution pattern. This shift underscores a critical trade-off: while longer k-mers enhance the resolution of GC content variation detection, they also considerably impact the accuracy of the model's Gaussian goodness-of-fit. This trade-off highlights the necessity of carefully evaluating the effects of k-mer length on both the enhancement of resolution and the overall fitting performance of the model.

The GSB framework was initially applied by selecting a 50-mer sequence length for modeling, using data derived from the EMP1-211 sequences to establish critical parameters (as illustrated in Fig. 4a). These parameters were then used to define a Gaussian distribution function, which was subsequently applied to the 50-mer sequencing data. This approach revealed significant

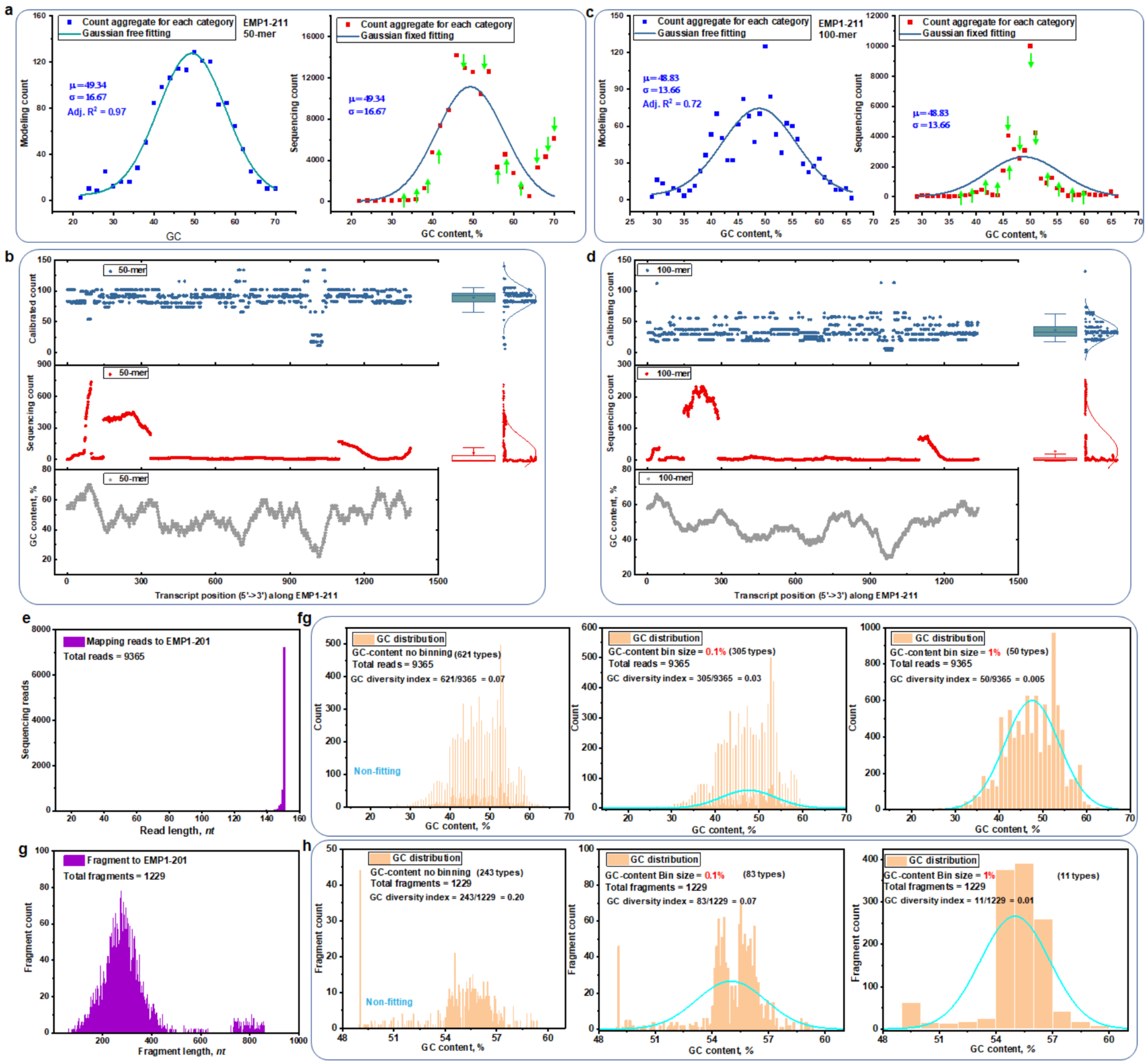

Fig. 4 The influence of k-mer length on GC-content-based sequencing data interpretation in GSB framework. **a**) For the 50-mer analysis, assuming even modeling coverage across the EMP1-211 transcript, data organized by GC content is modeled using a Gaussian distribution. This method allows for the accurate determination of key parameters such as mean and standard deviation. A Gaussian function with predefined parameters is utilized to fit the sequencing data categorized by GC content. **b**) A comprehensive stacked plot shows the relationship between GC content and sequencing counts (both original and calibrated) for 50-mers in the EMP1-211 isoform, providing a clear visualization of data adjustments. A box plot showcases the original versus calibrated sequencing counts of individual 50-mers across the entire transcript. **c**) The 100-mer dataset follows a similar pattern, with a Gaussian distribution used to the modeling counts across different GC content levels. The model's parameters outline its effectiveness, and adjustments based on 100-mer sequencing data are flagged for each GC category. **d**) Detailed plotting for each 100-mer of the EMP1-211 isoform shows GC content against sequencing counts before and after calibration. A box plot showcases the original versus calibrated sequencing counts of individual 100-mers across the entire transcript. **e**) A visualization presents the variability in the lengths of reads mapped to the EMP1-211 isoform. **f**) Sequencing reads are categorized by their GC content. Gaussian fits to 0.1% and 1% GC content bins are annotated. **g**) The analysis of fragment lengths from paired-end reads offers insight into the range of sizes observed. **h**) The cumulative sequencing reads across different GC content levels, with a focus on the proportions of G and C within the mapped paired-end fragment lengths, are summarized. A Gaussian curve fits the GC content distribution across 0.1% and 1% binning intervals

discrepancies between the observed sequencing counts and those expected across all GC-content categories, thereby highlighting biases inherent in the sequencing data. To address these biases for individual 50-mers along the EMP1-211 transcript, an averaging-driven calibration process was employed. This corrective measure successfully centralized the sequencing counts (as depicted in Fig. 4b), thereby enhancing the reliability and accuracy of the data. The study further extended this approach to the analysis of 100-mer sequencing data from the EMP1-211 transcript (as shown in Fig. 4c). After establishing key parameters and applying the Gaussian distribution function to the 100-mer sequencing data, the analysis revealed a substantial deviation from the expected Gaussian distribution when sequencing counts