OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf260
https://doi.org/10.1093/bib/bbaf260
Case Study

Deep learning reveals determinants of transcriptional infidelity at nucleotide resolution in the allopolyploid line by goldfish and common carp hybrids

Kaizhuang Jing[1], Tingchu Wei[2,3], Xuedie Gu[2], Guoliang Lin[2], Lin Liu[1,*], Jing Luo[2,3,*]

[1]School of Information, Yunnan Normal University, Kunming 650500, China
[2]State Key Laboratory for Conservation and Utilization of Bio-resource, School of Ecology and Environment, School of Life Sciences, Yunnan University, Kunming 650091, China
[3]Southwest United Graduate School, Kunming 650092, China
*Corresponding authors. Lin Liu, School of Information, Yunnan Normal University, Kunming 650500, China. E-mail: liulinrachel@163.com; Jing Luo, State Key Laboratory for Conservation and Utilization of Bio-resource, School of Ecology and Environment, School of Life Sciences, Yunnan University, Kunming 650091, China. E-mail: jingluo@ynu.edu.cn

Abstract

During DNA transcription, the central dogma states that DNA generates corresponding RNA sequences based on the principle of complementary base pairing. However, in the allopolyploid line by goldfish and common carp hybrids, there is a significant level of transcriptional infidelity. To explore deeper into the causes of transcriptional infidelity in this line, we developed a deep learning model to explore its underlying determinants. First, our model can accurately identify transcriptional infidelity sequences at the nucleotide resolution and effectively distinguish transcriptional infidelity regions at the subregional level. Subsequently, we utilized this model to quantitatively assess the importance of position-specific motifs. Furthermore, by integrating the relationship between transcription factors and their recognition motifs, we unveiled the distribution of position-specific transcription factor families and classes that influence transcriptional infidelity in this line. In summary, our study provides new insights into the deeper determinants of transcriptional infidelity in this line.

Keywords: transcriptional infidelity; the allopolyploid line by goldfish and common carp hybrids; transcription factors

Introduction

In 1956, Francis Crick proposed the central dogma [1], which illustrates the pathway of genetic information transfer among biological macromolecules. According to the central dogma, DNA serves not only as a template for its own replication but also as a template for RNA synthesis, which is the core process of transcription. Subsequently, RNA acts as a template guiding protein synthesis, forming the basis of the translation process [1, 2]. The central dogma indicates that the transfer of genetic information, whether between different generations or between different cells within the same organism, generally maintains a high level of fidelity, ensuring the stability of genetic information across species. However, during actual transcription, the DNA sequence should generate corresponding RNA sequences based on the principle of complementary base pairing, but transcriptional infidelity may still occur. This phenomenon refers to the generation of RNA sequences that do not completely match the template DNA sequence during transcription (i.e. RNA-DNA sequence differences), thereby violating the principle of complementary base pairing [3–5]. Previous studies on transcriptional infidelity in eukaryotic organisms (such as *Saccharomyces cerevisiae* and human B cells) have found that the transcriptional error rate can reach up to the order of 10^{-4} [6–8].

Transcriptional infidelity becomes increasingly pronounced in the allopolyploid line by goldfish and common carp hybrids. This line represents a vertebrate distant hybrid polyploidization model, where the maternal parent is the goldfish (*Carassius auratus* red var., $2n = 4x = 100$), and the paternal parent is the common carp (*Cyprinus carpio* L., $2n = 4x = 100$). The first two generations of hybrids are tetraploid ($2n = 4x = 100$), while subsequent generations are octoploid ($2n = 8x = 200$) [9, 10]. In previous studies, Professor Jing Luo's team from Yunnan University conducted resequencing and transcriptome analyses on progeny with different ploidy levels of the allopolyploid line, covering adult tissues from 12 individuals and 8 tissue types. The study measured transcriptional infidelity levels and found that transcriptional infidelity in this line reached a magnitude of 10^{-3}, which is significantly higher than the general level. Elevated transcriptional infidelity could lead to reduced proteostasis and affect various biological processes, including nucleotide synthesis, nitrogen metabolism, and tryptophan degradation [8]. Polyploidization occurs in only a small number of vertebrate species. The stable allopolyploid line used in this study provides a unique opportunity to investigate the causes of this rarity among extant animals. Analysis of the abnormal transcriptional infidelity observed in this line not only yields valuable insights into the molecular mechanisms underlying the

Received: October 29, 2024. **Revised:** May 05, 2025. **Accepted:** May 17, 2025

scarcity, low survival, and limited reproduction of polyploidized vertebrates, but also establishes a critical theoretical foundation for future research on vertebrate polyploidization.

The mechanisms underlying transcriptional infidelity are highly complex and likely influenced by multiple factors, including the classical RNA editing mechanisms [11] and DNA/RNA damage processes. Moreover, the allopolyploid line may experience "genome shock" induced by polyploidization, which itself can trigger frequent transcriptional infidelity events [12]. However, with the advancement of sequencing technologies, it has been discovered that many sites of transcriptional infidelity cannot be explained by classical RNA editing or similar mechanisms. This phenomenon may partly result from technical errors, but it may also reflect genuine biological phenomena [13–15]. Consequently, further experimental evidence is needed in this field to validate and support these findings, leading to a deeper understanding of the mechanisms driving transcriptional infidelity and its impact on biological systems. Previous studies [12] have utilized high-throughput sequencing technologies to uncover several patterns of transcriptional infidelity in this line. These include the characteristics of nucleotide substitutions—among which the four most prevalent types are A-to-G, T-to-C, G-to-A, and C-to-T—the genomic distribution of transcriptional infidelity events, the association between gene expression levels and transcriptional infidelity, and the functional enrichment of genes exhibiting both high expression and high transcriptional infidelity.

However, previous studies at nucleotide resolution have primarily relied on statistical approaches, which are limited in their ability to uncover position-specific pattern features of transcriptional infidelity. To investigate the causes of the abnormally high levels of transcriptional infidelity in this line, we aim to develop a computational method to identify regions of transcriptional infidelity at nucleotide resolution using contextual sequence information from the DNA. We also seek to study local genomic parameters that influence the occurrence of transcriptional infidelity, thereby providing molecular insights into the mechanisms behind transcriptional infidelity on a genome-wide scale. Deep learning methods, particularly genomic pre-trained models, have proven effective in capturing complex relationships and dependencies within DNA sequences [16, 17]. These methods have demonstrated superior performance in various genomics tasks, including polyadenylation site prediction [18] and single-strand break site prediction [19].

Therefore, we have developed the first deep learning model to study transcriptional infidelity at nucleotide resolution. This model can accurately identify DNA sequences within the genome of the allopolyploid line by goldfish and common carp hybrids that exhibit transcriptional infidelity and offers subregional predictions for these infidelity-prone DNA sequences. Additionally, we quantitatively analyze the contribution of motifs within the context of transcriptional infidelity sites and explore the relationship between transcription factors, their families, and classes with the binding motifs, revealing associations between transcriptional infidelity sites and position-specific transcription factor families and classes. In summary, our research offers new perspectives on understanding the mechanisms of transcriptional infidelity in the allopolyploid line by goldfish and common carp hybrids. All abbreviations are defined in the Nomenclature section of the Supplementary File.

Material and methods

Source of transcriptional infidelity site data

All transcriptional infidelity site data, along with the reference genome and annotation files, were provided by Professor Jing Luo's team at Yunnan University. The dataset we used had already undergone quality control and genome alignment processing, with classical RNA editing sites removed. We selected kidney cells from F_1-generation (Filial generation 1) individuals of the allopolyploid line by goldfish and common carp hybrids as the dataset for model training and analysis. In this sample, 3,610,001 transcriptional infidelity sites were detected, among which 118,702 sites exhibited multiple types of infidelity.

