A thesis submitted to the University of East Anglia
for the degree of Doctor of Philosophy

Investigate the functional importance of RNA secondary structure with computational biology approaches

Bibo Yang

September 2025

John Innes Centre, Norwich, United Kingdom

Acknowledgements

I want to thank many people for their support, help, and advice in the past four years. First and foremost, I would like to thank my brilliant primary supervisor Prof. Yiliang Ding for giving me the opportunity to get into the world of RNA biology and teaching me how to conduct scientific research. We are not only like supervisor and student but like friend and family as well. I also want to thank my supervision team, Prof. Richard Morris, my secondary supervisor, Dr. Haopeng Yu, my mentor, for their important guidance and helpful discussions. I also want to sincerely thank all the collaborators and other advisors, Prof. Caroline Dean, Prof. Zoë Waller, Prof. Huakun Zhang, Prof. Ke Li for their kind suggestions, guidance, and supports. I also want to show my appreciation to all the contributors of the projects in the thesis, Haopeng, Yueying, Jie, Xiaofei, Zongyun, Fan, Dilek, Haidan, Elisé, Yiman, Wenqian, Yuchen, Wenqing. Thank you so much for the help and support!

Next, I would like to thank every single member of the Ding and Dean groups. It is a wonderful team. I want to thank: Haopeng, Yueying, Xiaofei, Zhen, Zongyun, Susan, Jin, Maria, Wenqing, Jie, Ling, Yiman, Ran, Mingming, Qianqian, Shaoli, Jianfang, Jianhua, Haifeng, Miguel, Anna, Govind, Shuqin, Geng Jen. Thank you so much!

I also want to thank my friends. Thanks to friends in JIC: Jie Li, Shouchao, Honghao, Zhenglin, and Xuanjie. I also want to appreciate my old friends who gave me huge support in my hard period: Sissi, thank you for the support. Yunzhen, we had a wonderful trip in Turkey, Judy and Yanshen, we spent a wonderful Christmas in London, Yuanfei, he always shares new experiences, knowledge, thoughts to me, Yu Hu, Huiliang, Jiaqi, Chenjie as well. We have known each other for over ten years, and I am lucky to have you with me.

The biggest love and appreciation to my parents and my grandparents in the heaven. Thanks a lot for your love and support! I love you so much.

Abstract

RNA structures are fundamental regulators of diverse biological processes, yet the identification of functional structures remains a major challenge. Advances in deep sequencing–based probing have generated large-scale, high-precision datasets, reshaping the field and providing a strong foundation for computational approaches to discover the functional importance of RNA structures. This thesis presents three computational biology studies addressing RNA structural features and their functional roles from two complementary perspectives: structure motifs and structure dynamics.

The first project focused on i-motifs (iMs), a known cytosine-rich structural motif. A machine learning framework, iM-Seeker, was developed using iM-specific datasets to identify putative iM-forming sequences in DNA and RNA, predict folding status, and estimate folding stability. Application across over 400 plant transcriptomes revealed strong enrichment of RNA iMs in 5′UTRs, particularly in monocots, with positive correlation to environmental temperature. Translatome analyses indicated that 5′UTR iMs act as translational repressors, independent of folding stability.

The second project investigated RNA structure motifs underlying RNA stability in wheat. RNA structure was identified as the dominant determinant of stability. The functional structure discovery pipeline pyTEISER (python-implemented Tool for Eliciting Informative Structural Elements in RNA) was optimized by expanding motif seeds and improving computational efficiency, enabling the identification of stability-associated motifs. Subgenome-specific preferences highlighted the regulatory role of structural motifs in polyploid plants.

The third project addressed RNA structural dynamics during tomato fruit development. High-resolution structuromes were generated at two developmental stages, and RSDE (RNA Structural Dynamic Elements)-Tool was introduced to integrate ensemble-based structure analysis, statistics, and machine learning. Over 12,000 RSDEs were identified, predominantly in 5′UTRs and CDS regions, many significantly associated with translational efficiency changes in tomato development.

Collectively, these studies establish computational strategies to investigate crucial RNA structures, highlighting their diverse and critical regulatory roles in biological systems.

List of contents

List of Abbreviations

AI – artificial intelligence

HTS – high-throughput sequencing

NAI – methylnicotinic acid imidazolide

2A3 – 2-aminopyridine-3-carboxylic acid imidazolide

DMS – dimethyl sulphate

lncRNA – long non-coding RNA

GQS – G-quadruplex

G4 – G-quadruplex

RG4 – RNA G-quadruplex

TE – translation efficiency

SCFG – stochastic context-free grammar

CFG – context-free grammar

CLLM – conditional log-linear model

FM – foundation model

PDB – Protein Data Bank

TEISER – Tool for Eliciting Informative Structural Elements in RNA

RBP – RNA Binding Protein

MSA – Multiple Sequence Alignment

iM – i-Motif (intercalated-motif)

bp – base pair

nt – nucleotide

AutoML – automated machine learning

pH_T – transitional pH

XGBoost – Extreme Gradient Boosting

AUROC – Area Under the Receiver Operating Characteristic Curve

PCC – Pearson correlation coefficient

SPCC – Spearman correlation coefficient

MI – mutual information

BPP – base-pairing probability

cAI – codon adaptation index

tAI – tRNA adaptation index

ORF – Open Reading Frame

MFE – minimum free energy

RSDE – RNA structural dynamic element

t-SNE – t-distributed stochastic neighbour embedding

UTR – untranslated region

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Supplementary figures:

List of Tables

Chapter 1 Introduction

RNAs as one of the most vital essential molecules play crucial roles in gene expression regulations. In the central dogma, RNA contains the nucleotide sequences and transfers the genetic information from DNA to protein. RNAs carry genetic information and form complex secondary and tertiary structures. RNAs consist of four bases: adenine (A), uracil (U), cytosine (C), and guanine (G). Three kinds of canonical base pairs (A-U, C-G, G-U) and other uncanonical base pairs (e.g. C-C, G-G, and so forth) can be formed via the hydrogen bonds and other chemical interactions (Danaee, et al., 2018; Gautheret, et al., 1995). The canonical base pairs allow the RNA molecules to fold into complex secondary structures consisting of stems, hairpin loops, bulges, internal loops, external loops, and multiloops (Danaee, et al., 2018) (**Figure 1.1**). These secondary structures further form into more complex tertiary structures such as tRNA structure, G-quadruplex etc. RNA structure plays important roles in many functionalities, including transcription, RNA maturation, translation, and RNA degradation (Cao, et al., 2024). In the past several decades, biophysical approaches such as X-ray, nucleus magnetic resonance (NMR) and cryo-electron microscopy have already been used to investigate RNA structures. However, these methods can not reveal the RNA structures in cellular environments, and the low-throughout nature of these methods also limits the exploration of diverse RNA structures. Recently, with the advances of next-generation sequencing platforms, revolutionary high-throughput *in vivo* RNA structure profiling methods have been developed, revealing new insights into RNA structure functionality. These methods generated a huge amount of high-quality structurome data, which triggers the integrating of new data and classical computational concepts and

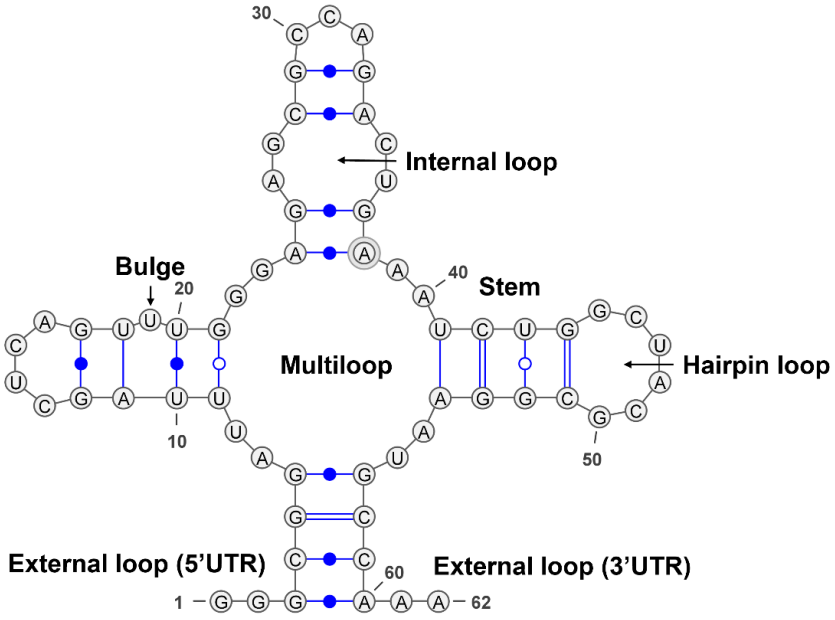


Figure 1. 1 Schematic of RNA secondary structures.

designing novel algorithms with interdisciplinary technologies from other fields, including applied mathematics, computer science, and artificial intelligence (AI). The mutual reinforcement of the wet and dry technologies has dramatically advanced the development of RNA biology from a structural perspective, transforming RNA biology into a new era.

1.1 Recent advances in RNA structure profiling methods

With the rise of high-throughput sequencing (HTS), the integration of RNA structure probing and HTS can be used to analyse tens of thousands of RNAs in one single experiment. The RNA structure probing methods can be divided into enzymatic probing, chemical probing, and proximity ligation. In 2010, Kertesz et al. developed a genome-wide *in vitro* RNA secondary structure measurement strategy, parallel analysis of RNA structure (PARS), and applied the method to yeast (Kertesz, et al., 2010). This enzymatic probing uses several ribonucleases (RNases), which cleave double-strand sites or single-strand sites. Besides PARS, other enzymatic methods such as FragSeq (Underwood, et al., 2010), dsRNA-seq (Decker, et al., 2019), and ssRNA-seq (Parkhomchuk, et al., 2009) have been developed for various species. However, all these enzymatic methods have a common limitation: they are only for *in vitro* analyses because the RNases cannot penetrate the cells. Nowadays, chemical probing-based RNA structure profiling methods and proximity ligation are more commonly used. The proximity ligation is the type of the crosslink-based methods for studying RNA structures. RNAs have RNA both inter- and intramolecular interactions, which can be captured by cross-linking and ligation followed by deep sequencing of the ligated chimaeras (Aw, et al., 2016; Cao, et al., 2024; Lu, et al., 2016). After mapping the chimaeras on references, the interacted sequences can be identified, which can be treated as the straightforward pairing constraints to guide the RNA structure prediction model (Aw, et al., 2016; Cao, et al., 2024; Lu, et al., 2016). The disadvantages of these methods are the low ligation efficiency and high noise backgrounds.

The chemical probing-based methods are the other type of RNA structure profiling methods. The chemical probing genome-wide RNA structure profiling approaches do not have the *in vitro* limitation because the chemical probes are more miniature so that they can be used in cells. The chemicals are dimethyl sulphate (DMS) and Selective 2'-Hydroxyl Acylation analysed by Primer Extension (SHAPE). DMS was originally only used to modify Watson-Crick faces of unpaired adenine (A) and cytosine (C). More recent improvement on the buffer condition allows the modifications of all four bases (Mitchell III, et al., 2023). SHAPE (i.e., NAI and 2-aminopyridine-3-carboxylic acid imidazolide, abbreviated as 2A3, are

widely used) can probe the 2′-OH group of unpaired four kinds of nucleotides (Manfredonia, et al., 2020; Marinus, et al., 2021; Spitale, et al., 2015). After the chemical treatment of living cells, the chemical probes modify the unpaired nucleotides, which can lead to the reverse transcription (RT) stop or introducing nucleotide mutations (RT-mutation) in the reverse transcription step. Next, the truncated cDNAs or cDNAs with mutational profiles are subjected to the high-throughput sequencing. Following the HTS, the corresponding computational pipeline processes the RT-stop or RT-mutation reads into RNA structure information called chemical reactivity. The chemical reactivity can be further used to estimate the probability of all the nucleotides in paired status or unpaired status. Higher chemical reactivities indicate the high probabilities of the corresponding nucleotides as single-stranded. To estimate the RT random stop or mutation, the experiments also set up the control group (non-chemical treatment) to measure the background noise.

The first two genome-wide *in vivo* secondary structure profiling methods based on DMS and RT-stop strategy were DMS-seq (Rouskin, et al., 2014) and Structure-seq (Ding, et al., 2014) in yeast and *Arabidopsis thaliana*, respectively. In 2015, a genome-wide SHAPE-based method, icSHAPE, utilised NAI (methylnicotinic acid imidazolide)-N_3 as the chemical probe was developed for the mouse (Spitale, et al., 2015). After that, many modified chemical methods based on either RT-stop or RT-mutation were developed for more specific biological questions and uncovered many basic rules across multiple species including HIV-1 virus, SARS-CoV-2 virus, rice, wheat, and human (Bohn, et al., 2023; Deng, et al., 2018; Manfredonia, et al., 2020; Siegfried, et al., 2014; Sun, et al., 2019; Yang, et al., 2021). For example, icSHAPE was applied to chromatin, nuclear and cytosolic RNA populations to study the structure heterogeneity in subcellular locations (Sun, et al., 2019). CAP-STRUCTURE-seq (Yang, et al., 2020) for capturing intact mRNA structuromes was developed to study the role of RNA structure on miRNA-mediated RNA degradation. Nuc-SHAPE-Structure-Seq (Liu, et al., 2021) for capturing pre-mRNA structuromes was used to study the functional role of RNA structure in RNA processing. SPET-seq (Incarnato, et al., 2017) and tNET-Structure-seq (Saldi, et al., 2021) for capturing the structural signals in nascent RNAs was developed to study the RNA structural dynamics during the co-transcriptional process in *E. coli* and human, respectively. Additionally, RNA G-quadruplex (RG4), a specific tertiary structure motif with the G-rich sequence content, is associated with various biological functions (Cao, et al., 2024; Yang, et al., 2020). To identify the folding status of RG4s across transcriptomes, the *in vitro* method rG4-seq (Kwok, et al., 2016) and *in vivo* method SHALiPE-Seq (Yang, et al., 2020) were designed to detect RG4s at the

transcriptome-wide scale. Furthermore, more methods were developed to advance the current methods and extend their applications using new sequencing platforms. Most methods were based on the short-read sequencing platform. These sequencing reads are ~150 nt long, that cannot distinguish structures from different RNA isoforms. The long-read sequencing technologies (e.g. PacBio and Nanopore sequencing) with chemical probing can overcome the limitations caused by the short-read methods. smStructure-seq (Yang, et al., 2022), Nano-DMS-MaP (Bohn, et al., 2023), and DMS-FIRST-seq (Begik, et al., 2025) are the successful applications of long-read sequencing to dissect RNA structure heterogeneity on PacBio and Nanopore sequencing, respectively. Other RNA structure profiling approaches combined with RNA direct Nanopore sequencing can directly determine the chemical modifications on RNAs without reverse transcription (Aw, et al., 2021; Bizuayehu, et al., 2022). In addition, Single cell sequencing is one of the most revolutionary technologies in the past decade. Structure profiling has also been integrated into single-cell sequencing to study the structural heterogeneity at single-cell resolution (Wang, et al., 2024). The fast development of the chemical probing-based methods has opened the door to getting high-quality structure information in different biological contexts, facilitating the investigation of RNA structure functionality.

1.2 The functional roles of RNA structure in plants

In Eukaryotes, the nascent mRNAs need a series of processes to be mature mRNAs, including 5′ capping, 3′ polyadenylation and RNA splicing co-transcriptionally or post-transcriptionally. The mature mRNA will then be exported to the cytosol and translated into proteins. Following the translation, mRNA will undergo the degradation process. Emerging evidence have showed that RNA structures influence every step of RNA life cycles.

One key stage where structure plays a crucial role is splicing. Splicing is a crucial step in deriving different RNA isoforms that can lead to different protein products. Splicing is a complex process involving the removing introns and ligation of exons, by interaction between pre-mRNA and the spliceosome which contain various proteins and snRNAs (Wan, et al., 2020). In this process, three conserved intron elements contribute to the spliceosome recognition of pre-mRNA in different stages. The first element is the 6-nucleotide sequence immediately downstream of the 5′ splice site (5′ SS). The canonical 5′ SS consensus is **GU**RAGU, and the first two nucleotides GU are significantly conserved in splicing among species (Kastner, et al., 2019). The second element is branch point sequence (BPS), a 7-

nucleotide segment located in the central part of the intron near the 3′ SS with YNCUR**A**C in plants as a typical consensus element and the bold A as the branch point adenosine site. The third element segment is the 3-nucleotide sequence Y**AG,** immediately upstream of the 3′ splice site (3′ SS) with bold AG as critical positions. These conserved sequence content of these three elements depends on their complementarity to the sequence content of the snRNA binding motif (e.g. the 5′ splice site complements the sequence of U1 snRNA). As another important feature besides sequence content, RNA structure also plays an essential role in splicing. *In vivo* RNA structure profiling in *Arabidopsis thaliana* revealed that several structure features are associated with splicing: (1) An unpaired, single-stranded state in the two-nucleotide region upstream of the 5′ SS is required for efficient splicing; (2) The single-strandedness of the branch point site is associated with splicing and the unpaired status of the branch site influences the 3′ splice site recognition; (3) The four-nucleotide region upstream of AG motif at the 3′ SS tends to be double-stranded in both spliced and unspliced transcripts, suggesting that there might be a potential structure motif functioning between the branch point site and the 3′ SS (Liu, et al., 2021). In the same study, a novel structure feature impacting polyadenylation and alternative polyadenylation was discovered: Two single-stranded regions (from -28 to -17 nt upstream of the poly(A) site and from -4 to +1 nt across the poly(A) site) contribute to the recognition of polyadenylation. Sequence analysis of the upstream region (-28 to -17) showed enrichment of the plant-preferred polyadenylation signal (PAS) motif such as UGUA, UAUA or occasionally AAUAAA, with AAAUAAA being more characteristic of and conserved in mammals (Loke, et al., 2005). However, the identified structure features did not show a strong correlation with their underlying sequence content and are likely to interact with RNA binding proteins.

Beyond splicing, RNA structures also critically impact translation. Translation is a crucial and fundamental process in life. RNA structure might influence the interaction between the ribosome and mRNA from ribosome assembly loading and scanning to elongation. In general, single-stranded regions or certain specific structures are often more favourable for protein binding. However, this relationship may vary depending on the biological context, as some stable structures might block ribosome assembly loading and scanning (Leppek, et al., 2018). Previous results showed that a single-stranded region immediately upstream of the start codon positively impacts translation efficiency. In contrast, high translation efficiency (TE) was associated with mRNAs exhibiting greater structural flexibility (such as single-strandedness) and a triplet-periodic structural pattern across CDS regions in wheat (Yang, et al., 2021). Additionally, some RNA structure motifs in the 5′UTR also influence

translation. G-quadruplex in UTRs region generally inhibit translation (Kwok, et al., 2015; Yang, et al., 2020). Furthermore, AI-driven analysis identified enrichment of high TE RNA secondary structure motifs formed by AGCU-repeat stems and low TE motifs formed by GC-rich stems, suggesting that very stable structures might hinder translation during elongation, while moderately stable structures might facilitate ribosomal progression (Yu, et al., 2024).

Extending beyond its influences on splicing and translation, RNA structure can also affect RNA stability in plants. Previously, RT-stop structure profiling revealed that strong RNA structures (e.g. GQS) in the 3′UTR can stabilize RNAs (Su, et al., 2018; Yang, et al., 2022), while another RT-based mutation reported apparently conflicting regulatory roles of RNA structures in *Arabidopsis thaliana* (Zhang, et al., 2024). The potential shielding of single-stranded regions from exonucleases might explain the stability-enhancing role of strong structures (Chlebowski, et al., 2013). However, the single-strandedness of RNA might influence other RNA processing (e.g. RNA polyadenylation which primarily promotes degradation) in plants. Thus, it is still not entirely clear how RNA structure comprehensively influences RNA stability. RNA structure can also regulate RNA stability by influencing microRNA (miRNA)-mediated cleavage. miRNAs are 20-24 nt non-coding RNAs involved in post-transcriptional regulation, including transcript cleavage and translation repression (Yu, et al., 2017). miRNA-mediated cleavage is involved in various cellular processes. miRNAs are incorporated into ARGONAUTE proteins (AGO); the resulting miRISC cleaves complementary RNA targets guided by perfect sequence complementarity between the miRNA and its binding site on target RNAs (Yang, et al., 2020; Yu, et al., 2017). Therefore, the sequence context around miRNA binding sites may contain signals for efficient miRNA-mediated cleavage. Previous studies suggested that local RNA structure accessibility impacts silencing efficiency (Kertesz, et al., 2007). Disrupting secondary structures can reduce target accessibility (Kertesz, et al., 2007). Supporting evidence includes the discovery of the Target-adjacent Nucleotide Motif (TAM) by Yang et al (2020) a single-stranded dinucleotide region adjacent to the cleavage site that promotes efficient cleavage. However, the comprehensive role of RNA structure in miRNA-mediated cleavage remains incompletely understood.

RNA structures also play vital roles in plant growth, development, and stress responses (e.g. temperature, light, and so forth) (Zhang and Ding, 2025). In plant development, flowering is one of the most important stages. RNA structures can regulate flowering time regulation by

modulating expression of *FLC* (*FLOWERING LOCUS C*), which encodes the key transcriptional repressor of flowering. During vernalization, prolonged cold exposure promotes flowering through epigenetic repression of *FLC*. This repression is mediated by an antisense long non-coding RNA (lncRNA) *COOLAIR*, transcribed antisense to *FLC*, which recruits chromatin modifiers to suppress *FLC* transcription. Structural flexibility of the *COOLAIR* II.i isoform facilitates chromatin binding (Yang, et al., 2022).

The plant vascular system enables long-range transport of molecules and nutrient, essential for sessile organisms. Phloem-mobile RNAs act as long-distance signals to synchronize physiological processes (Spiegelman, et al., 2013; Thieme, et al., 2015; Zhang, et al., 2016). Specific RNA structural features contribute to long-distance mRNA mobility: Transfer RNA-like structures (TLS) serve as phloem loading signals; Structure motifs within mRNA untranslated regions (UTRs) (Zhang, et al., 2016); Stem-loop structures in pri-miRNA transcripts also associate with systemic transport (Wang, et al., 2021). G-quadruplexes, a well-characterized RNA tertiary structure motif with high stability, confirms native existence in *Arabidopsis* by *in vivo* probing (Yang, et al., 2020). Plant RNAs containing G4 structures regulate: vascular tissue specification (Cho, et al., 2018); lateral root initiation (Yang, et al., 2020; Zhang, et al., 2019) and Cold-stress responsive transcription (Yang, et al., 2022).

1.3 Current computational approaches to discover functional RNA structure

Computational approaches play crucial roles in RNA structural investigation. Before the emergence of *in vivo* RNA structural profiling methods, computational pipelines enabled diverse applications. The core objectives include RNA structure prediction (the foundation of structural studies), followed by structural characterization. A given transcript sequence can adopt multiple structural conformations. Functionally significant RNA motifs (2~200 nt) often drive biological mechanisms more critically than whole-transcript folds (see Chapter 1.2). Thus, motif discovery represents another key computational challenge. An ensemble-based structural analysis provides enhanced insights into structure-function relationships, particularly when integrated with chemical profiling constraints. Based on this framework, further exploration covers: (1) RNA secondary structure prediction and characterizations and analysis; (2) structure motif detection; (3) modelling of RNA structural dynamics.

1.3.1 RNA secondary structure prediction and characterization

RNA secondary prediction tools can be divided into three categories: thermodynamics-based models, statistic-based models, and deep-learning-based models. Thermodynamic methods are classical and widely used software like RNAstructure (Reuter and Mathews, 2010), RNAfold (Hofacker, 2003), and RNAFramework (Incarnato, et al., 2018) employ dynamic programming with empirical nearest-neighbour parameters (e.g. Turner rules) to predict minimum free energy structures, typically the most stable folding. Modern approaches integrate chemical probing reactivity data as folding constraints to improve accuracy (Deigan, et al., 2009; Low and Weeks, 2010).

Statistical methods such as Monte Carlo simulations (e.g. MC-Fold) (Parisien and Major, 2008), stochastic context-free grammars (SCFGs, e.g. Pfold) (Knudsen and Hein, 2003), and conditional log-linear models (e.g. CONTRAfold) (Do, et al., 2006) also enable structure prediction. SCFGs model the joint probability of sequences and structures from training data, requiring no thermodynamic parameters. They also underlie algorithms for detecting conserved RNA motifs (Chapter 1.3.2).

Advance in deep learning have propelled RNA secondary structure prediction. Models like SPOT-RNA (Singh, et al., 2019), MXfold2 (Sato, et al., 2021) and UFold (Fu, et al., 2022) demonstrate excellent performance. For example, SPOT-RNA trained a deep learning model with Residual Networks and long short term memory network (LSTM) on the bpRNA database followed by fine-tuning on RNA sequences from Protein Data Bank (PDB) and achieved the good performance (Singh, et al., 2019). MXfold2 integrated both deep learning and thermodynamic parameters and achieved better performance than other traditional methods. Additionally, the burst of large models based on scaling law also promotes RNA structure studies (Sato, et al., 2021). The RNA foundation models (FMs) usually have pre-training and fine-tuning stages. In pre-training stage, a huge dataset of biological information (e.g. transcriptomes of large scale of species, multi-omics dataset, and so forth) is fed to the model with multiple pre-training strategies to enable model to capture features via learning through diverse layers. The established pre-trained model can be applied to various downstream tasks in diverse species and biological purposes. These models can be fine-tuned on the datasets with different scales and species. A few of RNA foundation models, including RibonanzaNet (He, et al., 2024), RNA-FM (Chen, et al., 2022) and PlantRNA-FM (Yu, et al., 2024) have shown the capability in the RNA secondary structure prediction tasks

with outstanding performances on benchmark of standard dataset (e.g. bpRNA data resource) than traditional methods such as RNAFold (Hofacker, 2003).

Following the RNA structure prediction, the characterization of RNA structure features is an important process. For example, previous study developed the bpRNA, a graph-based method coupling with RNA secondary structure annotation. bpRNA divides the elements into fundamental categories involving stems, hairpin loops, bulges, internal loops, external loops, multiloops, unpaired regions and pseudoknots. This type of characterization is very comprehensive, serving as the most fundamental RNA secondary structure motif elements (Danaee, et al., 2018). rnaConvert in Forgi package utilized a similar strategy in describing RNA structures divided into their five prime, three prime, stem, hairpin loop, multiloop, and interior loop (Thiel, et al., 2019). Another example is FOREST, which can be used to search either canonical or uncanonical hairpin structures (i.e., hairpins with bulges and internal loops) (Komatsu, et al., 2020). It could also search those complex RNA structures with multi-arm RNA junctions (Komatsu, et al., 2020). These methods mentioned above were developed to provide comprehensive descriptions based on the structural elements (e.g. stem, hairpin and so forth). These methods highly depend on the accuracy of RNA structure prediction models where there is limitation of the folding software leading to certain bias due to the RNA length, sequence contents and experimental constraints. In contrast, both tree model and graph model are two commonly-used strategies in transferring RNA structures into simplified representations. An example of tree model is ordered into the rooted tree with root nodes (fictive nucleotide pairing representing whole structure), leaf nodes (nucleotides), and internal nodes (nucleotide pairing) (Marchand, et al., 2024). The RNA tree representation serves as the base of some structure-based analyses such as the RNA structure prediction (e.g. Pfold) (Knudsen and Hein, 2003), the RNA structure similarity measurement (e.g. RNAforester) (Hochsmann, et al., 2003) and the RNA structure phylogeny analysis (e.g. Infernal) (Nawrocki and Eddy, 2013). For the graph model, the RNA-as-Graphs (RAG) framework was developed to map RNA structures using the graphic representation, where the double-stranded stems ranging from 2 to 9 represented as nodes and the single-stranded regions as edges (Zhu, et al., 2022). Using RNA-as-Graphs, the ‘RNA-Like’ graphic representations were obtained for RNA structural studies (Zhu, et al., 2022). Taken together, diverse methods of the RNA secondary structure annotation can be utilized for characterizing RNA structural features.

1.3.2 RNA functional structure motif detection

RNA secondary structure motifs can be regarded as the basic structural elements for RNA structural annotation. These motifs could be associated with specific biological functions (described in Chapter 1.2). Up-to-date, there are two main strategies for identifying functional structure motifs: (1) Identify RNA structure motifs and associate them to diverse biological functions: The RNA structure motifs with known features (e.g., sequence pattern, conserved structures, structure segments with predefined rule) are identified in either target RNAs or transcriptomes. Then these RNA structure motifs were further associated with diverse biological functions. (2) Train the AI models to detect the sequence and/or RNA structure motifs with specific biological functions to derive the functional RNA structure motifs. For both strategies, different types of data can be incorporated. For instance, the *in vivo* RNA structure probing dataset could be integrated into the identification of RNA structure motifs. For translational measurements, either polysome-associated translatome or ribosome-profiling dataset could be used as the functional measurements.

In the past, the first strategy is widely used. Many efforts were made in defining RNA structure motifs using diverse RNA structural properties. The definition of RNA structure motifs can be roughly divided into three categories: (1) Define RNA structure motifs based on their known sequence features; (2) Define RNA structure motifs based on the conservations of either sequences or structures; (3) Define RNA structure motifs based on its structural diversity and stability.