were categorized by GC content. This deviation confirmed the presence of complex biases within the sequencing data. However, the introduction of the GSB framework, with its systematic calibration function, effectively mitigated these biases. By calibrating the sequencing count aggregates for each GC content category and recalibrating the count averages for each 100-mer individually along the transcript (as illustrated in Fig. 4d), the procedure successfully centralized the counts, producing results consistent with those observed in the 50-mer data. Both the 50-mer and 100-mer analyses demonstrated the GSB framework's robust ability to effectively calibrate original sequencing counts. This calibration was crucial not only for correcting biases across GC content categories but also at the individual k-mer level. A notable aspect of this analysis was the variance observed in the determination of key parameters between the two datasets. Specifically, the analysis of the 100-mer data resulted in a higher variance ($R^2 = 0.72$) compared to the 50-mer data ($R^2 = 0.97$), underscoring differences in the precision and accuracy of bias correction between the two datasets.

Our analysis of GC-content distribution has progressed beyond the conventional use of fixed k-mer lengths, delving into the effects of variable k-mer lengths on the conformity of GC-content calculations to a Gaussian distribution within the GSB framework. The accuracy of transcript mapping analysis is inherently dependent on sequencing coverage profiles, which chart the depth of reads at specific positions on reference transcripts. These profiles result from aligning 'clean' reads-those that have undergone rigorous preprocessing to remove artifacts, adapters, and low-quality bases-with the reference transcripts. Consequently, the length of these processed reads can differ from their original 150nt size. For the purpose of GC-content analysis, we used the length of each processed read mapped to the reference transcript EMP1-211 as the k-mer unit for calculating GC content (as referenced in Fig. 4e). Although the majority of read lengths were standardized at 150 nucleotides (nt), a subset of reads were slightly shorter. A big challenge emerged when using these processed reads as the foundation for GC-content calculations. Specifically, the distribution of read counts across varying levels of GC content did not conform to a Gaussian distribution (as shown in Fig. 4f, left panel). This non-conformity arose from the extensive range of GC contents, with up to 621 distinct types spanning from 25 to 65%, leading to unpredictable variations in distribution. Due to these variations, there were no predetermined key parameters available for the GSB framework, necessitating the adoption of a more flexible and adaptable approach. To address this issue, we implemented a GC-content binning strategy, categorizing the data into discrete intervals on a 0.1% scale (illustrated in Fig. 4f, middle panel). This method effectively reduced the diversity of GC contents, resulting in a more structured distribution. Additionally, by adjusting the binning scale to 1%, we observed that the GC-content distribution aligned more closely with a Gaussian distribution model, specifically tailored to EMP1-211 (as depicted in Fig. 4f, right panel). In summary, by fine-tuning the approach to GC-content binning with an appropriate scale, the distribution of GC content-when influenced by variable k-mer lengths-can be realigned with a Gaussian model, thereby improving the robustness of the analysis within the GSB framework.

