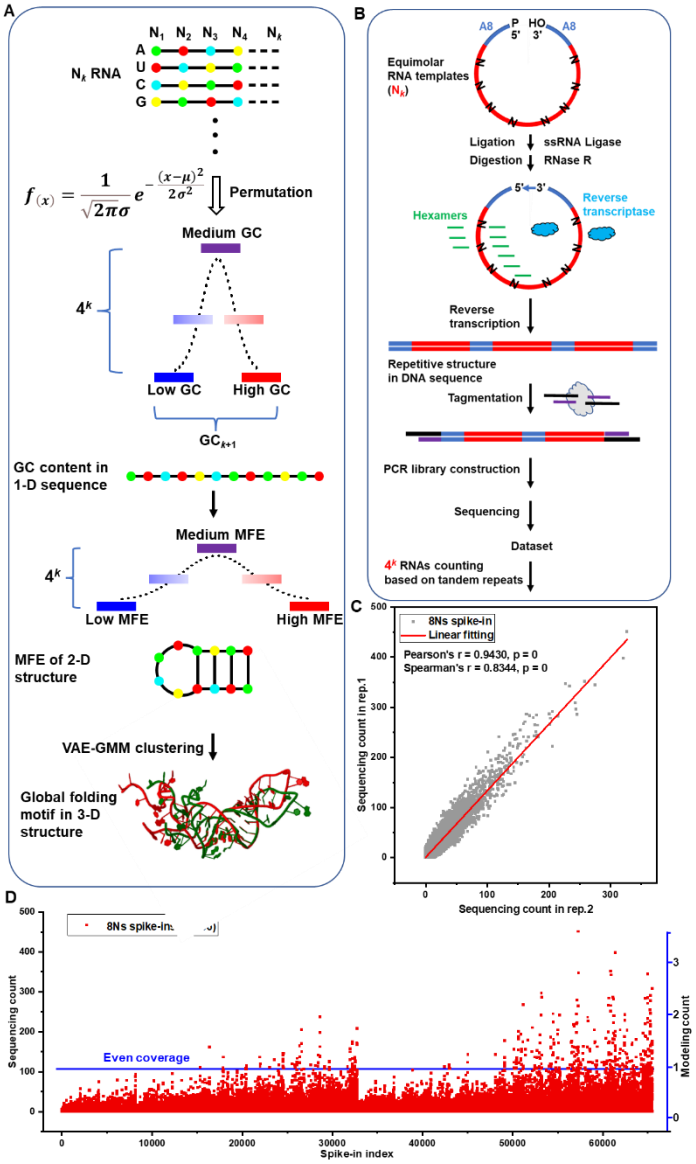
Supplemental material for
Deep structural clustering reveals hidden systematic biases in RNA sequencing data

Qiang Su[1,4,‡,*], Yi Long[2,‡], Deming Gou[3], Junmin Quan[4], Xiaoming Zhou[5,*] and Qizhou Lian[1,6,7,*]

Qiang Su, Xiaoming Zhou, and Qizhou Lian.
Email: su@chemie.uni-siegen.de, zhouxm@scnu.edu.cn and qz.lian@siat.ac.cn

This PDF file includes:
Figures S1 to S28

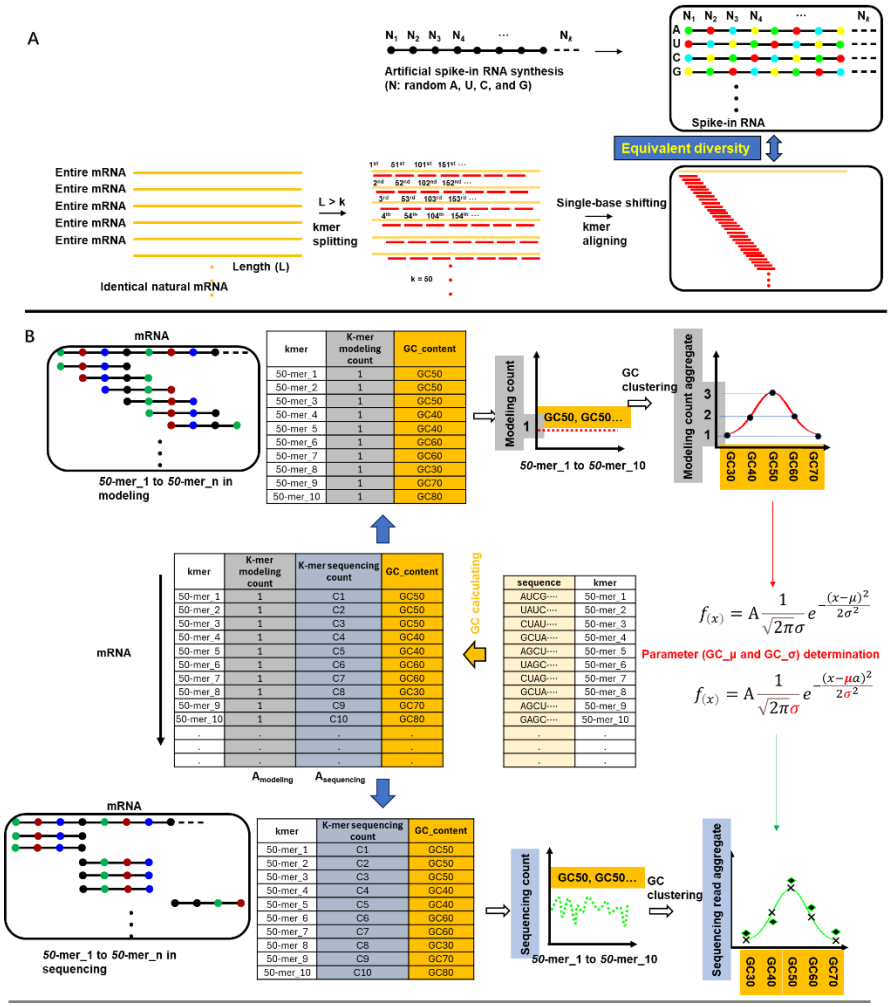
Fig. S1.



Supplemental Fig. S1. Multi-dimensional RNA structural features from primary to tertiary levels associated with sequencing biases. **a**) A multi-dimensional RNA structural model analyzing random base arrangements in spike-in RNA sequences. Spike-ins are categorized based on their GC-content (a primary, one-dimensional structural feature) and their MFE (a secondary, two-dimensional structural feature). The count distributions across GC-content and MFE categories follow Gaussian distribution functions that underpin a stepwise theoretical model. Additionally, spike-ins are clustered based on direct nucleotide sequences using an unsupervised VAE-GMM algorithm. **b**) Schematic representation of the sequential workflow for validating multi-dimensional RNA structure analysis. An RNA construct containing N=8 random nucleotides (4^8 = 65,536 unique sequences) is flanked by two identical poly(A) arms (see "Spike-in circularization" section). The construct is ligated using single-stranded RNA ligase, and unligated RNA is removed by RNase R digestion. The resulting circular RNA undergoes

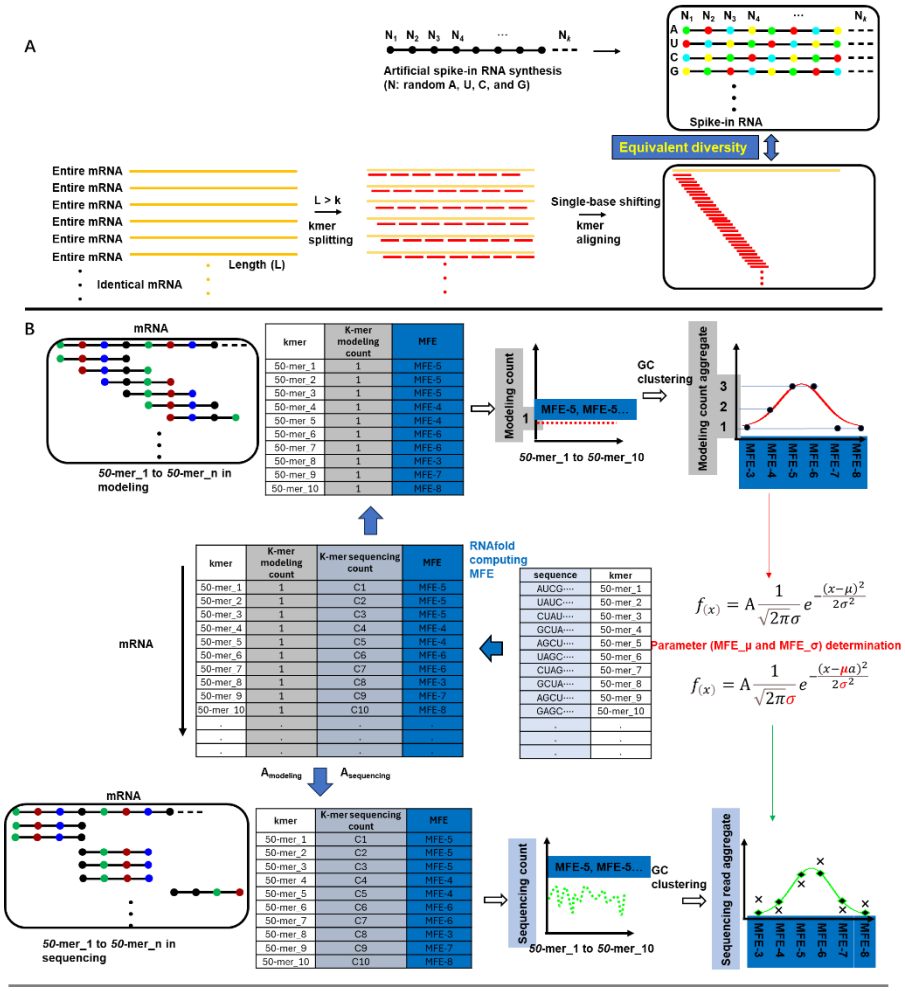
reverse transcription, primed with random hexamers. Sequencing adaptors are then added by tagmentation, followed by PCR amplification. Post-sequencing data are used to determine the frequency of each spike-in sequence. **c**) Linear regression analysis of sequencing read counts for the 65,536 unique spike-in RNA templates, based on biological replicates. **d**) Sequencing count profile illustrating distribution variability among the 65,536 unique spike-in sequences, encompassing all possible eight-base combinations. The observed distribution is compared with a theoretical count of 1, representing idealized conditions based on the model's assumptions.