Certain RNA structure motifs can be searched based on the sequence contents. One example of is the RNA G-quadruplex (GQS) motifs that can be defined by the sequence contents. For instance, Quadparser (Huppert and Balasubramanian, 2005) is the original program that was designed to search putative G-quadruplex (GQS). Although G-quadruplex is a complicated structure motif, the putative sequence formation follows the pattern $(G_{X+}L_{1-N})_{3+}G_{X+}$ where G means guanine and L represents four bases while X is 2 or 3 or more and N can be 1-7. Another specific secondary structure motif, i-motif (iM) with C-rich sequence contents, follows similar formation rules with GQS by the editing from G to C (described in 2.1.2). Apart from those RNA structure motifs with specific sequence contents, RNA secondary motifs can also be defined via the combination of both folding features and sequence contents. A computational pipeline TEISER (Tool for Eliciting Informative Structural Elements in RNA) and its advanced python-implemented version pyTEISER (Fish, et al., 2021; Goodarzi, et al., 2012) utilise context-free grammars (CFGs) and the information theory to define a structural seed library with ~70 million basic RNA hairpin structure

elements regarding both sequences and structures, followed by the associations with specific functions.

The second tactic to define RNA structure motifs is based on the similarity of sequences or structures in searching for the conserved RNA structure motifs across species. This method has the assumption that the essential RNA structural features must be conserved during evolution in serving specific functions. For the models that relied on sequence alignments, the algorithms of stochastic models and alignment-based methods have been widely used (Achar and Sætrom, 2015). One of these models is the covariance model (CM). This model constructs a guide tree containing the information of both sequences and structures with stochastic context-free grammars (SCFGs) from multiple aligned sequences. This model produces a parse tree according to the different statuses of the guide tree. The CM was used in quite a few of software, including tRNAscan-SE (Lowe and Eddy, 1997), RSEARCH (Klein and Eddy, 2003), CMfinder (Yao, et al., 2006), CaCoFold (Rivas, 2020), and Infernal (Nawrocki and Eddy, 2013). The CM was particularly used for finding specific motifs (like tRNAscan-SE for tRNA-like structure searching) and conservative secondary structure in the homologs based on the aligned multiple sequences from species with relatively close phylogenetic relationships (RSEARCH, Infernal and CMfinder). Another phylogeny-based method is R-scape (Rivas, et al., 2017), which integrated various statistical methods to measure pairwise covariations. Graph mining ideas (RNAmine) (Hamada, et al., 2006) and genetic programming (GPRM) (Hu, 2003) were applied to search for the RNA secondary structure motifs based on the sequence similarity. All these methods have a common limitation that the identification of RNA structure motifs is highly dependent on the similarity of aligned sequences from the chosen species. If the sequence similarity is very low, it is not possible to identify RNA structure motifs. In contrast, if the sequence similarity is very high, it is also not possible to distinguish conserved RNA structure motifs from conserved sequences. In addition to these methods relying on the sequence similarity, a new SHAPE-guided homology search and motif discovery algorithm called SHAPEWarp was developed based on the SHAPE-guided structural similarity by direct comparison of the chemical reactivity profiles and addressed the lack of sequence similarity (Morandi, et al., 2022).

The third way for identifying RNA structure motifs is based on the features of RNA structural diversity and stability. Software using this method detects relatively more stable structural elements with less RNA structural diversity based on the assumption that the

important RNA structure motifs tend to be more stable than those folded with randomly shuffled sequences with the same sequence contents. The typical tools (e.g. RNAz and ScanFold) utilize the sliding window strategy to search for the relatively stable motifs followed by multiple statistical measurements in determining the significances of RNA stabilities (Andrews, et al., 2018; Washietl, 2007). Superfold takes additional account of both low chemical reactivity and low per-nucleotide Shannon entropy (Siegfried, et al., 2014). Shannon entropy measures the structural diversity of an RNA sequence by assessing the probability of different base pairings within an ensemble of possible structures (Wang, et al., 2011). RNAvigator integrates both ScanFold and SuperFold (Delli Ponti, et al., 2022). Another example is MEMERIS (multiple expectation maximisation for motif elicitation in RNAs including secondary structures), an algorithm designed for searching for the single-stranded regions as the binding sites of diverse single-stranded RNA binding proteins (Hiller, et al., 2006). The method firstly calculates the probabilities in the Boltzmann ensemble of sub-strings being single-stranded via RNAfold (Hofacker, 2003) and then predicts the single-stranded regions via the expectation maximisation. Notably, some tools built on the similarity of sequences and/or structures also combine the RNA structural sampling and/or thermostability assessments into their algorithms (e.g. CMfinder and RNAz) (Washietl, 2007; Yao, et al., 2006).

After the identification of RNA structure motifs, the next step is to link the motifs with their functionalities (functional constraints). Usually, the functional constraints can be divided into the following categories: (1) Transcriptome-wide one-to-one labels describing the target function (e.g., RNA decay rate, RNA translation efficiency, RNA splicing efficiency, and so forth). One example is pyTEISER (Pythonic Tool for Eliciting Informative Structural Elements in RNAs), a computational framework for identifying RNA structure motifs that govern various functions (Fish, et al., 2021; Goodarzi, et al., 2012). After the generation of the pre-defined, large-scale motif library described above, the discretization strategy of functional measurement labels was taken to divide RNAs with sorted labels into different bins with ranked functional labels. Then the functional RNA structure motifs were selected based on the mutual information. This software showed strong generalizability and was used to successfully identify the functional motifs relative to RNA decay (Goodarzi, et al., 2012) and RNA splicing (Fish, et al., 2021). Currently, pyTEISER is one of two completed pipelines to search for RNA structure motifs related to different functional contexts. The other one is PlanRNA-FM, which will be described later. (2) Functional labels for a group of RNAs containing certain RNA structure motifs (e.g., the climate variables for different

species associated with the density of certain RNA structure motifs). For instance, a recent study found a significant correlation between the transcriptome-wide GQS densities across over 1,000 plant species and the corresponding habit temperature variables (Yang, et al., 2022). This result indicated that plant species growing in cooler areas tend to adopt more GQSs in their transcriptomes. This interesting observation led to the further explorations on the functional roles of RNA GQSs in plants where RNA GQSs stabilize the RNAs in plant response to cold. (3) Evolutionary constrains for identifying the functional RNA structure motifs. For instance, the phylogenetically conserved RNA structures of homologous transcripts across diverse species were suggested to be of functional importance. rRNAs and tRNAs contain very conserved RNA structure motifs derived from the phylogenetical evolution that were essential for translational regulations.

With the advance of computer powers, the other strategy is based on AI. The deep learning methods were used to uncover the hidden relationships between RNA sequences or/and structures and various functions such as the RNA-protein interaction (Cao, et al., 2024). For instance, the iDeepS (Pan, et al., 2018) is a deep learning-based method using a convolutional neural network (CNN) and bidirectional long short term memory network (BLSTM). In the model, both sequence information and structure information serve as input dataset for predicting RNA binding protein (RBP) binding sites and corresponding motifs. In addition, PrismNet is also designed for predicting protein-RNA interactions using a deep neural network, but with *in vivo* RNA structure information (Sun, et al., 2021; Xu, et al., 2023). Compared with other deep learning models, PrismNet integrated the *in vivo* RNA structure data and used an explainable AI method to dissect sequence motifs affecting RNA RBP binding affinity (Sun, et al., 2021). Another typical method is GraphProt2 utilising a graph convolutional neural network (GCN), which can incorporate variable length inputs to predict RNA RBP binding affinity and binding sites (Uhl, et al., 2019). In contrast, foundation models show the generalisation capability across lots of tasks including transcription initiation site, miRNA-target site prediction, splicing sequence features, and so forth (Brixi, et al., 2025; Saberi, et al., 2024). For instance, LoRNA foundation model built on the StripedHyena architecture along with long-read sequencing datasets across 26 human cell lines was able to identify certain sequence motifs associated with both splicing and polyadenylation selection (Saberi, et al., 2024). Up to date, the only foundation model with the capability to obtain the generalized RNA structure motifs associated with RNA functions is PlantRNA-FM (Yu, et al., 2024). PlantRNA-FM was trained on the Transformer architecture using ~1000 plant transcriptomes via three rounds of pre-training steps to learn

the RNA sequence features, genic region features, and RNA secondary structure features. Both RNA sequence and structure information were captured in these pre-training steps by both the masked language model and the sequence-structure alignment supervised learning. PlantRNA-FM achieved excellent performance on multiple downstream tasks, including genic region annotation prediction, RNA secondary structure prediction, and translation efficiency prediction. Notably, PlantRNA-FM also designed a comprehensive framework to identify the functional RNA structure motifs where the framework generates the attention matrix after the fine-tuning step on individual downstream tasks (e.g. translation efficiency classification). By subtracting the attention matrices of the background model from those of the true model, an attention contrast matrix can be obtained that highlights the importance of individual nucleotides on individual RNAs contributing to translation efficiency. Similarly, the attention contrast matrix for all the other RNA functions can be extracted in the same way. By extracting the seed library of RNA structure motifs and overlapping with the attention matrices, the functional RNA structure motifs associated with translation efficiency can be determined by the hierarchical clustering and statistical filtration (Yu, et al., 2024). This interpretation framework can be adopted by all the transformer-based AI as one of the explainable AI strategies for identifying functional RNA structure motifs.

pyTEISER and PlantRNA-FM are the two latest methods for identifying and generalizing the functional RNA structure motifs. A detailed comparison between the two methods is as follows. The advantages of pyTEISER include: (1) By discretizing continuous functional measurements (labels), pyTEISER reduces its reliance on the value distribution of the input data, thereby improving robustness of applications; (2) There are no length limitations of input sequences or genomes; (3) The huge seed library of pre-defined RNA structure motifs covers a wide variety of RNA structure types, increasing the likelihood of identifying functional RNA structure motifs. (4) The pipeline can be applied to different species and diverse RNA functions. pyTEISER also has certain limitations: (1) These pre-defined RNA structure motifs tend to be simple and short hairpins without complicated structure features (e.g. hairpins with multiple loops or internal bulges); (2) Although there is no requirement of the sequence length for the input data, the genic position of these RNA structure motifs was not taken into the consideration. In comparison with pyTEISER, PlantRNA-FM is more advanced: (1) PlantRNA-FM was trained on the transcriptome sequences over ~1000 plant species and corresponding RNA structures. The foundation model is known to sensitively and broadly capture sequence and structural features across the pre-training dataset. The attention contrast matrix derived from the model is more informative in measuring the effects

of individual nucleotides on individual RNAs; (2) PlantRNA-FM takes account of the genic locations of these RNA structure motifs. (3) The types of RNA structure motif are much more diverse than those in the pyTEISER. The limitations of the AI foundation models are: (1) The input length is limited to certain lengths based on the pre-training models. For example, the PlantRNA-FM allows the maximum length of 512 bp. (2) If the regression or classification tasks (e.g. translation efficiency classification) cannot achieve high accuracy of prediction, it is not possible to capture the function-specific information. The model is very sensitive to the value distribution. (3) The foundation model is pre-trained on the transcriptome/genome data of certain species with huge diversity. It is unknown whether certain genome-specific features may impact on the pre-training outcome. For instance, across plant kingdom, some crops such as wheat and rice have GC-rich genomes while most dicot species tend to have AT-rich genomes. It is unknown how these diverse features affect the pre-training step. Thus, there are a large amount of work to be done in evaluating and interpreting the foundation model. There will be considerable potential to improve and refine current methods for identifying functional RNA motifs.

1.3.3 Computational tools to investigate RNA structure dynamics

RNA structures are highly dynamic where one single RNA sequence can fold into diverse structural conformations. The diverse RNA structural conformations allow the potential switches between different conformations. Previous extensive studies have discovered diverse RNA structure switches (Khoroshkin, et al., 2024) such as riboswitches and RNA thermoswitches (Borovská, et al., 2025). Certain cellular or environmental conditions are capable of triggering the switches of RNA structural conformations. For instance, riboswitches are one type of regulatory RNA structure motifs that can directly bind small metabolites or ions and, in response, change their structural conformations to control gene expression. Furthermore, single nucleotide polymorphisms (SNPs) can also alter the RNA structural conformations. These SNPs are termed as RiboSNitches (Halvorsen, et al., 2010). Before the emergence of RNA structure chemical probing methods, software such as Stochastic and RNAsubopt was developed to sample various RNA structural conformations using Boltzmann sampling, which generates suboptimal structures in proportion to their thermodynamic probability (Harmanci, et al., 2009; Zuker, 1989). With the development of chemical probing methods, the *in vivo* RNA structure profiles provide more information for studying RNA structure dynamics in living cells. The tools integrated *in vivo* chemical probing data for studying RNA structure dynamics can be categorised into four types. The first type directly compares chemical reactivity profiles between different conditions.

Software including dStruct (Choudhary, et al., 2019), diffBum-HMM (Marangio, et al., 2021) and DiffScan (Yu, et al., 2022) was developed to identify the regions with significant chemical reactivity alterations between different conditions. The inputs are the chemical reactivities, reverse transcription stalling or mutation counts under different conditions with biological replicates, while the outputs are the regions with significant differences. The output hotspots are usually short, ranging from 10 nt to 25 nt.

Another type of methods directly utilizes chemical reactivities as constraints to generate RNA structural conformations. Recently, there were several software including Rsample (Spasic, et al., 2018), R2D2 (Yu, et al., 2021), and REEFFIT (Cordero and Das, 2015), that was designed to use chemical reactivities as constraints to derive a more accurate RNA structure ensemble. In these methods, the inputs are chemical reactivities, and the outputs are RNA structure ensembles or representative structures from ensembles. The application of using chemical reactivity data can provide additional experimental data into the RNA folding algorithm (usually thermodynamic RNA folding rules) to uncover the RNA structure information close to the cellular status. The chemical reactivity also enhances the sensitivity in detecting the alterations of RNA structural conformations. One example of this method is one recent work from Laederach's group. To study the influence of one single SNP on the structure of the *MAPT* (Microtubule Associated Protein Tau) RNA, which encodes Tau protein related to neuron disease, the DMS-MaP structural profiling of the *MAPT* RNA was performed for both wildtype and two mutation (SNP) types (Kumar, et al., 2022). Boltzmann sampling using Rsample guided by the DMS reactivities for three types, respectively was performed. The RNA structure ensemble of three types was clustered into different groups. Based on individual cluster, the representative structural conformations for each cluster are derived and the corresponding proportions are determined. In three types, the clusters of the RNA structural conformations are very distinct. These differences are functionally linked with the splicing events (Kumar, et al., 2022). Similarly, this type of methods using chemical probing data can be also used to detect RNA switches in living cells. For instance, SwitchFinder is one of the software developed to study RNA structural switches using *in vivo* RNA structure information (Khoroshkin, et al., 2024). SwitchFinder generates RNA structure ensembles via Boltzmann sampling with the constraints of chemical reactivities. Then it introduced a concept termed as "conflicting folded stems" to dissect two structural conformations with the same sequence as the RNA switches (Khoroshkin, et al., 2024).

The third type of methods is to deconvolute RNA structural conformations from ensemble RNA structure libraries. Most RNA structure chemical profiling methods are based on short-read sequencing platforms. The reverse transcription stalling-based methods for capturing chemical modifications generate averaged measurements of RNA structure information across all the molecules that are not able to derive diverse RNA structural conformations. In contrast, the reverse transcription mutation-based methods for measuring chemical modifications can be dissected into different types of mutation profiles, indicating distinct RNA structural conformations. Recent studies have made lots of efforts in dissecting the RNA structural conformations by deconvoluting the mutation profiles from all the reads containing chemical induced mutation profiles. These methods derived diverse novel computational technologies including RING-MaP (Homan, et al., 2014), DREEM (Tomezsko, et al., 2020), Seismic-RNA (Allan, et al., 2024), DRACO (Morandi, et al., 2021), and DANCE-MaP (Olson, et al., 2022). The methods mentioned above were developed based on either DMS or SHAPE induced mutational profiles derived from short-read sequencing platforms. These methods were designed to deconvolute the average chemical reactivity profiles into multiple distinct chemical reactivity profiles using different statistic models and/or clustering approaches. All these methods have been successfully applied on the short RNA segments. DRACO were applied to the long RNAs such as SARS-CoV-2 (~30,000 nt) while DREEM functioned in Human immunodeficiency virus-1 (HIV-1) RNA (Aviran and Incarnato, 2022). DREEM using multivariate Bernoulli mixture model (MBMM) and expectation-maximization (EM) algorithm was applied to reveal several RNA structural conformations (default=2) of the HIV-1 RNAs. The user-defined maximum number of RNA conformations is two to avoid the over-clustering (Morandi, et al., 2021). In contrast, DRACO, another novel computational method, was developed and tested on multiple datasets using the base co-mutated graph model from chemical induced mutation profiles. Then DRACO utilizes the spectral clustering and fuzzy clustering to dissect the mutations profiles without the limitation on the numbers of RNA structural conformations (Morandi, et al., 2021). Furthermore, DRACO takes use of the sliding window strategy to allow itself to be applied in target RNAs with long lengths, enabling the application for the transcriptome-wide analysis (Borovská, et al., 2025). DREEM and DRACO make different assumptions about mutational profiles, which results in distinct model fitness outcomes. DREEM assumed the raw sequencing reads follow a Bernoulli distribution and the probability of nucleotides being chemically modified is independent from other nucleotides across the RNAs, while DRACO built the model on the observed frequency of the co-mutation assumption of two nucleotides on the RNAs. The outputs of both DREEM and

DRACO are deconvoluted chemical reactivity profiles, further deriving the distinct RNA structural conformations. Once the deconvoluted chemical reactivity profiles are generated. The clustered chemical reactivity profiles can be further used to derive RNA structural conformations where the methods in the first two types can be applied.

The other type of methods is to directly cluster RNA structural conformations from structural ensembles. Two recent methods, IRIS (Zhou, et al., 2020) and DaVinci (Yang, et al., 2022) were developed. IRIS was developed based on the use of the Psoralen Analysis of RNA Interactions and Structures (PARIS) datasets that map RNA-RNA interactions and structures within living cells. The method can provide the *in vivo* RNA structure base-pairing information, which is used as the constraints in generating RNA structure ensembles. The other method, DaVinci is developed based on the single-molecule Structure sequencing (smStructure-seq) which generated the RNA structure chemical profiles at the single-molecule level using the PacBio long-read sequencing platform. The chemical induced mutations on individual reads are used to generate individual RNA structural conformations with hard constraints via CONTRAfold (Do, et al., 2006). DaVinci firstly generated all the

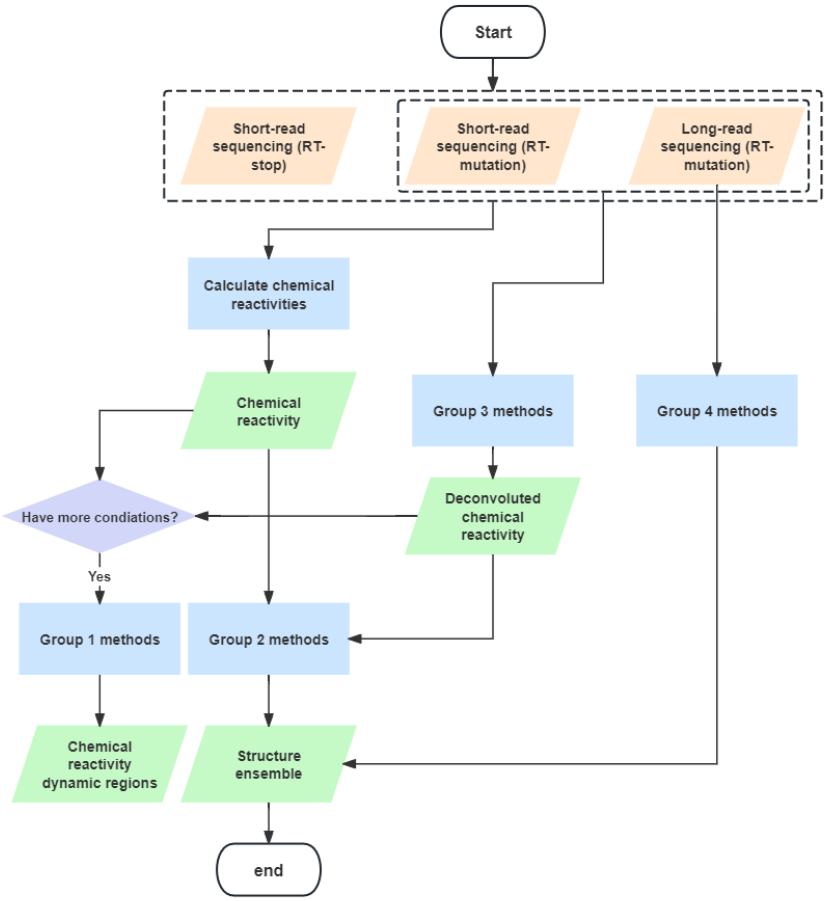


Figure 1. 2 The flow chart of computational methods to study RNA structure dynamics.

individual RNA structural conformations for each unique mutation profiles as the whole population of RNA structural conformations. Then all these RNA structural conformations are clustered into different structural conformations based on the similarity among all these RNA structural conformations. The input of DaVinci is the individual reads, while the output is an RNA structure ensemble with different clusters of structural conformations (Yang, et al., 2022).

All the methods are summarised in **Figure 1.2**. The key features of the methods described above are as follows: (1) If there are multiple reactivity profiles of the same RNAs, methods in group 1 can detect the regions with significant RNA structural changes. (2) The input data of methods in either group 1 or group 2 is the chemical reactivity profiles (RT-stop or RT-mutation counts for some tools). Reactivities can be derived from either RT-stop methods or RT-mutation methods using both short-read and long-read sequencing technologies. (3) Group 3 firstly generates deconvoluted chemical reactivity profiles, further deriving the RNA structural conformations. In contrast, group 4 directly utilize the individual reads (for DaVinci) to derive RNA structural conformations for further clustering of RNA structural conformations. Group 3 may lead to extremes of RNA structural conformations whilst group 4 relies on the sequencing quality. However, there are three limitations of these methods: (1) There is no integrated platform for multiple types of chemical profiling datasets (e.g. short-read RT-stop, short-read RT-mutation, and long-read RT-mutation) to perform RNA structure ensemble analysis. (2) Following obtaining the RNA structure ensembles, the characterization of RNA structures was further explored using the methods (described in 1.3.1). Both dimensionality reduction (e.g. Principal Component Analysis) and clustering (e.g. K-means clustering) are applied afterwards in achieving the major RNA structural conformations. In each cluster, the RNA structural conformation near the centroid of the cluster or the most stable structure is chosen as the representative RNA structural conformation. Sometimes, the representative RNA structural conformations of different clusters can be quite similar, but the RNA structural conformations within the same cluster can be quite different. This is likely to due to the bias of clustering. This issue might be resolved by using the RNA structural ensembles under different conditions where the alterations of RNA structural conformations could indicate the mostly likely representative RNA structural conformations. This issue can also be addressed through optimization of RNA structure characterization, dimensionality reduction, clustering, and representative structure selection. (3) Further comparison of the RNA structure dynamics of one RNA sequence under varying conditions requires a systematic pipeline from the identification of

the RNA structural dynamical regions and distinct RNA structure ensembles to the selection of representative RNA structural conformations.

1.4 Aims and objectives

RNA structures play crucial roles in gene regulation. However, identifying functional RNA structure motifs is a complex challenge that strongly depends on approaches capable of linking RNA structure with RNA function. In this thesis, I developed three projects that leverage diverse RNA structural and functional datasets as case studies to demonstrate how computational methods can be used to uncover the functional importance of RNA structures.

The first project is to identify i-motif (iM), a specific DNA/RNA secondary structure motif with a unique sequence context and the biophysical nature. We developed the first iM prediction platform using a traversal strategy based on graph, ensemble learning, and automated machine learning (AutoML) for determining putative iMs and iM-specific biophysical properties. Furthermore, we revealed the prevalence of RNA iMs in plants and found that 5′UTR iMs might function as translation suppressors. Notably, it is the first comprehensive and systematic study of RNA iMs. This project is also a showcase of the utilization of multiple computational methods for studying specific structure motifs with certain known sequence contents.

The second project is to apply one of the two RNA functional structure motif identification pipelines (described above), pyTEISER in determining the RNA stability-associated structure motifs across the wheat transcriptome. To address the limitation that pyTEISER only counts on standard hairpin motifs (described in 1.3.2), we expanded the motif library by including additional types of RNA structure motifs (e.g. hairpin with bulges and/or internal loops). Moreover, we also utilize regular expression and parallel computing strategies to accelerate the software. Our improved pyTEISER not only identified RNA stability-associated structure motifs but also revealed subgenomic distribution of these motifs across the wheat transcriptome. This project is one example in determining general RNA structure motifs associated with certain functions across the transcriptome.

The third project is to develop a new computational pipeline, RSDE-Tool, for identifying RNA structural dynamic elements (RSDEs) under different developmental conditions in tomato. To overcome the limitation described in 1.3.3, RSDE-Tool integrated DiffScan for

searching for the hotspots of chemical reactivity alterations, applied multiple statistic methods in comparing the changes of RNA structure ensembles, and combined the concept of conflicted folded structures with machine learning approaches in determining the representative RNA structural conformations. RSDE-Tool was demonstrated in characterizing the changes of the RNA structures in tomato fruits at two different developmental stages. In addition, we detected the RSDEs associated with the translational regulation at the different tomato fruit developmental stages. This project develops a new method for studying how RNA structure changes in regulating translation at different plant developmental stages as one case demonstrating the RNA structural dynamics in plant developments.

1.5 Outline of thesis

This thesis demonstrates the use of multiple computational approaches to discovery regulatory RNA structure motifs associated with different functions in different plant species.

Chapter 1 provides an overview of current knowledge in the RNA structure field, including the functional roles of RNA structures in plant, advances in deep-sequencing-based RNA structure profiling methods, and computational approaches for their analysis. It also outlines the scientific aims and objectives of this work and introduces the overall structure of the thesis.

Chapters 2, 3, and 4 present the core research projects of this thesis, each comprising an introduction, detailed methods and materials, results, discussion, and acknowledgements of contributors.

Chapter 2 focuses on a systematic study of i-motifs (iMs). The first iM-specific computational tool, iM-Seeker, was developed for iM prediction and used to map genome- and transcriptome-wide iM landscapes across the plant kingdom. Additionally, iMs located in 5′UTRs were identified as translational suppressors in plants. Parts of this work have been published in *Yang et al., 2024* and *Yu et al., 2024*.

Chapter 3 investigates the major mRNA features that influence RNA stability in the wheat transcriptome. RNA structures within the 3′UTRs were identified as dominant determinants of RNA decay. Furthermore, RNA-stability-associated structure motifs were discovered and

linked to sub-genome-specific differences in RNA stability. This study has been published in *Wu et al., 2024*.

Chapter 4 describes the development of RSDE-Tool, a computational pipeline for detecting RNA structural dynamic elements (RSDEs) and their representative structures. The tool was applied to the tomato fruit transcriptomes at different developmental stages, revealing RSDEs that act as translational regulators and may contribute to fruit development.

Chapter 5 provides a general discussion that summarizes the main findings and limitations of the thesis and outlines potential directions for future research.

Chapter 6 contains the appendix, including supplementary materials, code repositories, and publications completed during the PhD period.

Chapter 2 i-Motif: from prediction to biological function

2.1 Introduction

An intercalated motif (i-motif, iM) is a four-stranded, non-canonical DNA/RNA secondary structure motif that is prevalent in genomes and transcriptomes. The four-stranded structure is folded by cytosine (C)-rich sequences via a series of C tracts formed by hemi-protonated C-neutral C base pairs (C^+:C), as shown in **Figure 2.1**. The stability and formation of iMs are highly dependent on plenty of features. The most fundamental feature is the stability of C^+:C base pairs. The energy of C^+:C base pairs (169.7 kJ/mol) is much higher than that of G.C base pairs (96.6 kJ/mol), which makes iM a stable structure (Abou Assi, et al., 2018). Besides, many previous studies revealed that the formation and stability of iMs can be also affected by many additional features in the iM regions, including the length of C-tracts, the length of loops, the combinations of four nucleotides in loops, the nucleotide modifications (e.g. nucleotides, sugar backbone, and phosphates) etc., (Abou Assi, et al., 2018). In addition to the properties of the iM structures, the folding conditions are important for the formation and stability of iMs. The nature of C^+:C leads to the iMs folding under acidic pH conditions, which is the main trigger used in the approaches for the iM-related detections. Moreover, a number of biophysical studies on the DNA iM folding also showed the low temperature (e.g. 4°C) and the silver cation under the neutral pH condition can induce the folding of iMs (Abou Assi, et al., 2018; Day, et al., 2013). Additionally, both potassium cation and sodium

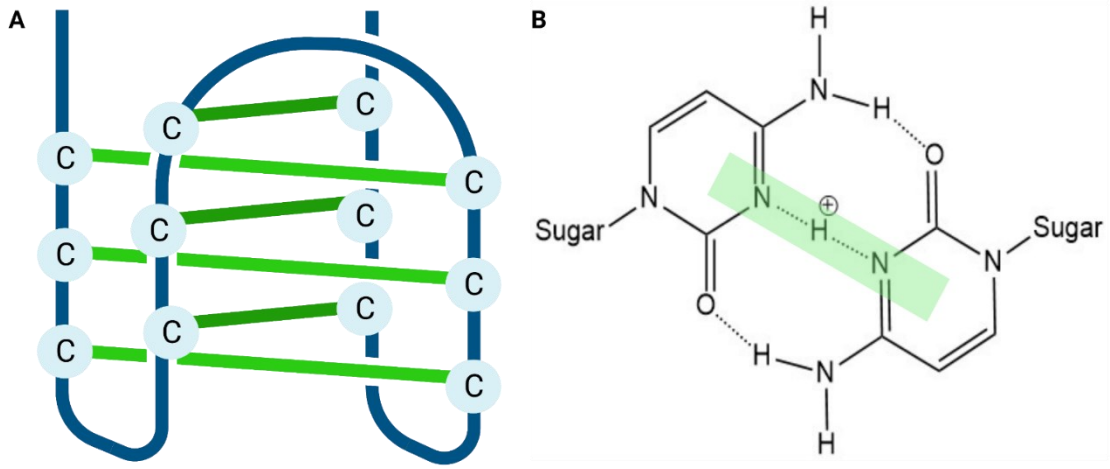


Figure 2. 1 The schematic of i-motif.
(**A**) The structure of i-motif. (**B**) The demonstration of hemi-protonated C-neutral C base pairs. Created with BioRender.com

cation at acidic pH can strengthen the folding status of iMs (Iaccarino, et al., 2019; Tao, et al., 2024). These factors have been used to alter the folding status of iMs in studying the functional roles of these iMs (Collin and Gehring, 1998; Zeraati, et al., 2018).