We further explored the impact of varying k-mer lengths on GC-content distribution, with a specific emphasis on fragments obtained through paired-end sequencing. Paired-end sequencing, which involves sequencing both ends of a DNA fragment, was applied to EMP1-211 transcript fragments. These fragments displayed a significant variation in length (as illustrated in Fig. 4g), which, in turn, led to notable differences in their GC-content (as depicted in Fig. 4h, left panel). Unlike the narrow length distribution observed in clean-read data, the paired-end fragments exhibited a considerably broader range of lengths, contributing to a higher level of GC-content variability. For this analysis, the length of each paired-end fragment was used as the unit for calculating GC-content. This approach revealed that the distribution of modeling counts across different GC-content values deviated markedly from the expected Gaussian distribution. This deviation was particularly pronounced in the raw, un-binned data, which exhibited a substantial increase in GC-content diversity. Specifically, the diversity index increased threefold, from 0.07 in clean-read data to 0.20 in paired-end reads, underscoring the enhanced diversity in GC-content types relative to the total fragment count (as shown in Fig. 4h, left panel). To address the challenges posed by this increased variability in GC-content distribution, we introduced a binning strategy. This method organized the data into more interpretable segments, using both 0.1% and 1% scales, with the objective of guiding the distribution closer to a Gaussian pattern. While this binning approach proved effective for sequences derived from shorter fragments (less than 150 nt), resulting in a Gaussian-like distribution, its effectiveness diminished for bins corresponding to longer fragments. Consequently, predetermining key parameters for this analysis became impractical. Our findings underscore the considerable influence of k-mer

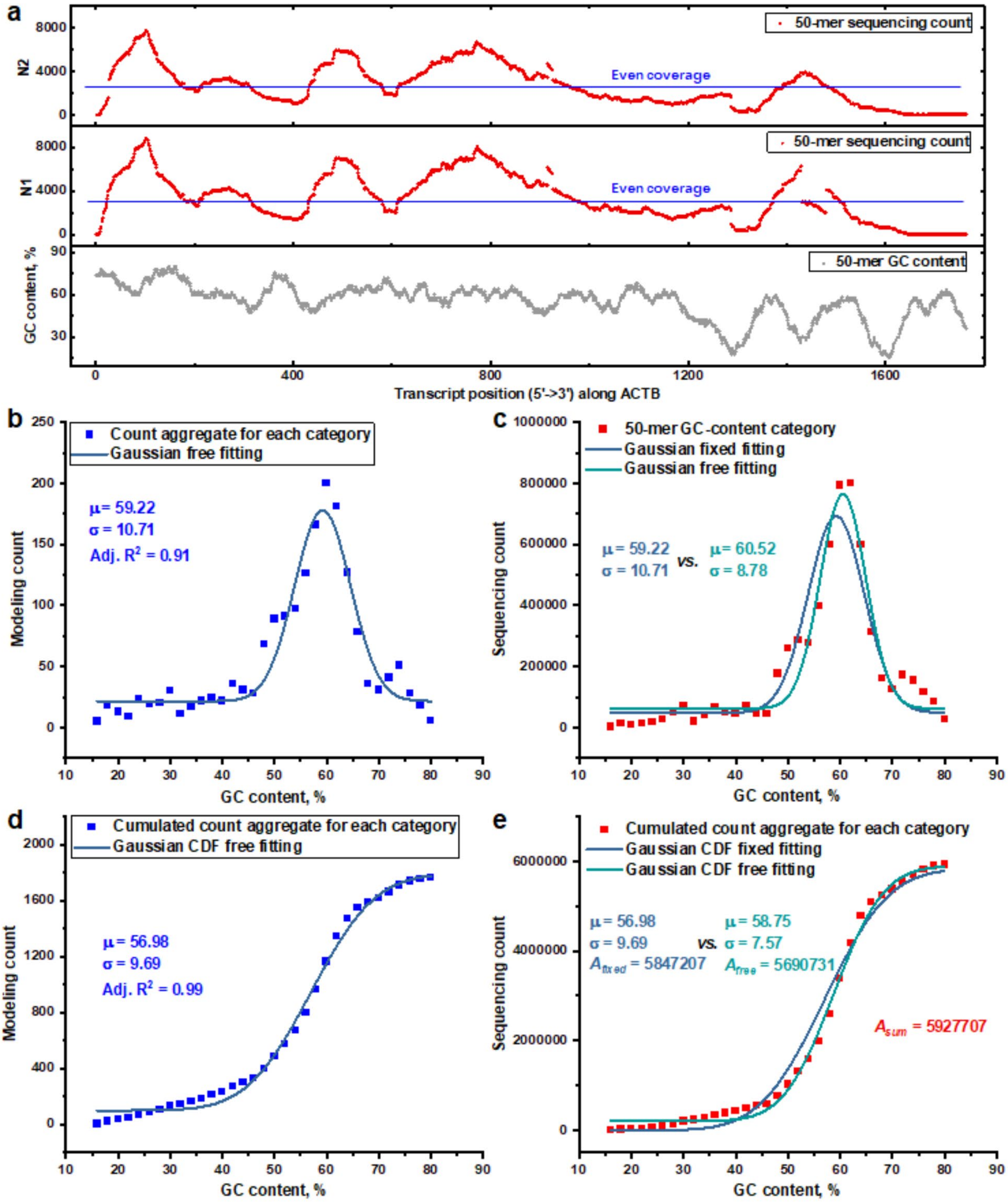

Fig. 5 (See legend on next page.)

length selection on the distribution of GC-content. This insight highlights the complexities involved in predicting GC-content distribution, particularly within the framework of GSB modeling. The results suggest that implementing a binning strategy for GC-content analysis can be beneficial. However, the success of this approach is inherently dependent on the careful selection of appropriate k-mer lengths.