Fig. S2.



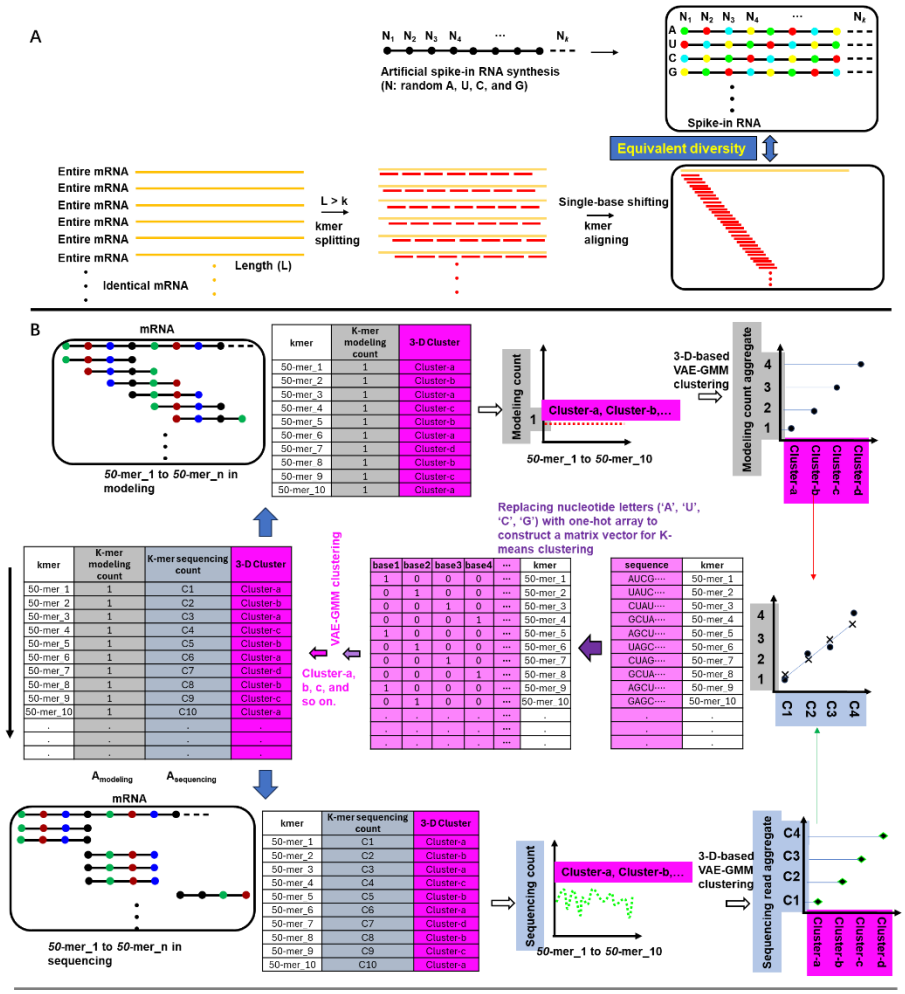
Supplemental Fig. S2. GC-based bias identification model. A) The spike-in RNAs are synthesized by randomly incorporating A, T, C, and G bases. All transcript sequences (Homo_sapiens.GRCh38.cdna.all.fa) were downloaded from www.ensembl.org. Each mRNA was split into kmers starting from the 5' end. These transcript-specific kmer sets were then aligned across the entire transcript. B) Transcript-specific benchmarking parameters are ascertained by categorizing evenly distributed 50-mers or global k-mer overlapping frequency throughout the transcript based on their GC content, and integrating individual occurrence counts within each GC category. The table shows the data acquisition from the sequence data and transcript sequence. The columns K-mer modeling count and GC content are used to determine the key parameters. The resultant GC content-enumerated counts are then fit to a Gaussian distribution function. This procures the two fundamental parameters: mean (μ) and standard deviation (σ). Uneven k-mer coverage exhibiting position-specific occurrence counting facets of the sequencing data form the columns of kmer-sequencing count and associated GC content. The k-mers, categorized by GC content, are summed up within each category, and plotted. Once the parameters derived from the modeling data are set in place, the Gaussian distribution function aptly acknowledges the intrinsic biases in the sequencing data.

Fig. S3.



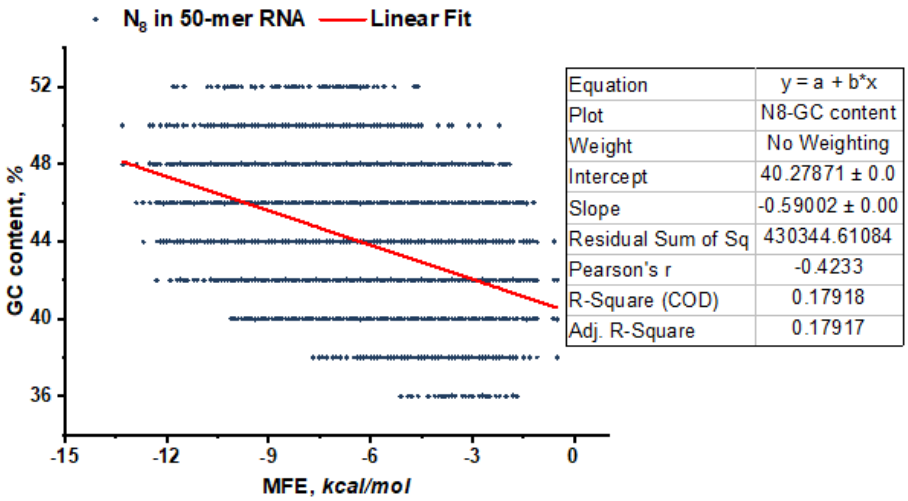
Supplemental Fig. S3. MFE-based bias identification model. A) The spike-in RNAs are synthesized by randomly incorporating A, T, C, and G bases. All transcript sequences (Homo_sapiens.GRCh38.cdna.all.fa) were downloaded from www.ensembl.org. Each mRNA was split into kmers starting from the 5' end. These transcript-specific kmer sets were then aligned across the entire transcript. B) Transcript-specific benchmarking parameters are ascertained by categorizing evenly-distributed 50-mers or global k-mer overlapping frequency throughout the transcript based on their MFE, and integrating individual occurrence counts within each MFE. The table show the data acquisition from the sequence data and transcript sequence. The columns K-mer modeling count and MFE are used to determine the key parameters. The resultant MFE-enumerated counts are then fit to a Gaussian distribution function. This procures the two fundamental parameters: mean (μ) and standard deviation (σ). Uneven k-mer coverage exhibiting position-specific occurrence counting facets of the sequencing data form the columns of kmer-sequencing count and associated MFE. The k-mers, categorized by MFE, are summed up within each category, and plotted. Once the parameters derived from the modeling data are set in place, the Gaussian distribution function aptly acknowledges the intrinsic biases in the sequencing data.

Fig. S4.