Currently, biophysical measurements, deep-sequencing approaches, and computational approaches are used to identify iMs. For studying individual iMs, biophysical assessments have been generally applied to determine the folding probability and the thermodynamic strength of certain iMs (Abou Assi, et al., 2018; Wright, et al., 2017). The biophysical approaches can be applied to measure the iMs folding status and strength. For instance, the circular dichroism spectroscopy (CD) can measure the iM folding and strength under different pH conditions, deriving the transitional pH values in measuring the iM stability. If the transitional pH is high, it means that the iM is able to form under relatively high pH, suggesting the strong folding status of the iM. In contrast, the low transitional pH indicates the weak folding status of the iM (Iaccarino, et al., 2019). The UV spectroscopy for measuring the melting and annealing temperatures can also be used to determine the thermal stability of iMs. Thermal difference spectra (TDS) is an analytical technique for measuring the differences between the UV absorption spectra at temperatures above and below the melting temperatures. TDS derived from UV melting can determine the folding status and strength of iMs. Although the biophysical approaches are able to determine the folding status and strength of iMs directly, the low-throughput nature of these approaches limits their applications at the genome-wide level. It is mostly used to assess and validate the folding status and strength of a small number of iMs. In recent years, with the booming of advanced deep-sequencing technologies, several iM-specific sequencing-based methods have been developed to capture iMs across genomes (Ma, et al., 2022; Zanin, et al., 2023). One of the methods, iM-IP-seq, integrates an iM-specific antibody, iMab, cooperated with the immunoprecipitation and the deep-sequencing platform to detect the *in vitro* iM profiles across the rice genome (Ma, et al., 2022). Another *in vivo* method used in the human genome also applied iMab to detecting the iMs which combined the Cleavage Under Targets and tagmentation (CUT&Tag) method for detecting iMs in nucleus (Zanin, et al., 2023). A large amount of iMs were detected via the deep sequencing. However, both the resolution and sufficiency of these methods in detecting iMs are limited due the specificity of the iMab antibody (Tao, et al., 2024). The use of the iMab antibody may also trigger the folding equilibrium of iMs. Nevertheless, these methods can still provide the folding landscape of iMs at the genome-wide level. Computational approaches also play important roles in iM studies. To predict putative i-motif (iM) sequences *in silico*, previous studies typically search

for the complementary G-quadruplex (GQS) sequences using the GQS prediction tools as the complementary sequence patterns between iMs and GQSs (Zanin, et al., 2023). Existing GQS prediction algorithms can be categorised into two main groups: one is based solely on sequence patterns without GQS-specific experimental data while the other incorporates GQS-specific experimental data. The first category includes tools like Quadparser (Huppert and Balasubramanian, 2005), Quadruplexes (Todd, et al., 2005), AllQuads (Kudlicki, 2016), and QuadBase2 (Dhapola and Chowdhury, 2016) , which rely on pattern-matching algorithms such as the regular expression. For instance, iMs follow the sequence pattern like $(C_{\geq 3}N_{1-12})_3C_{\geq 3}$ and there are ~770000 putative iM forming sequences followed the pattern across human genome. The sequence context meeting the sequence pattern above is termed the putative iM. Some methods, such as QGRS Mapper (Kikin, et al., 2006), G4P (Eddy and Maizels, 2006), and G4Hunter (Bedrat, et al., 2016), employ scoring systems to predict the likelihood of GQS formation or stability based on the features of GQSs (e.g. the number of G-tracts and the length of loops) (Puig Lombardi and Londoño-Vallejo, 2020). Since iMs share complementary sequence contents with GQSs, these tools can be adapted for iM predictions. For instance, the tools mentioned above can be used to search G-rich sequences with the complementary sequence pattern as iMs by replacing G with C. The second type of methods integrates experimental data (e.g., G4 ChIP-seq, CUT&Tag, or G4-seq), enhancing the prediction accuracy with genomic relevance (Elimelech-Zohar and Orenstein, 2023). Tools like PQSfinder (Hon, et al., 2017), G4boost (Cagirici, et al., 2022), Quadron (Sahakyan, et al., 2017), and DeepG4 (Rocher, et al., 2021) utilise machine learning or deep learning on the GQS-specific experimental data to predict the folding status of GQS, that could not reflect the folding status of iMs. G4Hunter, one exampled software, is used for the iM prediction with the scoring system accounting the C-rich regions with negative scores (because G represents positive scores) in excluding non-iM-forming sequences with the evaluation of both DNA strands (Bedrat, et al., 2016). Another tool, G4-iM Grinder (Belmonte-Reche and Morales, 2020), was initially designed for GQS identification and integrates multiple scoring systems (e.g., G4Hunter, PQSfinder, cGcC), but allows flexible parameter adjustments as changing G to C. In summary, the first category of methods described above primarily relies on sequence pattern of iMs. Consequently, any sequences matching these patterns are identified as putative iMs. As a result, when the same pattern criteria are applied, the number of putative iMs predicted by these methods is generally similar, with relatively minor variations. However, the vast majority of these predicted putative iMs has an extremely high likelihood of being false positive. This is an unavoidable compromise in the absence of comprehensive, high-confidence experimental data for i-

motifs. A similar situation existed for RNA GQS prior to the development of high-throughput *in vivo* RNA GQS mapping techniques like SHALiPE-seq. In *Arabidopsis thaliana*, tens of thousands of putative RNA GQSs were computationally predicted, yet SHALiPE-seq validated fewer than two hundred as credible (Yang, et al., 2020). Furthermore, another major limitation of these sequence-based methods is that they can only identify sequences containing a putative motif but cannot determine which specific segments function as C-tracts and which form loops. Within a C-rich region, a single putative i-motif-forming sequence could theoretically adopt multiple different iM conformations based on sequence analysis alone. Currently, there is no effective experimental method to determine which conformation, if any, actually exists *in vivo*. Existing computational detection approaches cannot even provide information about these potential alternative conformations. The second category comprises detection methods developed based on experimental data from GQS, which similarly represents a compromise in the absence of iM experimental data. Although iMs and GQSs share some structural similarities, their formation is governed by distinct physicochemical properties (Wright, et al., 2017). Consequently, their detection methods are fundamentally different and should not be conflated. Thus, a dedicated iM prediction method incorporating iM-specific experimental validations is needed for studying iMs. Thus, we developed a new method to specifically predict iMs with the integration of existing iM-specific experimental data.

The previous genome-wide iM scanning revealed that iMs were significantly enriched in the promoter regions, suggesting a potential regulatory role of iMs in the transcriptional regulation in human cells. Additionally, evidence from individual cases has also shown that iMs can fold in other genic regions (e.g., telomeres and transposable elements, TEs) and play diverse roles in different biological functions, including DNA replication, TE functions, and cell cycle regulations (Abou Assi, et al., 2018; Tao, et al., 2024; Zeraati, et al., 2018). Compared to DNA iMs, RNA iMs remain unexplored. Although there was evidence in the NMR studies that ribonucleotides can form iM structures (Campos Carrillo De Preciado, 2018; Snoussi, et al., 2001), RNA iMs were in general assumed to be less stable than DNA iMs (Campos Carrillo De Preciado, 2018; Snoussi, et al., 2001). However, the transcriptome-wide prevalence and functional roles of RNA iMs are still unknown.

In contrast to other organisms, plants are more likely to form stable DNA/RNA structures due to the low habitat temperatures (e.g., ~20 ℃). Furthermore, due to the lack of mobility compared to animals, plants have evolved complex and efficient mechanisms in responding

to dramatic environmental changes, including drought, flooding, temperature fluctuations, and salinisation, further impacting the subsistence and development of plants. Environmental stress (e.g., temperature stress, salinisation, and so forth) can alter the intracellular pH and change the concentrations of many ions, which may affect the folding status of iMs. RNA structure has been suggested to function as an important regulator under environmental stress, including temperature, salinity, light and so forth (Cao, et al., 2024). For instance, RNA GQS was found to form much stabler in the cold (4 ℃) to prevent RNAs from the degradation in *Arabidopsis thaliana* (Yang, et al., 2022*)*. With these advantages, plants are the ideal biological system for studying iMs and their functions. Therefore, we studied the occurrence of iMs across plant transcriptomes and the potential biological functions of these iMs.

This project resolved two challenges above: one is to specifically predict iMs across genomes and the other is to explore the global landscapes of RNA iMs across the plant kingdom and their potential functions.

To address the first challenge, we developed iM-Seeker, an innovative computational pipeline that integrates genome-wide iM profiling data (Zanin, et al., 2023), literature-reported and proprietary biophysical iM-stability data to predict both the formation and stability of iM structures. The core of iM-Seeker consists of a newly developed graph-based algorithm, a classification model, and a regression model. The graph-based algorithm was designed to comprehensively identify all putative iMs within given DNA or RNA sequences. The classification model of iM-Seeker employs a Balanced Random Forest model, trained using the identified iMs across human genomes using the iMab-based CUT&Tag sequencing platform (Zanin, et al., 2023). This model was developed to determine whether a given putative iM is likely to fold into an iM structure. Additionally, iM-Seeker used an Extreme Gradient Boosting (XGBoost) regressor to estimate the iM stability that is based on the iM-specific experimental data. These experimental data are the transitional pH values corresponding to different types of DNA iM sequences. The regression model is primarily designed for DNA iMs but can also serve to predict RNA iMs. This method has also provided new insights into the effects of the nucleotide composition one the iM structure formation and stability.

To solve the second challenge, we conducted the transcriptome-wide iM scanning across 433 plant transcriptomes with available habitat climate information using our iM-seeker. We

generated comprehensive iM landscapes across the plant kingdom. Our analysis revealed significant enrichments of iMs within the 5′UTR regions across diverse plant species, particularly in monocot species. Meta-analysis further indicated that plants growing in relatively warmer climates tend to retain much higher numbers of iMs in their 5′UTRs. Further analysis associated with the translation efficiencies across rice, wheat, tomato, and maize suggest that iMs in the 5′UTR regions may act as the suppressor of translation in plants. However, iM structural stability alone does not appear to be the primary determinant of its translational repression. In summary, these findings offer new insights into the functional importance of iM structures across plant transcriptomes, suggesting a general role of iMs in the translational regulation.

2.2 Materials and methods

2.2.1 Development of Putative-iM-Searcher

Putative i-motifs can be identified based on their sequence patterns, typically following the consensus motifs $(C_{\geq 3}N_{1-12})_3C_{\geq 3}$ where C denotes the cytosine tracts and N represents any four nucleotides (Guneri, et al., 2023; Williams, et al., 2023). Traditionally, putative iMs are inferred by searching for the sequences complementary to the G-quadruplexes (GQSs), using basic pattern-matching approaches. However, such assumptions and methodologies are limited in scope, as they do not adequately capture the structural diversity and variation caused by C^+:C base-pairing or topological configurations (Guneri, et al., 2023; Williams, et al., 2023). For instance, many functional iM sequences are more complex than those only with four cytosine tracts. In the sequences containing five or more C-tracts with varying lengths, it is difficult to assign which four cytosine tracts contribute to the iM formation. For instance, the sequence ‘CCCUCCCUCCCUCCCUCCC’ can be detected by the greedy pattern $(C_{\geq 3}N_{1-12})_3C_{\geq 3}$ using regular expression and contains five C-tracts. This sequence could theoretically form three distinct putative iM structures (using C-tracts numbered 1 to 5): 1235, 1345, or 1245. However, previous methods cannot distinguish all putative iMs from this sequence. The complexity can be further increased when the C-tract lengths are unequal. For example, the sequence ‘CCCCUCCCUCCCCCUCCC’ contains four C-tracts with different numbers of cytosines. This sequence can possibly form six different iMs based on the different arrangements of both the loop lengths and C-tracts: CCC-CU-CCC-U-CCC-CCU-CCC, CCC-CU-CCC-UC-CCC-CU-CCC, CCC-CU-CCC-UCC-CCC-U-CCC, CCC-U-CCC-U-CCC-CCU-CCC, CCC-U-CCC-UC-CCC-CU-CCC, and CCC-U-CCC-UCC-CCC-U-CCC.

To overcome the limitations above, we developed a more general and flexible framework using the directed graph traversal strategy. In this model, each cytosine tract (with length ≥3) is treated as a node, and the loops (spanning 1 to 12 nucleotides) are treated as the directed edges between the nodes. Based on this strategy, the first step is to identify all the C-tracts as the graph nodes on one sequence, and the directed edges added to the position where two C-tracts are separated by a loop ranging from 1 to 12 nucleotides. Once this directed graph is constructed, all putative iMs are derived by traversing the graph starting from each node. Traversal path is defined as subgraphs containing at least four nodes and three directed edges connecting these nodes. The iM candidates are identified by selecting the subgraph with the first four nodes in each traversal path.

To select representative iMs from the pool of possibilities, we implemented four strategies: greedy overlapping, greedy non-overlapping, non-greedy overlapping, and non-greedy non-overlapping, which is aligned with the nomenclature established in QuadBase2 (Dhapola and Chowdhury, 2016). The overlapping strategy allows multiple iMs to share the same start coordinate, while the non-overlapping strategy ensures that no iMs are overlapped. In the greedy mode, iMs with the longest C-tracts and maximum loop lengths are prioritized, whereas the non-greedy mode favors the longest C-tracts with the shortest loops. It is important to note that a single iM sequence may yield multiple iMs with different C-tracts and loops. Among these, two conformations of iMs are highlighted based on their stability criteria: (conformation A) the iM structure with the lowest standard deviation of loop lengths; (conformation B) the iM structure with the shortest total length of the two side loops. For instance, the sequence 'CCCAAACCCCGGGCCCCUUUCCC' is a putative iM sequence. The conformation A is 'CCC-AAAC-CCC-GGG-CCC-CUUU-CCC', which ensures that the lengths of all loops are kept as uniform as possible. The conformation B is 'CCC-AAA-CCC-CGGGC-CCC-UUU-CCC'. The two side loops are 'AAA' and 'UUU', which ensures the shortest total length of the two side loops. Additionally, users can also configure the pipeline to output all potential iM structures for a given sequence. This algorithmic framework has been implemented using Python and we termed it as Putative-iM-Searcher (**Figure 2.2 A**). Putative-iM-Searcher in command-line version is available at https://github.com/YANGB1/Putative-iM-Searcher.

2.2.2 Development of iM-Seeker

2.2.2.1 Data collection and preprocessing

We collected the published iMab CUT&Tag sequencing data in the human genome, which contains *in vivo* information of iMs (Zanin, et al., 2023). The data was retrieved from the NCBI GEO database under accession number GSE220882. The data, provided in BigWig format, encompasses iM-forming regions experimentally identified in two human cell lines: 93T449 (WDLPS) cell line and HEK293T cell line (human embryonic kidney cells). Each cell line contains three biological replicates. We used HEK293T cell data for our downstream analysis due to its significantly higher number of iM-forming regions with high confidence compared to that in the 93T449 cell line, as previously noted (Zanin, et al., 2023). The BigWig files were converted into bedGraph format, after which peak calling was performed using SEACR v1.3 with the code "0.01 non-stringent" used in earlier studies (Meers, et al., 2019; Zanin, et al., 2023). Finally, iM-forming regions with high confidence were determined by identifying the peak regions consistently present across all three biological replicates using bedtools (Quinlan and Hall, 2010).

In addition, we collected many iM sequences, along with their corresponding transitional pH values measured by CD, from the published literatures. We used this information to build the model to predict the iM stability (**Supplementary Table 1**). In parallel, we also generated a large amount of experimental dataset ourselves to expand the diversity of the dataset for modelling (**Supplementary Table 2**). The detailed methods are described in 2.2.4.

We applied our Putative-iM-Searcher to identify putative iM sequences in both the iM-forming regions determined with high confidence in the experimental dataset and the interval regions at both Watson and Crick strands across the human reference genome (GRCh38). Within this framework, putative iM sequences located in the iM-forming regions with high confidence were classified as folded iMs, while putative iMs situated in non-peak intervals were considered as unfolded iMs. To ensure an unbiased evaluation of the classification model, we adopted a non-overlapping selection strategy for all datasets. Four distinct classification datasets were constructed based on different configuration strategies:

- Dataset 1: non-overlapping, greedy selection, conformation A.
- Dataset 2: non-overlapping, greedy selection, conformation B.
- Dataset 3: non-overlapping, non-greedy selection, conformation A.
- Dataset 4: non-overlapping, non-greedy selection, conformation B.

To support the development of the regression model, we built a reliable dataset of iM sequences with corresponding experimentally transitional pH (pH_T) values from both the

literatures (**Supplementary Table 1**) and our in-house biophysical analysis (**Supplementary Table 2**, described in 2.2.4). Using our Putative-iM-Searcher, we also select putative iMs from this dataset. Sequences with the same nucleotide composition but associated with varying pH_T values were also excluded to prevent redundancy. The multiple iMs with the same predicted putative iM sequence and pH_T were combined into one set to avoid bias. Two models were trained using 33 distinct iM-related features extracted from the classification (iMab CUT&Tag sequencing data) and regression (pH_T data) datasets. The 33 iM-related features are as below:

- Length-related features (9 features in total): C-tract length, total iM length, overall loop length, middle loop length, lengths of the longest and shortest side loops, the sum of side loop lengths, as well as the length of longest and shortest individual loops.
- Nucleotide composition features (24 features in total): nucleotide densities (A, C, G, T/U) across the full iM, total loops, middle loop, longest side loop, shortest side loop, and the combined side loops.

These processed features were used to construct models for classification (folded vs. unfolded iMs) or regression (diverse folding strength based on pH_T values). For the regression model, the folding strength of each iM was quantified by transitional pH value after standardization and min-max scaling. Specifically, we first performed standardization using the equation below:

$$x' = \frac{x - \mu}{\sigma}$$

where x, μ and σ denote the original transitional pH, mean and standard deviation of the transitional pH, respectively. Next, to scale the standardized values into the range [0,1], we applied min-max normalization using the equation below:

$$x'' = \frac{x' - min(x')}{max(x') - min(x')}$$

Here, min(x′) and max(x′) are the minimum and maximum values of the standardized transitional pH across all samples. x″ represents the final iM strength.

2.2.2.2 Development of the classification model to predict iM folding status in genome

To evaluate the classification performance of our models across the four constructed datasets, we implemented a five-fold cross-validation strategy in combination with nine widely-used and established machine learning algorithms. An overview of each classifier is detailed in **Table 2.2**. To identify the combination of model and dataset with the best performance, the classification performance was evaluated using two key metrics: Area Under the Receiver

Operating Characteristic Curve (AUROC) and Balanced Accuracy, which is designed for imbalanced datasets. The model, along with the dataset that achieved the highest values in these metrics was selected as the optimal classifier. We randomly divided the entire dataset into two parts: 90% of the dataset was allocated as the training & validation set, while the remaining 10% was used as the test set. The training & validation set was used to search for the best hyperparameters by the grid search combined with the five-fold cross-validation strategy. The model performance was evaluated on the test set using accuracy, recall, specificity, and AUROC. We extract the feature importance directly via the code 'model.feature_importances_' The algorithm was implemented via Python packages: Scikit-learn (Pedregosa, et al., 2011), Imbalanced-learn (Lemaître, et al., 2017), and XGBoost (Brownlee, 2016; Chen and Guestrin, 2016). The schematic is shown in **Figure 2.2 B**.

2.2.2.3 Development of the regression model to predict iM folding strength

The regression model employed the same iM identification approach and structural assignment as used in the classification model. To evaluate the performance of the regression model, a five-fold cross-validation strategy was also adopted across thirteen widely recognized regression algorithms, summarized in **Table 2.2**. The entire dataset was randomly split, with 80% of items used as the training & validation set, where the hyperparameter optimization was performed through the five-fold cross-validation strategy combined with the grid search. The remaining 20% served as the test set to assess the performance. The regression models were evaluated using three key metrics: coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE). To interpret the contributions of iM features within the model, we extracted feature importance scores from the regressor using the code 'importance_type = gain'. The algorithm was

implemented via Python packages: Scikit-learn (Pedregosa, et al., 2011) and XGBoost (Chen and Guestrin, 2016). The schematic is shown in **Figure 2.2 B**. iM-Seeker in command-line version is available at https://github.com/YANGB1/iM-Seeker.

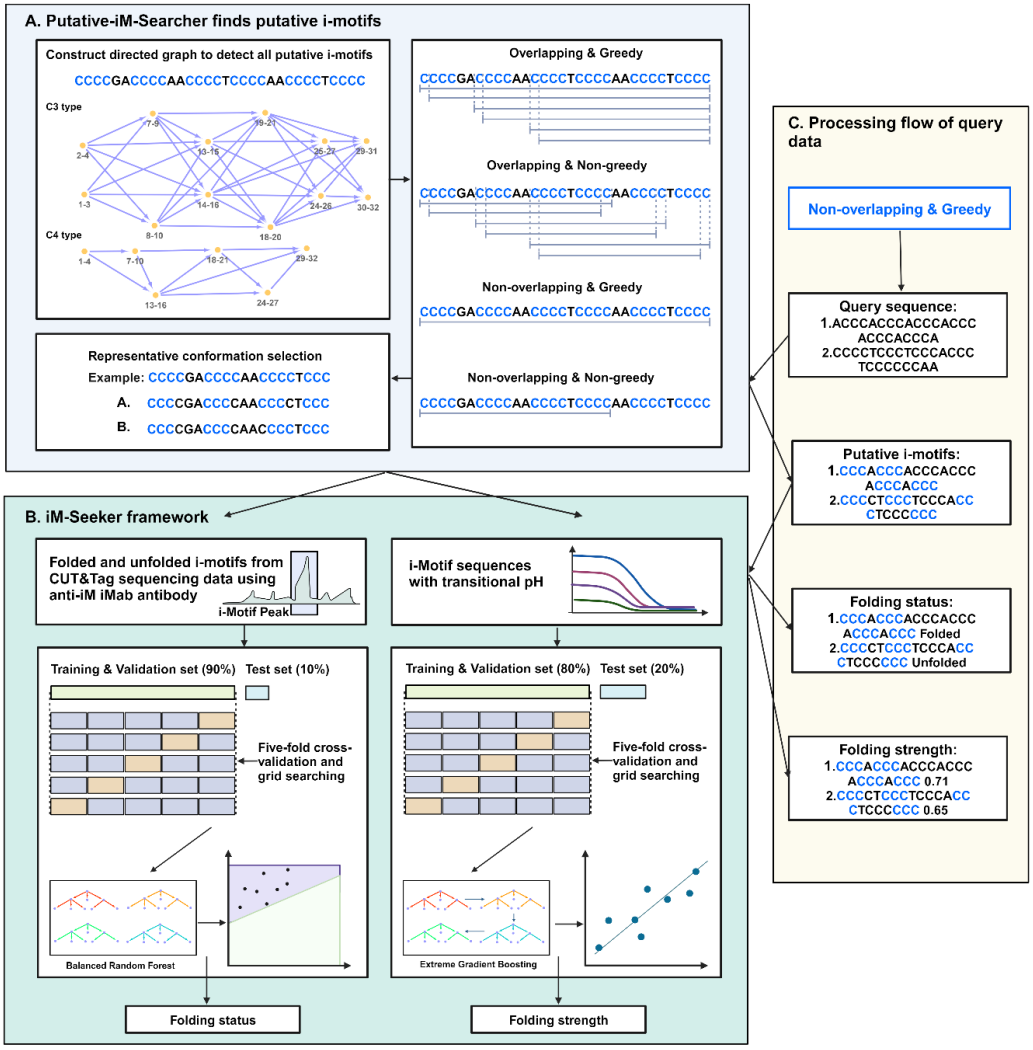


Figure 2. 2 Overview of the iM-Seeker pipeline.
(**A**) Putative-iM-Searcher module. This module identifies all putative iM conformations. The users can choose the representative structures, using combinations of overlapping or non-overlapping schemes, greedy or non-greedy strategies, and two representative conformation options. (**B**) Architecture of iM-Seeker. The iM-Seeker integrates the Putative-iM-Searcher with CUT&Tag sequencing datasets to systematically locate putative iMs across the human genome. Based on the CUT&Tag profiles, putative iMs are annotated as folded or unfolded. For model training, the dataset is divided into training&validation (90%) and test (10%) sets. Model tuning is performed via five-fold cross-validation and grid search to search best hyperparameters on the training/validation set. Final evaluation is conducted on the test set. Folding status is predicted using a Balanced Random Forest classifier. For folding strength estimation, Putative-iM-Searcher is used to search putatvie iMs in biophysical dataset, followed by similar model constuction strategies with folding status model. The dataset is divided into training&validation (80%) and test (20%) sets and the folding strength estimation is performed via an XGBoost regressor. (**C**) Workflow for processing new input sequences using iM-Seeker. Created with BioRender.com.

Table 2. 1 The Overview of machine learning methods tested in iM-Seeker

Algorithms	Tested categories	Description
Decision Tree	Folding status prediction Folding strength estimation	A non-parametric supervised model by building a tree-like statistic model.
Random Forest	Folding status prediction Folding strength estimation	Ensemble learning model integrated ensemble of decision trees.
Balanced Random Forest	Folding status prediction	Adjusted Random Forest with under-sampling strategy to avoid overfitting imbalanced labeled dataset.
Naive Bayes	Folding status prediction	Supervised model based on Bayes theory to assume the conditional independence of every pair of variables given a label.
Linear Discriminant Analysis	Folding status prediction	Supervised model based on Fisher's Linear Discriminant to project data point to lower dimensions to maximize the difference between labels.
Easy Ensemble	Folding status prediction	Ensemble learning model integrated ensemble of AdaBoost learners trained on different balanced samples from random under-sampling to avoid overfitting imbalanced labeled dataset.

Balanced Bagging	Folding status prediction	Ensemble learning model using bagging strategy to integrate ensemble of weak learners trained on different balanced samples from random under-sampling to avoid overfitting imbalanced labeled dataset.
Random Undersampling Boosting	Folding status prediction	Adjusted Adaptive Boosting ensemble learning with random under-sampling to avoid overfitting imbalanced labeled dataset.
Extreme Gradient Boosting	Folding status prediction Folding strength estimation	Adjusted Gradient Boosting integrated ensemble of weak learners. Compared with Gradient Boosting, Extreme Gradient Boosting can be better in both engineering perspective and performance.
Linear Regression	Folding strength estimation	Supervised model to investigate the linear relationship between response variable and explanatory variables.
Ridge Regression	Folding strength estimation	Adjusted linear regression with L2 regularization to reduce training data overfitting.
Lasso Regression	Folding strength estimation	Adjusted linear regression with L1 regularization to reduce training data overfitting.
Elastic Net Linear Regression	Folding strength estimation	Adjusted linear regression with both L1 and L2 regularization to reduce training data overfitting.

Linear Support Vector Regression	Folding strength estimation	Supervised model based on trying to find a hyperplane in searching space to separate the labels via linear kernel.
Radial Basis Function Support Vector Regression	Folding strength estimation	Supervised model based on trying to find a hyperplane in searching space to separate the labels via Radial Basis Function kernel.
K-Nearest Neighbors Regression	Folding strength estimation	A non-parametric supervised model using K-Nearest Neighbors to predict the outcomes.
Adaptive Boosting	Folding strength estimation	Ensemble learning model integrated with weak learners using adaptive boosting strategy.
Gradient Boosting	Folding strength estimation	Ensemble learning model integrated weak learners using gradient boosting strategy.
Random Sample Consensus	Folding strength estimation	Supervised model using Random Sample Consensus to estimate the model parameters.

2.2.3 Development of the iM-Seeker webserver with the automated machine learning function

2.2.3.1 Development of iM-Seeker webserver

The iM-Seeker webserver was designed to accommodate a diverse range of functionalities, broadly classified into several categories: (1) The basic functions of the iM-Seeker are to predict putative iM sequences and features. (2) The automated machine learning (AutoML) function allows users to upload their own in-house biophysical data to build regression models automatically for improving the prediction. (3) iM-seeker hosted the DNA iM landscapes across 30 species: the blue whale, cow, dingo, cat, dog, dromedary, giant panda, goat, guinea pig, horse, house mouse, human, pig, rabbit, rat, *Arabidopsis*, barley, maize, moss, rapeseed, rice, wheat, chicken, mallard, *C. elegans*, *E. coli*, frog, fruit fly, yeast, and zebrafish. These iMs were predicted and are available for the online download. In building

this webserver, a collection of advanced frontend and backend tools were utilized. Vue 3, which is a widely used JavaScript framework, was applied to build the frontend architecture, including the user interface. The backend was implemented using Python 3, leveraging the FastAPI framework for web services, the Celery for task management, and the Redis for the message broker and lightweight database. The whole webserver was designed with data privacy. The no user data will be stored permanently. Any files generated by the model are automatically deleted after 30 days.