(See figure on previous page.)

Fig. 5 Validation of a GC content based GSB framework for abundance calibration. **a**) Representation of sequencing count distribution against GC content for 50-mer sequences from the 5′ start of the ACTB human transcript (accession ENST00000646664), in two different human samples. The graph reflects a consistent sequence coverage across the transcript. **b**) The alignment of 50-mer sequences, categorized by their GC content, with a Gaussian distribution model. With the assumption of even coverage throughout the transcript, this modeling approach enables the precise assessment of key parameters, specifically the mean and standard deviation, tailored to the transcript. **c**) Empirical sequencing data, classified by GC content, is aligned and calibrated using a Gaussian distribution function with predefined, specific parameters. Additionally, this data is refined through a flexible fitting process employing the Gaussian distribution function, tailored with distinct key parameters. **d**) The Gaussian cumulative distribution function (CDF) is employed to align 50-mer modeling count categorized by GC content, to determine mean and standard deviation. **e**) The sequencing data, organized by its GC content, undergoes an alignment and calibration using the Gaussian Cumulative Distribution Function (CDF), executed with predefined parameters that yield a specific amplitude indicative of abundance levels. Furthermore, the data undergoes a flexible fitting process with the Gaussian CDF, outputting a free-fitting amplitude. Additionally, the figure illustrates the unadjusted aggregate of 50-mer counts, providing a base reference for the data prior to calibration and fitting processes

Improving the quantification of transcript with the GSB framework

In a detailed examination of the GSB framework, the research delved further into an analysis of transcript data, utilizing samples from diverse human sources. This stage encompassed a meticulous comparison of sequencing counts with GC content, specifically targeting 50-mer sequences derived from the 5’-start position of the human ACTB transcript (ENST00000646664). The study was conducted across two distinct human samples, aiming to unveil nuanced insights through this focused approach. Both samples exhibited a comparable coverage pattern in the distribution of 50-mer counts (as illustrated in Fig. 5a). Although the actual coverage across the transcript varies, the uniform distribution of each 50-mer count was established a priori to define critical parameters for the dual-model GSB analysis. To quantify these parameters, a Gaussian distribution function was employed, allowing for flexible fitting. The key parameters within this model were predetermined (as shown in Fig. 5b), setting the stage for the subsequent step of aligning the GC-based sequencing counts with the Gaussian distribution. At this juncture, a significant discrepancy emerged between the actual sequencing counts and those predicted by the model in two replicates (as shown in Fig. 5c and supporting information Fig. S8). Furthermore, the sequencing data were subjected to fitting with a Gaussian distribution function in a free manner, diverging from the fixed-parameter approach initially utilized. This free-fitting approach generated transcript-specific key parameters based directly on the sequencing data. A quantitative comparison between the fixed and flexible fitting approaches highlighted a notable variation in parameters, emphasizing the intricate biases embedded within sequencing data.

To further enhance the GSB framework, we have integrated a Gaussian Cumulative Distribution Function (Gaussian-CDF) that draws on Gaussian distribution principles. This integration is meticulously carried out within the GSB framework, where essential parameters are determined through an intensive convergent process. This involves the free-fitting of the Gaussian CDF, with a focused assessment on the cumulative k-mer modeling count aggregates, specifically targeting even 50-mer distributions in the ACTB transcript (as detailed in Fig. 5d). Leveraging this enriched foundation, the GSB framework is adept at employing these parameters to fit cumulative sequencing k-mer count aggregates across a range of GC content categories using the Gaussian CDF. This detailed fitting process generates an amplitude profile, marking a crucial indicator of bias-free abundance and providing a nuanced representation of the collective contributions from individual k-mers within the transcript (as illustrated in Fig. 5e and supporting information Fig. S9). A key feature of this approach is the derivation of critical parameters independently from any empirical sequencing data, enabling a strategic separation. This differentiation ensures the genuine abundance contributions are captured by the fixed-parameter Gaussian CDF, eliminating embedded data biases. In addition, this study contrasts a free-style fitting of cumulative k-mer sequencing count aggregates with a fixed fitting approach. This comparative analysis between the amplitude results from both fitting strategies, alongside the summation of individual k-mer counts, highlights meaningful differences. Such distinctions underscore the robustness of our self-benchmarking strategy, affirming its capability to deliver an unbiased quantification of k-mer abundance by systematically circumventing potential biases, thereby ensuring a more coherent and refined analysis.