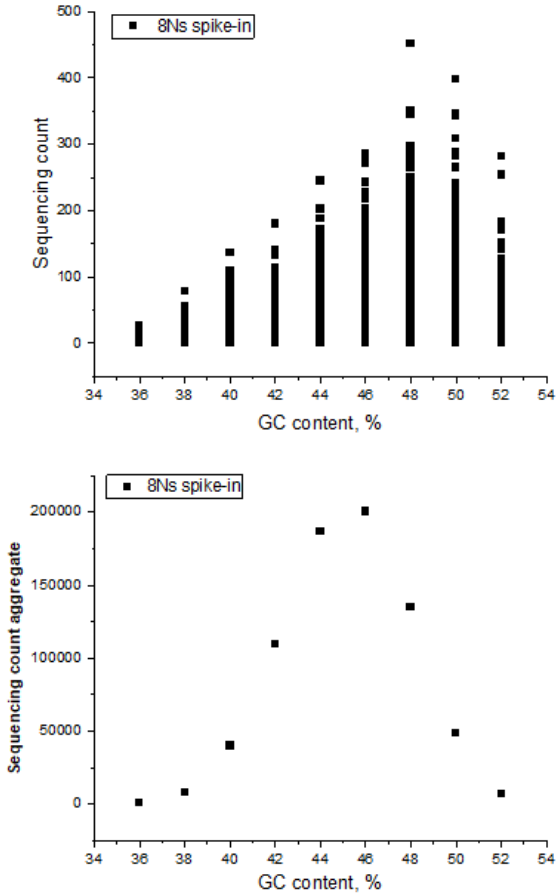
Supplemental Fig. S4. 3-D structure-based bias identification model. A) The spike-in RNAs are synthesized by randomly incorporating A, T, C, and G bases. Transcript sequences from Homo sapiens (GRCh38.cdna.all.fa) were obtained from www.ensembl.org. These mRNA sequences were divided into k-mers starting from the 5' end, and alignments were performed across each transcript to create transcript-specific k-mer sets. B) For each transcript, benchmarking was conducted by identifying evenly distributed 50-mers or global k-mer overlapping frequency, which were then categorized by their VAE-GMM cluster labels. The modeling and sequencing counts of kmer occurrences were aggregated for each cluster label to assess variability across the transcript. The derived data is presented in a table, showing modeling and sequencing counts of k-mers and their associated cluster labels for each transcript. This information was used to plot the data, highlighting disparities in k-mer coverage which revealed position-specific biases. The aggregation of k-mers within each cluster was computed and visualized. A linear fit line was then drawn based on model counts and actual sequencing data for each cluster to simulate theoretical values. Discrepancies between the simulated data and the actual plotted data were analyzed to identify and quantify bias within each cluster.

Fig. S5.



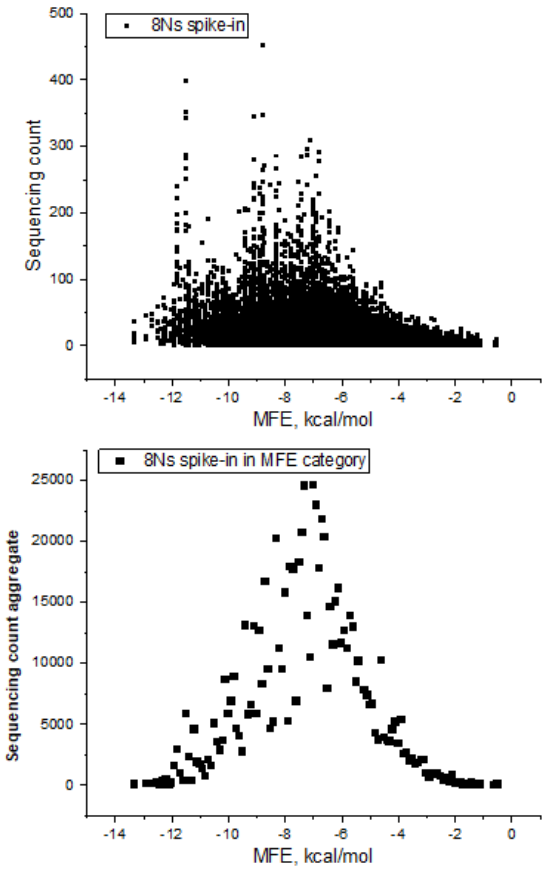
Supplemental Fig. S5. Linear regression analysis correlating GC content with MFE across 65,536 unique spike-in RNA templates. This analysis illustrates the relationship between nucleotide composition and the thermodynamic stability of RNA structures.

Fig. S6.



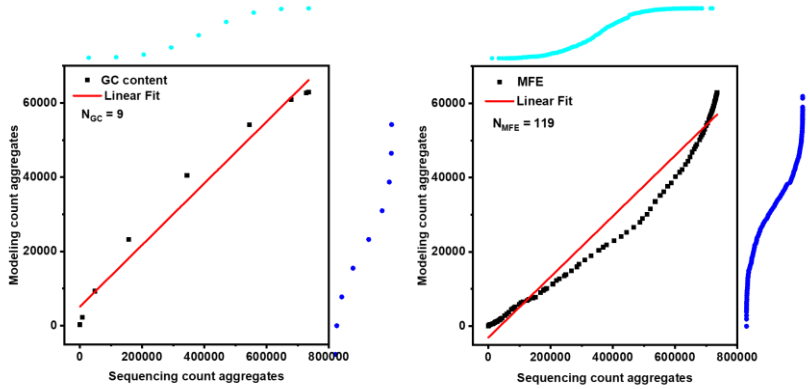
Supplemental Fig. S6. Aggregate the sequencing counts of individual spike-in RNA sequences, denoted as 4^8, by categorizing them according to their GC content, as shown in the upper panel. This approach reveals the distribution and collective tally of counts within various GC content categories, as detailed in the lower panel. It emphasizes the method of grouping sequences that share the same GC-content values, thereby illustrating the relationship between GC content and count distribution.

Fig. S7.



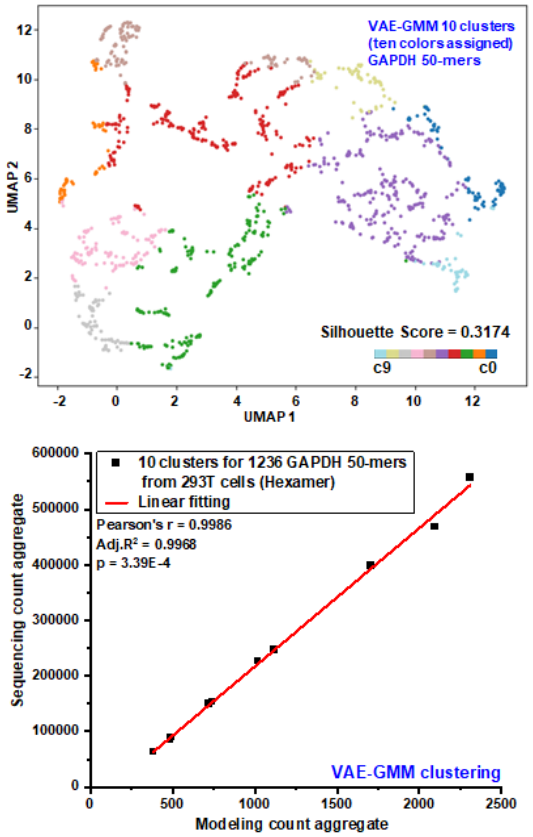
Supplemental Fig. S7. Aggregation of sequencing counts for individual spike-in RNA sequences, denoted as (4^8), categorized by their MFE. The upper panel presents the overall distribution of counts across different MFE categories, showcasing the collective total within each category. The lower panel provides a detailed breakdown of the data, emphasizing how sequences are grouped based on similar MFE values.

Fig. S8.



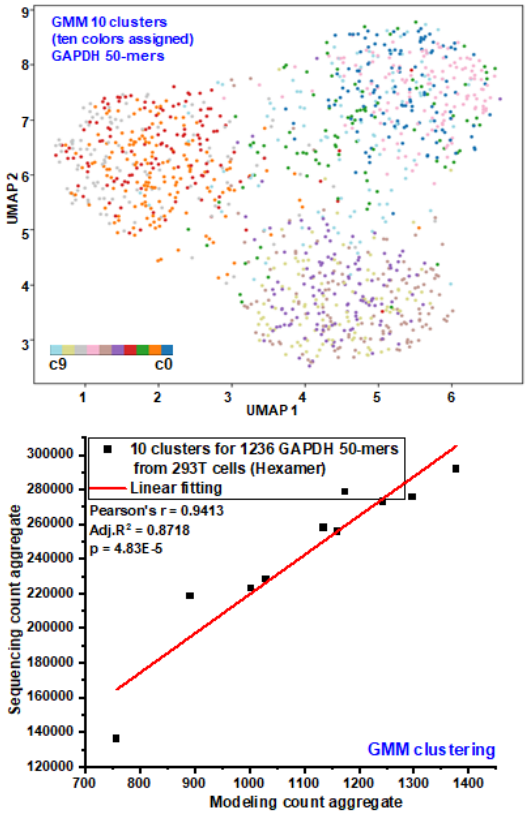
Supplemental Fig. S8. Probability-Probability (P-P) plots comparing observed sequencing counts to modeled theoretical counts. These plots utilize data derived from Figure 3a and Figure 3b in main manuscript. (A) P-P plot for data corresponding to Figure 3a, with a fitted linear regression line yielding a Pearson's correlation coefficient of 0.9876. (B) P-P plot for data corresponding to Figure 3b, where the linear regression shows a Pearson's correlation coefficient of 0.9850.

Fig. S9.