Developing the TTIRI model

Input dataset

To enable the model to effectively learn information related to transcriptional infidelity from the DNA sequence context, we constructed the candidate positive samples by sliding a window of 200 nt with a step size of 10 nt across the reference genome, excluding sequences that did not contain transcriptional infidelity sites. For constructing the candidate negative samples, we applied the same sliding method within coding regions, excluding sequences with transcriptional infidelity sites. The candidate positive samples did not use coding regions as a sliding template to avoid a potential loss in the distribution of infidelity site samples due to coding regions being shorter than the specified sample interval length. Finally, we selected sequences containing infidelity sites from the candidate positive samples as positive samples and sequences without infidelity sites from the candidate negative samples as negative samples.

Through these steps, we obtained 20,647,506 positive samples and 2,650,319 negative samples. To avoid introducing unnecessary complexity from a small number of samples containing non-determined bases and to reduce the storage and computational burden during model training, we first filtered out all samples containing the "N" character, retaining only sequences composed of the four determined bases "ATGC". Considering the significant difference in data volume between positive and negative samples, which could bias the model during training and affect its generalization ability, we randomly selected 10,000 samples from each group to create a balanced dataset. We then split this dataset into training, validation, and test sets in an 8:1:1 ratio to ensure the reliability and stability of model evaluation.

Model architecture

The tokenization and embedding learning foundation of the TTIRI model adopts the approach of the DNABERT-2 model [20], utilizing the pre-training and fine-tuning paradigms provided by this model. The concept of DNABERT-2 is inspired by the BERT model [21] from the field of natural language processing (NLP), which is a contextual language representation model based on the Transformer architecture [22]. The model is initially pre-trained on a large amount of unlabeled data to develop general understanding and then fine-tuned under the guidance of specific tasks to address various task-specific data applications. The effectiveness of DNABERT-2 has been validated across multiple bioinformatics problems, with its performance comparable to the current state-of-the-art models [17, 23–25].

For an input sequence of 200 nt, the model first tokenizes the DNA sequence using a combination of SentencePiece [26] and Byte-Pair Encoding (BPE) [27] (Fig. 1A). SentencePiece is a language-agnostic tokenizer that treats each input as a raw stream without assuming any pre-tokenization. BPE is a compression algorithm that learns a fixed-size, variable-length token vocabulary based on the co-occurrence frequency of characters and is widely used as a tokenization strategy in NLP. The specific process of BPE tokenization begins with initializing the vocabulary with all unique characters in the corpus. Then, in each iteration,

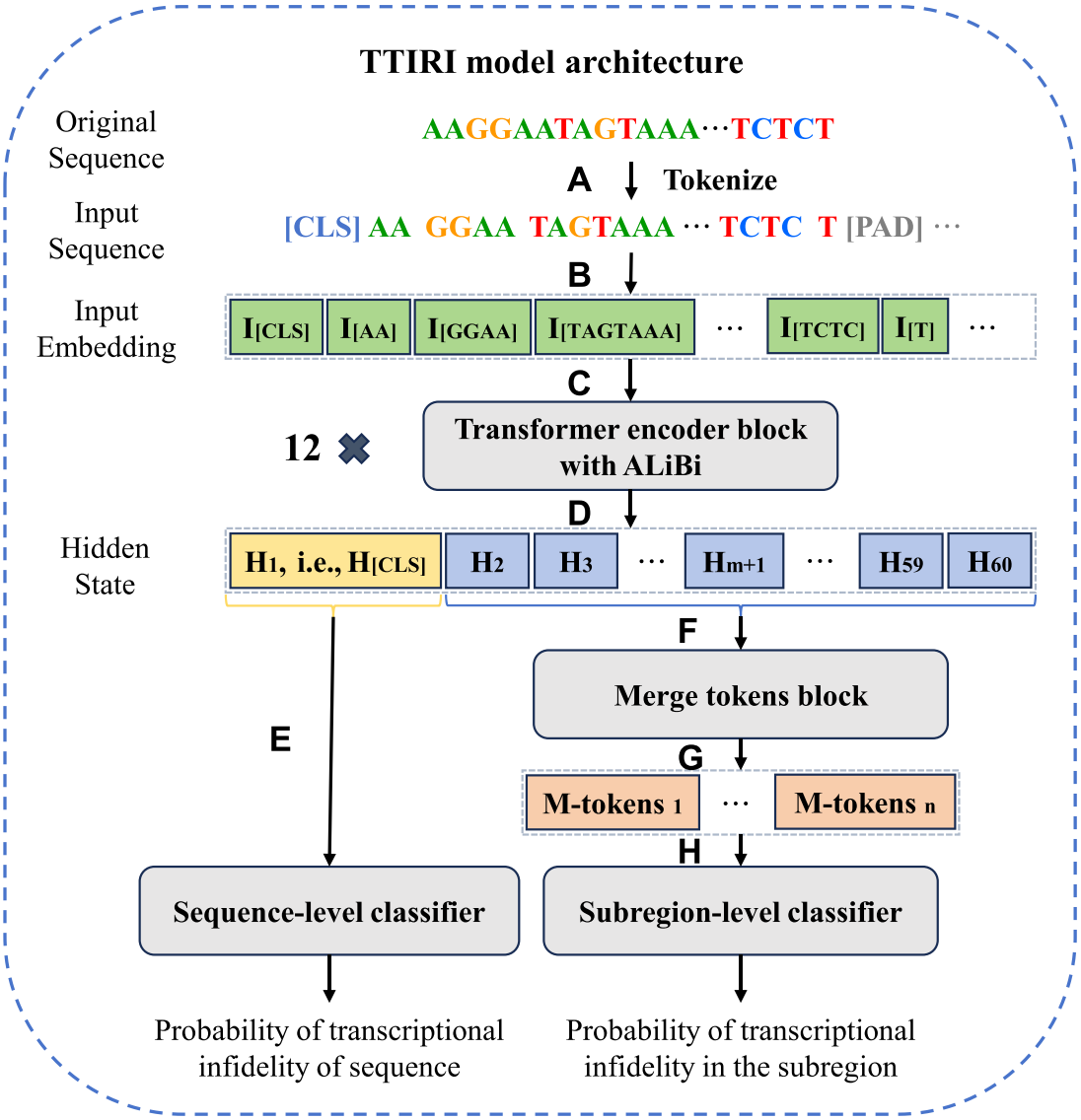

Figure 1. TTIRI model architecture. (A) Tokenization of the original sequence into tokens. (B) Conversion of tokens into numerical vectors. (C) The numerical matrix is fed into the Transformer encoder block with ALiBi for contextual learning. (D) Acquisition of new hidden representations enriched with contextual information. (E) The CLS token is passed to the sequence-level classifier. (F and G) The remaining tokens are passed to the merging token block to obtain m-tokens representing subregions. (H) The m-tokens are passed to the subregion-level classifier.

the most frequently occurring a pair of characters is identified and added to the vocabulary as a new token, replacing all identical character pairs in the corpus with this new token. This process is repeated until the desired vocabulary size is constructed. Through this method, BPE effectively captures patterns and structures within DNA sequences, providing richer representations for downstream tasks. The vocabulary was empirically derived using the DNABERT-2 model, with a vocabulary size of 4096. Given the variable length of the tokens in the vocabulary, a 200 nt sequence, after tokenization, is approximately reduced to one-fifth of its original length, or about 40 effective tokens, as evaluated by the median token count across all samples in the dataset. After obtaining the effective tokens, special tokens such as CLS, SEP, and padding tokens are added to facilitate subsequent sequence classification and to ensure consistency in the total number of tokens per sequence. We denote the total number of tokens as n_t ($n_t = 60$).