2.2.3.2 Development of iM-Seeker automated machine learning session

iM-Seeker webserver integrates an automated machine learning (AutoML) session, which enables users to upload their own biophysical data to build their machine learning model automatically for improving the prediction. The service was designed for developing regression models with continuous labels such as transitional pH and other biophysical numeric indicators. The construction of a machine learning model typically involves several steps including selecting features based on the input data for building models (feature engineering), selecting the types of models for learning, optimizing the hyperparameters in the chosen models, and estimating the model performance. The AutoML was designed to automate the entire process described above. The construction of machine learning models can be regarded as an optimization task. To realize this task, we used a bi-level programming framework that is a mathematical modelling framework characterized by a hierarchical structure. This framework consists of one optimization problem (lower-level optimization)

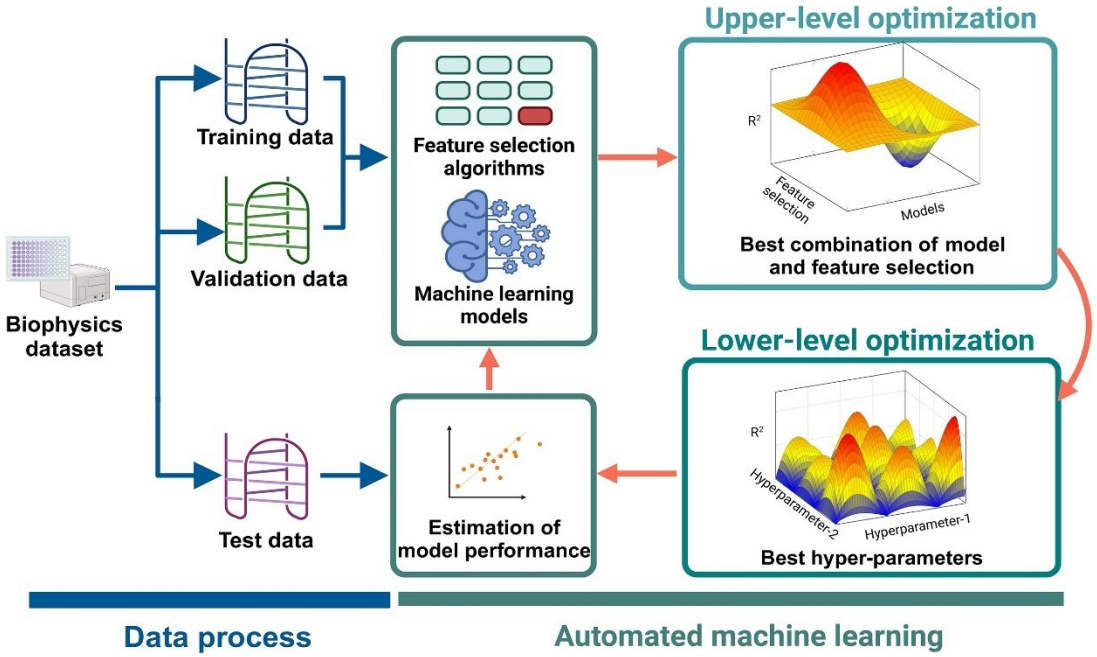


Figure 2. 3 The schematic of iM-Seeker automated machine learning service.

embedded within another optimization problem (upper-level optimization). The upper-level problem aims to optimize its objective function while anticipating the optimal response of the lower-level problem that is an independent optimization task constrained by the upper-level problem. The iM machine learning process was divided into the upper-level optimization tasks (i.e., feature engineering and model selection) and the lower-level optimization tasks (i.e., hyperparameter optimization). The schematic is shown in **Figure 2.3**.

The upper-level optimization tasks were set to find the best combination of available feature engineering methods (**Table 2.2**) and available machine learning models (**Table 2.3**). The searching space is all the combination of available models (12 in our case) and feature engineering methods (seven in our case, six feature engineering methods along with one option without any feature selection). In our case, there are 84 available combinations that serve as the options in the searching space for the upper-level optimization process. The Tabu search was set as the optimization algorithm for the upper-level optimization process. The lower-level optimization task was set to search for the best hyperparameters in the models. The searching space is all the hyperparameters of selected combinations from the upper-level optimization process. The optimization algorithm for the lower-level optimization process is a Tree-structured Parzen Estimator (TPE) based on Bayesian optimization. The objective function is R^2, measuring the divergence between predicted and observed values.

Table 2. 2 The overview of feature engineering methods for iM-Seeker AutoML

Algorithm	Parameter	Description
Generic Univariate Feature Selector	Mode, Score_func_name, Percentile, K_best	Selects features based on univariate statistical tests
Select From Model	Estimator_choice, N_estimators, Max_depth, Threshold, Max_features	Selects features based on importance weights from a model
Sequential Feature Selector	Estimator, Scoring_choice, Forward_choice,	Adds or removes features to form a feature subset

	K_features_value, N_estimators, Max_depth	
Recursive Feature Elimination	Estimator_name, N_features_to_select, Max_depth	Recursively removes features to minimize the feature set
Recursive Feature Elimination with Cross-Validation	Min_features_to_select, Scoring, Estimator_name, Max_depth, N_estimators, Estimator	Performs RFE in a cross-validation loop to find the optimal number of features
Variance Threshold Feature Selector	Threshold	Removes low-variance features

Table 2. 3 The overview of machine learning methods for iM-Seeker AutoML

Algorithm	**Parameter**	**Description**
Ridge Regression	Alpha	Adjusted linear regression with L2 regularization to reduce training data overfitting
Decision Tree	Max_depth, Min_samples_split, Min_samples_leaf	A non-parametric supervised model by building a tree-like statistic model.
Random Forest	N_estimators, Max_depth, Min_samples_split, Min_samples_leaf	Ensemble learning model integrated ensemble of decision trees.
Gradient Boosting	N_estimators, Learning_rate, Max_depth, Min_samples_split, Min_samples_leaf	Ensemble learning model integrated weak learners using gradient boosting strategy.
Support Vector Regression	C, Kernel	Supervised model based on trying to find a hyperplane in searching space to separate the labels via different function kernels.

Multi-Layer Perceptron	Batch_size, Alpha, Learning_rate, Learning_rate_init, Momentum, Hidden_layer_sizes, Activation, Solver	A very classical feedforward artificial neural network (ANN)
Extremely Randomized Trees	N_estimators, Max_depth, Min_samples_split, Min_samples_leaf	Ensemble learning model integrated weak learners using randomized decision trees on various sub-samples of the dataset and using averaging to improve the predictive accuracy and avoid overfitting
Bagging Regression	N_estimators, Max_samples, Max_features, Base_estimator	Ensemble learning model using bagging strategy to integrate ensemble of weak learners
Adaptive Boosting	Base_estimator, Max_depth, N_estimators, Learning_rate	Ensemble learning model integrated weak learners using adaptive boosting strategy.
Stacking Regression	Final_estimator_name, N_estimators, Max_depth, Max_iter, Learning_rate, Min_samples_leaf	Ensemble learning model using stacking strategy to integrate ensemble of weak learners
Histogram-Based Gradient Boosting	Max_iter, Learning_rate, Max_depth, Min_samples_leaf	A gradient boosting-based ensemble learning technique that operates on histograms to allow for faster learning
Extreme Gradient Boosting	Booster, Lambda, Alpha, Subsample, Colsample_bytree, Max_depth, Min_child_weight, Eta, Gamma, Grow_policy	Adjusted Gradient Boosting integrated ensemble of weak learners. Compared with Gradient Boosting, Extreme Gradient Boosting can be better in both engineering perspective and performance.

2.2.4 Biophysical characterisations of putative iMs

To expand the dataset and enhance the diversity of training material for iM-seeker, we generated a substantial amount of experimental data through biophysical characterization. The oligonucleotides containing diverse putative iM-forming sequences was synthesized and HPLC purified by Eurogentec (Belgium). The oligonucleotides were resuspended in ultrapure water, and their concentrations were quantified using a Nanodrop. DNA samples were prepared at 10 μM in 10 mM sodium cacodylate (NaCaco) and 100 mM KCl, across a pH range of 4–8. Annealing was achieved by heating the DNA for 5 min at 95 °C, followed by slow cooling to room temperature overnight.

Circular dichroism (CD) spectra of the annealed putative iM oligonucleotides were acquired on a JASCO 1500 spectropolarimeter (JASCO UK Ltd.) under a constant nitrogen flow. Four scans were collected from 200–320 nm at 20 °C, with a 0.5 nm pitch, 200 nm/min scan speed, 1 s response time, 1 nm bandwidth, and 200 mdeg sensitivity. Spectra of DNA samples and buffer at corresponding pH were subtracted prior to zero correction at 320 nm. The transitional pH (pH_T) was determined by plotting the ellipticity at 288 nm against pH.

For UV spectroscopy, CD-annealed samples at pH 5.5 were diluted to 2.5 μM in the same buffer and analysed using a V-750ST UV/VIS spectrophotometer (JASCO UK Ltd.). Thermal difference spectra (TDS) and melting indicators, including melting temperature (T_M), annealing temperature (T_A), and hysteresis (T_H), were determined. Absorbance at 295 nm was measured at 1 °C intervals during three cycles of denaturation and reannealing (4 °C for 10 min, then heating at 0.5 °C/min to 95 °C, holding for 10 min, and cooling back at the same rate). Melting and annealing temperatures were extracted by the first derivative method, as previously described (Wright, et al., 2020). After the final reannealing cycle, samples were stored at 4 °C prior to further analysis.

For TDS measurements, absorbance spectra were collected from 230–320 nm following incubation at 4 °C for 10 min (folded state) and again after 10 min at 95 °C (unfolded state). TDS signatures were obtained by subtracting unfolded from folded spectra, applying zero correction at 320 nm, and normalizing to a maximum absorbance of 1 (Mergny, et al., 2005).

2.2.5 Identification and characterization of the transcriptome-wide iM landscapes

To explore the iM landscapes across plant kingdom and their associations with the corresponding habitat environments, we collected 433 transcriptomes of land plants from the 1000 plants initiative (1KP) following the previous paper (Leebens-Mack, et al., 2019;

Yang, et al., 2022) that contain habitat information with available bioclimatic variables. The bioclimatic variables were obtained following our previous paper in the lab (Yang, et al., 2022). The geographic distribution information (latitude and longitude of representative habitats) of 433 land plants was collected from the Global Biodiversity Information Facility (www.gbif.org). According to the obtained latitude and longitude, nineteen environmental variables of 433 species with over 100 observations were collected from the WorldClim database (Fick and Hijmans, 2017). For each variable per plant species, the 10th, 25th, 50th, 75th, and 90th quartiles were calculated to summarize the climatic distribution for each species. The details of these environmental variables are shown as follows:

BIO1: Annual mean temperature

BIO2: Mean diurnal temperature range (monthly mean of max temp − min temp)

BIO3: Isothermality (BIO2/BIO7 × 100)

BIO4: Temperature seasonality (standard deviation × 100)

BIO5: Maximum temperature of the warmest month

BIO6: Minimum temperature of the coldest month

BIO7: Annual temperature range (BIO5 − BIO6)

BIO8–BIO11: Mean temperatures of the wettest, driest, warmest, and coldest quarters, respectively

BIO12–BIO19: precipitation variables in different condition, including annual, wettest month, driest month, seasonality (coefficient of variation), wettest quarter, driest quarter, warmest quarter, and coldest quarter.

Our iM-Seeker (command-line version) was firstly applied to search putative iMs on these plant transcriptomes using non-overlapping and non-greedy strategy '--overlapped 2 --greedy 2 --representative_conformation 1'. Only iMs with '+' as the last letter of the iM id were selected. The search of iMs on the complementary sequences is only applied in the DNAs, not RNAs. The Pearson Correlation Coefficient (PCC) analysis between iM densities and bioclimate variables was performed by in-house Python scripts, and the corresponding *P* values were adjusted by the Benjamini–Hochberg method (FDR). The lowest FDR among the different quartiles of climate features was used to represent the corresponding variables following our previous study (Yang, et al., 2022).

2.2.6 Calculation of translation efficiency

To explore the impact of iMs on RNA functions, we collected the polysome-seq and RNA-seq datasets of both rice (BioProject ID number PRJNA1112739) and tetraploid wheat

(Kronos, BioProject ID number PRJNA723219) data. We also collected the ribosome-seq and RNA-seq datasets of tomato (BioProject ID number PRJNA514710) and maize (BioProject ID number PRJNA822292). The transcriptome references were downloaded from the Phytozome database (rice v7.0; tomato ITAG4.0; maize RefGen_V4) and the Ensembl Plants database (Kronos wheat Svevo RefSeq 1.0). For these four species, the corresponding adapters were trimmed by Trim Galore v0.6.7 (Krueger, 2015). The Salmon software 'quant mode' was used to map the raw reads to the transcriptome references and quantify the read counts with the code '--validateMappings' followed by the calculation of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) (Patro, et al., 2017). Translation efficiencies (TE) of individual genes are calculated by dividing the mean FPKM values of three biological replicates of the polysome-seq/ribosome-seq datasets by the mean FPKM values of the corresponding three RNA-seq replicates (Yang, et al., 2021). To avoid the complications caused by different isoforms, we only counted the longest isoform for each gene locus. We also excluded the transcripts with mean FPKM values of the three RNA-seq replicates less than 1 for further analysis.

2.2.7 Analysis between translation efficiency and iM-related features

Thirty-four iM-related properties were extracted based on diverse iM sequences, including 33 iM features described in 2.2.1 and the iM folding strength (iM stability score) in the iM-Seeker. Spearman correlations were calculated by in-house Python scripts. The measurements of the Mutual Information (MI) between TE and all the iM features were performed by function mutual_info_regression in Python scikit-learn package (Pedregosa, et al., 2011).

2.2.8 Gene Ontology enrichment

GO annotations information of four species (rice, tomato, maize, tetraploid wheat) was obtained from the Ensembl Plants. clusterProfiler was used to enrich the GO terms for those genes containing iMs (Yu, et al., 2012).

2.2.9 Statistical hypothesis test

All the statistical analysis and significance test were performed by in-house Python scripts, and the individual tests were indicated in figure legends.

2.2.10 Use of AI during the writing of the thesis

The Grammarly web (https://app.grammarly.com/) was used for grammar check.

2.3 Results

2.3.1 Overview of iM-Seeker

iM-Seeker is an integrated computational framework that leverages a newly developed graph-based algorithm, a classification model, and a machine learning model to predict putative iM sequences, the iM folding status, and iM folding strength. iM-Seeker can be used for both DNA and RNA. A schematic overview of the iM-Seeker pipeline is illustrated in **Figure 2.2**.

The first module, Putative-iM-Searcher, is designed to identify putative iM sequences (**Figure 2.2 A**). This module constructs a directed graph based on the input DNA/RNA sequences, where the C-tracts are represented as the graph nodes and the loop regions as the edges. From this graph, representative conformations are selected using a combination of overlapping vs. non-overlapping, greedy vs. non-greedy, and iM conformation selection strategies.

For the iM prediction, two machine learning models are employed (**Figure 2.2 B**). A Balanced Random Forest classifier is trained on the genome-wide iM profiling data obtained through iMab-based CUT&Tag sequencing in determining DNA iM folding status. In parallel, an XGBoost regressor is trained using the biophysically validated iMs with associated pH_T values for measuring the iM folding strength.

The operational workflow of iM-Seeker upon receiving query sequences is shown in **Figure 2.2 C**. The Putative-iM-Searcher scans the input sequences to identify putative iM sequences. Each putative iM sequence is then evaluated by the classifier to determine whether it is likely to fold. These putative iMs are further processed by the regression model for estimating their folding strengths.

2.3.2 The prediction of the iM folding status

The recently published iMab CUT&Tag sequencing data including three biological replicates provided a robust foundation of datasets for distinguishing between folded iMs and unfolded iMs (Zanin, et al., 2023). In this analysis, putative iM sequences located within the intersected high-confidence iM-peak regions were annotated as folded iMs, while those from adjacent regions lacking iM-peak signals were classified as unfolded ones. To build

classification model, four datasets were built using either greedy or non-greedy selections, and either representative iM conformation strategies A or B under the non-overlapping setting. The two greedy datasets included 8,837 folded iMs and 733,115 unfolded putative iM sequences, while the two non-greedy datasets contained 9,641 folded iMs and 755,747 unfolded iM sequences.

A total of 33 iM-related features were extracted from the datasets. Nine different machine learning classifiers were evaluated using the five-fold cross-validation strategy across all four dataset configurations to determine the best combination of model and dataset. Based on the average AUROC (Area Under the Receiver Operating Characteristic Curve) and the balanced accuracy across folds, the Balanced Random Forest classifier outperformed others. Its good performance to handle imbalanced dataset made it the most suitable model for the classification task (**Figure 2.4 A**). Among all dataset types, the greedy strategies had the better performance metrics compared to the non-greedy configurations (**Figure 2.4 A**). Although differences between conformations A and B were not statistically significant, conformation A achieved slightly higher AUROC and balanced accuracy scores than conformation B. Then we selected the greedy, non-overlapping, conformation A dataset for the final classification (**Figure 2.4 A**).

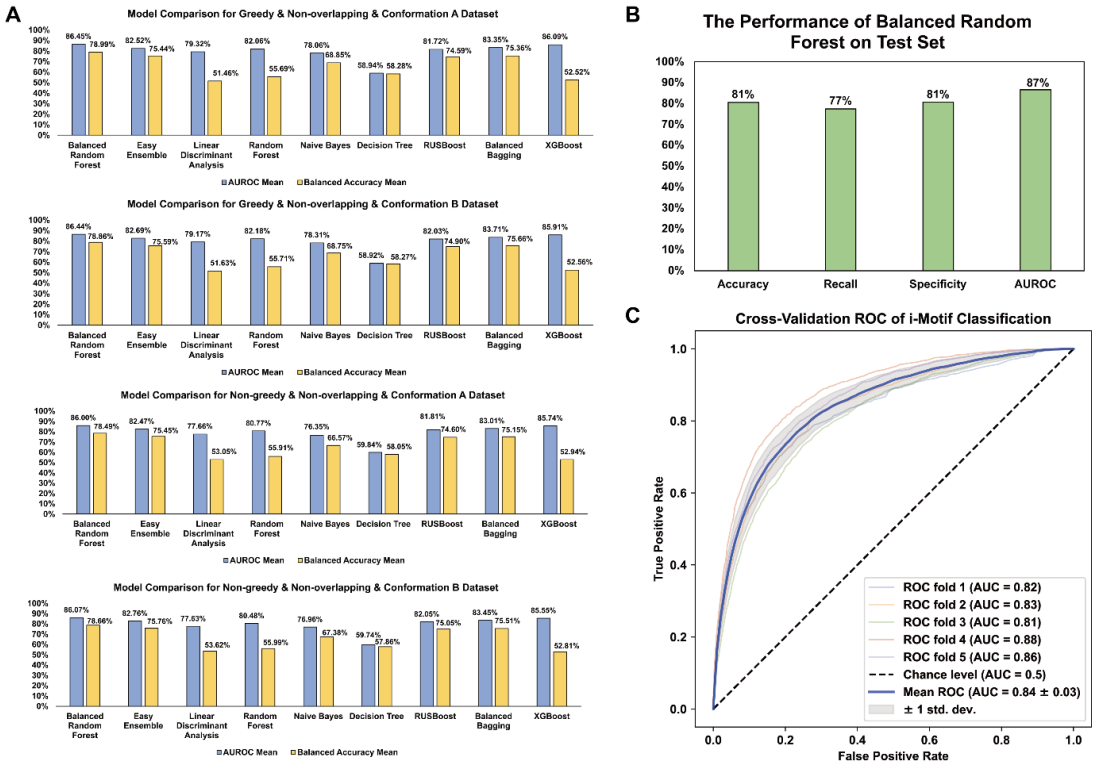


Figure 2. 4 Model selection and evolution for the classification task.
(**A**) Comparative analysis of classification algorithms. Performance was assessed across nine machine learning classifiers, including Decision Tree, Random Forest, Balanced Random Forest, Naive Bayes, Linear Discriminant Analysis, Easy Ensemble, Balanced Bagging, RUSBoost, and XGBoost on four datasets with different representative iM structure extraction strategies. Based on AUROC and balanced accuracy, the Balanced Random Forest model trained on the greedy, non-overlapping, and conformation A dataset outperformed others. (**B**) The performance evaluation of the selected classifier on test set. The Balanced Random Forest model achieved excellent performance on test set, with an accuracy of 81%, recall of 77%, specificity of 81%, and AUROC of 0.87. (**C**) Receiver Operating Characteristic (ROC) analysis. ROC curves from five-fold cross-validation are illustrated, with each fold represented by different colours. The corresponding AUROC values are indicated, and the mean ROC curve is highlighted in blue. A black dashed line marks the baseline performance representing random probability.

To evaluate the performance, the entire dataset (~740,000 data items) was split into the training & validation set (90%) and the test set (10%), where the test set (~74,000 items) was large enough for the model assessment. The hyperparameter optimization of the Balanced Random Forest model was performed on the training & validation set using the five-fold cross-validation strategy along with the grid search. On the test set, the final model achieved an accuracy of 81%, a recall of 77%, a specificity of 81%, and an AUROC of 0.87 (**Figure 2.4 B**), indicating the excellent performance in distinguishing putative iM sequences between folded and unfolded states. Additionally, the five-fold cross-validation strategy across the entire dataset confirmed high generalisability, with AUROC scores higher than 0.8 in all folds (**Figure 2.4 C**).

We extracted the feature importance and conducted the independent *t*-test of individual features between unfolded and folded iMs (**Table 2.4**). Features with greater importance were considered more influential in model construction and more critical for the iM folding *in vivo*. G densities in the folded iMs (e.g. G density in iM and G density in loop regions) are significantly higher than those in the unfolded putative iM sequences, suggesting that *in vivo* folded iMs containing more G in their loop regions. In contrast, the biophysical analysis found that G inside iMs can destabilise iM folding (Brooks, et al., 2010; Fojtík and Vorlícková, 2001; Wright, et al., 2017). The C densities in the loop regions and the length of C-tracts in the folded iMs are higher than those in the unfolded putative iM sequences. Interestingly, higher T and A densities are enriched in the unfolded putative iM sequences. Our findings suggest that the *in vivo* folded iMs may not represent the most biophysically stable conformations, potentially enabling structural flexibility necessary for their functions.

Table 2. 4 The feature importance derived from the classification model

Feature	**Feature Importance**	**Mean value in folded iMs**	**Mean value in unfolded iMs**	***P* value by *t*-test**	**folded vs unfolded**
G density in iM	0.062	0.16615	0.114187102	0	folded > unfolded
G density in loop regions	0.055	0.28186	0.188014227	0	folded > unfolded
A density in iM	0.052	0.0859	0.122959016	0	folded < unfolded
A density in loop regions	0.050	0.14176	0.207166992	0	folded < unfolded
T density in loop regions	0.047	0.17923	0.265357567	0	folded < unfolded
T density in iM	0.046	0.10745	0.156991489	0	folded < unfolded
C density in iM	0.045	0.6405	0.605862393	2.64E-291	folded > unfolded

C density in loop regions	0.040	0.39715	0.339461214	0	folded > unfolded
T density in side loops	0.039	0.17231	0.265348517	0	folded < unfolded
A density in side loops	0.038	0.13766	0.206388777	0	folded < unfolded
C density in side loops	0.035	0.40913	0.340290585	0	folded > unfolded
G density in side loops	0.033	0.28091	0.187972121	0	folded > unfolded
G density in middle loop	0.031	0.2938	0.175505924	0	folded > unfolded
C density in longest side loop	0.029	0.41642	0.342950885	2.23E-285	folded > unfolded
G density in shortest side loop	0.029	0.31268	0.171500246	0	folded > unfolded
C density in middle loop	0.027	0.38254	0.338834673	6.74E-82	folded > unfolded
T density in longest side loop	0.026	0.17145	0.258264081	0	folded < unfolded
T density in middle loop	0.026	0.18295	0.272065245	0	folded < unfolded
A density in middle loop	0.025	0.14071	0.213594158	0	folded < unfolded
A density in longest side loop	0.024	0.1393	0.205745642	0	folded < unfolded
T density in shortest side loop	0.024	0.17264	0.292961105	0	folded < unfolded

C density in shortest side loop	0.023	0.38026	0.322243323	8.25E-109	folded > unfolded
G density in longest side loop	0.023	0.27283	0.193039392	0	folded > unfolded
A density in shortest side loop	0.021	0.13442	0.213295326	2.17E-268	folded < unfolded
iM length	0.021	32.7969	32.06350982	1.31E-22	folded > unfolded
Two side loop length	0.020	13.447	13.35493476	0.112787856	folded > unfolded
Middle loop length	0.019	6.87779	6.46448238	2.67E-31	folded > unfolded
Loop length	0.019	20.3248	19.81941714	1.69E-11	folded > unfolded
Shortest loop length	0.018	4.25574	4.156267434	0.000616409	folded > unfolded
Shortest side loop length	0.017	5.02116	4.956969916	0.050585188	folded > unfolded
longest side loop length	0.016	8.42582	8.397964849	0.389706999	folded > unfolded
Longest loop length	0.014	9.24194	9.06190434	8.20E-11	folded > unfolded
C-tract length	0.003	3.11803	3.061023168	2.06E-37	folded > unfolded

2.3.3 The prediction of the iM folding strength

We curated a dataset including 206 putative DNA iM sequences with their corresponding pH_T values by integrating both literature-reported data items (**Supplementary Table 1**) and experimentally obtained biophysical measurements (**Supplementary Table 2**). The comparative analyses using CD spectroscopy, UV absorbance, and thermal difference

spectra (TDS) between the iM folded and unfolded sequences validated the reliability of our experimental data (**Supplementary Table 2**). To avoid the potential bias, we also excluded one published dataset that included 196 sequences composed solely of cytosine and thymine, which might distort the influence of loop nucleotide composition. From the initial 206 sequences, 171 were determined as high-confidence iM-forming candidates based on criteria including TDS values. Furthermore, we removed the data items corresponding to identical putative iM structures with differing pH_T values and merged those with identical iM structures and matching pH_T values. This process resulted in 120 distinct putative iMs with pH_T, which were further confirmed using our Putative-iM-Searcher configurations (i.e., greedy, non-overlapping, and conformation A) as the classification model, followed by the extraction of 33 iM-related features. The corresponding pH_T values were normalized to a 0-1 range using the standardization and min-max scaling to represent iM folding strength.

Table 2. 5 Model comparison of multiple regressors on the prediction of iM folding strength.

Index	**R^2 Mean**	**Root Mean Squared Error Mean**	**Mean Absolute Error Mean**
Linear Regression	-0.158	0.195	0.139
Ridge Regression	-0.012	0.182	0.132
Lasso Regression	-0.027	0.185	0.146
Elastic Net Linear Regression	-0.027	0.185	0.146
Decision Tree	-0.002	0.181	0.134
Random Forest	0.434	0.138	0.105
Support Vector Regression	-0.043	0.185	0.130
Radial Basis Function Support Vector Regression	0.187	0.165	0.120

KNN	0.111	0.173	0.128
AdaBoosting	0.355	0.147	0.113
Gradient Boosting	0.379	0.144	0.110
RANSAC	-2.184	0.315	0.222
XGBoost	0.458	0.134	0.103

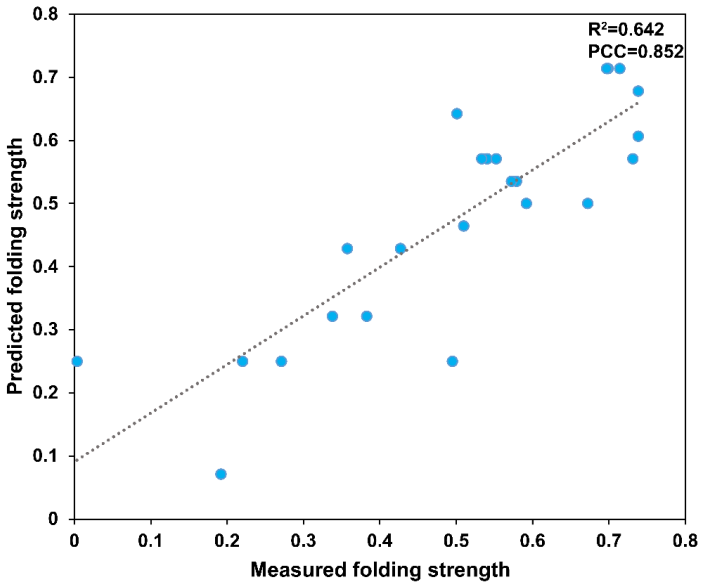


Figure 2. 5 Evaluation of regression performace of XGBoost on test set.
The predictive model has a good predicative performance on test set (n=24). Specifically, the Pearson correlation coefficient reached 0.852 ($P < 1.25 \times 10^{-7}$), while the coefficient of determination ($R^2 = 0.642$) further confirmed the model's ability to predict iM folding strength

To identify the optimal regression model, thirteen different algorithms were evaluated using the five-fold cross-validation strategy. The model performance was assessed based on the average values of three metrics: R^2, RMSE, and MAE, across the five folds. Among the evaluated models, the XGBoost regressor demonstrated the best performance and was selected as the final regression model (**Table 2.5**). The dataset was divided into the training & validation set (80%) and the test set (20%). After optimizing hyperparameters with the grid search combined with the five-fold cross-validation strategy on the training & validation set, the optimized XGBoost model was then tested on the test set. The model demonstrated strong predictive performance, with an R^2 of 0.642, an RMSE of 0.104, and an MAE of 0.08. Additionally, the Pearson correlation coefficient (PCC) revealed a significant correlation

between the predicted and observed folding strengths ($p < 1.25 \times 10^{-7}$), further supporting the reliability and outperformance of the model (**Figure 2.5**).