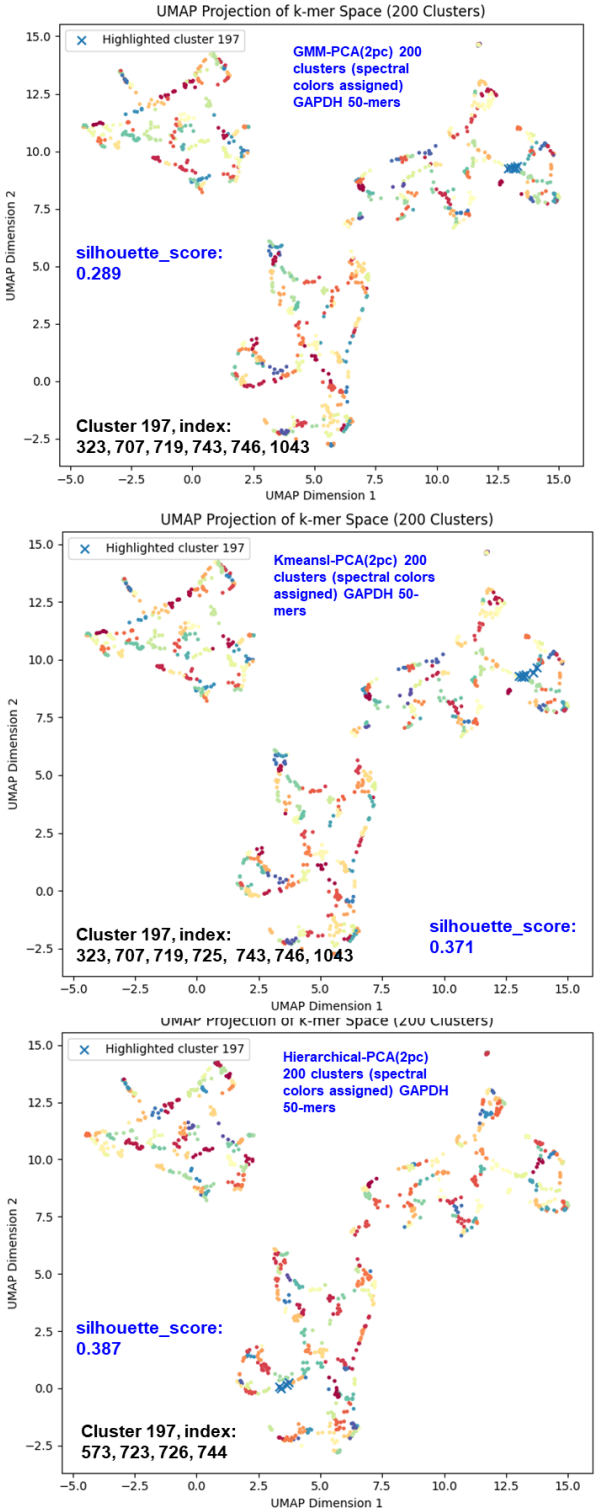
Supplemental Fig. S9. UMAP plot of GAPDH 50-mers clustered into 20 clusters using VAE-GMM. Clusters are color-coded. Modeling predictions and actual counts are aggregated for regression analysis.

Fig. S10.



Supplemental Fig. S10. UMAP plot of GAPDH 50-mers clustered into 20 clusters using GMM-only. Clusters are color-coded. Modeling predictions and actual counts are aggregated for regression analysis.

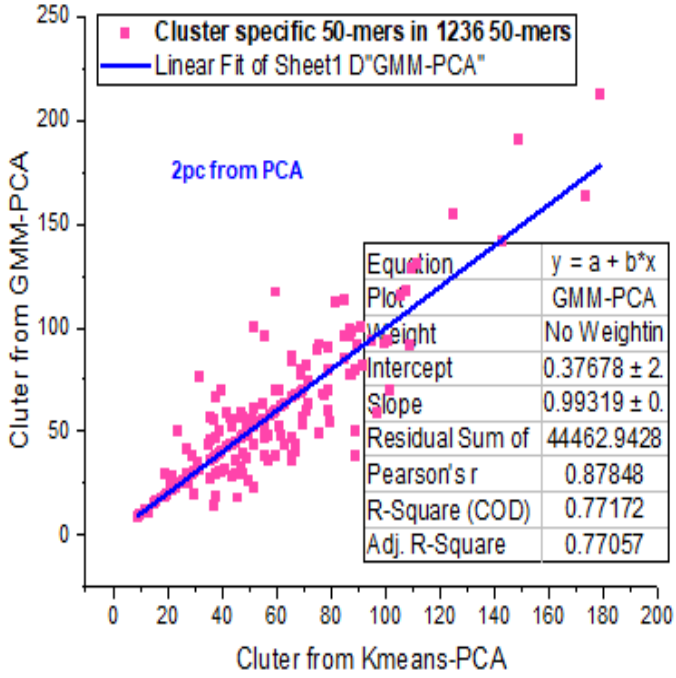
Fig. S11.



Supplemental Fig. S11. UMAP visualizations of 50-mer sequence clusters obtained using GMM-PCA, K-means-PCA, and Hierarchical-PCA methods. The sequences were initially reduced to two dimensions by principal component analysis (PCA), selecting the top two principal components. The reduced datasets were then clustered into 200 clusters

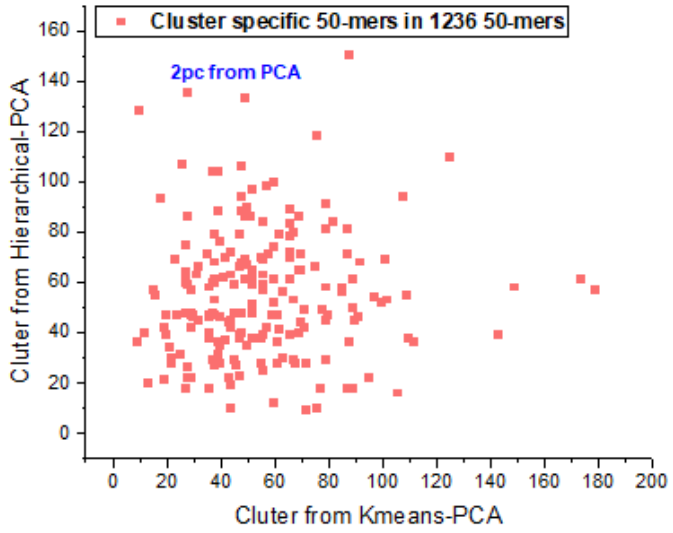
using one of three methods: Gaussian Mixture Model (GMM), K-means, or Hierarchical clustering. In each UMAP plot, Cluster 197 is highlighted to demonstrate that each clustering method yields a different composition of 50-mer sequences within this cluster. The specific indices of the 50-mers contained in Cluster 197 are listed for each method, illustrating the variability in cluster membership depending on the clustering algorithm used.

Fig. S12.



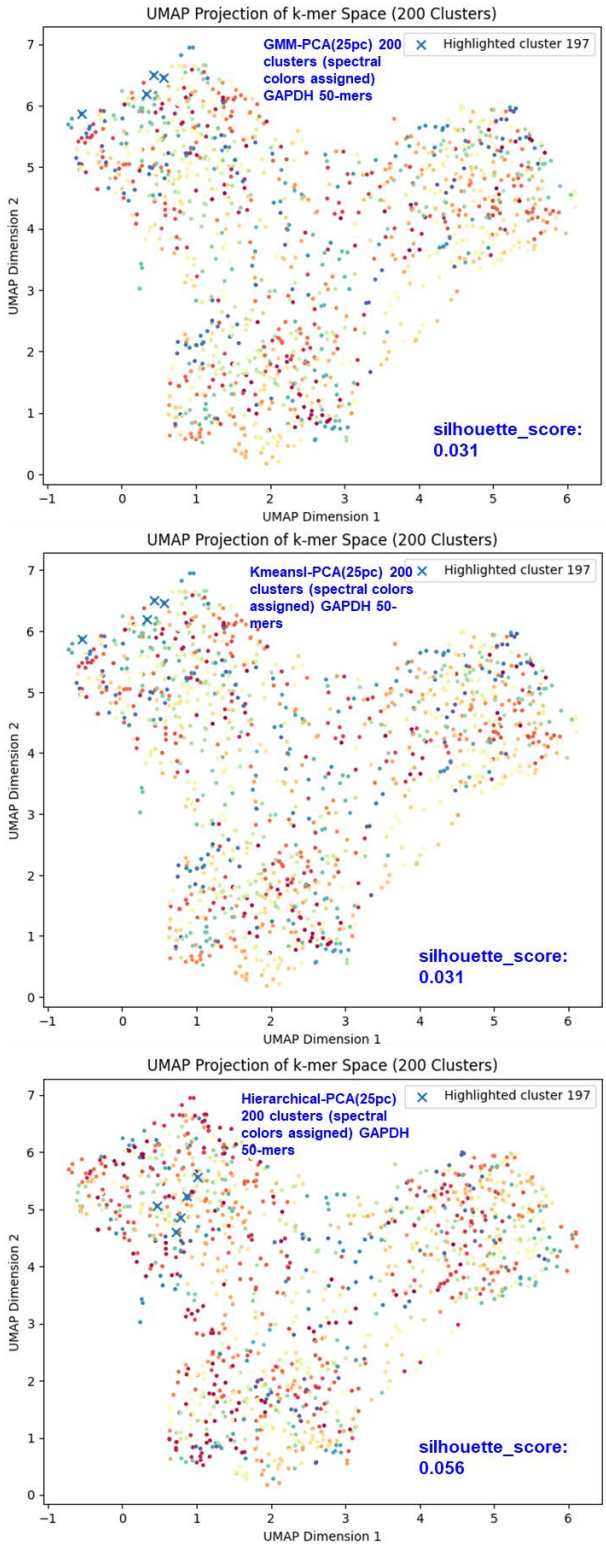
Supplemental Fig. S12. Linear regression analysis comparing cluster sizes from GMM-PCA and KMeans-PCA clustering of 1,236 GAPDH 50-mer sequences. The sequences were one-hot encoded, and dimensionality was reduced to 2 principal components using Principal Component Analysis (PCA). Clustering into 200 clusters was then performed using both Gaussian Mixture Model (GMM) and KMeans algorithms. The resultant cluster sizes from each method were compared using linear regression to illustrate the relationship between these two clustering approaches.