Following tokenization, samples are transformed into token representations. These tokens are then passed to the embedding layer (Fig. 1B), which converts each token in the sequence into a corresponding numerical vector, representing each sequence as a matrix $M \in \mathbb{R}^{n_t \times d_h}$, where d_h ($d_h = 768$) denotes the size of the hidden layer, i.e. the dimension of the embedded numerical vectors. Next, matrix M is passed to a stack of 12 Transformer encoder blocks (Fig. 1C), each containing 768 hidden units and 12 attention heads. Positional embeddings are incorporated through the attention with linear biases (ALiBi) mechanism [28]. Formally, each encoder block captures the contextual information of M through the multi-head self-attention mechanism:

$$MultiHead(M) = \text{Concat}(Head_1, \ldots, Head_h)W^O$$

$$Head_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}} + B_i\right)V_i$$

$$Q_i = MW_i^Q; K_i = MW_i^K; V_i = MW_i^V$$

where $W^O \in \mathbb{R}^{d_h \times d_h}$ and $\{W_i^Q, W_i^K, W_i^V\}_{i=1}^h \in \mathbb{R}^{d_h \times d_k}$ are learnable weight matrices, h ($h = 12$) represents the number of self-

attention heads, d_k ($d_k = 64$) is the column vector dimension of Q_i and K_i, and $B_i \in \mathbb{R}^{n_t \times n_t}$ denotes the linear bias matrix calculated using the ALiBi method. When computing the next hidden state of M, each attention head first calculates the attention scores between every pair of tokens, then uses these scores as weights to perform a weighted sum over the rows of V_i, thereby generating a new hidden representation for each token in the sequence that reflects the contextual tokens (Fig. 1D).

After learning through the 12 stacked encoder blocks, the model passes the CLS token of each sequence to the sequence-level classifier (Fig. 1E) to predict transcriptional infidelity phenomena at the sequence-level. Simultaneously, the model merges the remaining tokens of each sequence (denoted as $M^t \in \mathbb{R}^{(n_t-1)\times d_h}$ for any given sequence) and passes them to the subregion-level classifier (Fig. 1F–H) to predict transcriptional infidelity at the subregion-level. In the sequence-level classifier, we optimize using the binary cross-entropy loss function:

$$\ell_{sequence} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$$

where N denotes the batch size, $y_i \in \{0,1\}$ represents the true label of the sample (0 for normal transcription sequence, 1 for transcriptional infidelity sequence), and $p_i \in [0,1]$ is the predicted probability that the sample is a transcriptional infidelity sequence.

In the subregion-level prediction, adjacent tokens are first merged through non-overlapping max-pooling and average-pooling, then the results of these two methods are concatenated and dimensions are restored to obtain improved hidden representations of subregions (containing one or more tokens). The new representation of any sequence after token merging is denoted as $M^m \in \mathbb{R}^{\widehat{n_t}\times d_h}$, where $\widehat{n_t} = \lceil \frac{n_t}{m} \rceil$, m $\in$ {1,2,...,n_t-1} denotes the pooling window size (Fig. 1G). The merging process is formalized as follows:

$$M^m = \mathbf{Concat}(\mathbf{Mean}(M^t, m), \mathbf{Max}(M^t, m))W^C$$

where $\{\mathrm{Mean}(M^t, m), \mathrm{Max}(M^t, m)\} \in \mathbb{R}^{\widehat{n_t}\times d_h}$ denote non-overlapping (i.e. stride equal to m) average and max pooling over M^t with a window size of m, and $W^C \in \mathbb{R}^{2d_h \times d_h}$ is a learnable weight matrix for restoring the hidden layer dimensions. Each new token in M^m is referred to as an m-token (i.e. a subregion), where m indicates the number of merged tokens. M^m is then passed to the subregion-level classifier (Fig. 1H) to predict whether each subregion exhibits transcriptional infidelity. Due to the class imbalance between positive and negative categories in subregion-level prediction, we optimize using a weighted binary cross-entropy loss function:

$$\ell_{token} = \frac{-\sum_{i=1}^{N}\sum_{t=1}^{\widehat{n_t}} V_{i,t}\left(w \cdot y_{i,t}\log p_{i,t} + (1-y_{i,t})\log(1-p_{i,t})\right)}{\sum_{i=1}^{N}\sum_{t=1}^{\widehat{n_t}} V_{i,t}}$$

$$w = \begin{cases} \dfrac{\sum_{i=1}^{N}\sum_{t=1}^{\widehat{n_t}} V_{i,t}(y_{i,t}=0)}{\sum_{i=1}^{N}\sum_{t=1}^{\widehat{n_t}} V_{i,t}(y_{i,t}=1)}, \text{if} \sum_{i=1}^{N}\sum_{t=1}^{\widehat{n_t}} V_{i,t}(y_{i,t}=1) > 0 \\ \sum_{i=1}^{N}\sum_{t=1}^{\widehat{n_t}} V_{i,t}(y_{i,t}=0), \text{otherwise} \end{cases}$$

where $V_{i,t} \in \{0,1\}$ indicates whether a subregion is valid (i.e. the merged original tokens are not all padding tokens), $y_{i,t} \in \{0,1\}$ represents the true label of the subregion, and $p_{i,t} \in [0,1]$ denotes the predicted probability of transcriptional infidelity in the subregion.

The final model loss is composed of both sequence and subregion predictions, formally expressed as:

$$\mathcal{L} = \lambda \ell_{sequence} + (1-\lambda)\ell_{token}$$

where $\lambda \in (0,1)$ is a hyperparameter used to balance the two loss components. In our experiments, we chose λ to be 0.2 based on the evaluation results of the combined predictions.

Model evaluation

In this study, we chose AUROC and AUPRC as evaluation metrics due to their significant effectiveness in assessing classification model performance. These metrics provide an overall measure of performance across all possible classification thresholds, ensuring a comprehensive evaluation of the model's performance.

First, we selected AUROC as the evaluation metric for sequence-level predictions, mainly for the following reasons: AUROC measures the overall ability of the model to distinguish between positive and negative samples, providing a summary of the model's performance across various decision thresholds. For balanced datasets, AUROC accurately reflects the model's overall discriminative ability, with a range from 0 to 1, where 0.5 indicates a random model and 1 indicates a perfect model. A higher AUROC value indicates better performance in distinguishing between positive and negative classes.

Secondly, for subregion-level predictions, we selected AUPRC as the evaluation metric due to the class imbalance in our dataset. AUPRC, which is based on precision and recall, better reflects the model's predictive ability on minority class samples. Precision measures the proportion of true positive samples among all samples predicted as positive, while recall measures the proportion of correctly predicted positive samples among all actual positive samples. AUPRC is particularly useful when the costs of false positives and false negatives differ, as it can identify the optimal classification threshold that minimizes cost. By using AUPRC, we can more comprehensively evaluate the model's performance when handling imbalanced datasets.

Quantitative analysis of the contribution of position-specific motifs to transcriptional infidelity sites

To interpret the regulatory parameters involved when the model identifies transcriptional infidelity at the genomic level, we first quantified the contribution of position-specific hexamers in Section 2.3.1. However, given the complexity and inefficiency of directly analyzing all 4096 possible hexamers, an effective motif screening or classification method is necessary. Therefore, in Section 2.3.2, we further screened the contributions of position-specific hexamers using transcription factor motifs to extract key motifs closely related to transcriptional infidelity. Finally, in Section 2.3.3, we categorized and analyzed these motifs according to transcription factor families and classes, thereby revealing the regulatory roles of different transcription factor families and classes in transcriptional infidelity.