We then analysed the relative contributions (feature importance) of the iM-related features derived from our regression model (**Figure 2.6**). Features with greater importance were considered more influential in the model construction and more critical for the iM folding strength. Based on the Pearson correlation coefficients (PCCs) between the feature measurements and iM folding strengths, all features were categorized into two groups: negatively correlated features, which were considered to suppress the iM folding strength (**Table 2.6**), and positively correlated ones, which enhance the iM folding strength (**Table 2.7**). The ten most significant features from each group are illustrated in **Figure 2.6**. The nucleotide composition was found to have strong associations with the iM folding strength. More stable iMs tend to have higher frequencies of C and T, particularly T within the side loops (**Figure 2.6 A**). Conversely, G and A densities were associated with low iM folding strengths (**Figure 2.6 B**). Among all length attributes, longer C tracts coupled with shorter loops were associated with more stable iM structures. This finding is consistent with previous reports showing that, under the same experimental conditions, iMs with extended C have greater stability than those with shorter C tracts (Fleming, et al., 2017; Fojtík and Vorlícková, 2001).

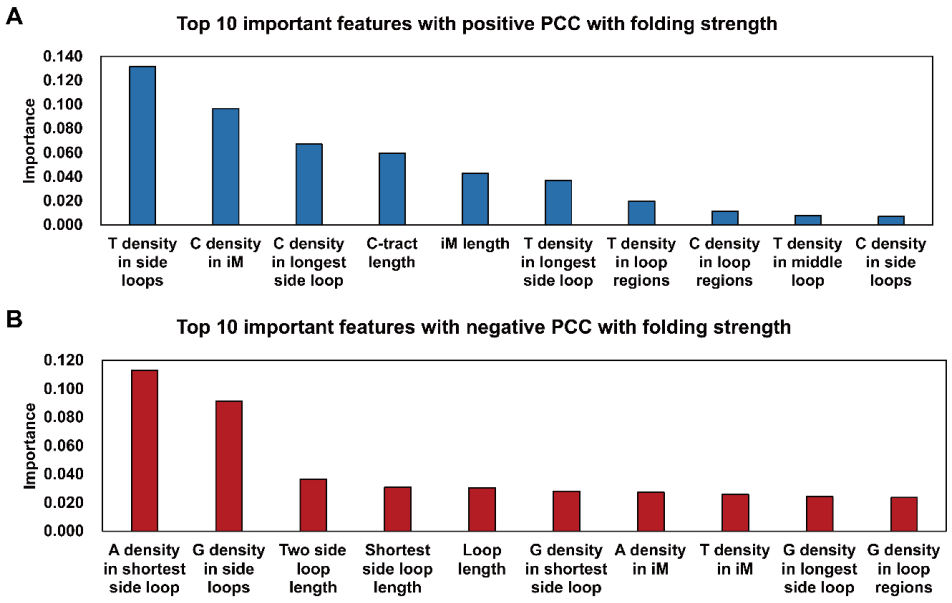


Figure 2. 6 The iM feature importance obtained from the regression model.
(**A**) The 10 most important features showing a positive Pearson correlation coefficient (PCC) with folding stability. (**B**) The 10 most important features exhibiting a negative Pearson correlation coefficient (PCC) with folding stability

Table 2. 6 Important features with negative Pearson Correlation Coefficient (PCC) between features and folding stability

Feature	Feature Importance	Pearson Correlation Coefficient (PCC)	*P* value of PCC
A density in shortest side loop	0.113	-0.149	0.103
G density in side loops	0.091	-0.138	0.133
Two side loop length	0.037	-0.084	0.361
Shortest side loop length	0.031	-0.108	0.242
Loop length	0.030	-0.103	0.265
G density in shortest side loop	0.028	-0.091	0.323
A density in iM	0.027	-0.295	0.001
T density in iM	0.026	-0.044	0.635
G density in longest side loop	0.024	-0.141	0.125
G density in loop regions	0.024	-0.155	0.091
G density in iM	0.021	-0.235	0.010
A density in loop regions	0.020	-0.212	0.020
G density in middle loop	0.011	-0.075	0.418
longest side loop length	0.011	-0.053	0.568
Middle loop length	0.007	-0.119	0.195
A density in side loops	0.004	-0.228	0.012

Shortest loop length	0.003	-0.116	0.209
A density in middle loop	0.002	-0.078	0.394
Longest loop length	0.001	-0.075	0.416
A density in longest side loop	0.000	-0.228	0.012

Table 2. 7 Important features with positive Pearson Correlation Coefficient (PCC) between features and folding stability

Feature	**Feature Importance**	**Pearson Correlation Coefficient (PCC)**	***P* value of PCC**
T density in side loops	0.131	0.174	0.058
C density in iM	0.097	0.391	0.000
C density in longest side loop	0.067	0.138	0.133
C-tract length	0.059	0.457	0.000
iM length	0.043	0.199	0.030
T density in longest side loop	0.037	0.155	0.091
T density in loop regions	0.020	0.138	0.133
C density in loop regions	0.011	0.124	0.178
T density in middle loop	0.008	0.058	0.530
C density in side loops	0.007	0.119	0.194
C density in middle loop	0.003	0.050	0.586

T density in shortest side loop	0.003	0.152	0.098
C density in shortest side loop	0.003	0.051	0.582

2.3.4 The iM-Seeker webserver with the automated machine learning function

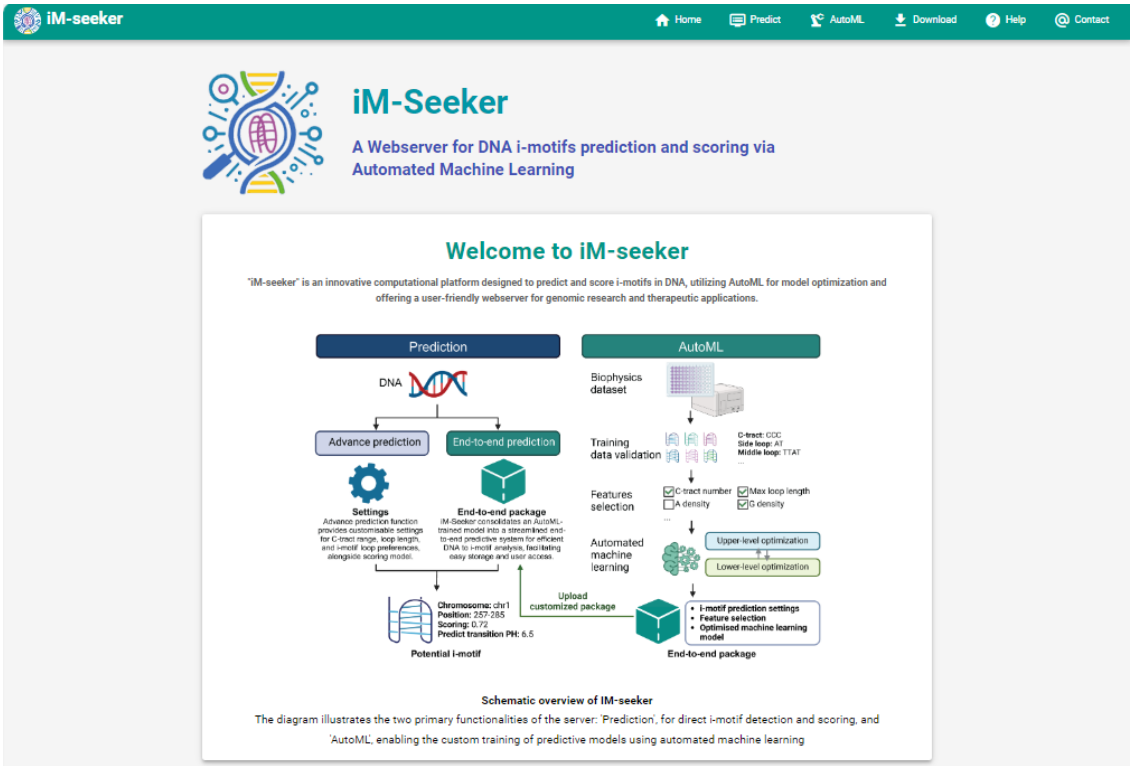


Figure 2. 7 The main interface of the iM-Seeker webserver.

The iM-Seeker webserver is a well-designed and user-friendly platform, which is available at https://im-seeker.org/. The initial interface of the website is shown in **Figure 2.7**. It contains three core functions: 'Predict', 'AutoML', and 'Download'. User instructions can be found under the 'Help' section. In the 'Predict' session, the users can upload or paste their query sequences in the FASTA format and have two options to set up the configurations. The first option is the end-to-end prediction with the default configuration. The other option is more advanced where the users can set up the configuration (i.e., C-tract length, loop length, greed or non-greedy strategy, overlapped or non-overlapped strategy, and conformation A & B strategy) by themselves. After the processing of the submitted task, the webserver offers the information including the location of putative iMs within the query sequences, the predicted iM sequences, the predicted properties, and the general statistics (i.e., the length of sequences, C density, and average loop length so forth). The outcome will be kept on the server for 30 days before the automatic deletion. The second session is the

'AutoML', which is designed to offer users the function to build the machine learning model automatically with their own in-house data. To use this session, users need to upload or paste their data in the CSV format, including query sequences containing putative iMs and corresponding biophysical measurements (e.g. transitional pH). Next, the users can set up the configurations for the Putative-iM-Searcher to identify the putative iMs and select the corresponding features (33 features in total, consistent with the default iM-Seeker mode). In the next step, the webserver will provide users five options to run AutoML for building the machine learning model, including four simple settings and one advanced setting. Simple settings are consisted of 'Swift Basic', 'Fast Essential', 'Precise Advanced', and 'Best Performance' for the diverse demands on the computing resources. Four simple settings provide increased numbers of the iteration rounds of upper- and lower-level optimisations, and more options for the feature engineering methods and machine learning models. For example, 'Swift Basic' only uses XGBoost without the feature engineering, while 'Best Performance' counts all the feature engineering approaches along with all ensemble machine learning methods. In addition to these four simple settings, the advanced setting enables users to customize the configuration manually according to their needs. After configuring and submitting settings, the webserver will search for the best combination of feature engineering approaches, models, and hyperparameters through the backend bi-level optimisation framework. Different settings acquire various processing time. For instance, the 'Swift Basic' option completes in just a few minutes, whereas 'Best Performance' may take several days, depending on the dataset size. Once the model is successfully built, the webserver presents the model's performance evaluation (R^2) and allows users to download the model in the pickle format. The final section, 'Download', provides access to pre-computed iM data—including iM sequences and their biophysical properties—in the CSV format for 30 species, covering animals, plants, yeast, and bacteria.

2.3.5 The transcriptome-wide iM landscapes in plant kingdom

The 1000 plants initiative (1KP) generated comprehensive and systematic transcriptomic resources in plant kingdom (Leebens-Mack, et al., 2019). To obtain the corresponding plant habitat information, we extracted the transcriptomes of 433 land plants across six plant clades, including dicots (N=277), monocots (N=43), gymnosperms (N=30), ferns (N=30), lycophytes (N=10), and bryophytes (N=43) from the 1KP datasets. The iM-Seeker was applied to identify putative iM sequences across 433 plant transcriptomes. The configuration of 'non-overlapped' and 'non-greedy' was used to count the putative iMs without the redundancy. The putative iMs were identified in different genic regions (i.e., 5′UTR, CDS,

and 3′UTR) across all the plant transcriptomes, respectively (**Figure 2.8**). The iM densities (iM numbers per megabase) were defined to represent the iM abundance (Mullen, et al., 2010). The iM densities in the three genic regions were calculated across plant species, respectively (**Figure 2.9**). Among the three genic regions, 5′UTRs tend to enrich more iMs than CDSs and 3′UTRs across most plant species (**Figure 2.9**). Monocots have overall higher iM densities in both 5′UTR and CDS regions than the other five plant clades (**Figure 2.10**), indicating that monocots adopt more iMs across their transcriptomes. The representative species of monocots and dicots are rice (*Oryza sativa*) and *Arabidopsis thaliana*, respectively (**Figure 2.11**). Consistent with the global trends, both iM densities and numbers in the rice transcriptome are much higher than those in the transcriptome of *Arabidopsis*, especially in the 5′UTRs. The iM density in the 5′UTRs of rice genes reaches 578.70 iMs/Mb, while the iM density in *Arabidopsis* is only 11.82 iMs/Mb. To avoid the potential bias caused by the cytosine (C) contents across different plant species, we also calculated the C densities (the percentage of C in four nucleotides) and defined another indicator, the iM enrichment in relations to the cytosine enrichment (iM numbers per megabase cytosine). We then compare both iM density and iM enrichment to investigate whether the difference of iM densities is due to the diverse cytosine densities. Although we observed significant divergences of C densities across the three genic regions between monocots and plant species in the other five clades, the iM enrichment shows highly significant differences compared to the differential C densities (**Figure 2.12**; **Figure 2.13**). Our results indicate that the higher iM density in the monocots may be partially contributed by the C density.

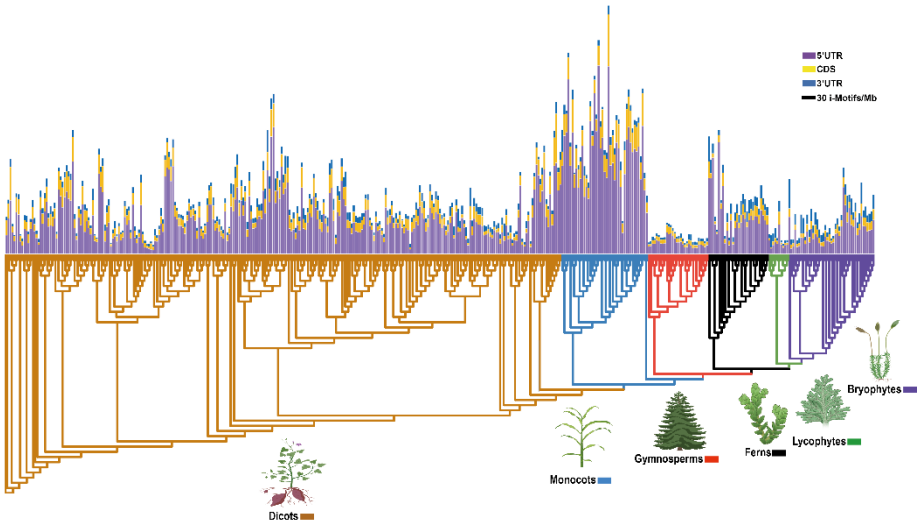


Figure 2. 8 The landscape of transcriptome-wide i-motif in plant kingdom.
The iM densities (iMs/Mb) of various genic regions (5′UTR, CDS, 3′UTR) across 433 land plants. The plants are in six clades. N= 277, 43, 30, 30, 10, 43 for dicots, monocots, gymnosperms, ferns, lycophytes, and bryophytes, respectively.

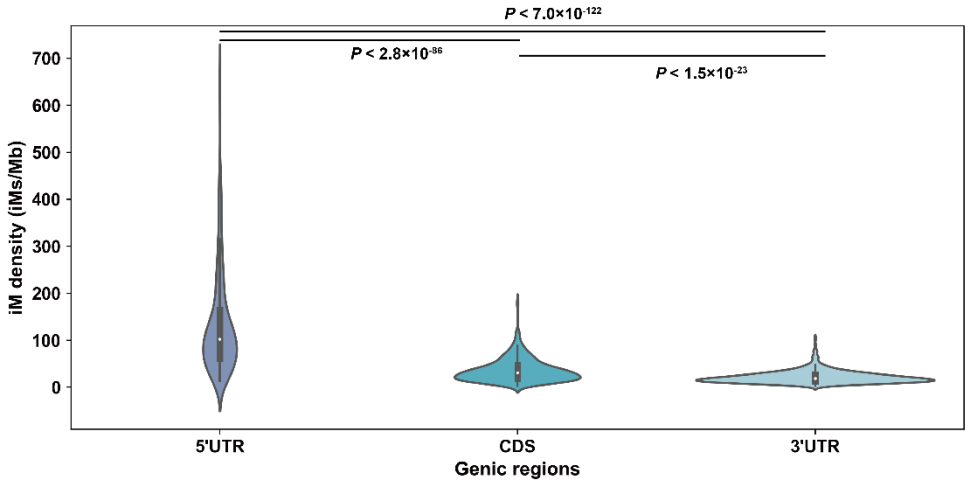


Figure 2. 9 Comparison of iM of genic regions.
The comparison of iM densities among three genic regions (5′UTR, CDS and 3′UTR) across 433 land plants with significance tested by Mann-Whitney *u*-test with Bonferroni correction.

To characterize different types of iMs across plant kingdoms, we divided iMs into five categories based on the layer of C-tracts and the longest loops. Type 1 refers to the iMs with three-layer C-tracts and longest loops ranging from one to four nucleotides; Type 2 refers to the iMs with three-layer C-tracts and longest loops ranging from five to eight nucleotides; Type 3 refers to the iMs with three-layer C-tracts and longest loops ranging from nine to

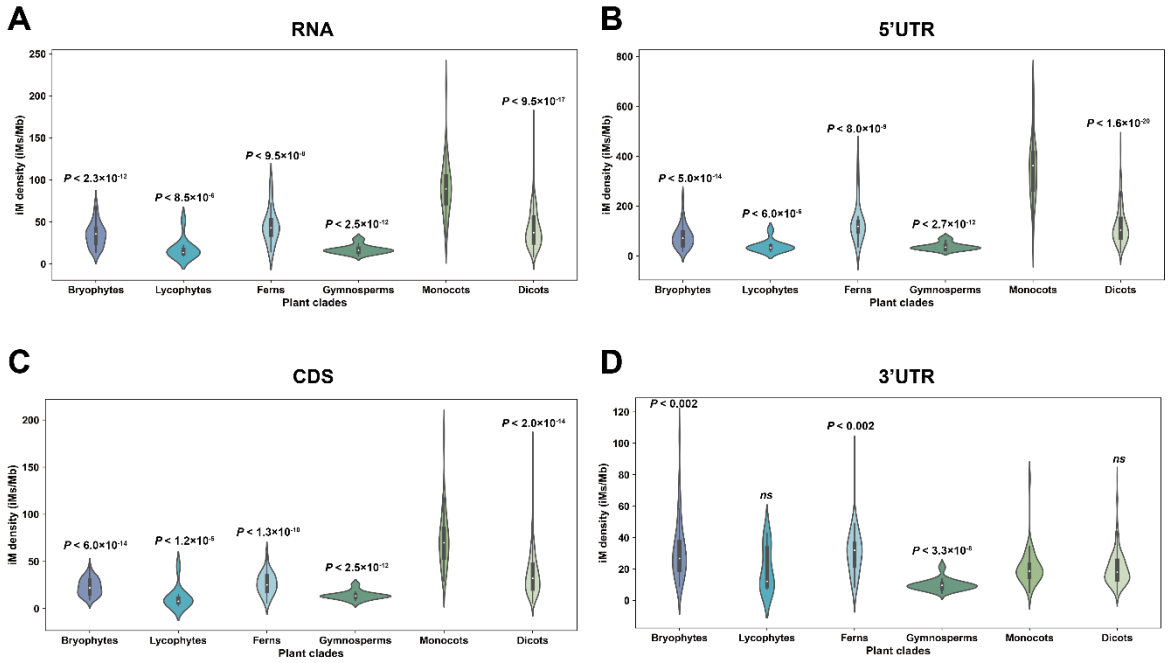


Figure 2. 10 The distribution of iM density in genic regions.
(**A**) Whole transcriptomes (**B**) 5′UTRs (**C**) CDSs (**D**) 3′UTRs. Statistical analysis was performed between monocots and other five plant clades with significance tested by Mann-Whitney *u*-test with Bonferroni correction.

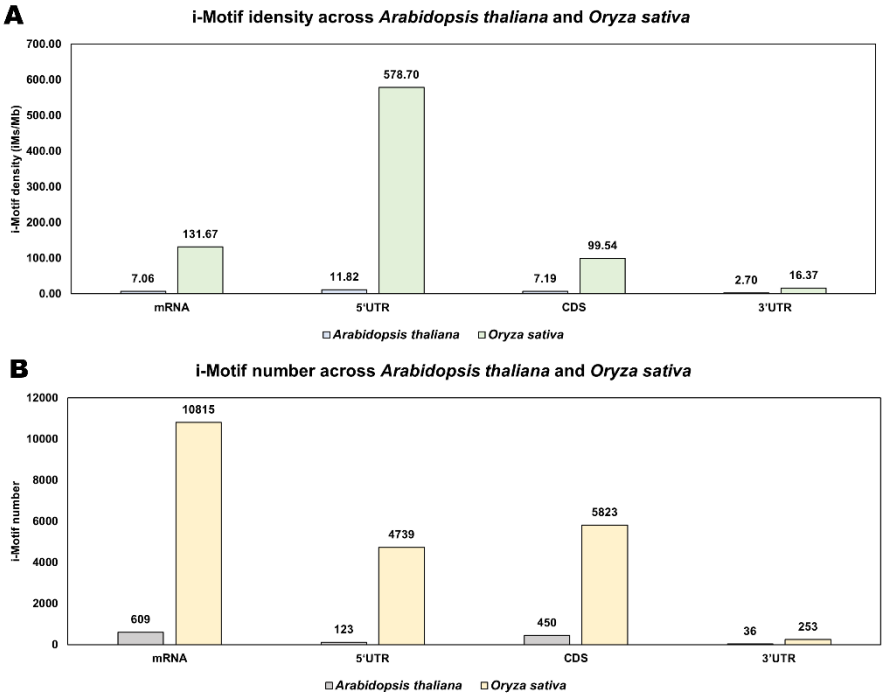


Figure 2. 11 iM comparison between rice and *Arabidopsis thaliana*.
The comparison of iM densities (**A**) and iM number (**B**) between rice and *Arabidopsis thaliana* among three genic regions (5′UTR, CDS and 3′UTR).

twelve; Type 4 refers to the iMs with four-layer C-tracts; Type 5 refers to the iMs whose the layer of C-tracts exceed four. We counted different types of iMs across the 433 plants and found both type 2 and 3 are the predominant forms across all plant clades (**Table 2.8**; **Figure 2.14**). Across all plants, the five iM types accounted for 11.67%, 29.29%, 55.84%, 2.98%, and 0.22%, respectively, with the majority corresponding to the iMs with three-layer C-tracts and longest loops ranging from five to twelve nucleotides (**Table 2.8**). We further applied iM-Seeker to estimate the iM folding strength (stability) for individual iMs across six plant clades. Although the prediction of the iM folding strength was designed using experimental biophysical data of the DNA iMs due to limited RNA iM experiments, it provides a useful reference for RNA iMs. Notably, we found that ferns trend to adopt relatively stronger iMs compared to other clades (**Figure 2.15**). In summary, these findings highlight the significant enrichment of iMs in 5′UTR regions and their prevalence in monocots among plant species.

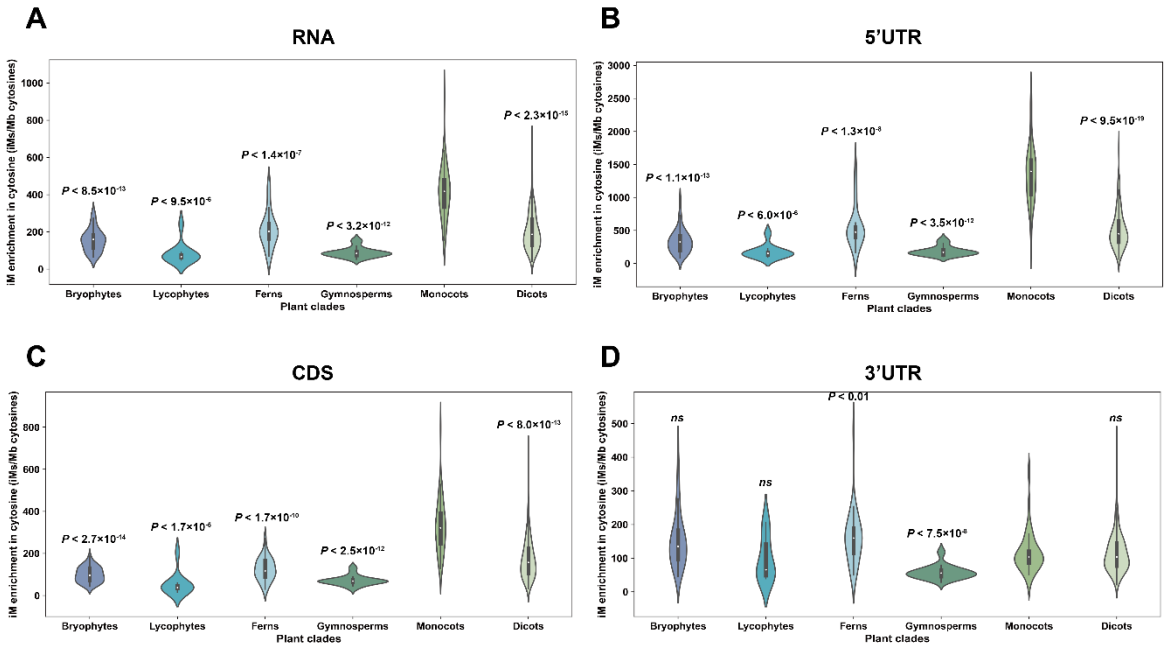


Figure 2. 12 The distribution of iM enrichment in genic regions.
(**A**) Whole transcriptomes (**B**) 5′UTRs (**C**) CDSs (**D**) 3′UTRs. Statistical analysis was performed between monocots and other five plant clades with significance tested by Mann-Whitney *u*-test with Bonferroni correction.

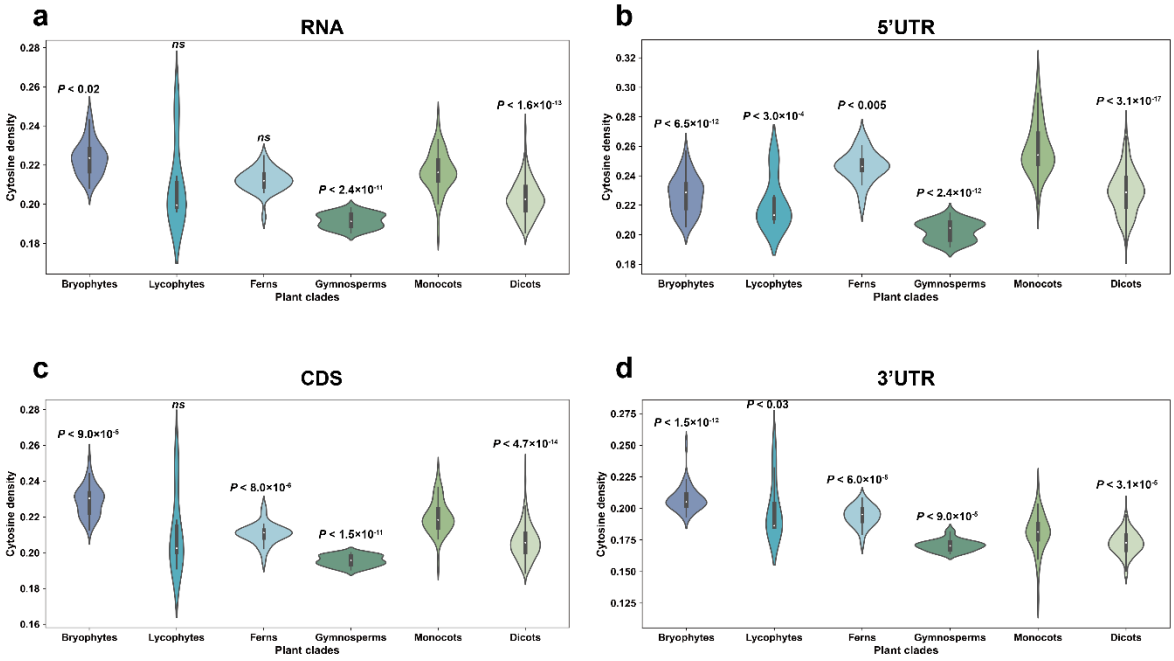


Figure 2. 13 The distribution of C density in genic regions.
(**A**) Whole transcriptomes (**B**) 5′UTRs (**C**) CDSs (**D**) 3′UTRs. Statistical analysis was performed between Monocots and other five plant clades with significance tested by Mann-Whitney *u*-test with Bonferroni correction.

Table 2. 8 The percentage of iMs of different types in six plant clades

iM type 1 indicates iMs with three-layer C-tracts and longest loops ranging from one to four

nucleotides; iM type 2 indicates iMs with three-layer C-tracts and longest loops ranging from five to eight nucleotides; iM type 3 indicates iMs with three-layer C-tracts and longest loops ranging from nine to twelve; iM type 4 indicates iMs with four-layer C-tracts; iM type 5 indicates iMs whose the layer of C-tracts exceed four.