Fig. S13.



Supplemental Fig. S13. Linear regression analysis comparing cluster sizes from Hierarchical-PCA and KMeans-PCA clustering of 1,236 GAPDH 50-mer sequences. The sequences were one-hot encoded, and dimensionality was reduced to 2 principal components using Principal Component Analysis (PCA). Clustering into 200 clusters was then performed using both Gaussian Mixture Model (GMM) and KMeans algorithms. The resultant cluster sizes from each method were compared using linear regression to illustrate the relationship between these two clustering approaches.

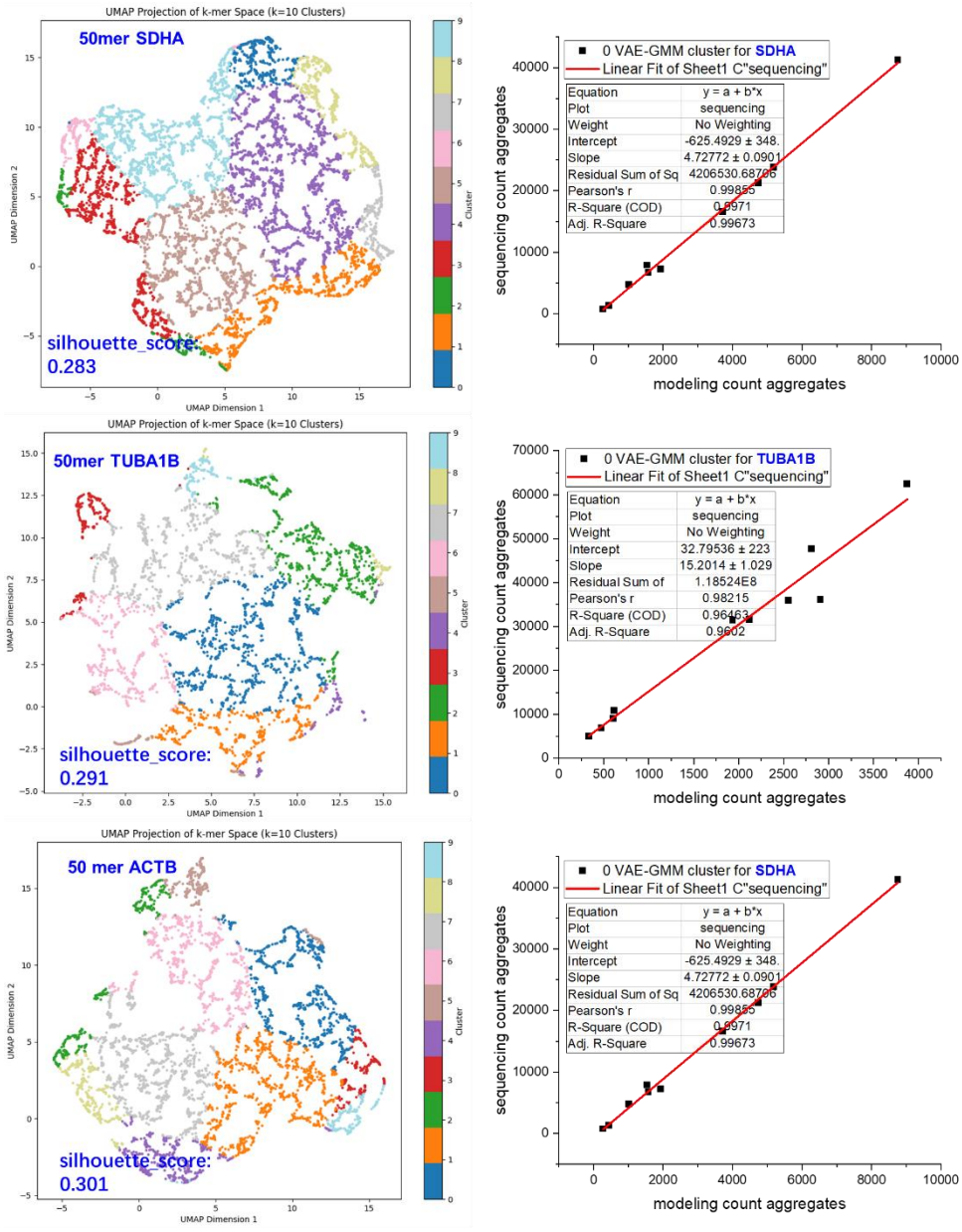
Fig. S14.



Supplemental Fig. S14. UMAP plots of 50-mer sequence clustering using GMM-PCA, K-means-PCA, and Hierarchical-PCA methods. The sequences were first reduced to the top 25 principal components (25 PCs) via PCA and then clustered into 200 clusters using

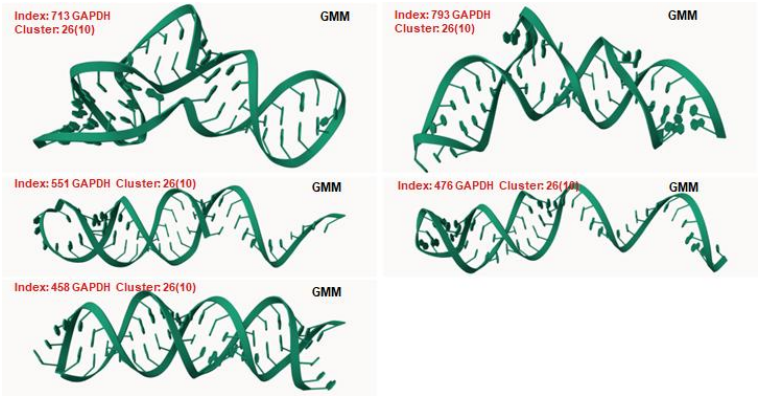
Gaussian Mixture Model (GMM), K-means, or Hierarchical clustering methods. In each plot, cluster 197 is highlighted to illustrate that this cluster contains different sets of 50-mers specific to each clustering method.

Fig. S15.



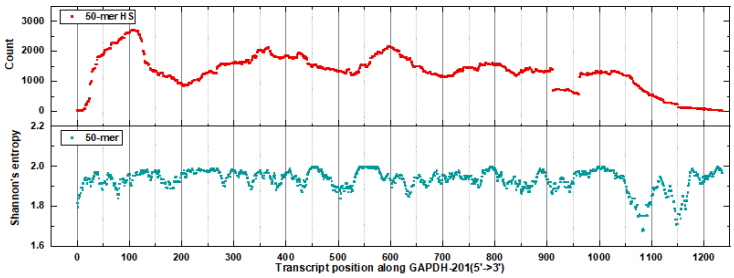
Supplemental Fig. S15. Robustness of VAE-GMM clustering and bias modeling across additional individual transcripts. a) VAE-GMM clustering of 50-mers from individual transcripts. UMAP projections showing 50-mers derived from ACTB, SDHA, and TUBA1B transcripts. Each point represents a 50-mer, colored according to its assignment to one of 10 clusters generated by VAE-GMM (using 2 latent dimensions). Clear separation between clusters is observed for each transcript. b) Correlation between modeling and sequencing k-mer counts within VAE-GMM clusters. Scatter plots illustrating the relationship between aggregated modeling k-mer counts (x-axis) and aggregated sequencing k-mer counts (y-axis) for each of the 10 clusters identified by VAE-GMM for: ACTB, SDHA, and TUBA1B. Each point represents a cluster.

Fig. S16.



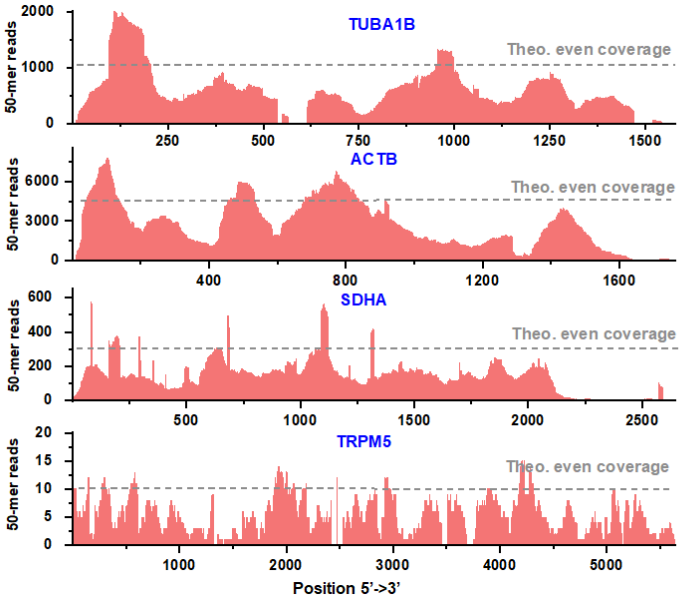
Supplemental Fig. S16. AlphaFold-predicted Structural Diversity in GMM Cluster 26. Cluster 26 from the GMM-only approach contains ten GAPDH 50-mers (g), including index 467 from panel. The structural diversity within this cluster is showcased by 50-mers with indices 458, 476, 551, 713, and 793.

Fig. S17.