Quantifying the impact of position-specific hexamers on transcriptional infidelity sites

To identify the motifs driving the formation of genome-wide transcriptional infidelity sites and to study the impact of position-specific hexamers on the occurrence of transcriptional infidelity,

we conducted the following experiments. We randomly sampled 10,000 transcriptional infidelity sites and obtained the sequences 100nt upstream and downstream of these sites, forming 201 nt-long transcriptional infidelity sequences (with the central position confirmed to have transcriptional infidelity). Next, we constructed 10,000 201nt-long normal transcription sequences using the same method as the TTIRI model input dataset and performed model training and prediction at the sequence-level only. Finally, we analyzed these 10,000 transcriptional infidelity sequences.

We employed motif perturbation to systematically quantify the contribution of position-specific hexamers to the model's prediction of transcriptional infidelity in a sequence. For each hexamer appearing in a transcriptional infidelity sequence, we replaced it 100 times with a random hexamer, keeping the rest of the sequence unchanged. Predictions were made for both the original and perturbed sequences, and the impact of motif perturbation was quantified using the median change in the log-odds of predicted classification probabilities from 100 substitutions. Importance scores for position-specific hexamers in all correctly predicted transcriptional infidelity sequences were calculated as follows:

$$IS_{i,j}^{\text{hexamer}} = V_i \cdot \text{median}\left(\Delta\,\text{logit}_{i,j}^{(1)}, \ldots, \Delta\,\text{logit}_{i,j}^{(C)}\right)$$

$$\Delta\,\text{logit}_{i,j}^{(k)} = \text{logit}(P(S_i)) - \text{logit}\left(P\left(S_{i,j}^{(k)}\right)\right)$$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$i \in [1, N]; j \in [1, L-5]; k \in [1, C]$$

where $N = 10,000$ denotes the number of sequences analyzed, $L = 201$ represents sequence length, $C = 100$ is the number of perturbations, S_i is the i transcriptional infidelity sequence, $S_{i,j}^{(k)}$ is the sequence with the hexamer at the j position of the i sequence randomly perturbed, $P(\cdot) \in (0, 1)$ denotes the model's prediction probability, and $V_i \in \{0, 1\}$ indicates whether the i sequence was correctly predicted as a transcriptional infidelity sequence (i.e. $P(S_i) \geq 0.5$, thus $V_i = 1$; otherwise, $V_i = 0$).

Quantifying the impact of position-specific transcription factor binding motifs on transcriptional infidelity sites

To obtain high-quality information on transcription factors and their binding motifs, we selected the latest 2024 version of publicly available, high-quality vertebrate transcription factor data from the Jaspar database [29], which includes 828 transcription factors from 105 families and 44 classes. First, we normalized each transcription factor's position frequency matrix provided by the database using a pseudocount of 0.5 to avoid zero frequencies. Then, we converted the normalized result into a position weight matrix (PWM) using the following formula:

$$PWM_{i,j} = \log\left(\frac{PPM_{i,j}}{p_i}\right)$$

$$PPM_{i,j} = \frac{PFM_{i,j} + \alpha}{\sum_k (PFM_{k,j} + \alpha)}$$

where $i, k \in \{A, C, G, T\}$ represent the four bases, $j \in \{1, 2, \ldots, L\}$ represents the position, L is the motif length, and $\alpha = 0.5$ is the pseudocount. p_i denotes the background probability, which we assume is equal for the four bases, setting p_i to 0.25.

Next, we used each PWM to align and score each transcriptional infidelity sequence. For a given PWM and sequence to be aligned, we used a sliding window approach with a stride of 1nt to extract subsequences s of the same length as the PWM, and their scores on the PWM were calculated as follows:

$$Score(PWM, s) = \sum_{j=1}^{L} PWM_{s_j, j}$$

where L is the motif length, and s_j denotes the base at the j position of the subsequence. To further standardize these scores, we applied a sigmoid function to map the scores to the range of 0 to 1 and set a threshold of 0.999 to filter the standardized scores to ensure alignment reliability.

Finally, we calculated the impact scores of position-specific transcription factor binding motifs on transcriptional infidelity sites for the correctly identified transcriptional infidelity sequences, formally represented as follows:

$$IS_{i,j,n}^{\text{TF}} = \frac{1}{L_n - 5} \sum_{k=0}^{L_n - 6} IS_{i,j+k}^{\text{hexamer}}$$

$$i \in [1, N]; j \in [1, L-5]$$

where N is the number of sequences, L is the sequence length, n denotes a transcription factor, and L_n is the length of the binding motif for transcription factor n.

Quantifying the macro impact of transcription factor families and classes in model predictions

To observe the contribution of position-specific motifs to the model's prediction of transcriptional infidelity sequences from a macro perspective, we aggregated the results of the analysis of 10,000 sequences and classified transcription factors according to their families and classes to reduce quantification errors and information redundancy introduced by similar transcription factor binding motifs. The impact scores of position-specific transcription factor families and classes on model predictions were calculated as follows:

$$IS_{j,fn/cn}^{\text{Family/Class}} = \frac{\sum_{i=1}^{N} \sum_{n \in fn/cn} V_{i,j,n} \cdot IS_{i,j,n}^{\text{TF}}}{\sum_{i=1}^{N} \sum_{n \in fn/cn} V_{i,j,n}}$$

$$j \in [1, L-5]; V_{i,j,n} \in \{0, 1\}$$

where *fn/cn* represents a set of transcription factors of a family or class, and $V_{i,j,n}$ indicates that the i sequence is a valid prediction and the j position is bound by transcription factor n (in this case, $V_{i,j,n} = 1$, otherwise $V_{i,j,n} = 0$).

Results

Development of a deep learning model named TTIRI to identify transcriptional infidelity regions

The original data on transcriptional infidelity sites that we used underwent standard processing, including quality control and genome alignment, and classical RNA editing sites were excluded. To ensure that the model could fully learn the real environment where transcriptional infidelity occurs, we constructed negative samples by sliding and extracting 200 nt of normal transcription sequences in the coding region. Positive samples, on the other hand, were created by sliding and extracting 200 nt sequences containing transcriptional infidelity sites on the reference genome. The choice to use the reference genome as the

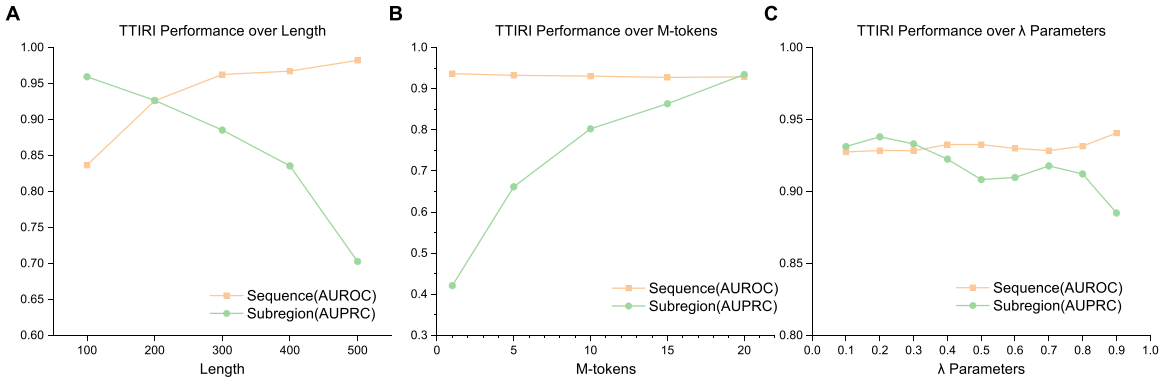

Figure 2. Comparison of TTIRI model performance under different settings. (A) Evaluating the model performance variation with sequence length (from 100 to 500 nt) to determine the optimal sequence length. (B) Evaluating the model performance variation with subregion size (from subregions containing 1 token to 20 tokens) to determine the optimal subregion size. (C) Evaluating the impact of the two-layer classifier loss adjustment hyperparameter λ on model performance.

template for extracting positive samples is due to the finding that about 79% of the coding regions are shorter than the specified 200 nt, but transcriptional infidelity sites within these regions should not be ignored.