Species	iM type1	iM type2	iM type3	iM type4	iM type5
Dicot	11.74%	28.94%	56.47%	2.65%	0.20%
Monocot	10.83%	30.67%	54.97%	3.37%	0.16%
Gymnosperm	9.88%	26.91%	60.19%	2.87%	0.15%
Fern	13.76%	29.44%	51.85%	4.47%	0.48%
Lycophyte	13.07%	27.90%	55.22%	3.56%	0.25%
Bryophyte	11.68%	29.66%	55.11%	3.29%	0.26%
All plants	11.67%	29.29%	55.84%	2.98%	0.22%

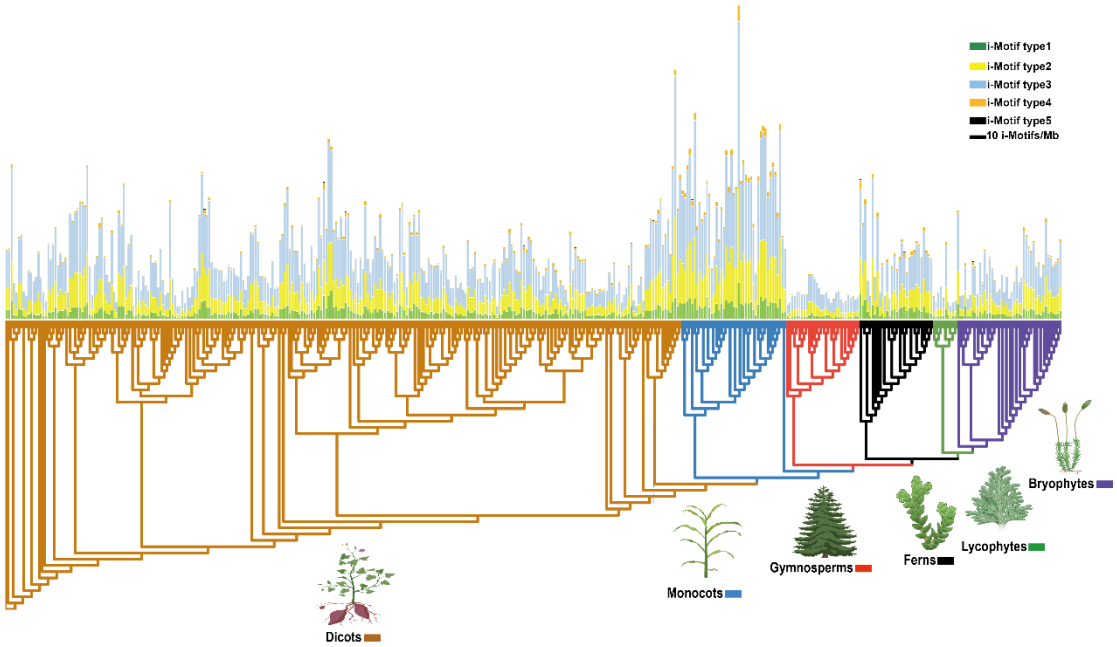


Figure 2. 14 The landscape of transcriptome-wide i-motif types in plant kingdom.
The landscape of transcriptome-wide i-motifs of different types across 433 land plants. The plants are in six clades. N= 277, 43, 30, 30, 10, 43 for dicots, monocots, gymnosperms, ferns, lycophytes, and bryophytes, respectively. The iMs are divided into 5 categories. Type 1 indicates iMs with three-layer C-tracts and longest loops ranging from one to four nucleotides; Type 2 indicates iMs with three-layer C-tracts and longest loops ranging from five to eight nucleotides; Type 3 indicates iMs with three-layer C-tracts and longest loops ranging from nine to twelve; Type 4 indicates iMs with four-layer C-tracts; Type 5 indicates iMs whose the layer of C-tracts exceed four.

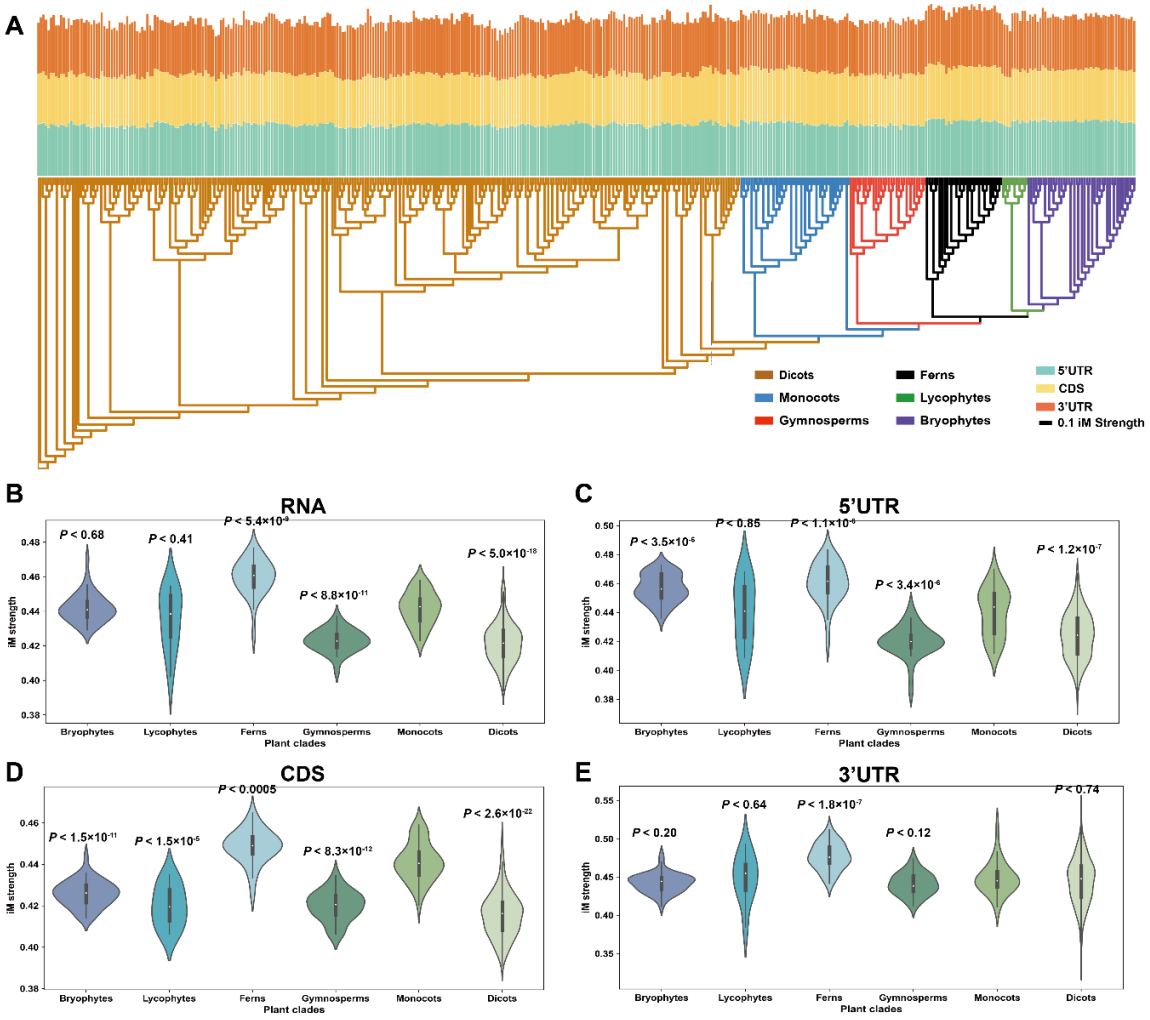


Figure 2. 15 The distribution of iM strength in genic regions.
(**A**) The landscape of transcriptome-wide i-motifs mean folding strength of different genic regions (5′UTR, CDS, and 3′UTR) predicted by iM-Seeker across 433 land plants. The plants are in six clades. n = 277, 43, 30, 30, 10, 43 for dicots, monocots, gymnosperms, ferns, lycophytes, and bryophytes, respectively. The distribution of iM strength predicted by iM-Seeker in whole transcriptomes (**B**), 5′ UTR (**C**), CDS (**D**), 3′ UTR (**E**) across six clades. Statistical analysis was performed between Monocots and other five plant clades with significance tested by Mann-Whitney *u*-test.

To further characterize these iMs landscapes, we collected the 19 bioclimatic variables (i.e., 11 temperature variables and 8 precipitation variables) according to the habitats of these 433 plant species following our previous study (Yang, et al., 2022). We measured the Pearson Correlation Coefficient (PCC) between the iM density and the associated bioclimatic variables (**Figure 2.16**). The PCCs between the iM density and the temperature variables were significantly high positive in 5′UTR and CDS regions (**Figure 2.16 A**). In contrast, overall weaker positive or negative correlations were observed between the iM density and the precipitation variables (**Figure 2.16 A**). We also calculate the PCCs between the C

density and the environmental variables, and between the iM enrichment and the environmental variables (**Figure 2.16 B-C**). Overall, either more negative correlations or less positive correlations were observed between the C density and the temperature variables, while stronger positive correlations were found between the iM enrichment and the temperature variables (**Figure 2.16 B-C**). Among the eleven temperature variables, BIO5 (max temperature of the warmest month) and BIO10 (mean temperature of the warmest quarter) are the two variables with the highest correlations between the iM density and the temperature variables in the 5′UTR regions. Then we separately calculated PCCs for BIO5 and BIO10 separately in monocots and in other plant species (**Figure 2.16 D-G**). The PCCs of monocots are much higher than those in other plants (Monocots: BIO5=0.45, BIO10=0.32; Other plants: BIO5=0.27, BIO10=0.27), suggesting that the enrichment of iMs in 5′UTRs of monocots have much stronger associations with temperatures. We also compared the BIO5 and BIO10 values between monocots and other plants. The mean values of both BIO5 (monocots: 25.86°C; other plants: 24.35°C) and BIO10 (monocots: 24.17°C; other plants: 22.43°C) of monocots are significantly higher than those of other plant species (**Figure 2.17**). This result indicates monocots generally habit in relatively warmer areas than other plant species. The results above show that plants growing in relatively warmer areas tend to adopt more iMs across their 5′UTRs. This pattern is more evident in monocots. The enrichment of iMs in monocots might be the molecular marker during evolution, highlighting their habitat temperatures.

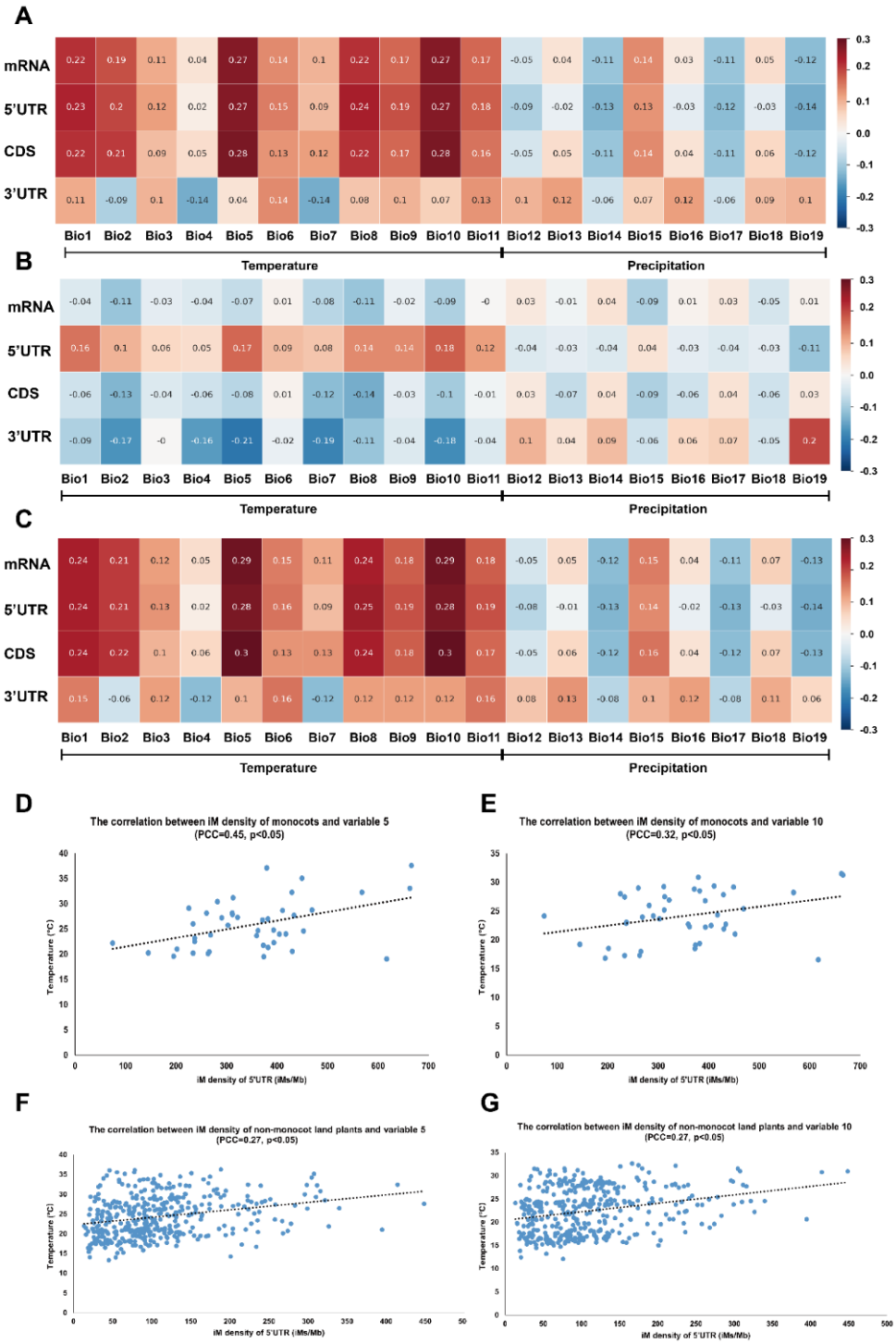


Figure 2. 16 The association between iM and environmental variables.
(**A**) The Pearson Correlation Coefficients (PCCs) between iM density and associated bioclimatic variables related to the habitats 433 land plants. (**B**) The heat plot showing the PCCs between cytosine density and associated bioclimatic variables related to the habitats of 433 land plants. (**C**) The heat plot showing the PCCs between iM enrichment (iMs/Mb cytosines) and associated bioclimatic variables related to the habitats of 433 land plants. (**D**) The PCCs between iM density and temperature-significant variables BIO5 (max temperature of the warmest month) across monocot land plants. (**E**) The PCCs between iM density and temperature-significant variables BIO10 (mean temperature of the warmest quarter) across monocot land plants. The PCCs between iM density and two temperature-significant variables BIO5 (**F**) and BIO10 (**G**) (BIO5: max temperature of the warmest month; BIO10: mean temperature of the warmest quarter) across non-monocot land plants.

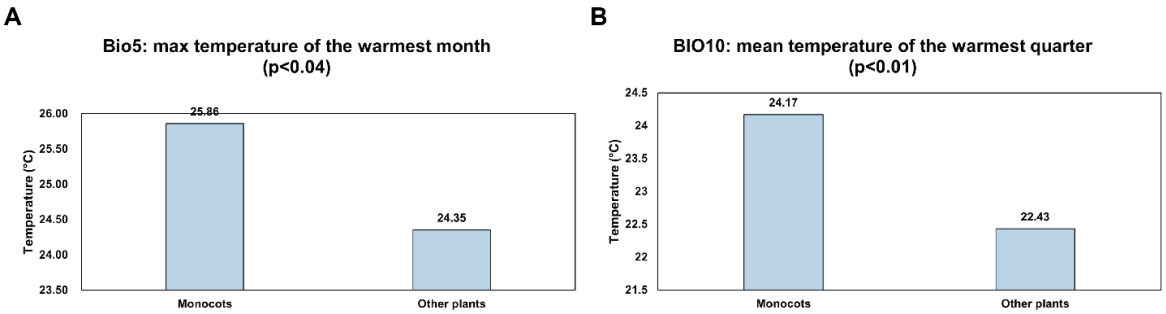


Figure 2. 17 The comparison of BIO5 and BIO10.
The mean value of BIO5 (**A**) and BIO10 (**B**) between monocots and other plants. Statistical analysis was performed between monocots and other plants with significance tested by Mann-Whitney *u*-test.

2.3.6 The molecular functions of iMs in plant 5′UTRs

Due to the significant enrichment of iMs in 5′UTRs, we further investigated the functional role of these iMs highly enriched in plant 5′UTRs. Previous studies revealed the huge impact of RNA structure in 5′UTRs on translation by blocking the ribosome loading, affecting pre-initiation complex scanning and start codon selection (Cao, et al., 2024; Zhang and Ding, 2025). The translation-related structures include the hairpin motifs and RNA GQSs, another non-canonical RNA structures (Yang, et al., 2020; Yang, et al., 2021). These priori knowledges inspired us to hypothesize that iMs may also affect translations. To validate our hypothesis, we collected the published ribosome profiling data for tomato (*Solanum lycopersicum*) and maize (*Zea mays*) and polysome profiling data for rice (*Oryza sativa* L. ssp. *japonica*) and Kronos wheat (*Triticum turgidum* ssp. *durum*). We re-analysed these data via mapping the raw reads to their corresponding transcriptomes, counting the FPKM and calculated the translation efficiency (TE). After the data processing, we then compared the TEs between genes with and without iMs in their 5′UTRs (**Figure 2.18**). In all four species,

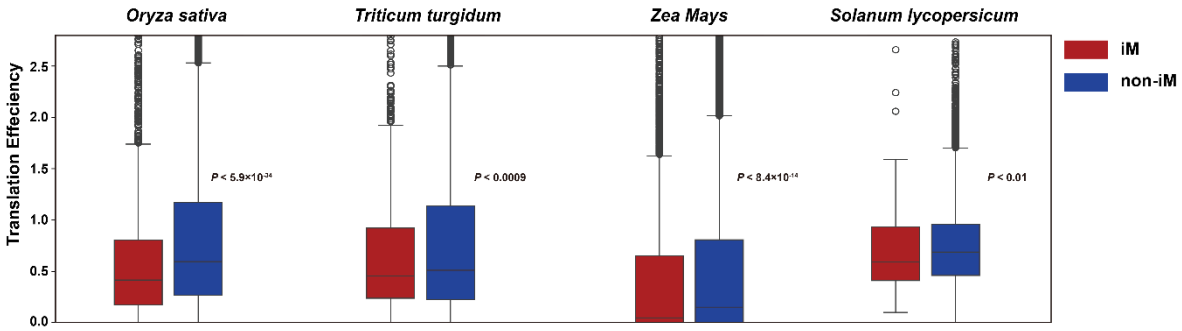


Figure 2. 18 The distribution of translation efficiencies across four plants.
The transcript group with iMs in 5′UTRs is marked in red while group without iMs in 5′UTRs in blue. Statistical analysis was performed with significance tested by Kolmogorov-Smirnov test.

the genes with iMs in their 5′UTR have significantly lower TEs than those without iMs, suggesting that iMs in 5′UTRs suppressing translation. The gene ontology analysis was further performed on these genes with iMs in their 5′UTRs across four species (**Figure 2.19**; **Figure 2.20**). We found that genes with iMs in 5′UTRs tend to be associated with fundamental biological functions related to protein phosphorylation and dephosphorylation,

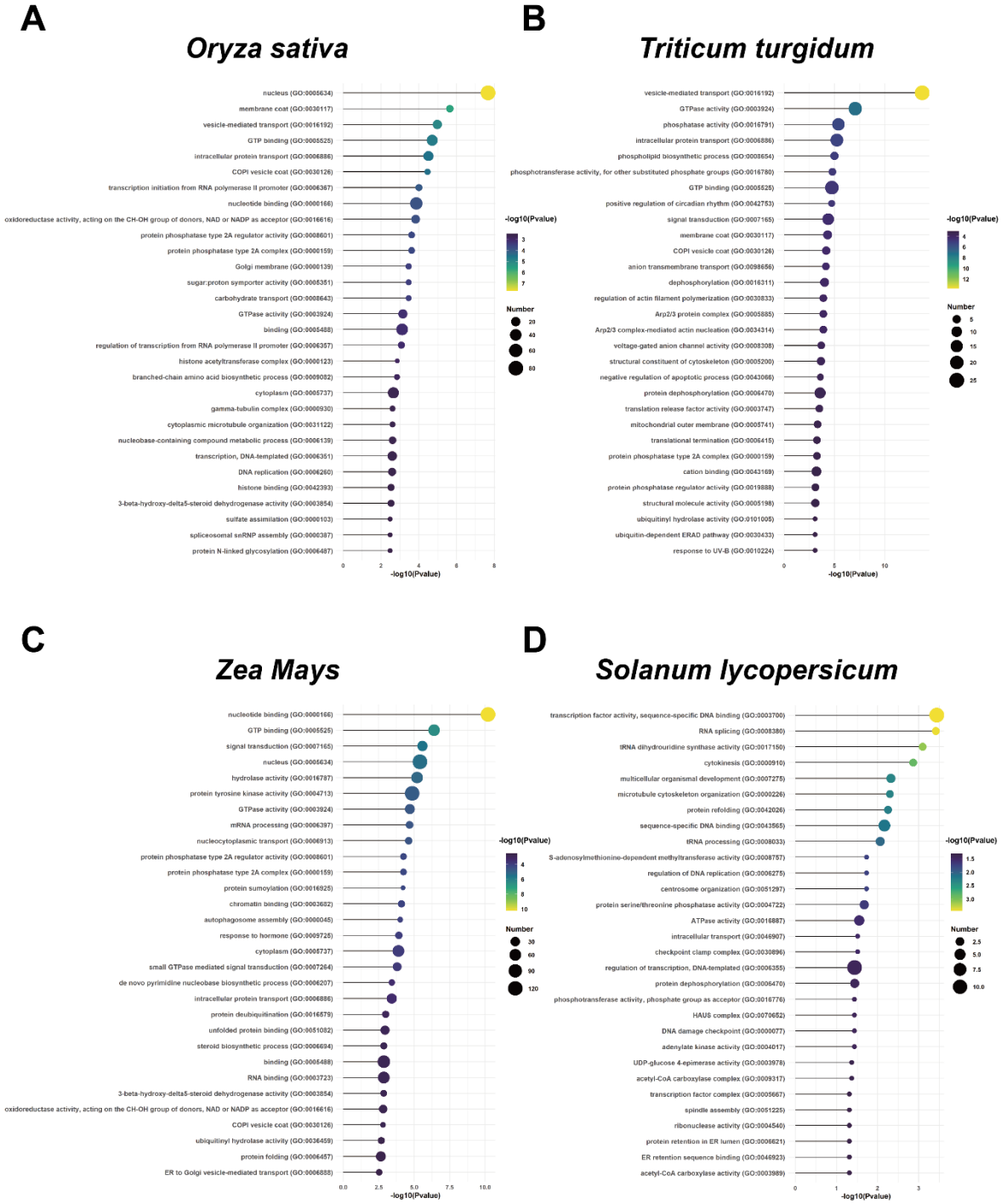


Figure 2. 19 The gene ontology items of RNAs with iMs in 5′UTRs across four plants. The top 30 enriched GO items from RNAs with iMs in 5′UTR for four species including rice (**A**), Kronos wheat (**B**), maize (**C**), and tomato (**D**).

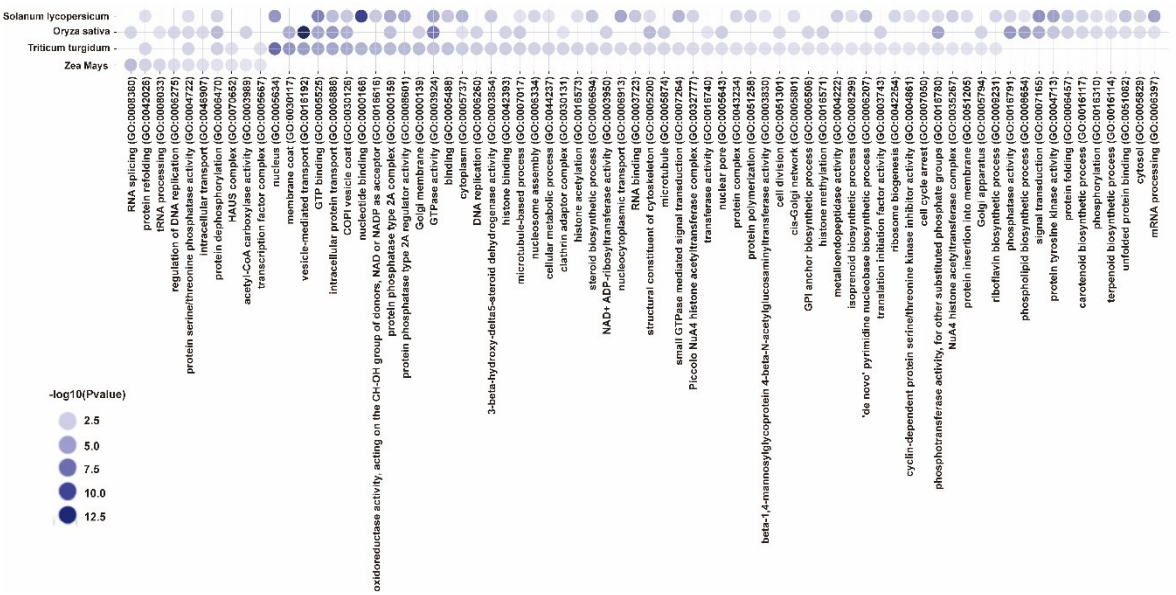


Figure 2. 20 The common gene ontology items of RNAs with iMs in 5′UTRs across four plants.

GTP binding and GTPase activity, intracellular protein transport, RNA splicing and so forth across four species (**Figure 2.20**). To characterize the iM-related features which may be important for the iM-mediated suppression of translation, we derived 34 different iM-related features according to individual iMs in the 5′UTRs in four species including 9 length-related features, 24 features regarding the densities of four nucleotides, and the iM folding strength (stability). We then calculated the mutual information (MI) and the spearman correlation coefficient (SPCC) for individual features and corresponding TEs. Moreover, we performed *u*-test between genes with high TE and low TE for all features to determine the iM features that are important for regulating translation. The iM features with MI (MI > 0.001) and *u*-test (P<0.05) are presented in **Figure 2.21**. Our results suggested that both the nucleotide density and loop length of iMs are important for their effects on translation. In details, the Adenine (A) and guanine (G) densities in the loop regions are more important than other feature in regulating translation. Notably, the nucleotide composition in the loops have been suggested critical to stabilize iM structures in previous studies (Fojtík and Vorlícková, 2001; Wright, et al., 2017) and our results in 2.3.3. Our results suggested that certain iMs with different A, G and C densities may affect their regulations of translation.

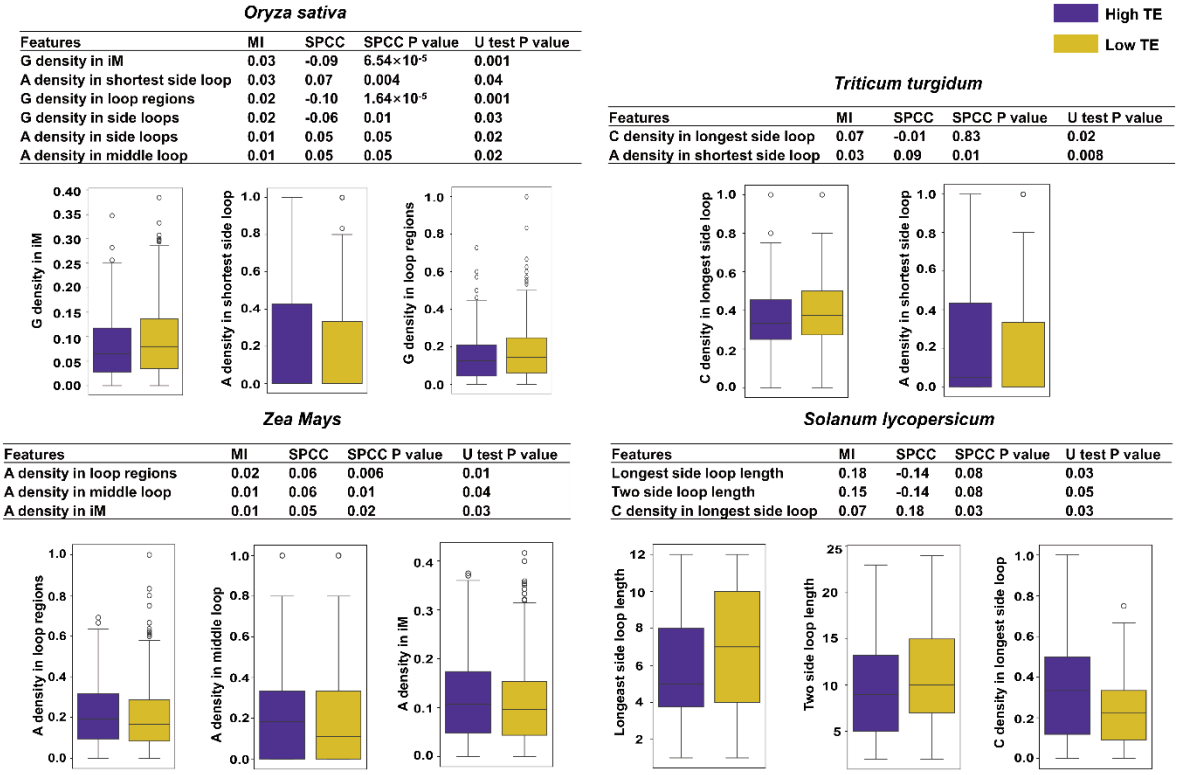


Figure 2. 21 The information of TE-associated iM features across four plants.
The assotiation between TE and iM features of four species (MI > 0; *P* value of Mann-Whitney *u* test <0.05). Statistical analysis of each iM features was performed between high TE group (purple) and low TE group (yellow) with significance tested by Mann-Whitney *u* test.