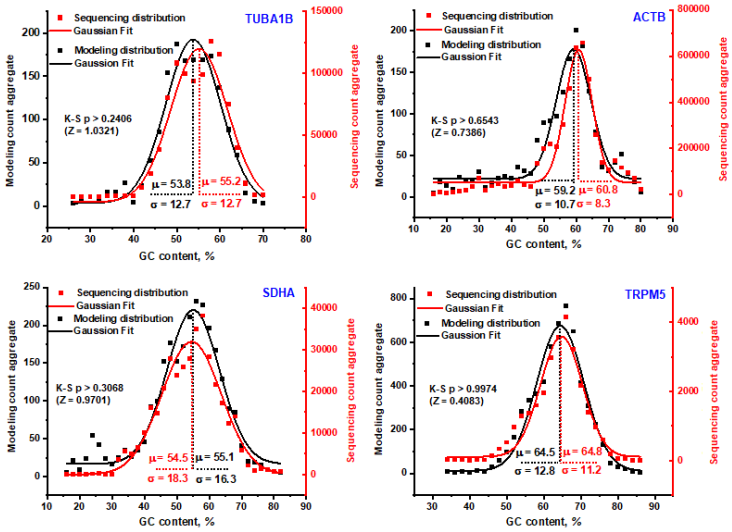
Supplemental Fig. S17. The stack plot represents the sequencing counts for 50-mer sequences derived from GAPDH-201, accompanied by the Shannon entropy for each 50-mer. Counts for 50-mers grouped into different clusters are distinctly labeled.

Fig. S18.



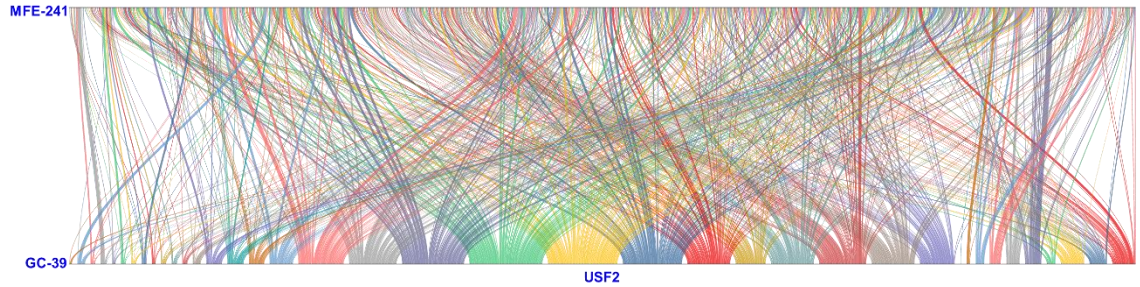
Supplemental Fig. S18. Distribution of 50-mer sequencing counts across the entire regions of genes TUBA1B, ACTB, SDHA, and TRPM5, shown alongside a uniform distribution for comparative analysis. This figure highlights the variation in sequencing count distribution relative to an even distribution model.

Fig. S19.



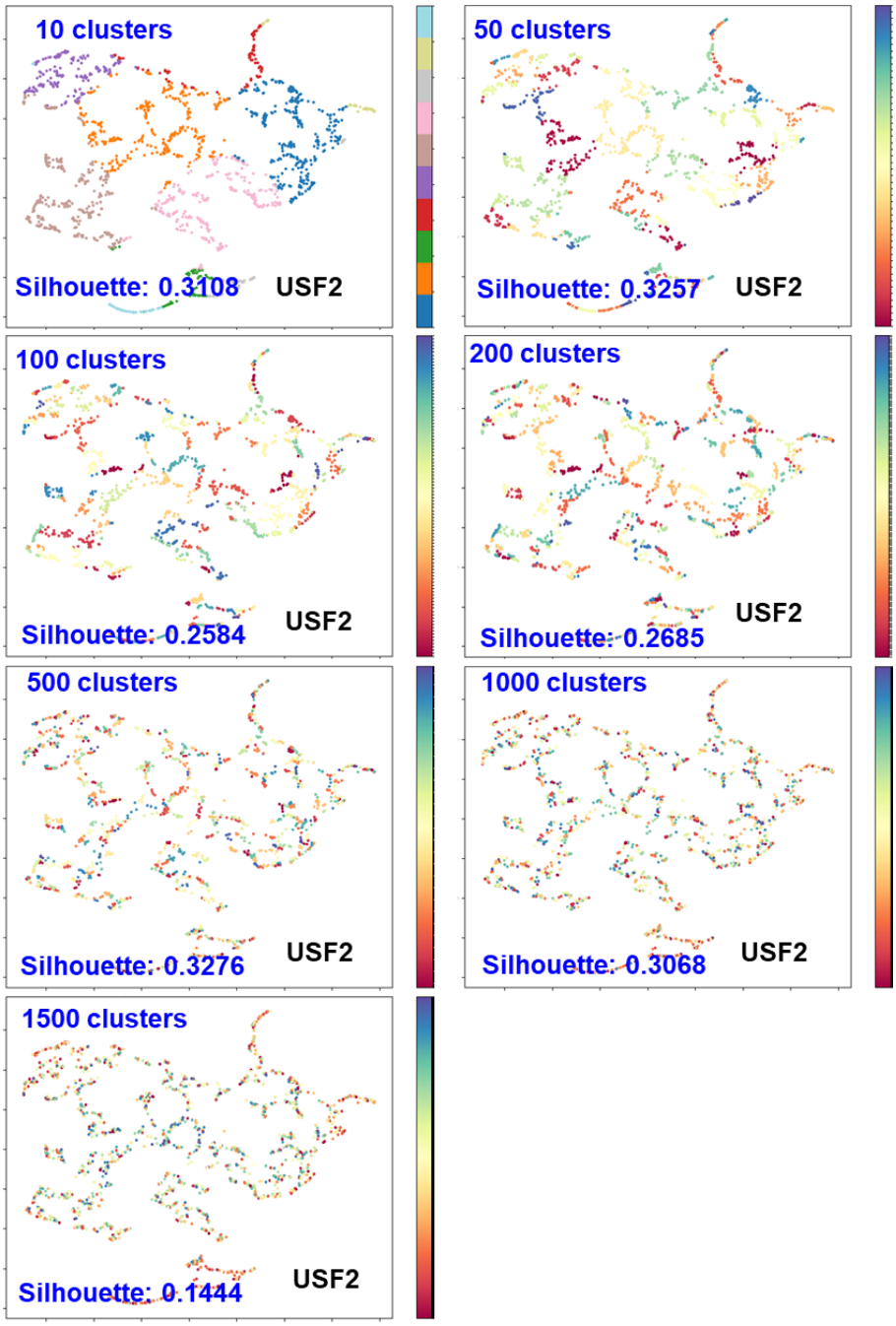
Supplemental Fig. S19. Quantitative analysis of 50-mer count aggregates from the genes TUBA1B, ACTB, SDHA, and TRPM5 utilizing uniform distribution models and actual sequencing data. Gaussian distributions were employed in a free-fitting approach to uncover discrepancies between theoretical models and observed sequences, illustrating the variability in parameter estimates across different gene sequences.

Fig. S20.



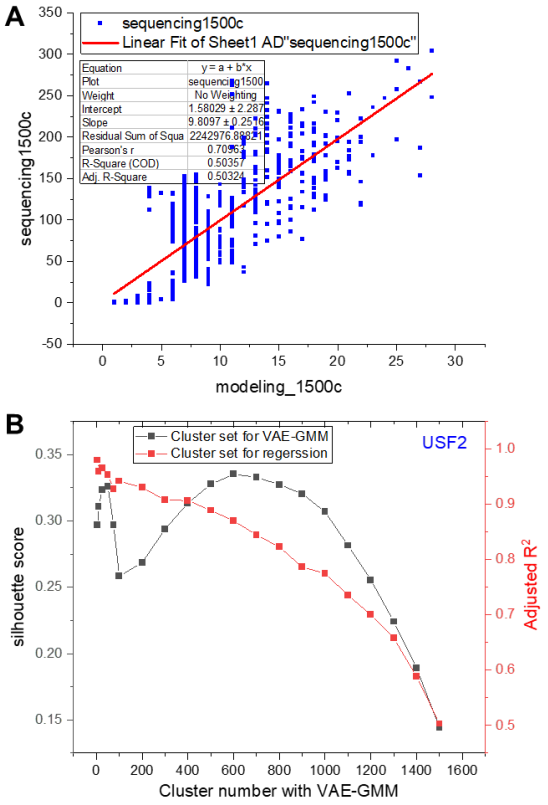
Supplemental Fig. S20. Illustration of the relationship between GC content and minimum free energy (MFE) categorizations for 50-mer segments in the USF2 transcript. This figure depicts the 1696 50-mers grouped into 39 distinct categories based on GC content and 241 categories based on MFE. The visualization highlights the complex and non-linear connections between these categorizations, showing that k-mers within a single GC content category exhibit a wide range of MFE values, and conversely, k-mers sharing the same MFE category display diverse GC contents. This demonstrates the inherent structural diversity and lack of a simple linear relationship between GC content and MFE, supporting the argument that models based on one feature cannot be appropriately evaluated using the other, as elaborated in the main text.

Fig. S21.