Our model constitutes a two-layer transcriptional infidelity region identifier (TTIRI) computational framework (Fig. 1). The backbone component of the model employs the pre-training and fine-tuning structure of the DNABERT2 language model [20], utilizing its excellent embedding learning features to achieve better DNA sequence representation. The model has two output branches: one for predicting the classification probability that a sequence undergoes transcriptional infidelity (sequence-level prediction), and another for predicting the classification probability of transcriptional infidelity occurring in subregions of the sequence (subregion-level prediction). The length of the subregions is not fixed but divided according to the number of tokens. In our default setting, a 200 nt sequence is approximately divided into two subregions (details provided in the methods section). Subregion-level prediction is a further refinement of sequence-level prediction; when the model predicts that a sequence will exhibit transcriptional infidelity, it then proceeds to predict for its subregions. This design allows us to identify regions of transcriptional infidelity at a more detailed nucleotide resolution.

In the sequence-level prediction component, given the balanced nature of the dataset, we use the area under the receiver operating characteristic curve (AUROC) as the evaluation metric. In contrast, for the subregion-level prediction component, where there is significant imbalance in label classes, we employ the area under the precision-recall curve (AUPRC) as the evaluation metric.

TTIRI model accurately predicts transcriptional infidelity regions

We randomly divided the constructed dataset into training, validation, and test sets in an 8:1:1 ratio. During training, only the training set is used to train the model, and all evaluation results are based on the test set. TTIRI is the first computational method designed specifically for predicting transcriptional infidelity sites. Given that its foundation model, DNABERT2, has demonstrated superior performance over baseline machine learning models (such as convolutional neural networks (CNN), long short-term memory networks (LSTM), gated recurrent units (GRU), and their common combinations) across various biological problems, our study focuses on the rationality of the model's architecture and parameter design and their effectiveness in predicting transcriptional infidelity regions. The superiority of the proposed two-layer architecture is validated in the "Two-layer architecture design advantage" section of the Supplementary File. The results show that, across varying subregion sizes, the two-layer design achieves an average improvement of ∼6 percentage points in subregion-level AUPRC compared to the single-layer design. We next explore the impact of different parameter configurations on model performance and determine the optimal parameter settings.

First, we examined the impact of different sequence sample lengths on the model's predictive performance (Fig. 2A). To do this, we created variants of the model's dataset with sequence lengths ranging from 100 nt to 500 nt to identify the optimal sequence length. Figure 2A shows that as sequence length increases, sequence-level prediction performance improves, but subregional-level prediction performance decreases. This phenomenon may be due to the increased sequence length providing richer contextual information, but exacerbating class imbalance within the subregions, thus impairing predictive performance. The experimental results indicate that the model achieves the best overall performance with a sequence length of 200 nt in the two-layer prediction setup.

Next, with the sequence length fixed at 200 nt, we investigated the impact of different subregion sizes on model performance (Fig. 2B). We represented the subregion size by the number of tokens and analyzed cases where the number of tokens ranged from 1 to 20. When the token count was 20, a 200 nt sequence was approximately divided into two subregions. Figure 2B shows that as the number of tokens in a subregion increases, subregional-level prediction performance gradually improves, while sequence-level prediction remains largely unaffected. The model achieves optimal overall predictive performance when the token count is set to 20. However, as the number of tokens increases, although predictive performance improves, the model's ability to identify subregions with finer detail diminishes.

Finally, we studied the effect of the λ hyperparameter, which adjusts the relative weight of the losses between the two classifiers during model training, on model performance with a sequence length of 200 nt and a token count of 20 (Fig. 2C). The experimental results show that as the value of λ increases, sequence-level prediction performance improves, but subregional-level prediction performance significantly decreases. Considering overall performance, a λ setting of 0.2 yields the

best results. Under this setting, our model achieved an AUROC of ∼0.928 in sequence-level prediction and an AUPRC of ∼0.937 in subregional-level prediction. More comprehensive evaluation metrics and detailed numerical results are documented in Supplementary Data 1.

Quantifying the contribution of position-specific motifs to predicting transcriptional infidelity

Although empirical models can effectively capture the contextual environment in which transcriptional infidelity occurs, the black-box nature of deep learning models makes it challenging to provide biologically insightful analyses of these contexts. Therefore, we employed model interpretability methods to quantify the contribution of position-specific motifs to transcriptional infidelity prediction, aiming to offer biological insights into the occurrence of transcriptional infidelity. Initially, we reconstructed the dataset and conducted sequence-level training and prediction. The positive samples in the new dataset centered on empirically verified transcriptional infidelity sites, while the negative samples represented sequences with normal transcription throughout. This dataset reconstruction was intended to quantify the contribution of position-specific motifs to transcriptional infidelity sites, rather than merely capturing the occurrence region. Subsequently, we evaluated all correctly predicted transcriptional infidelity sequences using a hexamer perturbation approach. For each position in every sequence, we performed 100 perturbations of the hexamer and calculated the median change in the log-odds of the prediction probability before and after the perturbation to determine the contribution score of the position-specific hexamer.

Through the methods described above, we obtained contribution scores for position-specific hexamers within specified sequences. However, due to the large number of hexamer types ($4^6 = 4096$) and the significant influence of noise on individual sequence interpretation scores (including noise introduced by interpretative methods, model fit discrepancies, and inherent dataset noise), we faced challenges in meaningful motif selection. To address this issue, we utilized transcription factors closely associated with the transcription process to identify their binding motifs, thus facilitating more meaningful motif selection. Initially, we downloaded data on vertebrate transcription factors from the 2024 edition of the Jaspar database [29], which includes 828 transcription factors belonging to 105 families and 44 classes. We then scanned all correctly predicted transcriptional infidelity sequences, binding transcription factors to their reliable binding motifs, and combined this with the contribution scores of position-specific hexamers to determine the contribution of position-specific transcription factors (represented by their corresponding binding motifs) to transcriptional infidelity sites. Lastly, to identify the main drivers behind the formation of transcriptional infidelity sites across the transcriptome, we merged the analysis results of all sequences and eliminated the influence of frequency. Detailed scores for transcription factor motif recognition importance and frequency distributions are recorded in Supplementary Data 2. Additionally, we have provided a specific-sequence analysis example in the "Sequence-Specific Analysis Example" section of the Supplementary File.

We conducted a further classification analysis of transcription factors based on their families and classes using three analytical approaches: (i) Top individual sites: We identified the top six families and classes with the highest relative importance scores at specific individual sites (Fig. 3A and B). (ii) Cumulative importance scores: We identified the top seven families and top five classes based on the cumulative relative importance scores across all sites (Fig. 3C and D). (iii) Positive importance across all sites: We displayed families and classes that exhibit positive importance scores across the entire site distribution (Fig. 3E and F). To reduce noise in the data and facilitate the observation of distribution trends, we applied weighted smoothing with a window size of 10 during visualization. In Fig. 3A, despite the limited number of binding sites for transcription factors within these families, their binding motifs exhibited higher importance scores in specific relative regions of the infidelity sites. For instance, the HD-ZF factors family displayed higher contribution scores within the −40 nt to −60 nt range around the infidelity sites. In Fig. 3C, we observed that the positional contribution distributions for the HD-ZF factors and E2A families were quite similar, as were those for the TBrain-related factors, TBX6-related factors, TBX1-related factors, and TBX2-related factors families. Notably, the HD-ZF factors family ranked highly in both the first and second analytical approaches. In Fig. 3E, we observed transcription factor families that consistently contributed positively across the distribution of infidelity sites, including the NK, Paired-related HD factors, HOX, and HD-LIM families, which exhibited highly similar patterns across the entire distribution. Figure 3B, D, and F reveal similar distribution patterns when transcription factors were classified by class rather than by family. However, it is important to consider the value of *N* (as indicated in the legend, representing the total number of binding sites) when interpreting these distributions, as the *N* value can influence the scaling of the distribution. Detailed importance scores and frequency distributions for transcription factor families and classes are recorded in Supplementary Data 3.