2.4 Summary and discussion

2.4.1 iM-Seeker, the most comprehensive computational iM-specific prediction platform to date

Here, we developed the most comprehensive computational iM-specific prediction platform to date, iM-Seeker, including a putative iM identification tool, Putative-iM-Searcher, two ensemble learning models for prediction of DNA iM folding status and folding strength, and an iM-Seeker webserver providing online service with automated machine learning functions. Previously, the iM prediction relies on the GQS-related computational tools due to the complementary sequence pattern between iM and GQS. However, iMs and GQSs are different structures with distinct biophysical properties. The prediction tools trained on the experimental data on GQSs are likely to result in bias and errors. Therefore, it is important to develop specific tools for predicting iMs using the iM-specific experimental data. Up to date, iM-Seeker is the only prediction tool integrated the iM-specific experimental data. We firstly focused on the identification of putative iMs. Since existing methods are unable to distinguish between different iM structural conformations, we developed Putative-iM-

Searcher, which treats input DNA/RNA sequences as direct graphs and systematically search all possible iM conformations via the graph traversal. Users can customize parameters such as the C-tract layer and the lengths of the first, second, and third loops. To extract representative conformations, the tool supports multiple strategies, including overlapping vs. non-overlapping, greedy vs. non-greedy, and specialized approaches for representative conformation identifications. Additionally, users have the option to retrieve all putative iM conformations. Compared to previous computational tools for identifying putative iMs, Putative-iM-Searcher provides clear and explicit information about all potential iM conformations within a given sequence, including the specific composition of C-tracts and loops. Although both prior methods and Putative-iM-Searcher can detect genomic regions containing putative iMs, Putative-iM-Searcher goes a further step by extracting detailed structural annotations. This additional information effectively assists users in defining and categorizing i-motif subtypes and characteristics, which also serves as a foundational basis for constructing most non-deep-learning predictive models. Beyond the sequence patterns, both folding status and folding strength are another two important indicators for studying iMs. We leveraged existing resources: genome-wide CUT&Tag sequencing data, which identifies folded iMs in human genome. These datasets were used to build a machine learning model to classify the folding status. We also utilized the biophysics experimental data such as pH_T to build a machine learning model to predict the iM folding strength. Our work represents the first application of machine learning to predict iMs. For each iM, thirty-three features including C-tract layers and nucleotide composition were extracted to build a machine learning model. While deep learning methods tend to have better performances in capturing unknown features, they require large, labelled datasets. Given the relatively small scale of available iM data, classical machine learning methods (e.g. tree-base models) are better suited due to their interpretability and robustness. iM-Seeker integrates ensemble learning on selected features to achieve strong performance in both classification and regression tasks. The choice of models typically depends on the dataset characteristics and how well the model aligns with its distribution. For classification, the dataset was quite imbalanced, with far fewer folded iMs (8,837) than unfolded motifs (733,115), leading to potential bias in standard classifiers. To overcome this, we employed the Balanced Random Forest algorithm, which uses the under-sampling strategy within a decision-tree-based ensemble learning framework to mitigate overfitting to the majority class. The model achieved recall and specificity rates of 77% and 81%, respectively, demonstrating strong predictive accuracy for both positive and negative samples. For regression, we utilized XGBoost, a widely adopted ensemble learning regressor in computational biology. The

model provided reliable predictions of iM folding strength, achieving an R^2 of 0.642, RMSE of 0.104, and MAE of 0.08 on the test set, highlighting its strong generalization capability. In summary, as iM-Seeker incorporates all currently available experimental data related to iMs, it provides information and predictive capabilities that are inherently incomparable to GQS-based tools, which are designed solely around G-quadruplex data and methods. We leveraged CUT&Tag data from the human genome to distinguish genomic regions with folded i-motifs from regions without folded iMs and subsequently searched these regions for putative iM sequences to train a classification model. This represents the first computational tool specifically developed for i-motif prediction. However, a potential limitation lies in the resolution of CUT&Tag data: the iMab antibody may bind relatively extended genomic regions containing long C-rich segments. To ensure that the model learns well-defined features, we used Putative-iM-Searcher to obtain clearer structural annotations of iM conformations. As a trade-off, our current model does not explicitly account for potential influences from flanking sequences upstream or downstream of the motif. Future work will focus on employing deep-learning architectures that integrate longer C-rich sequences and their surrounding contexts to build more accurate models and uncover deeper sequence–structure relationships. Unlike previous tools, iM-Seeker can also provide estimation of iM stability (folding strength), a feature enabled by iM-specific biophysical data. The current version is trained on over one hundred high-confidence data points. In the future, when more experimental data become available, the diversity of iM composition will be better embedded into the model parameters, which is expected to improve predictive performance.

In addition to predictive performance, model interpretation also highlighted several informative features. From the classifier trained on *in vivo* DNA iM folding status, we observed that folded iMs are enriched in G and C, whereas unfolded sequences show higher T and A content. Previous biophysical studies showed that A and G are negatively associated with iM stability, while C and T corelate with iM stability positively. In contrast, the regressor based on the *in vitro* physical experimental data, we found that the more C-tract layers, the shorter loops, and the higher numbers of C and T were associated with stronger iM folding strength. The presences of G and A nucleotides across iMs was associated with weaker iMs. These findings align well with previous biophysical studies. The apparent paradox—where guanine (G) is known to destabilize iM formation *in vitro*, yet folded iMs *in vivo* do not minimize G content and even exhibit elevated G levels relative to unfolded ones—highlights the conceptual distinction between iM folding status and iM folding strength. Here, folding status refers to the predicted *in vivo* foldability derived from

experimental maps of folded versus unfolded iMs, whereas folding strength reflects thermodynamic stability measured by *in vitro* biophysical assays. These two properties are not necessarily correlated: iMs that fold *in vivo* are not always the most stable *in vitro*, and highly stable *in vitro* candidates (e.g., those with long C-tracts and T/A-rich loops) may not represent the functional forms present in cells. This disconnect may be explained by the proposed role of iMs as structural switches in DNA systems (Kang, et al., 2014; Niu, et al., 2018). For an iM to participate in dynamic regulatory transitions—folding and unfolding in response to cellular cues—excessive stability could be disadvantageous, as it would hinder the conformational plasticity required for switching. Thus, the presence of G, which modestly reduces iM stability, may reflect a biological requirement for moderate, rather than maximal, stability *in vivo*. In this context, iMs need only fold sufficiently to perform their function, without being locked in an overly rigid conformation. This balance between foldability and reversibility could be key to their physiological regulation and represents an important direction for further functional characterization.

Our iM-Seeker webserver has been released to the public. Our iM-Seeker webserver is a well-designed and user-friendly webserver, which provides users with functions to predict putative iMs and extract corresponding biophysical properties. We also predicted iMs across the genomes of 30 species and revealed the diverse iM landscapes across different species. We made these data publicly available on the website for users to access and download. Furthermore, our webserver also integrates an AutoML session to offer an opportunity for users to upload their own biophysical data to build their own model automatically. With the increase of iM-related data, the incorporation of more diverse datasets will significantly improve the predictive accuracy. The AutoML function will provide the easy access to the users, particularly wet-bench users to build their own model by simply inputting their experimental results. The users do not need the training background in bioinformatics or AI. The uploaded data can be either DNA or RNA iMs with their corresponding biophysical labels. In summary, iM-Seeker represents the most comprehensive iM-specific prediction platform to date and is expected to offer new opportunities for studying DNA and RNA iM functions.

2.4.2 The functional role of iMs in plants

RNA is composed of four bases —A, U, C, and G— that fold into different RNA structures, facilitating diverse functions. A systematic study across plant kingdom revealed that higher AU densities than GC densities mostly appear across 5′UTR, CDS, and 3′UTR regions

(Yang, et al., 2022). Interestingly, some clades shown unusual patterns. For instance, both lycophytes and bryophytes exhibit the highest G densities in the 5' UTR and CDS regions, while C displays markedly higher density in the 5′UTR regions than in CDS and 3' UTR regions. Specifically, C is the least abundant nucleotide in CDS and 3′UTR, yet in the 5′UTR its relative abundance increases, even surpassing G in dicots and becoming the most frequent nucleotide in many monocots. These observations suggest that C enrichment in the 5′UTR regions may be evolved to play a unique regulatory role.

Our study indicated that iMs may help explain this phenomenon. Transcriptome-wide predictions revealed strong enrichment of iMs in 5′UTR regions across all plant species, with monocots showing the highest abundance among six clades (dicots, monocots, gymnosperms, ferns, lycophytes, and bryophytes). Previous studies have demonstrated that iMs are enriched in promoter regions at the DNA level (Zanin, et al., 2023). Although DNA and RNA are distinct molecules, promoter sequences and 5′UTR sequences can share similarities in nucleotide composition, which may introduce a potential bias in RNA analyses. However, it should be noted that in DNA studies, only a small subset of regions containing putative iMs has been experimentally validated as capable of folding *in vivo* (Zanin, et al., 2023). This implies that even if promoter and 5′UTR sequences exhibit compositional resemblance, the enrichment observed for most putative iMs cannot be functionally attributed to promoter-specific DNA activities. Investigating this distinction represents an important direction for future work. The iM enrichment results in 5′UTR align with the nucleotide-level patterns described above. Correlation analyses further suggest that iM density in 5′UTR regions is associated with the corresponding temperature variables of individual plant species. Notably, the RNA G-quadruplexes (GQS) is suggested to be evolved in plant adaptation to cooler environments (Yang, et al., 2022). Our results suggest that iMs might be evolved in plant adaptation to warm environments.

Strong RNA secondary structures in 5′UTRs tend to suppress translation (Kwok, et al., 2015; Yang, et al., 2021). Our analysis on the existed polysome profiling and ribosome profiling dataset revealed that the presence of iMs in 5′UTRs tends to repress translation, consistent with effects of other strong RNA secondary structures and tertiary structures such as GQS. Notably, translation suppression did not strongly correlate with iM folding strengths. Previous DNA studies have shown that iMs can act as structural switches, dynamically interconverting with alternative structures to regulate transcription (Kang, et al., 2014). For example, the BCL2 (B-Cell Leukemia/Lymphoma 2) promoter in human contains an iM

whose transition between folded and alternative hairpin structures modulates the transcriptional activity (Kang, et al., 2014). Another example is found in the promoter region of the lepidopteran gene *BmPOUM2* from *Bombyx mori*. iM formation in the C-rich region is triggered only after the transcription machinery has been assembled, driven by negative supercoiling. The resulting i-motif then recruits a transcription factor or DNA-binding protein, which activates the transcription complex and promotes the initiation and elongation phases. Upon transcription completion, the DNA reverts to its canonical duplex state (Niu, et al., 2018). RNA iMs, as one strand without the force from the complementary stand, are even more prone to folding, further supporting their role as dynamic regulatory elements (Deep, et al., 2025). Taken together, our findings suggest that plants may regulate gene expression through iM folding in their adaptation to the distinct environmental stresses.

Translation is as one of the most essential biological processes at the molecular level. Previous studies have identified a wide range of factors influencing translation efficiency (Schuller and Green, 2018). Within the 5′UTR alone, these features include upstream open reading frames (uORFs), sequence motif (e.g., the Kozak sequence), GC content, RNA modifications, RNA structure, and so forth (Schuller and Green, 2018). These elements are often interconnected and may influence one another. For instance, regions with high GC content are more likely to form stable RNA structures, including hairpin structures with more G-C pairs, as well as non-canonical structures such as GQSs and iMs. Both hairpin structures and GQSs have been shown to affect translation (Yang, et al., 2020; Yu, et al., 2024). Although our meta-analyses suggest that iMs may play a repressive role in translation, it remains difficult to disentangle their effects from those of other overlapping regulatory features. Therefore, an important future direction is to develop sensitive experimental approaches capable of identifying high-confidence RNA i-motifs in functional contexts.

2.5 Other contributors to the work described in this chapter

I acknowledge all the internal and external collaborators in this chapter. The project was in the collaborations with Prof. Zoë Waller's group [University College London, UK], and Prof. Ke Li's group [University of Exeter, UK]. All the talented collaborators who have contributed to work in this chapter are list below:

- Dr. Dilek Guneri [University College London, UK] contributed to generating biophysical data for iM model development (mainly).

- Dr. Elisé P Wright [University College London, UK, now University of Western Sydney, Australia] contributed to generate the biophysical data for iM model development (partly).
- Wenqian Chen [University College London, UK] contributed to collect the published biophysical data for iM model development (partly).
- Dr. Haopeng Yu [Join Innes Centre, UK] contributed to design the Automated machine learning session and build the webserver (partly).
- Dr. Fan Li [University of Exeter, UK] contributed to design the Automated machine learning session (partly).
- Dr. Yiman Qi [Join Innes Centre, UK] contributed to build the iM-Seeker webserver (partly).

Chapter 3 Discover RNA-stability-related RNA structure motifs in Kronos wheat

3.1 Introduction

mRNA stability is a key regulator of post-transcriptional activity and plays a central role in shaping gene expression. A variety of factors are known to influence mRNA decay in different organisms, including RNA length, sequence contents, RNA methylation, codon usage, and RNA secondary structures (Chen and Shyu, 1995; Feng and Niu, 2007; He and He, 2021; Mauger, et al., 2019; Mishima and Tomari, 2016; Presnyak, et al., 2015). For instance, in *E. coli*, shorter mRNAs generally exhibit enhanced stability (Feng and Niu, 2007). In contrast, in zebrafish, the long 3′UTR can reduce codon-mediated deadenylation, thereby promoting transcript stability (Mishima and Tomari, 2016). Additionally, sequence features can also impact RNA degradation, including the AU-rich elements (AREs) in 3′UTR, the miRNA binding site, polyA length, GU-rich elements, and so forth (Chen and Shyu, 1995; Presnyak, et al., 2015; Yang, et al., 2020). These features modulate interactions with RNA-binding proteins and other RNAs, which can either promote or inhibit stability. RNA methylation is another important factor. For example, N^6-methyladenosine (m^6A)-modified RNAs interact with YTHDF2 proteins, that promote both deadenylation and endoribonucleolytic cleavage (He and He, 2021). Codon optimality also influences decay, with mRNAs enriched in non-optimal codons showing reduced stability in yeast (Mishima and Tomari, 2016; Presnyak, et al., 2015). Finally, RNA secondary structures, particularly within the 3′UTRs, strongly influence RNA stability (Mauger, et al., 2019; Mortimer, et al., 2014). Genome-wide RNA structure profiling has revealed that stable 3′UTR structures can enhance RNA stability across many species including human, yeast, and *Arabidopsis*, and rice (Geisberg, et al., 2014; Goodarzi, et al., 2012; Su, et al., 2018; Wu and Bartel, 2017; Yang, et al., 2020). It might be partially because strong RNA structures block decay-promoting proteins, such as STAU1 in humans, which recognizes stem-loop motifs (Kim, et al., 2014). However, contrasting evidence exists. For example, a study in *Arabidopsis* found that weak RNA structures may enhance RNA stability, highlighting the complexity of RNA structural impacts on RNA stability (Zhang, et al., 2024).

Wheat is one of the most crucial crops in the world, providing 20% of the world's food supply. It is cultivated on more land area than any other crop and is one of the most primary sources of carbohydrates. Therefore, wheat plays a crucial role in global food security, international

trade, and rural economies. It is very important to fundamentally understand how to regulate gene expression in wheat. The stability of RNA plays important role in the control of gene expression. However, the transcriptome-wide RNA stability landscape is still lacking, meaning any assumption that factors identified in other systems (described above) are relevant lacks direct experimental justification. Additionally, the direction of effect (stabilizing or destabilizing) for a given factor is not universally conserved and can be species-dependent. For instance, shorter transcript length correlates with increased stability in *E. coli* but decreased stability in zebrafish (Feng and Niu, 2007; Mishima and Tomari, 2016). This context-dependence makes it difficult to predict how these features regulate RNA stability in wheat. Therefore, a systematic investigation within wheat is necessary to define its unique RNA stability landscape and elucidate the precise role of sequence and structural features. Therefore, we raised the first biological question: what mRNA features determine RNA stability in wheat.

Wheat is also a model allopolyploid organism (Levy and Feldman, 2022). Each subgenome in an allopolyploid (e.g. wheat) derives from distinct ancestral species, and the hybridization process has driven extensive genomic alterations. Over evolutionary time, the subgenomes have become integrated and now function in a coordinated manner, though gene expression often shows subgenomic asymmetry. Wheat comprises multiple types, including diploid (AA), tetraploid (AABB), and hexaploidy (AABBDD) wheat. Previous studies revealed the imbalanced RNA abundances among homologous genes from different ancestral chromosomes in wheat (Ramírez-González, et al., 2018). Because RNA stability contributes to steady-state RNA levels, RNA degradation is also likely to occur in an asymmetric manner among different subgenomes. Therefore, we raised the second question: can the same stability-associated RNA features detected in question 1 also influence the asymmetry of RNA stability.

To address these questions, we used Kronos wheat, a tetraploid durum wheat cultivar (AABB) with two subgenomes (A and B), which provides a simplified genomic framework for analysis. We leveraged RNA degradation profiles from Kronos wheat generated by our collaborator (Prof. Huakun Zhang in Northeast Normal University, China) and previously published Kronos wheat RNA structurome dataset from our lab. Our analysis revealed that various mRNA features influence mRNA stability, with RNA secondary structures in the 3′UTRs emerging as the key regulator alongside RNA length, AU/GC content, codon usage and translation. Although meta-analyses have indicated that RNA secondary structure

influences RNA decay, the detailed mechanistic underpinnings of this relationship remain poorly understood. Previous research on the roles of RNA structure in other functional contexts—including transcription, translation, and splicing—has identified a set of functional structural motifs. These motifs have been instrumental in studying the molecular mechanisms in relevant biological contexts, as detailed in 1.2. Using an optimized pyTEISER approach, we identified multiple RNA secondary structure motifs associated with stability. Interestingly, these structure motifs showed asymmetrical enrichments between subgenomes, suggesting that RNA degradation is modulated by subgenome-specific structural features. Together, our findings indicate that homoeologous RNAs in wheat may be differentially regulated through asymmetric RNA structures, which influence degradation and contribute to subgenome asymmetry in gene expression.

3.2 Materials and methods

3.2.1 Analysis between RNA stability and RNA features.

Prof. Huakun Zhang in Northeast Normal University, China provided the degradation profile of Kronos wheat. The format is that each RNA has a corresponding RNA decay rate value. The RNA structure chemical reactivity of Kronos wheat was calculated using the previously published SHAPE-Structure-seq data in the lab, following the instruction (Yang, et al., 2021), in the method in 4.2.3. To explore the mRNA features which influence the decay rate in Kronos wheat, we select several factors according to related studies on other organisms described in 2.1 including RNA length, nucleotide composition (i.e., AU content and GC content), two RNA structure-based indicators per RNA (i.e., average SHAPE reactivity and average base-pairing probability), two codon-related indicators including cAI (codon adaptation index) and tAI (tRNA adaptation index), and translation efficiency (Chen and Shyu, 1995; Feng and Niu, 2007; He and He, 2021; Mauger, et al., 2019; Mishima and Tomari, 2016; Presnyak, et al., 2015). RNA methylation was not included in the analysis because corresponding data are not available for Kronos wheat.

The base-pairing probability was calculated using RNAfold with parameter '-p' and SHAPE constraint (Hofacker, 2003). The Codon Adaptation Index (cAI) is a widely used metric to assess the extent to which the codon usage of a gene matches the codon usage bias of a set of highly expressed reference genes within a given species (Carbone, et al., 2003). It provides an indirect estimate of the potential translational efficiency of a gene. To calculate cAI, a

codon weight w_i is first assigned to each codon i encoding an amino acid based on its relative frequency of usage in a set of highly expressed genes:

$$w_i = \frac{f_i}{f_{max}}$$

Where f_i is the observed frequency of codon i among synonymous codons in the reference set and f_{max} is the frequency of the most used synonymous codon for the corresponding amino acid. The cAI of an RNA with L codons is then computed as the geometric mean of the codon weights:

$$\text{cAI} = \left(\prod_{i=1}^{L} w_i \right)^{1/L}$$

A cAI value closer to 1 indicates a codon usage pattern more like that of highly expressed genes, suggesting higher potential translational efficiency. The tRNA Adaptation Index (tAI) quantifies the degree of adaptation of a gene's codon usage to the tRNA pool of the organism, reflecting the translational selection pressure on codon choice (Man, et al., 2005). It incorporates both the tRNA gene copy number (as a proxy for tRNA abundance) and wobble base pairing rules. The relative adaptiveness w_i of each codon i is calculated as:

$$w_i = \sum_{j=1}^{n_i} \left(n_{tRNA_j} \cdot s_j \right)$$

Where n_i is the number of tRNA species that can recognize codon i, n_{tRNA_j} is the gene copy number of the j tRNA species recognizing codon i, $s_j \in (0,1]$ is a selective constraint value reflecting the efficiency of wobble pairing between the tRNA and the codon (Sabi and Tuller, 2014). After normalizing w_i by the maximum w value among synonymous codons for each amino acid, the tAI of an RNA with L codons for a gene is computed analogously to cAI:

$$\text{tAI} = \left(\prod_{i=1}^{L} w_i \right)^{1/L}$$

tAI values range from 0 to 1, with higher values indicating greater adaptation to the cellular tRNA repertoire and potentially more efficient translation. The cAI was calculated using python CAI package (Lee, 2018) and tAI was calculated using stAIcalc software (Sabi, et al., 2017). The translation efficiency was calculated using the same data and methods as described in 2.2.6.

3.2.2 The overview of optimised pyTEISER

The pyTEISER (python-implemented TEISER) is designed to find the functional RNA secondary structure motifs (basic hairpin structure motifs) according to multiple transcriptome-wide measurements and has already been applied to RNA stability and RNA splicing studies (Fish, et al., 2021; Goodarzi, et al., 2012). This is also one of the two computational pipelines that can detect motif elements relative to different biological processes with quantitative transcriptomic measures as inputs.

Firstly, pyTEISER defined and generated the structure motif seeds utilising context-free grammar. A seed is a sequence representing a stem-loop structure following three criteria: (1) The stem length is between 4 and 7 nucleotides with the loop length between 4 and 9 nucleotides. (2) The non-degenerate bases (A, U, C, and G) in the left stem sequence and loop sequence need to be between 4 and 6. And correspondingly, all the nucleotides in the left stem were matched with nucleotides in the right stems (i.e., U-R, C-G, G-Y, A-U, N-N), where Y represents UC, R represents AG, and N represents AGCU. (3) An information-estimated parameter, information content needs to be between 14 and 20 bits and is defined as: $Information\ Content = -\sum_{i\ \in sequence} \log_2 P_i$ where i means nucleotides and sequence represents the left stem and loop of the seeds. The P_i is given by:

$$P_i = \begin{cases} 0.25 \times 0.5 & if\ i\ is\ U\ or\ G\ in\ left\ stem \\ 0.25 \times 0.25 & if\ i\ is\ C\ or\ A\ in\ left\ stem \\ 2 \times (0.25 \times 0.25 + 0.25 \times 0.5) & if\ i\ is\ N\ in\ left\ stem \\ 0.25 & if\ i\ is\ (A, G, C, U, N)\ in\ loop \end{cases}$$

An example seed would have a sequence of ‘NUUN-NGNG-NRRN’ and a secondary structure of ‘<<<<....>>>>’. The number of non-degenerate bases in the left stem and loop is 4 and the information content is 16.83, so it is a qualified seed. At last, there are about 70 million seeds resulting from these criteria. The limitation is only typical stem-loop structures were included in the library. To increase the diversity of the seed library, we expanded the library. We used the FOREST approach to define *in vivo* RNA structure motifs of durum wheat (Komatsu, et al., 2020). We performed local folding of the transcriptome 3′UTR sequences with both valid RNA degradation data and *in vivo* RNA structure (RPKM of *in vivo* SHAPE-seq > 1) and 3′UTR longer than 30 nt using RNAfold, constrained by *in vivo* SHAPE-Structure-seq data with the code ‘RNAfold --shape=SHAPE_FILE --maxBPspan=30’. After obtaining the local RNA structure of the 3′UTR sequence, the FOREST program was used to identify all the single-terminal RNA structure motifs across 3′UTRs (code: ‘python FOREST.py -L 30’) (Komatsu, et al., 2020). The FOREST method can derive the stem-loop structures with bulges or internal loops, like "<<<<....>>.>>",

expanding the range of RNA structural patterns. The length of pyTEISER motif seeds generally falls within a range of approximately 12 to 23 nucleotides. To encompass more complex, non-canonical stem-loop motifs that may contain bulges or internal loops—which often require additional sequence length in single-stranded regions—we set an upper length threshold of 30 nucleotides for motif identification. It is important to clarify that this threshold defines the maximum allowable length for a motif, not a fixed or required length. Excessively high length thresholds can lead to motifs with very low enrichment frequencies, thereby reducing the statistical confidence in their biological relevance. Based on the structure of each RNA structure motif, 10,000 seeds were extracted. These seeds were generated randomly, following the criterion that the bases in the left stem region must match the corresponding nucleotide in the right stem (i.e., U-R, C-G, G-Y, A-U, Y-R, R-Y, K-N, M-K, S-B, W-D, B-N, D-N, H-D, V-B, N-N. Y represents UC, R represents AG, K represents UG, M represents AC, S represents GC, W represents AU, B represents GUC, D represents GAU, H represents ACU, V represents GCA, and N represents AGCU). The seed generation process resulted in approximately 30 million seeds for subsequent statistical testing and mutual information measurement.

After the seed generation, pyTEISER uses seeds to scan all the input RNA sequences. After the scanning, we performed a first round of filtering to reassess whether the RNA structure represented by each seed would fold *in vivo*. RNA sequence totalling 100 nt upstream and downstream of the RNA structure seeds were presented and constrained with *in vivo* SHAPE reactivities to calculate the free energy ΔG_{raw}. Next, we forced the region where the RNA structure motif seed was located to fold into the corresponding RNA structure and calculated the free energy ΔG_{fold}. The location that RNA structure motif seeds were set as the real folding location if $\Delta G_{fold}/\Delta G_{raw} > 0.5$ (Fish, et al., 2021). Seeds which can be found in more than 30 transcripts were kept for further investigation. After the scanning and *in vivo* filtration, each seed will have a binary vector called ‘seed occurrence profile’ as output. If the seed can be found in the k-th sequence, the k-th position in the seed occurrence is set to ‘True’ otherwise, ‘False’. After that, the sequences are divided into 15 equal-sized bins (recommended by pyTEISER tutorial) according to their continuous genome-wide measurements (decay rate) so that the Mutual Information (MI) calculation can be performed to capture the dependency between seed occurrence profiles and transcriptomic measurements for every seed individual using the equation below:

$$MI(R;A) = \sum_{r\in R}\sum_{a\in A} p(r,a)log\left(\frac{p(r,a)}{p(r)p(a)}\right)$$

Where R represents the presence and absence of an RNA structure motif and A represents the bins, for each decay group. Then, a series of randomization-based tests is used for the significance calculation and seed selection. The RNA sequences were divided into an equal number of bins of equal size as in the last step randomly, and the MI values were calculated. The process is repeated 1000 times and calculates the the Z-score for individual seeds using the following equation:

$$Z - score = \frac{MI(R,A) - \frac{1}{n}\sum\left(MI_{shuffled}(R,A)\right)}{\sigma_{MI_{shuffled}(R,A)}}$$

Where σ denotes the standard deviation of 1000 shuffled MI values and n is the number of randomization (n=1000). Then, RNA structure motifs with an MI greater than 0.001 and a Z-score greater than 2 were selected as degradation-related RNA structure motifs. Among them, if the frequency of the RNA structure motif is positively correlated with RNA degradation rate, then it is regarded as an RNA structure motif that accelerates RNA

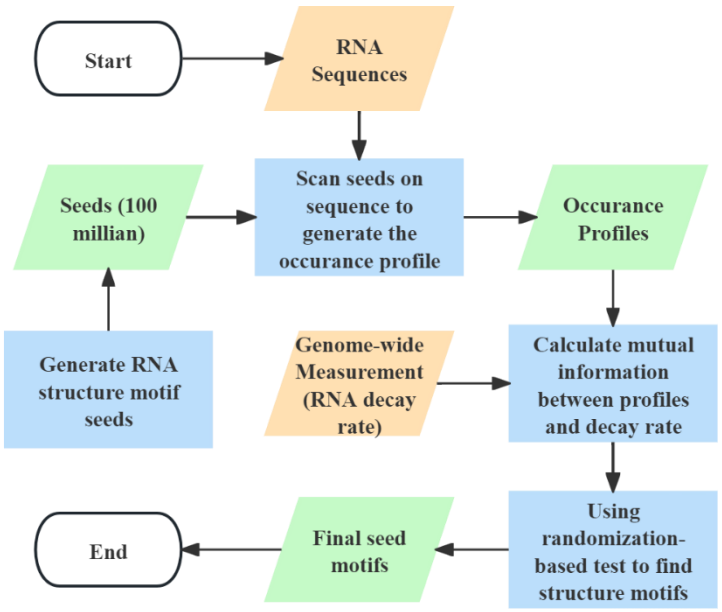


Figure 3. 1 The flow chart of optimised pyTEISER.

degradation, and vice versa for RNA stabilization. To expedite the program, we employed a parallel computing strategy and regular expressions to replace k-mer matching in the structural seed transcriptome-wide scanning step, which is very time-consuming. The flow chart of pyTEISER is shown in **Figure 3.1**.

To detect the structure motifs of over-representation or under-representation in each bin, the binomial distribution was used, following the instructions (Elemento, et al., 2007; Goodarzi, et al., 2012). Let N represent the total number of transcripts used in pyTEISER, and n denotes the number of transcripts containing a specific structure motif. For a target bin, let K be the total number of RNAs in the target bin, and x is the count of RNAs in the target bin that contain the motif. Assuming the null hypothesis that the motif is randomly and independently distributed in all the bins, the probability of observing at least x such RNAs with motif within the bin is computed using the binomial model with frequency $f = n/N$:

$$P(X \geq x) = \sum_{t=x}^{K} \binom{K}{t} f^t (1-f)^{K-t}$$

A motif is considered statistically over-represented in the target bin if the cumulative probability of observing x or more occurrences is below the Bonferroni-adjusted threshold $0.05/N_e$, where N_e is the number of bins (i.e., there are 15). Conversely, under-representation structure motifs are determined if the cumulative probability of observing up to x occurrences satisfies $P(0 \leq X \leq x) < 0.05/N_e$, applying the same binomial distribution principles. The calculation of binomial distribution was implemented via the binom function in the scipy package in Python.