Comparative analysis of GSB and other bias correction methods

To evaluate the effectiveness of the GSB model, it is important to compare it against other established bias correction methodologies employed by popular transcript quantification tools. Tools like Salmon, Cufflinks, and Kallisto each integrate specific strategies to address different types of biases-such as GC bias in Salmon [14], fragment bias in Cufflinks [15], and hexamer priming in Kallisto [19]. Our comparative analysis will focus on how the GSB framework performs in relation to these well-regarded methods, each recognized for their distinctive approaches to mitigating biases. To thoroughly examine how biases may disproportionately affect sequencing coverage across a reference genome, we look at the GAPDH-201 gene, which consists of nine exons. By plotting the sequencing depth during the alignment of paired-end fragments, we can observe noticeable variations in read coverage among different genomic regions (as shown in Fig. 6a). These differences underscore considerable inconsistencies in sequencing coverage throughout the gene. Using the GAPDH dataset, we will conduct an in-depth analysis, utilizing multiple bias mitigation models to achieve a thorough assessment.

Addressing GC content biases is key to correcting imbalances in read counts, which can vary due to differences in GC content across transcript fragments. LOESS

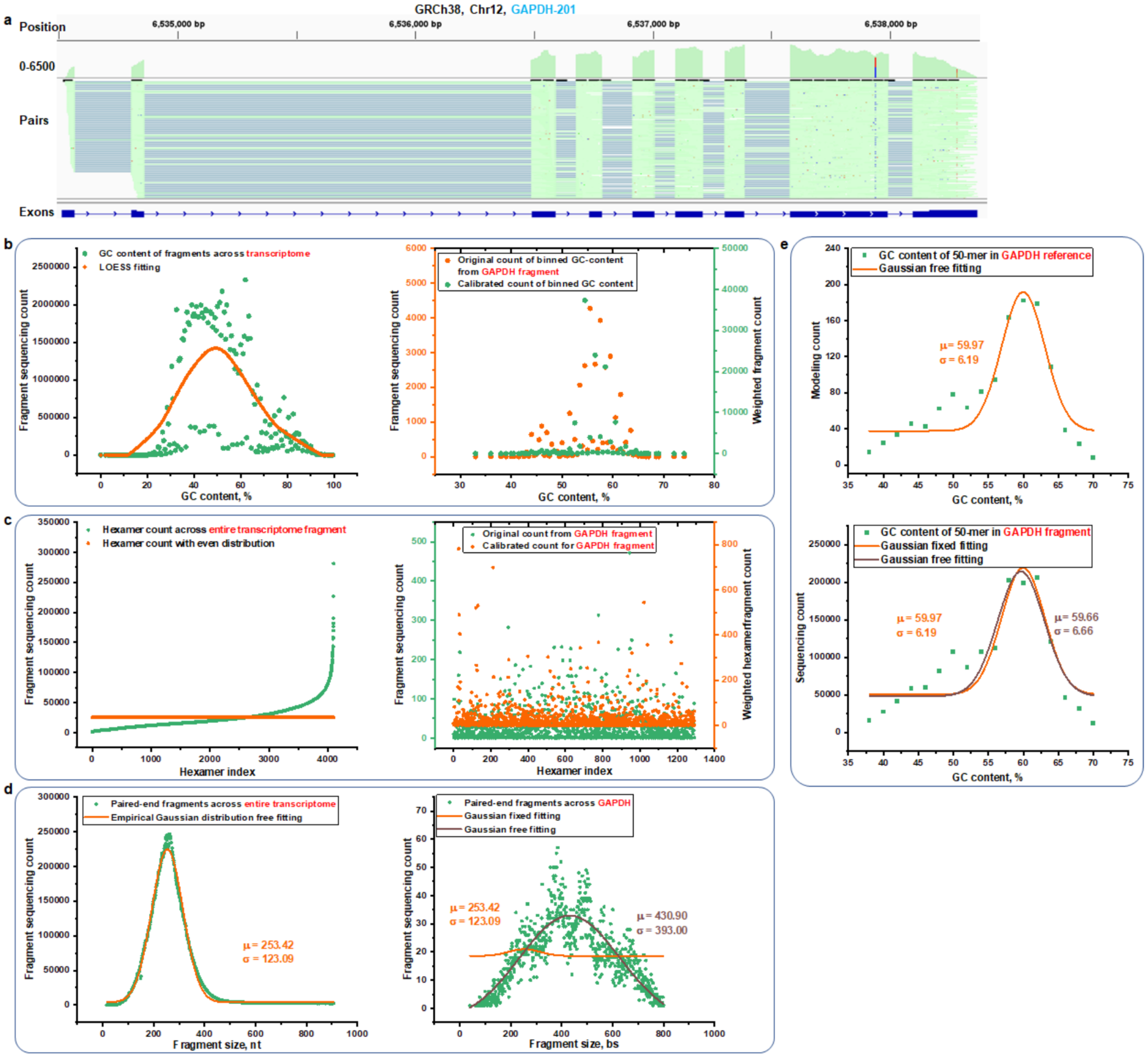

Fig. 6 Different bias correction models in RNA-seq for GAPDH-201 transcript analysis. **a**) Sequencing depth visualization demonstrates the variability in read coverage across the nine exons of GAPDH-201, showcasing the sequencing depth for paired-end fragments aligned with the genomic reference. **b**) It outlines a method to address GC bias by correlating the GC content, binned at 0.5% intervals, of paired-end fragments to their sequencing counts across entire transcriptome. Adjustments are made by applying GC-content-specific weighting factors to correct the counts of all GAPDH-201 mapped fragments. **c**) The correlation between sequencing counts and the presence of each possible hexamer at the start of the paired-end fragments across transcriptome is analyzed. Weighting factors derived from this empirical analysis adjust the sequencing counts for fragments starting with specific hexamers in GAPDH-201, addressing hexamer-related biases. **d**) Fragment size bias correction utilizes an empirically fitted Gaussian distribution to the size data of all paired-end fragments in the sequencing. The derived parameters from this distribution are then used to adjust the counts of GAPDH mapped fragments, thereby normalizing for the effect of fragment size variations. bs represent bases. **e**) GSB framework demonstrates a theoretical model-based approach using predetermined Gaussian distribution parameters to calibrate GC-organized sequencing counts of 50-mers

(Locally Estimated Scatterplot Smoothing), a non-parametric regression technique, effectively smoothens the relationship between GC content and fragment count throughout the transcriptome [23]. However, applying LOESS directly to unprocessed data can lead to suboptimal fits, especially when using original GC content data from various lengths of paired-end fragments (refer to supporting information Figure S10). To improve the fit quality, we organized the GC-indexed fragment count data into bins (as demonstrated in Fig. 6b). This binning strategy helps manage the variability in GC content seen across different fragment lengths, thereby enhancing the accuracy of the LOESS fit. It also establishes comprehensive and robust weighting factors for each GC content category (as depicted in supporting information Fig. S11). In specific cases, such as with the transcript for GAPDH, we adjust sequencing fragment counts by applying a weighting factor to counteract the GC bias. Although this adjustment aligns the counts more accurately, it also introduces additional variability into the calibrated data across various GC content categories. This observation suggests that further refinements are needed in the calibration process to achieve more consistent and reliable adjustments.