Finally, we analyzed the core consensus motifs for the transcription factor families mentioned above. We found that the core consensus motif for the HD-ZF factors and E2A families was CACCTG, while the core consensus motif for the TBrain-related factors, TBX6-related factors, TBX1-related factors, and TBX2-related factors families was AGGTGTGA. The core consensus motif for the NK, Paired-related HD factors, HOX, and HD-LIM families was [C/T]AATTA. Additional information on the contributions of transcription factor families and classes can be found in Supplementary Data 2 and 3.

Validation of the reliability of the interpretation scheme through basic sequence features

To verify the reliability of the proposed interpretation scheme, we further analyzed the basic features of the sequences used for training and interpretation, including those containing regions of transcriptional infidelity and regions of normal transcription. These features cover the proportions of single-nucleotides, as well as the ranking changes in the proportions of dinucleotides and trinucleotides, as shown in Fig. 4. In Fig. 4A, we observed that in sequences with normal transcription, the proportions of the four single-nucleotides were relatively uniform, with almost no significant differences. However, in sequences with transcriptional infidelity, the AT content was significantly higher than the GC content, with a difference of ∼9 percentage points. This difference may suggest a correlation between the occurrence of transcriptional infidelity and the nucleotide composition of the sequence. In Fig. 4B and C, we further examined the ranking changes in the proportions of dinucleotides and trinucleotides between sequences with transcriptional infidelity and those with normal transcription. We noticed that the dinucleotides and

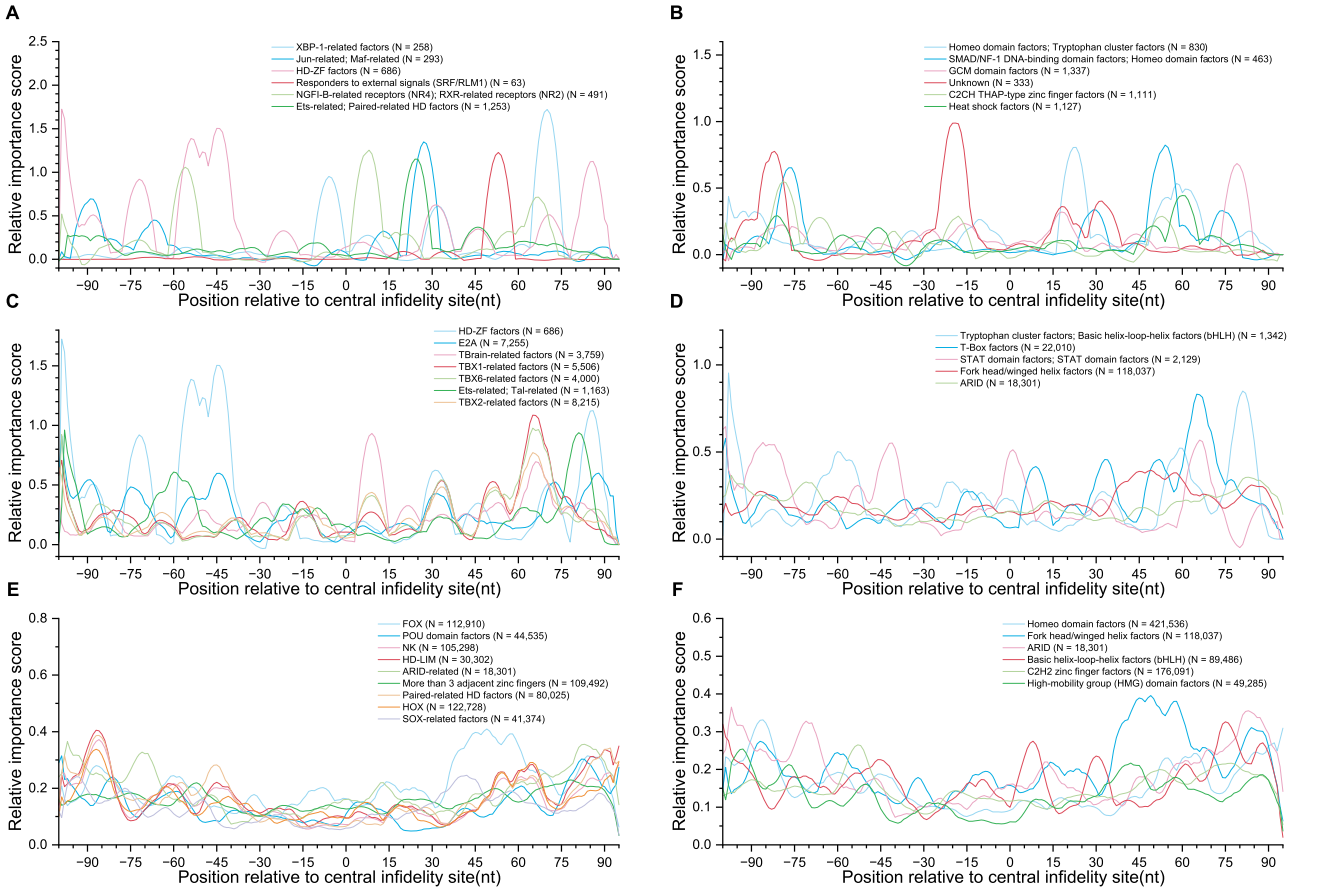

Figure 3. Analysis of the contribution of position-specific transcription factor families and classes to transcriptional infidelity sites. In all legends, the final "N" represents the total number of binding sites for the transcription factors within each family or class. (A and B) respectively show the top six categories within transcription factor families and classes that have the highest relative importance scores at specific individual sites. (C and D) respectively display the top seven families and top five classes based on the cumulative relative importance scores across all sites. (E and F) respectively illustrate the families and classes that exhibit positive importance scores across the entire distribution of sites.

trinucleotides whose rankings significantly increased in transcriptional infidelity sequences might be closely related to the occurrence of transcriptional infidelity.

Our deep learning framework and its interpretation scheme offer a deeper perspective for understanding the potential causes of transcriptional infidelity, and the interpretation results align closely with the findings from the basic sequence feature analysis. For example, the importance of dinucleotides TT, AA, AT, GT, and trinucleotides TGT, ATT, and AAT corresponds with the core recognition motifs of some key transcription factor families, such as AGGTGTGA (associated with the TBrain-related factors, TBX6-related factors, TBX1-related factors, and TBX2-related factors families) and [C/T]AATTA (associated with the NK, Paired-related HD factors, HOX, and HD-LIM families). Moreover, the interpretation results also revealed scenarios that contrasted with the trends observed in the basic sequence features. For instance, although the proportions of dinucleotides CA, TG, CT, CC, and the trinucleotide CTG decreased in transcriptional infidelity sequences, the interpretation results indicated that the motif CACCTG (a core motif for the HD-ZF factors and E2A families) played a significant role in transcriptional infidelity. Overall, the changes in basic sequence features further validate the reliability of the interpretation results. Additionally, these interpretative analyses can uncover specific phenomena that are challenging to capture through basic sequence feature analysis, providing a more comprehensive perspective for understanding the mechanisms underlying transcriptional infidelity.