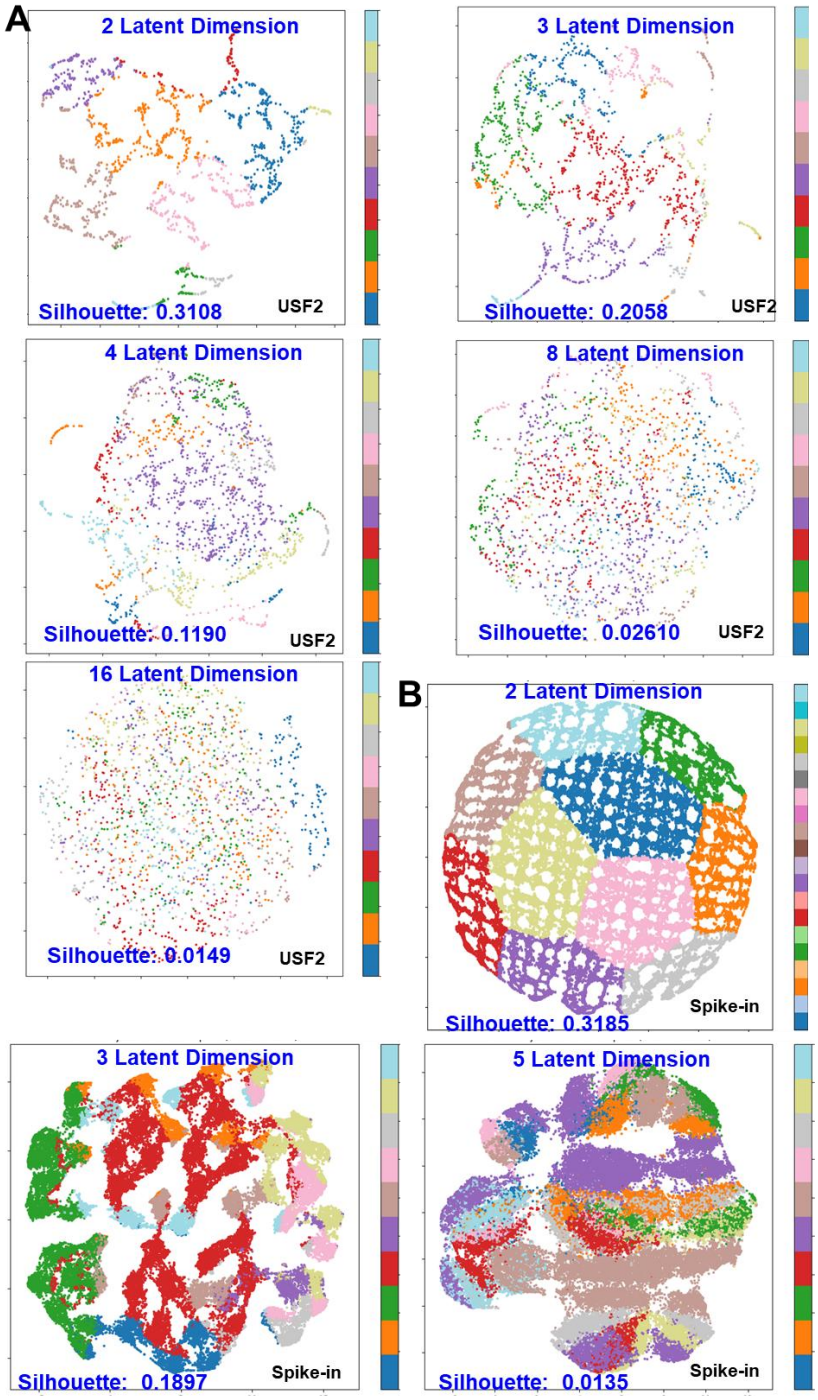
Supplemental Fig. S21. UMAP visualizations of VAE-GMM clustering for 1,697 USF2 50-mers at varying granularity. Clustering was performed using seven different target cluster numbers (k): 10, 50, 100, 200, 500, 1000, and 1500, with all other VAE-GMM parameters held constant. Each plot illustrates the distribution of 50-mers (points) in the UMAP-reduced space, colored by their assigned cluster, revealing the dataset's structural diversity at different resolutions.

Fig. S22.



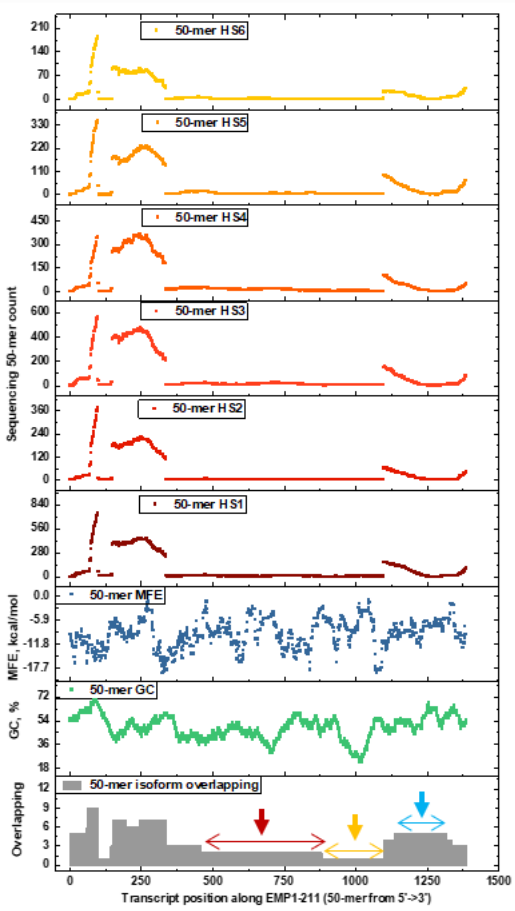
Supplemental Fig. S22. VAE-GMM Clustering Performance and Resolution Analysis for USF2 50-mers. **a**) Linear regression of aggregated modeling predictions versus actual sequencing counts for 1696 unique USF2 50-mers clustered into 1500 groups using VAE-GMM. Each point represents one cluster. The solid line indicates the linear fit, and the associated R^2 value quantifies the correlation, assessing the accuracy of the model at this high clustering resolution. **b**) Performance of VAE-GMM clustering as a function of the number of target clusters (clustering resolution) for the same 1696 unique USF2 50-mers. The x-axis shows the number of GMM clusters applied to the VAE-derived latent space, ranging from 5 to 1500. The left y-axis displays the Silhouette score, indicating clustering quality. The right y-axis displays the adjusted R^2 value from linear regressions (as shown in panel (a) for 1500 clusters) performed at each respective cluster resolution, assessing the relationship between modeled and sequencing counts.

Fig. S23.



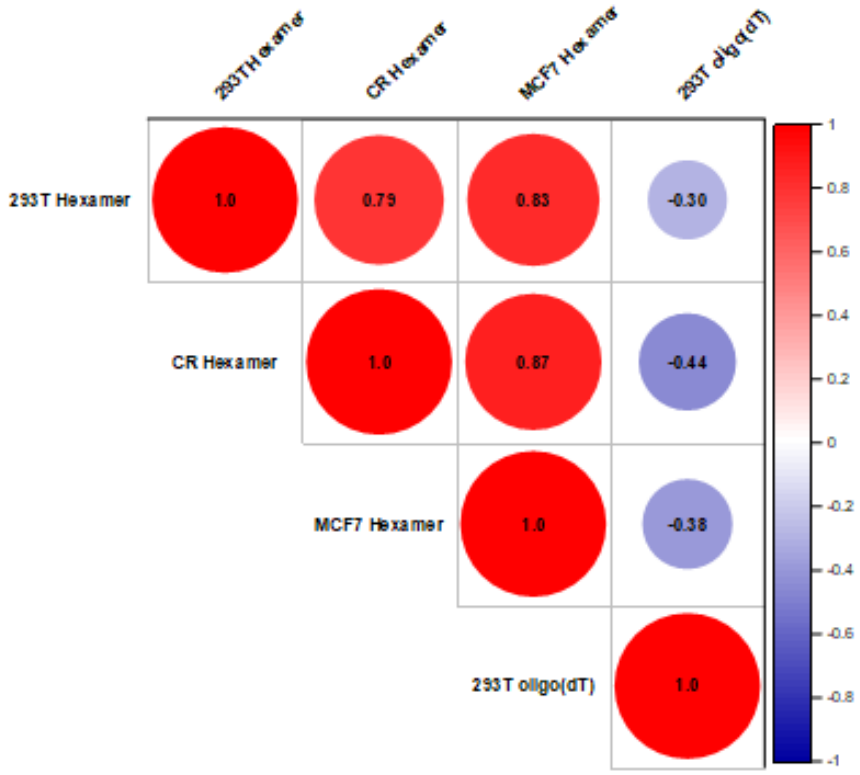
Supplemental Fig. S23. Effect of latent dimension on VAE-GMM clustering shown by UMAP visualization. (a) UMAP plots for 1,697 USF2 50-mers with the latent dimension of the VAE-GMM model set to 2, 3, 4, 8, and 16. (b) UMAP plots for 65,536 spike-ins with the latent dimension set to 2, 3, and 5. For all plots, the VAE-GMM was configured with 10 preset clusters.

Fig. S24.