Hexamer priming bias also impacts RNA-seq, in which random hexamer primers show a preference for certain RNA sequences during library preparation. This preference can distort the representation of different RNA types in the sequencing data. To address this, researchers analyze the frequency of each hexamer at the beginning of reads or fragments across the entire transcriptome [16]. In mitigation efforts, a theoretical uniform distribution is used as a baseline for all hexamers (as shown in Fig. 6c). The observed frequency of each hexamer is then normalized against this average to calculate a weighting factor (as detailed in the supporting information Fig. S12). These correction factors are applied to adjust the fragment counts that begin with each hexamer, thus rectifying the hexamer-associated biases in specific transcripts, including those like GAPDH. This approach ensures that bias corrections are applied accurately based on the observed discrepancies. However, it's important to acknowledge that these weighting factors are specific to each sample and can be affected by additional variables.

Fragment size bias presents another challenge in RNA-seq, as certain fragment lengths might be preferentially amplified, sequenced, or aligned due to methodological biases and enzyme preferences. To address this issue, the distribution of all mapped fragments within the transcriptome often be analyzed to detect empirical patterns [15, 24]. In our study, we identified a Gaussian distribution of fragment sizes (as illustrated in Fig. 6d). This global distribution helps define the typical behavior of fragment sizes across the transcriptome. To compensate for the disparities in fragment size representation, we adjust the fragment counts for specific transcripts, such as GAPDH, based on this Gaussian model. By aligning the fragment sizes with the parameters derived from global Gaussian distribution, we mitigate the discrepancies caused by unequal fragment sizes. Additionally, we employ a cumulative distribution function (CDF) established from this model to determine essential calibration parameters that represent the spectrum of fragment sizes observed throughout the transcriptome. Using this approach, the fragment counts for GAPDH are calibrated, aligning with the parameters established from the CDF (as shown in the supplementary Fig. S13). However, despite these adjustments, there are still notable deviations between the fragment size distribution of GAPDH and the global transcriptome distribution. This discrepancy indicates that while a transcriptome-wide reference can alleviate some of the biases associated with fragment sizes, further targeted analyses or interventions might be necessary to fully address these issues in RNA-seq data.

To mitigate biases associated with GC content, fragment size, hexamer binding, and other factors in sequencing data, the primary strategies rely on the empirical sequencing data to construct their initial models [16, 20–23]. However, it is challenging to find a dataset that is influenced by only one type of bias, as real-world datasets usually exhibit multiple overlapping biases. This creates a complex scenario where attempting to correct one bias could inadvertently affect other biases present in the data. This intricacy underscores the difficulty in achieving completely unbiased RNA-seq data. In contrast, the GSB framework diverges from reliance on empirical data and instead utilizes theoretical modeling (as depicted in Fig. 6e). This approach involves setting critical parameters based on a GC-content oriented Gaussian distribution and its cumulative distribution function. By fixing these parameters, the framework can systematically apply the cumulative distribution function to sort and analyze GC-influenced sequencing data, allowing for precise quantification of transcript levels, such as those of GAPDH (referenced in supporting information Fig. S14). By employing theoretical, data-driven parameters, this strategy ensures that the parameters accurately mirror the true characteristics of transcripts, reducing external biases and potentially leading to more unbiased data in RNA-seq analysis.