Discussion

In this study, we proposed a deep learning framework, TTIRI, designed to predict transcriptional infidelity regions within the allopolyploid line by goldfish and common carp hybrids. To the best of our knowledge, TTIRI is the first computational model to utilize genomic sequence information to locate regions of transcriptional infidelity. To accurately identify these regions, we employed a two-layer prediction strategy: an initial prediction at the sequence-level followed by a refined prediction for subregions. However, during the subregion prediction process, the challenge of data class imbalance could not be adequately addressed by common techniques of equal downsampling. Therefore, we introduced a weighted loss function to mitigate the negative impact of class imbalance on model performance. Through experimental validation, we identified the optimal balance between sequence length and subregion design, demonstrating that TTIRI can accurately identify transcriptional infidelity regions at both the sequence and subregion levels. To gain a deeper understanding of how the model captures transcriptional infidelity regions based on sequence information, we designed a systematic quantitative analysis scheme. Initially, we perturbed

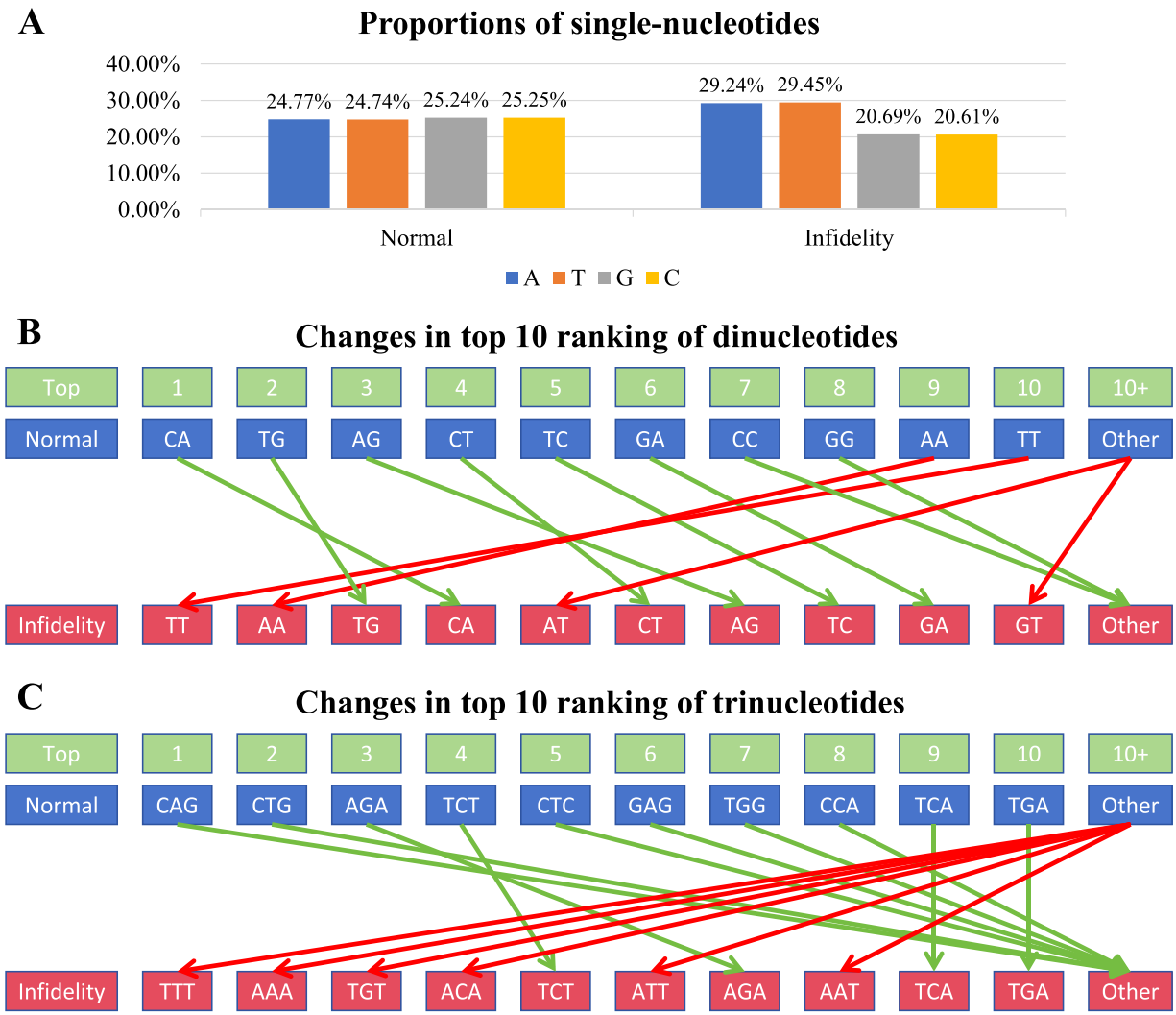

Figure 4. Analysis of basic sequence features for normal sequences and sequences with transcriptional infidelity used in training and interpretation analyses. (A) Proportions of single-nucleotides in normal and transcriptional infidelity sequences. (B and C) Changes in the top 10 ranking of dinucleotides and trinucleotides by frequency proportion, where green arrows indicate a decrease or no change in ranking from normal to transcriptional infidelity sequences, and red arrows indicate an increase in ranking, which may suggest a close association between these dinucleotides or trinucleotides and the occurrence of transcriptional infidelity.

the position-specific hexamers at transcriptional infidelity sites and quantitatively analyzed their contributions to the predictions. However, due to the existence of as many as 4096 types of hexamers, directly analyzing the contribution of all hexamers proved to be both complex and inefficient. To enhance the quality and efficiency of our analysis, we introduced transcription factor recognition motifs, utilizing motif screening to narrow down the scope of analysis. We then further classified and analyzed the predictive contributions according to transcription factor families and classes. Notably, the interpretive analysis revealed that TTIRI successfully captured key transcription factor recognition motifs, such as the core motif CACCTG of the HD-ZF factors and the E2A families, along with their position-specific contribution characteristics. We further validated the reliability and insights of the interpretation results by examining the consistency between changes in basic sequence features and the interpretation outcomes. These findings indicate that TTIRI not only effectively predicts transcriptional infidelity regions in the allopolyploid line by goldfish and common carp hybrids, but its predictive results and interpretive analysis also provide new perspectives and understanding for a deeper exploration of the transcriptional infidelity phenomenon.

Despite the satisfactory performance of TTIRI, there is still room for further optimization. Currently, TTIRI uses only DNA context as input, which might limit the model's ability to accurately characterize species genomes and their differentiated cellular states in the absence of genomic annotations, such as genomic regions, structures, and epigenetic information. For instance, studies have shown that integrating DNA structural information—such as minor groove width, propeller twist at base-pair resolution, roll, and helix twist—can significantly enhance model performance in DNA double-strand break prediction and reveal the critical role of DNA structure in lesion recognition [30]. Similarly, in CRISPR-induced double-strand break prediction, the inclusion of epigenetic features like CTCF, Dnase, H3K4me3, and RRBS as additional features has markedly improved predictive power, with ablation experiments further highlighting the importance of these chromatin status markers in the CRISPR-induced cleavage process [31, 32]. Moreover, TTIRI still has certain limitations regarding the granularity of subregion

prediction, potentially stemming from the constraints of the model design and the influence of data class imbalance in subregions. Future versions of TTIRI will attempt to integrate more genomic information and refine the model architecture to achieve more precise predictions, ideally reaching single-nucleotide resolution, thereby enabling a deeper understanding of the regulatory mechanisms underlying transcriptional infidelity. Although there is still room for improvement in the current study, our findings strongly support the understanding of transcriptional infidelity patterns within the allopolyploid line by goldfish and common carp hybrids, showcasing the potential application value of this tool in related research. Future work will focus on enhancing TTIRI's predictive accuracy and interpretative capabilities, laying the foundation for an in-depth exploration of the molecular mechanisms underlying transcriptional infidelity phenomena.