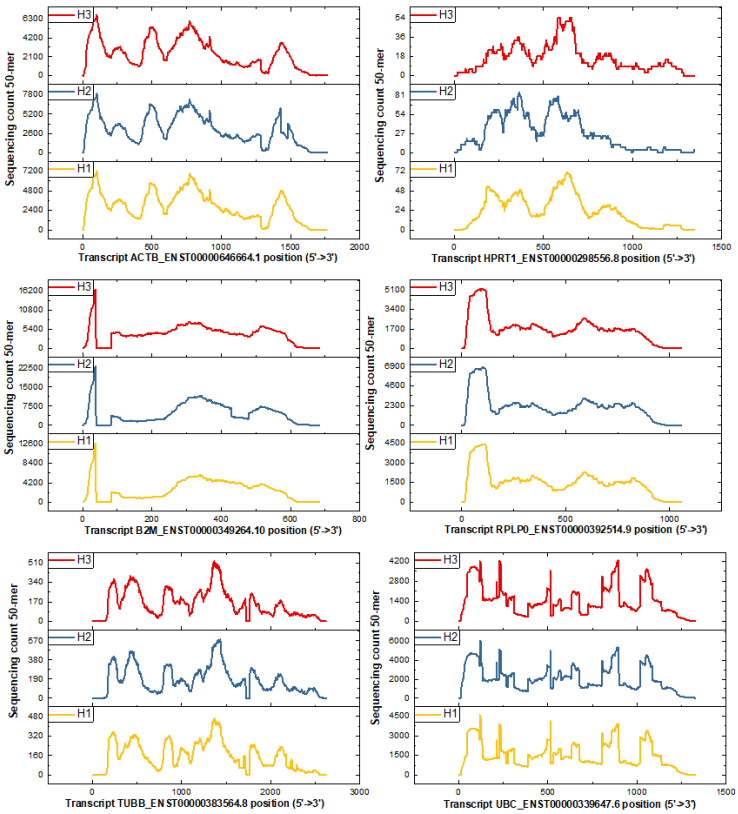
Supplemental Fig. S24. Stacking plot illustrating the distribution of 50-mer sequencing counts in the EMP1-211 region across six human colorectal samples. The plot includes additional layers indicating the GC content, minimum free energy (MFE), and the overlapping degree of each 50-mer. The arrow highlights the isoform set-specific region.

Fig. S25.



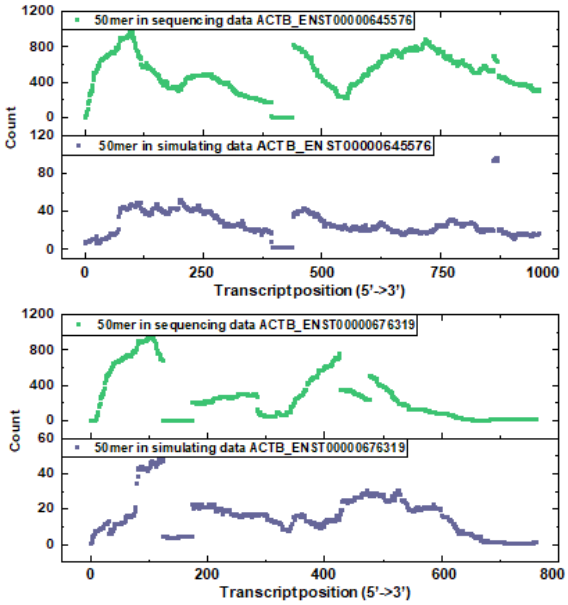
Supplemental Fig. S25. This correlation plot depicts the Pearson correlation coefficients among the 50-mer sequencing profiles of EMP1-211 across different sample preparations and cell lines, specifically HEK293T, MCF7, and a human colorectal sample. The sequencing methods compared include a hexamer-based workflow with Ribo-off rRNA depletion, a hexamer-based workflow with oligo(dT)-based mRNA enrichment, and an oligo(dT)-priming based workflow for total RNA extraction. This analysis highlights the degrees of similarity and differences between the various datasets, providing insights into the consistency of gene expression profiling across techniques and sample types.

Fig. S26.



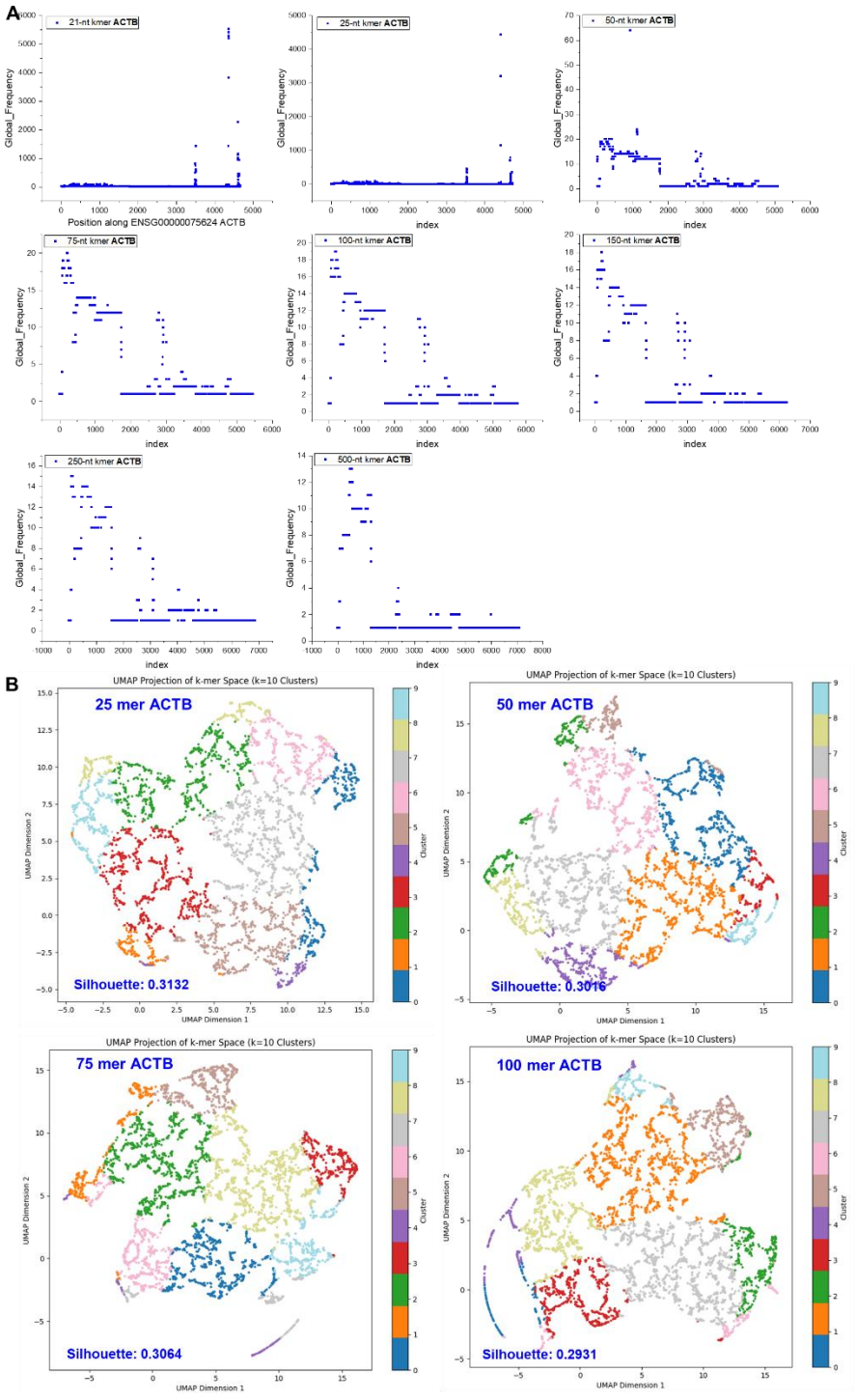
Supplemental Fig. S26. Consistent coverage pattern from sequencing RNA-seq data across different samples. The figure illustrates the consistent coverage patterns observed in RNA-seq data across three different samples (H1, H2, and H3). The analysis focuses on typical housekeeping genes at the isoform level, including ACTB (ENST00000646664.1), B2M (ENST00000349264.10), HPRT1 (ENST00000298556.8), RPLP0 (ENST00000392514.9), TUBB (ENST00000383564.8), and UBC (ENST00000339647.6).

Fig. S27.



Supplemental Fig. S27. Consistent coverage pattern in simulated and sequencing RNA-seq data. This panel compares simulated data with sequencing data by examining the 50-mer overlapping pattern for two ACTB isoforms: ACTB_ENST00000645576 and ACTB_ENST00000676319. The 50-mer count shows significant variation along the transcript lengths, illustrating the differences in sequence coverage. The total number of reads is set at 10,000,000 for both simulated and sequenced data.

Fig. S28.



Supplemental Fig. S28. Impact of k-mer length on frequency analysis and VAE-GMM clustering for ACTB. a) Global frequency of overlapping k-mers (lengths: 21, 25, 50, 75, 100, 150, 250, 500 nt) identified within the ACTB transcript. The displayed degree corresponds to the overall occurrence frequency. b) Comparative VAE-GMM clustering of ACTB k-mers using different input k-mer lengths (25, 50, 75, and 100 nt). All

clustering analyses were performed with a pre-set number of 10 clusters for the GMM, and the resulting cluster separations are visualized via UMAP.