Discussion

To mitigate RNA-seq biases, targeted bias mitigation methods can be employed, which are designed to neutralize the effects of specific types of distortions within the datasets. The integration of multiple bias mitigation strategies is essential for accurate transcript quantification; however, it presents the challenge of overfitting due to the complex interactions of various sources of bias, which may sometimes be interrelated [13, 25]. While the conventional methods based on expection-maxmization flow algorithm have advanced in the modeling and mitigation of specific biases, they are not comprehensive in addressing all potential biases, particularly for those that are more complex and not easily characterized. It is generally accepted that the complete elimination of all biases in transcript quantification is an elusive goal [26, 27]. In the field of bias modeling, distinct methodologies differentiate the GSB framework from other common approaches like Alpine, Salmon, or XAEM, especially in terms of identifying, correcting, and understanding biases and their foundational distributions [13, 14, 16, 28, 29]. A universally acknowledged technique in this area involves comparing observed k-mers or sequence alignments against an unbiased reference to detect over- or under-sampling biases. This method stands out for its effectiveness and precision and is a core strategy across different analysis techniques. However, the practical implementation of this strategy shows considerable variation between the GSB framework and other methods. Generally, alternative approaches concentrate solely on rectifying a particular bias and create an unbiased baseline directly derived from the empirical sequencing data. This presupposes that once a particular bias is corrected, the adjusted empirical data can potentially serve as an ideal, neutral benchmark for comparison. This is not achievable with actual sequencing data. To illustrate this, we processed the same raw data through different bias correction models-no bias correction, GC bias correction, and sequence bias correction-and compared the resulting transcript abundance profiles (see Supporting Information Fig. S15a). Our findings reveal meaningful deviations between the GC-bias corrected data and the sequence-bias corrected data, even after applying individual bias corrections. This demonstrates that individual bias correction models are insufficient for addressing the complex, intertwined biases present in sequencing data. Therefore, a more holistic approach is needed to effectively mitigate these biases and achieve accurate quantification. In contrast, the GSB framework introduces a novel approach by ensuring an unbiased foundation through the strategic management of GC content distribution (as shown Supporting Information Fig. S15b). This method is grounded in theoretical principles rather than relying on empirical data, which sets it apart. By prioritizing the control of GC content, the GSB framework offers an innovative solution for obtaining a more objective reference point for analysis. This distinctive contribution is particularly effective in correcting most forms of potential bias, enhancing the reliability and coherence of its analyses. To validate the GSB's robustness, we applied it across different scenarios, ranging from simple artificial spike-in RNA to RNA extracted from cell lines and human samples. This comprehensive testing across varied sample types demonstrates the framework's versatility and consistency, reinforcing its overall effectiveness.

The GSB framework directly constructs benchmarks using only the sequences of transcripts, providing a distinct advantage. It enables the targeted selection of any transcript region for model development, focusing solely on acquiring sequencing data for that specific area. This is particularly advantageous in the context of isoforms, which often share sequence segments. The overlap of these segments, combined with sequencing biases and the complex nature of isoform diversity [30–32], can lead to disparate coverage across the entirety of an isoform's sequence. The programmability feature of the GSB framework allows for precise mitigation of biases for any chosen regions, whether it be the whole transcript region or particular areas of isoform overlap. This strategic targeting and mitigation of biases enhance data analyzing accuracy and efficiency.

Following the GSB framework, sequencing data is organized according to GC content and subjected to a Gaussian fitting process. This process utilizes predetermined parameters from modeling data, with the resulting Gaussian-distributed counts serving as unbiased indicators for sequencing counts across each GC content category. To further refine this method, the unbiased sequencing counts obtained for each GC category are averaged across all corresponding k-mers within that category. This step systematically reduces bias at the single k-mer level. Nevertheless, the resulting calibrated k-mer count distribution along the transcript remains irregular, deviating from the even coverage observed in modeling data. This discrepancy can be attributed to the relatively low resolution of counts categorized by GC content. Despite the accurate calibration of overall counts for each GC category, these categories comprise numerous individual k-mers with identical GC content but varied high-dimensional structures. These structures likely influence sequencing efficiency, contributing to the non-uniform k-mer count distribution. However, it is important to note that the calibrated sequencing counts for each GC category aligns with the Gaussian distribution pattern, indicating the absence of bias effects. In summary, while the GSB framework largely reduces bias by calibrating k-mer count aggregates across GC content categories, variations in k-mer structure within these categories can still affect the uniformity of the k-mer count distribution. Future improvements may require addressing these high-dimensional structural variations to achieve a more even coverage across the transcript.

When normalizing the GSB framework, it is essential to fit the model counts, categorized according to GC-content, with a Gaussian distribution function.