Conclusion

In this study, we present the first application of deep learning at nucleotide resolution to uncover patterns of transcriptional infidelity in the allopolyploid line by goldfish and common carp hybrids. We developed a two-layer computational framework (TTIRI model) that accurately identifies regions of transcriptional infidelity and devised an accompanying interpretability scheme to systematically quantify the positional contributions of hexamer motifs, transcription factor families, and classes to model predictions. Our findings provide novel insights into the pattern features and molecular mechanisms of transcriptional infidelity at nucleotide resolution, and offer valuable leads for dissecting the deeper factors driving this phenomenon. Moreover, this work contributes important understanding toward the molecular basis underlying the rarity, reduced viability, and limited reproduction of polyploidized vertebrates, thereby establishing a robust theoretical foundation for future research on vertebrate polyploidization.

Key Points

- First application of deep learning at nucleotide resolution to uncover pattern features of transcriptional infidelity.
- Development and implementation of a two-layer computational framework (TTIRI model) that accurately identifies regions of transcriptional infidelity.
- Establishment of an interpretability scheme that systematically quantifies the contributions of position-specific hexamers, transcription factor families, and classes to model predictions, revealing their associations with transcriptional infidelity.
- Provision of novel insights into the pattern features and molecular mechanisms of transcriptional infidelity at nucleotide resolution.
- Offering valuable leads and references for understanding the drivers of transcriptional infidelity and for future research on vertebrate polyploidization.

Supplementary data

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

Conflict of interest: None declared.

Funding

This work was supported by the Applied Basic Research Project in Yunnan Province (202201AT070042), the project funding of the "Support Program of Xingdian Talents", the National Natural Science Foundation of China (61862067, 32293253), Yunnan Fundamental Research Projects (202301BF070001-019, 202301AU070193) and National Key R&D Program of China (2022YFC2602500, 2022YFC260250302).

Data availability

To facilitate replication of this study and further advancement in related research, we have made the data and code used for the models and interpretation methods developed in this work publicly available at https://github.com/KzJing/TTIRI.

References

1. Crick F. Central dogma of molecular biology. *Nature* 1970;**227**:561–3. https://doi.org/10.1038/227561a0
2. Cobb M. 60 years ago, Francis crick changed the logic of biology. *PLoS Biol* 2017;**15**:e2003243. https://doi.org/10.1371/journal.pbio.2003243
3. Bar-Yaacov D, Avital G, Levin L. *et al.* RNA–DNA differences in human mitochondria restore ancestral form of 16s ribosomal RNA. *Genome Res* 2013;**23**:1789–96. https://doi.org/10.1101/gr.161265.113
4. Li M, Wang IX, Li Y. *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. *science* 2011;**333**:53–8. https://doi.org/10.1126/science.1207018
5. Piechotta M, Wyler E, Ohler U. *et al.* Jacusa: site-specific identification of rna editing events from replicate sequencing data. *BMC Bioinf* 2017;**18**:1–15.
6. Wang IX, Core LJ, Kwak H. *et al.* RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase ii. *Cell Rep* 2014;**6**:906–15. https://doi.org/10.1016/j.celrep.2014.01.037
7. Wang IX, Grunseich C, Chung YG. *et al.* RNA–DNA sequence differences in saccharomyces cerevisiae. *Genome Res* 2016;**26**:1544–54. https://doi.org/10.1101/gr.207878.116
8. Verheijen BM, Van Leeuwen FW. Commentary: the landscape of transcription errors in eukaryotic cells. *Front Genet* 2017;**8**:219. https://doi.org/10.3389/fgene.2017.00219
9. Luo J, Chai J, Wen Y. *et al.* From asymmetrical to balanced genomic diversification during rediploidization: subgenomic evolution in allotetraploid fish. *Sci Adv* 2020;**6**:eaaz7677.
10. Liu S, Luo J, Chai J. *et al.* Genomic incompatibilities in the diploid and tetraploid offspring of the goldfish× common carp cross. *Proc Natl Acad Sci* 2016;**113**:1327–32. https://doi.org/10.1073/pnas.1512955113
11. Fresard L, Leroux S, Roux P-F. *et al.* Genome-wide characterization of rna editing in chicken embryos reveals common features among vertebrates. *PLoS One* 2015;**10**:e0126776. https://doi.org/10.1371/journal.pone.0126776
12. Lu B. *The phenomenon of RNA and DNA sequence differences at transcriptome in the nascent polyploidization of a polyploidy line by goldfish × common carp*. Master's thesis. Kunming, China: Yunnan University, 2016. Chinese.
13. Guo Y, Hui Y, Samuels DC. *et al.* Single-nucleotide variants in human RNA: RNA editing and beyond. *Brief Funct Genomics* 2019;**18**:30–9. https://doi.org/10.1093/bfgp/ely032
14. Ramaswami G, Li JB. Identification of human RNA editing sites: a historical perspective. *Methods* 2016;**107**:42–7.

15. Sijia W, Fan Z, Kim P. *et al.* The integrative studies on the functional a-to-i rna editing events in human cancers. *Genomics Proteomics Bioinf* 2023;**21**:619–31. https://doi.org/10.1016/j.gpb.2022.12.010
16. Li Z, Gao E, Zhou J. *et al.* Applications of deep learning in understanding gene regulation. *Cell Reports Methods* 2023;**3**:100384. https://doi.org/10.1016/j.crmeth.2022.100384
17. Ji Y, Zhou Z, Liu H. *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20. https://doi.org/10.1093/bioinformatics/btab083
18. Stroup EK, Ji Z. Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat Commun* 2023;**14**:7378. https://doi.org/10.1038/s41467-023-43266-3
19. Sheng X, Wei J, Sun S. *et al.* Ssblazer: a genome-wide nucleotide-resolution model for predicting single-strand break sites. *Genome Biol* 2024;**25**:46.
20. Zhou Z, Ji Y, Li W. *et al.* DNABERT-2: efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006. 2023.
21. Devlin J, Chang MW, Lee K. *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proc 2019 Conf North American Chapter Assoc Comput Linguistics: Human Language Technologies, Vol 1 (Long and Short Papers)*. 4171–86. Minneapolis, MN, USA: Association for Computational Linguistics, 2019.
22. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.), *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA: Curran Associates Inc., 2017, 5998–6008.
23. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025;**22**:287–97. https://doi.org/10.1038/s41592-024-02523-z
24. Nguyen E, Poli M, Faizi M, *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S (eds.), *37th Conference on Neural Information Processing Systems 37 (NeurIPS 2023)*. 43177–201. New Orleans, LA, USA: Curran Associates Inc., 2023.
25. Schiff Y, Kao C-H, Gokaslan A. *et al.* Caduceus: Bi-directional equivariant long-range DNA sequence modeling. arXiv preprint arXiv:2403.03234. 2024.
26. Kudo T. Sentencepiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226. 2018.
27. Sennrich R. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909. 2015.
28. Press O, Smith NA, Lewis M. Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409. 2021.
29. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R. *et al.* Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2024;**52**:D174–82. https://doi.org/10.1093/nar/gkad1059
30. Mourad R, Ginalski K, Legube G. *et al.* Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome Biol* 2018;**19**:1–14. https://doi.org/10.1186/s13059-018-1411-7
31. Chuai G, Ma H, Yan J. *et al.* Deepcrispr: optimized crispr guide rna design by deep learning. *Genome Biol* 2018;**19**:1–18.
32. Lin J, Zhang Z, Zhang S. *et al.* Crispr-net: a recurrent convolutional network quantifies crispr off-target activities with mismatches and indels. *Advanced science* 2020;**7**:1903562.

Briefings in Bioinformatics, 2025, **26(3)**, bbaf260
https://doi.org/10.1093/bib/bbaf260
Case Study