This fitting is used to determine the crucial parameters for empirically modeling GC-categorized data when the parameters are held constant. However, challenges arise when the available regions are too short, resulting in an insufficient number of k-mers for effective categorization. Under these circumstances, an alternative approach is adopted where the sum of counts from individual k-mers is used to estimate k-mer abundance. To address variability in data coverage and complexity, a specific threshold for the number of k-mers is established and applied during the operation of the GSB script. This threshold helps in managing scenarios with sparse data by providing a consistent measure for comparison. Additionally, the unique structural characteristics of k-mers within these regions can sometimes make it impractical to fit these counts using a typical GC-based distribution plot. To overcome this, counts associated with individual k-mers are aggregated, providing a composite measure of abundance. This aggregation aids in smoothing out the variability and allows for more stable statistical analysis. By aggregating k-mer counts and setting predetermined k-mer thresholds, the revised approach not only compensates for inadequate data but also strengthens the analytical framework. This meticulously adjusted methodology ensures a deeper and more accurate understanding of the statistical processes inherent in the GSB framework, enhancing the reliability and robustness of the results.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10814-0.

Supplementary Material 1

Acknowledgements

We thank Prof Junmin Quan's support.

Author contributions

QS and YL designed the experiments, QS and YL contributed to data analysis and writing the manuscript; QS, JMQ and QZL contributed to writing the manuscript. All authors reviewed the manuscript. DMG provided some of the raw data.

Funding

We acknowledge financial support from the Shenzhen-Hong Kong Collaborative Innovation Research (SGDX20201103095404019), National Natural Science Foundation of China (82170070, 82241022), Shenzhen Municipal Research Grant (JCYJ20210324120206017), Yunnan Provincial People's Hospital Cooperation Project (202201AY070001-224), Guangdong Provincial Key Laboratory of Regional Immunity and Diseases (2019B030301009). We acknowledge financial support from the Long Yi Postdoctoral Special Fund (S249201002). This project is in part supported by the National Key Research and Development Program of China (grant no :2023YFA0914904; 2022YFA1105601).

Data availability

The raw and processed sequencing data have been submitted to the NCBI Sequence Read Archive (SRA) and are publicly available via the link: https://www.ncbi.nlm.nih.gov/sra. These datasets can be accessed using the project accession number PRJNA999048. The RNA sequencing (RNA-seq) data for samples N1 and N2 can be found under accession numbers SRR25438891 and SRR25438890, respectively. For the N8 samples with 50-mer spike-ins, in both the first and second replications, the relevant SRA accession number is SRR25438885. Additionally, the SRA accession numbers for the HEK293T total RNA-based RNA-seq for the first and second replications are SRR25438884 and SRR25438883, respectively. Should further information or data be required, we invite requests directed to the corresponding authors.

Code availability

The code used for data analysis is available at https://github.com/QiangSu/N-sequence. The code is also available in methods.

Declarations

Ethics approval and consent to participate

This project was approved by the institutional review board at Shenzhen University General Hospital. Human colorectal samples were ethically acquired from the Shenzhen University General Hospital. Before the collection of samples, each donor provided informed consent, authorizing the retrieval of biopsies and enabling the conduct of extensive molecular profiling of their transcriptomes.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Faculty of Synthetic Biology, Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Shenzhen University of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]Institute of Chemical Biology, Shenzhen Bay Laboratory, Shenzhen, China
[3]Shenzhen Key Laboratory of Microbial Genetic Engineering, Vascular Disease Research Center, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China
[4]State Key Laboratory of Chemical Oncogenomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen, China
[5]Cord Blood Bank, Guangzhou Institute of Eugenics and Perinatology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[6]State Key Laboratory of Pharmaceutical Biotechnology, Department of Medicine, The University of Hong Kong, Hong Kong SAR, China

Received: 8 July 2024 / Accepted: 19 September 2024
Published online: 30 September 2024

References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56.
2. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature. 2022;608:353–9.
3. Kovaka S, Ou S, Jenike KM, Schatz MC. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. Nat Methods. 2023;20:12–6.
4. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31:1009–14.
5. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39:1348–65.
6. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
7. Zhenqiang S, Paweł PŁj, Sheng L, Jean TM, Danielle TM, Wei S, Charles W. and so on.: A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 2014, 32:903–914.

8. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
9. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9:171–81.
10. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
11. t Hoen PA, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, Laros JF, Buermans HP, Karlberg O, Brännvall M, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31:1015–22.
12. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
13. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12:R22.
14. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
15. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
16. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38:e131.
17. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902.
18. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8:469–77.
19. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
20. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci U S A. 2011;108:19867–72.
21. Love MI, Hogenesch JB, Irizarry RA. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nat Biotechnol. 2016;34:1287–91.
22. Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol. 2010;11:R50.
23. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.
24. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7:1009–15.
25. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014;15:709–21.
26. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11:220.
27. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013;10:71–3.
28. Jones DC, Ruzzo WL, Peng X, Katze MG. A new approach to bias correction in RNA-Seq. Bioinformatics. 2012;28:921–8.
29. Deng W, Mou T, Kalari KR, Niu N, Wang L, Pawitan Y, Vu TN. Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. Bioinformatics. 2020;36:805–12.
30. Turro E, Su SY, Gonçalves Â, Coin LJ, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 2011;12:R13.
31. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25:1026–32.
32. Gunady MK, Mount SM, Corrada Bravo H. Yanagi: fast and interpretable segment-based alternative splicing and gene expression analysis. BMC Bioinformatics. 2019;20:421.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.