To detect the structure motifs enriched in either A or B subgenomes, a hypergeometric distribution was used. Let N represent the total number of transcripts used in pyTEISER, n denotes the number of transcripts containing a specific structure motif, K represents the total number of transcripts with decay rate in subgenome A, and k donates the number of transcripts with target motifs in subgenome A. The significance (P-value) of observing at least k RNAs in subgenome A was computed using the function below:

$$P(X \geq k) = \sum_{x=k}^{K} \frac{\binom{n}{x}\binom{N-n}{K-x}}{\binom{N}{K}}$$

The same procedure was applied for group B. The calculation of the hypergeometric distribution was implemented via the hypergeom function in the scipy package in Python.

3.2.3 Plasmid construction and dual-luciferase reporter assay

The 3′UTR fragments containing three structural motifs (sRSM2, sRSM32, and sRSM15) were from *TRITD1Av1G039750* (sRSM2 and sRSM32) and *TRITD4Bv1G019060* (sRSM15), along with their disrupted and rescued mutant variants. These 3′UTRs were inserted into the inter2 expression vector and fused with Firefly luciferase reporter gene using the two digestion sites of BamHI and SmaI, with the In-Fusion cloning kit (Clontech).

Following the sequence confirmation using Sanger sequence platform, the recombinant constructs were transformed into *Agrobacterium tumefaciens* GV3101 and subsequently infiltrated into *Nicotiana benthamiana* leaves. In the tobacco (*Nicotiana benthamiana*) reporter assay, seedlings were harvested 48 h after agroinfiltration and sectioned into ~5 mm leaf disks, which were then incubated in cordycepin-containing buffer before infiltration. Gene expression was quantified using a dual-luciferase reporter assay according to the standard protocol (Yang, et al., 2021). Samples were collected at successive time points of 15, 30, 60, 120, 240, and 480 min.

3.2.4 Use of AI during the writing of the thesis

The Grammarly web (https://app.grammarly.com/) was used for grammar check.

3.3 Results

3.3.1 RNA structure might be the dominant factor for RNA stability

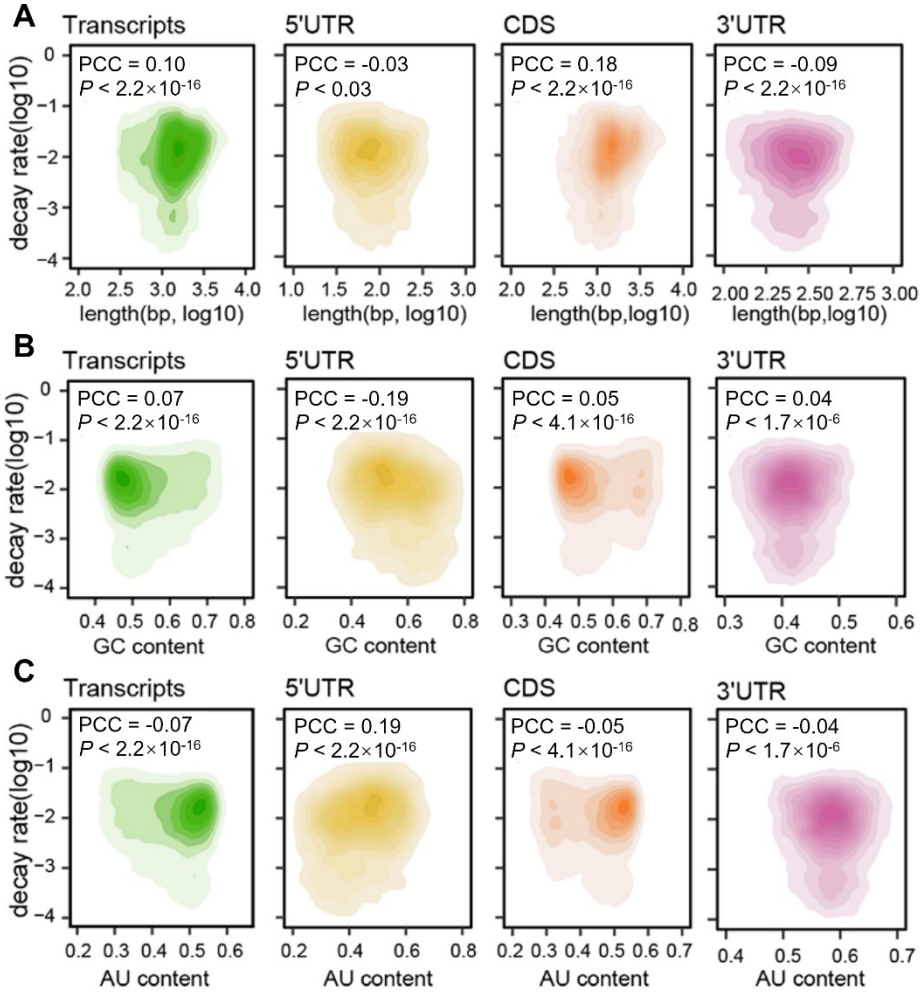


Figure 3. 2 The association between mRNA decay and sequence features in wheat. The Pearson correlation coefficient between RNA decay rate and sequence-based features including RNA length (**A**), GC content (**B**) and AU content (**C**) across different genetic regions.

We obtained RNA degradation profiles for Kronos wheat from our collaborators, including 33,430 transcripts with reliable decay rate measurements. Previous studies in various species have demonstrated that factors such as transcript length, GC/AU content, translation, codon, and RNA secondary structure can influence mRNA stability (Chen and Shyu, 1995; Feng and Niu, 2007; Hanson and Coller, 2018; Presnyak, et al., 2015; Sidorenko, et al., 2017; Zhang, et al., 2024). To investigate whether similar associations exist in Kronos wheat, we conducted a comprehensive analysis of mRNA features. Specifically, we computed Pearson correlation coefficients (PCCs) between mRNA decay rates and the lengths of distinct genic regions, including the 5′UTR, CDS, and 3′UTR. Our analysis revealed a relatively weak positive correlation between decay rate and transcript length, with the strongest association observed for coding sequence (CDS) length (**Figure 3.2 A**; PCC = 0.18, $P < 2.2 \times 10^{-16}$). Overall, we observed that longer mRNAs tend to degrade more rapidly, consistent with findings reported in other species (Geisberg, et al., 2014; Narsai, et al., 2007). In addition,

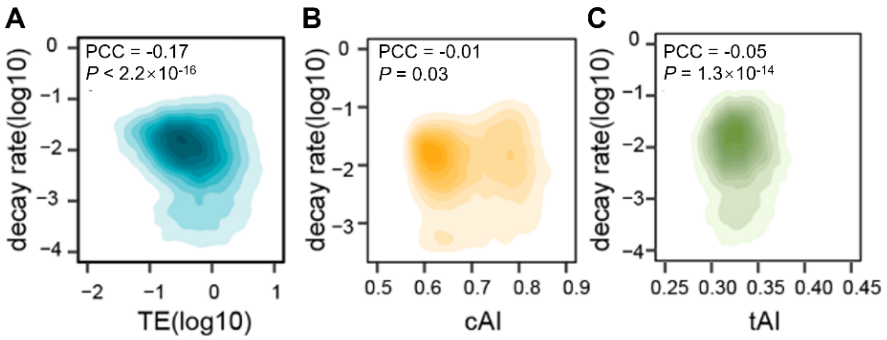


Figure 3. 3 The association between mRNA decay and translation in wheat.
The Pearson correlation coefficient between RNA decay rate and translation-related features including translation efficiency (**A**), cAI (**B**) and tAI (**C**).

GC content in the 5′UTR showed a weak negative correlation with RNA decay rates, whereas AU content in the 5′UTR exhibited a positive correlation (**Figure 3.2 B-C**), suggesting that sequence composition within the 5′UTR may influence transcript stability. Given that recent studies have indicated codon-dependent effects of translation on mRNA stability (Hanson and Coller, 2018), we next examined the relationship between translation efficiency (TE) and RNA decay. A weak negative correlation was observed, indicating that transcripts with higher TE tend to be more stable (**Figure 3.3 A**; PCC = -0.17, $P < 2.2 \times 10^{-16}$). To further assess codon-related influences, we analysed the codon adaptation index (cAI) and tRNA adaptation index (tAI) in relation to decay rates. While cAI showed no strong correlation with RNA stability (PCC = -0.01, $P = 0.03$), tAI exhibited the relatively stronger negative correlation than cAI (PCC = -0.05, $P = 1.3 \times 10^{-14}$) (**Figure 3.3 B-C**). Collectively, these results suggest that mRNA length, sequence features, translational efficiency, and codon usage contribute subtly to the regulation of mRNA stability in wheat.

Previous studies in various species have shown that RNA structures within the 3′UTR are often linked to mRNA stability (Goodarzi, et al., 2012; Wu and Bartel, 2017; Zhang, et al., 2024). To investigate whether this relationship also exists in wheat, we used *in vivo* RNA

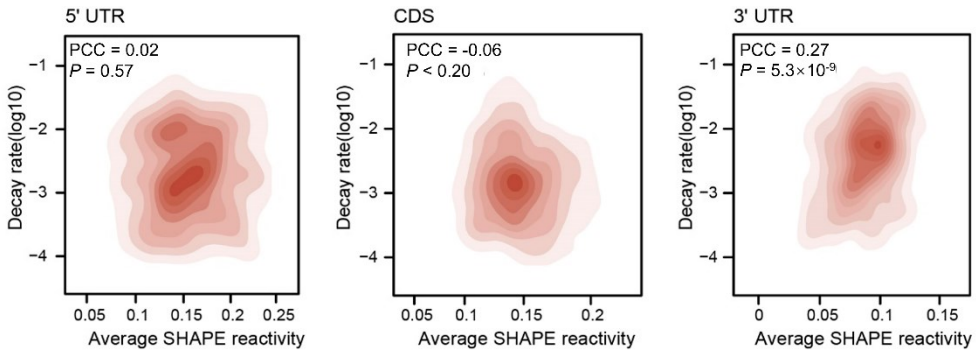


Figure 3. 4 The association between mRNA decay and chemical reactivity.

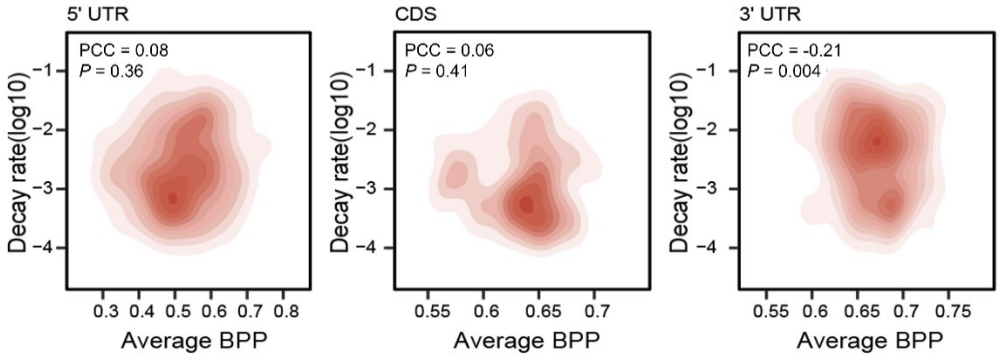


Figure 3. 5 The association between mRNA decay and base-pairing probability.

structurome data from Kronos wheat. SHAPE reactivities were derived from published SHAPE-Structure-seq datasets, yielding reliable structural information for 24,486 transcripts. Among these, 10,124 transcripts had both high-confidence SHAPE reactivities and corresponding RNA decay rates. SHAPE reactivities, which reflect the single-strandedness of individual nucleotides, were averaged across different genic regions and correlated with mRNA decay rates using Pearson correlation coefficients (PCCs). We observed a significant positive correlation between decay rates and average SHAPE reactivities specifically in the 3′UTRs (**Figure 3.4**; PCC = 0.27, $P = 5.3 \times 10^{-9}$), suggesting that more single-stranded regions in the 3′UTR are associated with faster mRNA degradation. To further validate this observation, we computed base-pairing probabilities (BPPs) for each nucleotide based on the *in vivo* structure profiles. Consistent with the SHAPE data, a significant negative correlation was found between average BPPs in the 3′UTRs and mRNA decay rates (**Figure 3.5**; PCC = -0.21, $P = 0.004$), further supporting the idea that structured regions within the 3′UTR contribute to mRNA stabilization.

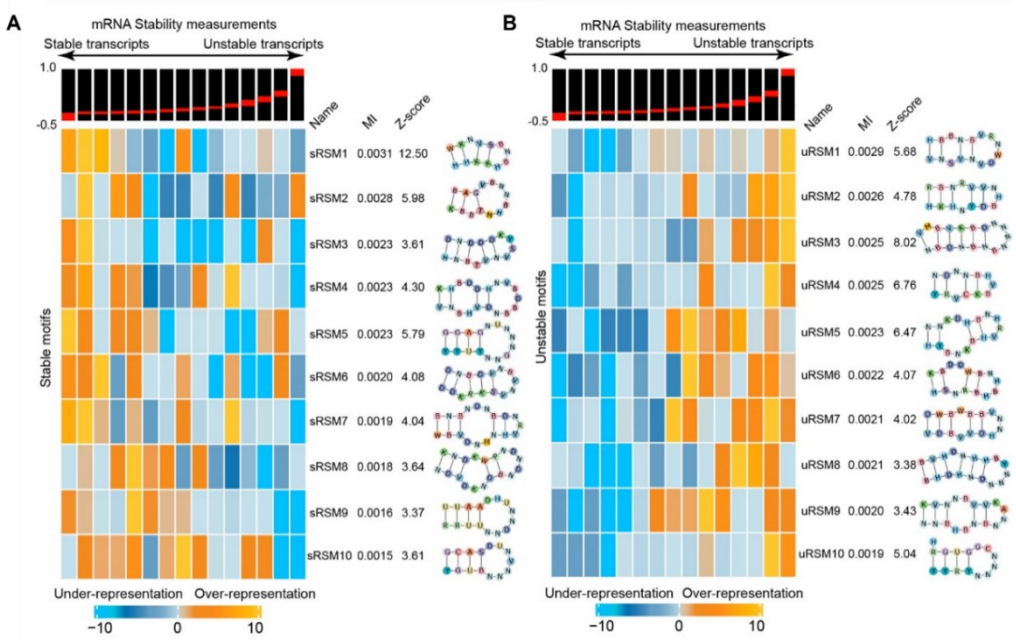


Figure 3. 6 The RNA secondary structure motifs contributing to mRNA stability. Representative stable RNA structure motifs (sRSMs) (**A**) and Representative unstable RNA structure motifs (uRSMs) (**B**). Transcripts were grouped into 15 equal-sized bins based on their stability scores, ranging from highly stable (left) to highly unstable (right), as indicated by the red trend line. The heatmap displays the enrichment patterns of ten selected motifs across these bins, where the colour intensity reflects their relative enrichment.

3.3.2 The discovery of RNA stability RNA structure motifs using optimised pyTEISER

Building on these findings, we conducted a systematic identification of RNA structure motifs linked to mRNA stability in wheat using the computational tool pyTEISER (Pythonic Tool for Eliciting Informative Structural Elements in RNAs) (Fish, et al., 2021). As introduced in Section 1.3.2, pyTEISER is a robust framework designed to uncover functional RNA secondary structure motifs. It initially generates a library of approximately 70 million canonical RNA stem-loop motifs, incorporating both non-degenerate and degenerate nucleotides. These motifs are subsequently subjected to multiple rounds of statistical filtering to identify those significantly associated with regulatory functions. However, a known limitation of pyTEISER is its exclusion of non-canonical stem-loop motifs, such as motifs containing bulges or internal loops, from the initial seed library. To address this gap, we implemented an additional pipeline to expand the motif library. Specifically, we employed RNAfold to predict local secondary structures within the wheat transcriptome and then used FOREST software to extract structural elements that include bulges or internal loops. From each motif category, 10,000 representative seeds were randomly sampled, resulting in a collection of approximately 30 million non-canonical RNA structural seeds. By integrating the canonical and non-canonical motif seed libraries, we applied mutual information (MI) and MI-based statistical filtering to identify RNA secondary structure

motifs associated with mRNA stability. This analysis led to the identification of 118 high-confidence structure motifs within the 3′UTRs that met the thresholds (**Figure 3.6**; **Supplementary Tables 3-4**). Among these, 64 motifs were designated as sRSMs (stable RNA structure motifs) due to their positive association with mRNA stability, while 54 were classified as uRSMs (unstable RNA structure motifs) based on their negative correlation with stability (**Figure 3.6**; **Supplementary Tables 3-4**). These results underscore the importance of RNA secondary structure in modulating mRNA stability in wheat.

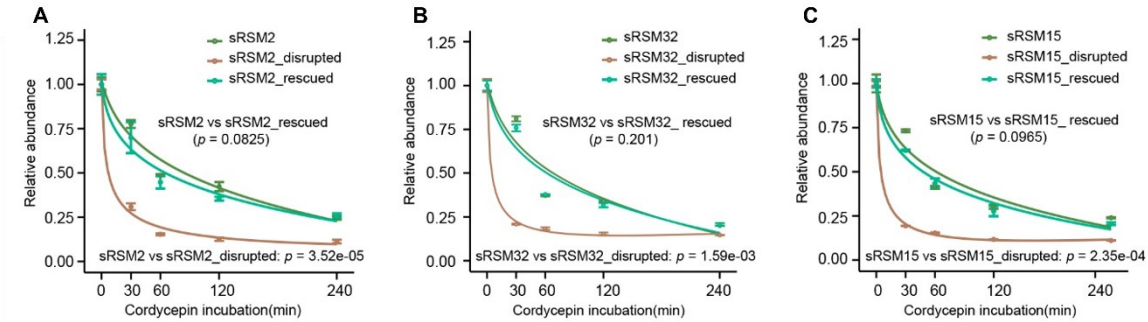


Figure 3. 7 The justification of the decay rate of original, rescued, or disrupted motifs of sRSMs.
For each motif, decay curves were compared between three construct types: those retaining the native motif structure (dark green), those in which the structure was restored through sequence redesign (light green), and those in which the motif was disrupted (brown) for sRSM2 (**A**), sRSM32 (**B**), and sRSM15 (**C**). The significance is calculated by one-sided repeated measures ANOVA test.

To further validate the functional relevance of the identified stable RNA structure motifs (sRSMs), we synthesized the 3′UTR fragments corresponding to three representative sRSMs (sRSM2, sRSM15, and sRSM32) and designed mutated variants in which the motifs were either structurally disrupted or rescued through sequence redesign. Specifically, mutations will be introduced within the motif-containing region. The first type of mutations is designed to disrupt motif formation, thereby preventing the structure from folding. The second type aims to restore the structural conformation of the motif. We performed the dual-luciferase reporter assay using these 3′UTR fragments on cordycepin-treated tobacco leaves to measure decay rate by measuring the RNA abundance at different time points. By constructing Firefly reporter fusions with the 3′UTRs fragments, together with an identical Renilla luciferase control, constructs carrying disrupted motif structures exhibited significantly accelerated decay compared to their wild-type counterparts (**Figure 3.7**). In contrast, the decay rates of the rescued constructs were markedly slower than those of the disrupted forms and largely comparable to the original sRSMs (**Figure 3.7**). This experimental justification provided additional evidence that the sRSMs can play a role in regulating RNA stability. Compared

to sRSMs, the experimental validation of uRSMs is more challenging. Previous studies generally revealed that a strong, stable structure (e.g., stem-loop) is more likely to inhibit RNA decay, which forms the basis for validating sRSMs (Geisberg, et al., 2014; Goodarzi, et al., 2012; Su, et al., 2018; Wu and Bartel, 2017; Yang, et al., 2020). The existence of uRSMs suggests that the mechanisms by which RNA structures influence stability may be diverse. The specific mechanism through which uRSMs promote decay remains unclear, which consequently constrains the design of validation experiments. For an uRSM, it is not yet known whether disrupting or reinforcing its structure would accelerate or suppress decay. Elucidating this direction represents a key objective for future work.

To investigate potential subgenome-specific preferences, we further analysed the enrichment of sRSMs and uRSMs across the A and B subgenomes. We observed distinct enrichment patterns, with certain sRSMs and uRSMs preferentially occurring in one subgenome over the other (**Figure 3.8**). In 64 stable motifs, 40 motifs were found to be enriched more in the A subgenome, while 24 were found to be more in the B subgenome. For unstable motifs, 29 were enriched more in A and 25 were distributed more in B. These results indicated the subgenome-level asymmetrical enrichment of the structure motifs, suggesting RNA structures are likely to regulate RNA stability in different subgenomes in wheat.

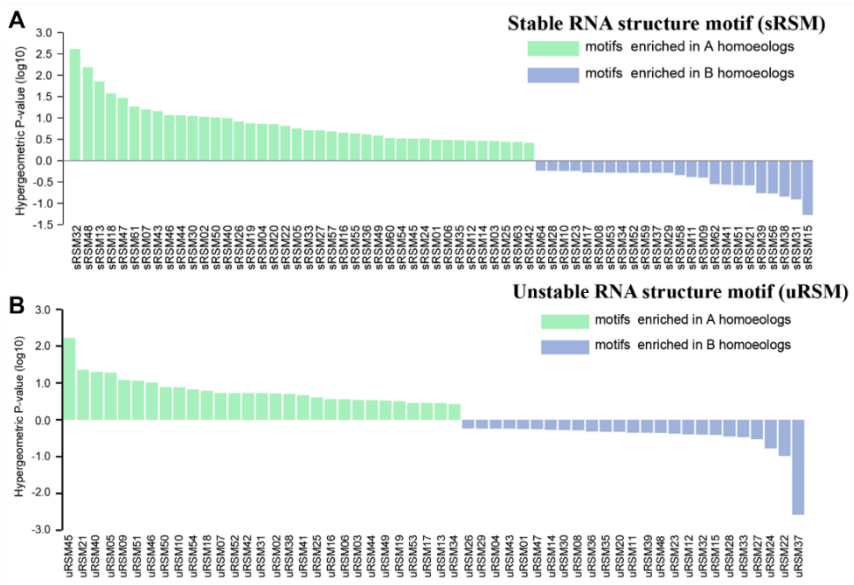


Figure 3. 8 The subgenome-level preference of RNA stability structure motifs.
The enrichments of the stable RNA structure motifs (A) and unstable RNA structure motifs (B) in the A and B subgenomes.

3.4 Summary and discussion

Our integrative analysis of diverse mRNA features, including transcript length, nucleotide composition, translation, codon, and RNA secondary structures, provides a comprehensive perspective on their roles in influencing mRNA stability. In wheat, we observed that transcripts with longer CDSs typically exhibit faster decay (**Figure 3.2 A**), likely due to an increased probability of premature termination codons (PTCs) that can activate nonsense-mediated decay pathways (Kurosaki and Maquat, 2016). Furthermore, nucleotide composition within the 5′UTR strongly impacts stability: transcripts with GC-rich 5′UTRs tend to be more stable, whereas those with AU-rich 5′UTRs degrade more rapidly (**Figure 3.2 B-C**). This observation contrasts with findings from *Arabidopsis*, where no significant link was reported between AU/GC content and mRNA decay across genic regions (Narsai, et al., 2007). Interestingly, our AU-related sequence findings are consistent with previous analyses in *Arabidopsis* that identified AU-rich motifs enriched in uncapped transcripts (Jiao, et al., 2008), suggesting that regulatory sequence elements may act synergistically rather than independently.

We also found a positive association between translational efficiency and mRNA stability, suggesting that ribosomes may shield transcripts from degradation enzymes (Roy and Jacobson, 2013). However, the relatively weak correlation suggests a more complex relationship between translation and decay mechanisms. In addition, *in vivo* RNA structure profiling in wheat revealed that SHAPE reactivities within 3′UTRs correlate positively with mRNA decay rates, implying that more single-stranded regions in the 3′UTR render transcripts less stable. This trend aligns with observations from *Arabidopsis*, rice, humans, and yeast (Geisberg, et al., 2014; Kharel, et al., 2023; Su, et al., 2018; Yang, et al., 2020). One likely explanation is that endonucleases preferentially bind single-stranded regions, making such RNAs more susceptible to cleavage (Chlebowski, et al., 2013; Houseley, et al., 2006). On the other hand, alternative *in vivo* structure mapping using chemical probing and mutation profiling has revealed an inverse relationship between 3′UTR structure and mRNA abundance in *Arabidopsis* and rice (Zhang, et al., 2024), suggesting that specific RNA structures may promote degradation via early polyadenylation. These inconsistencies may arise from methodological differences or reflect the functional diversity of 3′UTR structural elements involved in multiple regulatory events, including exonucleolytic decay, endonucleolytic cleavage, and processing (Jia, et al., 2021).

Notably, while the Pearson correlation coefficients (PCCs) between global structural or sequence features of the 5′UTR, CDS, and 3′UTR regions were relatively low in our analysis

(**Figures 3.2-3.5**), a critical consideration is the fundamental non-independence of these genic regions. They are not freely assorting variables but are co-linear, co-transcribed components of a single mRNA molecule. This intrinsic linkage creates multiple layers of interdependence that can drive observed results in ways not fully captured by linear correlation. Shared transcript-level labels (e.g., decay rate) could impose dependencies that are not captured by simple linear correlation. Therefore, the low PCC observed may reflect this complex, integrated biology rather than a complete absence of relationship. Our findings highlight associations within this constrained system, and future work employing models that account for the hierarchical structure of transcriptome could further partition the distinct contributions of each feature in different genic regions.

Using the statistical tool pyTEISER pipeline, which incorporates principles from information theory, we identified 118 RNA structure motifs significantly associated with mRNA stability. Among these, 64 motifs were classified as stability-promoting and 54 as destabilizing (**Figure 3.6**). Overall, we observed that mRNAs favour stable secondary structures over extended single-stranded conformations, likely to protect against unspecific endonucleolytic degradation. These regulatory structure motifs are distributed across the transcripts. Our genome-wide RNA structurome analysis in wheat thus defines a set of structural elements that directly modulate transcript stability.

Furthermore, we observed subgenome-specific enrichment of stability-related structure motifs (**Figure 3.8**). This indicates that, despite high sequence similarity between subgenomes, divergence in structural regulatory features may contribute to subgenomic differences in post-transcriptional control, highlighting a potential layer of evolutionary specialization.

3.5 Other contributors to the work described in this chapter

I acknowledge all the internal and external collaborators in this chapter. The project was in the collaboration with Prof. Huakun Zhang's group [Northeast Normal University, China]. All the talented collaborators who have contributed to work in this chapter are list below:

- Dr. Haidan Wu [Northeast Normal University, China] provided the RNA decay profile (decay rate) and contributed to perform luciferase reporter assay (mainly).
- Dr. Haopeng Yu [Join Innes Centre, UK] provided RNA structure profiles in previous research and contributed to data analysis (partly).

- Dr. Yueying Zhang [Join Innes Centre, UK] contributed to perform luciferase reporter assay (partly).

Chapter 4 Investigate RNA structure dynamics during tomato development

4.1 Introduction

RNA structures act as critical regulators and participate in a wide range of biological functions. Under different physiological environments, the same transcript can adopt distinct structural conformations in response to varying cellular signals, thereby fulfilling diverse functions. This dynamic property of RNAs, known as RNA structure switching, is a process that plays a crucial role in regulating many biological processes. A classic and well-studied example is the riboswitch. Riboswitches are cis-acting RNA structural elements that regulate gene expression by binding to metabolites (Pavlova, et al., 2019). They typically consist of two parts: the aptamer domain and the expression platform. The aptamer domain contains a structured region capable of specifically binding to a small molecule (e.g. thiamine pyrophosphate, cobalamin, flavin mononucleotide, and so forth). Ligand binding induces the changes of structural conformations in the expression platform, thereby leading to precise regulation of downstream biological functions (Kavita and Breaker, 2023). For example, the thiamine pyrophosphate (TPP) riboswitch, present across all three domains of life, regulates transcription and translation in bacteria and regulates RNA splicing in eukaryotes (Pavlova, et al., 2019; Wachter, et al., 2007). To date, more than 55 classes of riboswitches have been identified and experimentally validated across kingdoms (Kavita and Breaker, 2023).

Beyond riboswitches, other examples of RNA structure dynamics influence gene expression. Human 7SK snRNA, for instance, can toggle between two conformations that either bind or not bind P-TEFb, as a transcriptional switch. When 7SK binds P-TEFb, the low abundance of P-TEFb will suppress transcription globally (Olson, et al., 2022; Spitale and Incarnato, 2023). Many viruses also exploit RNA structural dynamics. Studies in viruses such as HIV-1 and SARS-CoV-2 have revealed that dynamic RNA structures impact many biological functions such as splicing, export, and translation (Manfredonia and Incarnato, 2021; Sherpa, et al., 2015; Spitale and Incarnato, 2023; Tomezsko, et al., 2020). Transcriptome-wide analyses on the RNA structure ensembles in *E. coli* identified RNA thermometers capable of switching RNA structural conformations in response to temperatures (Borovská, et al., 2025). In plants, G-quadruplexes (GQS) and i-motifs (iMs; described in Chapter 2) have also been proposed as temperature-sensitive RNA switches (Yang, et al., 2022).

Another well-studied aspect related to RNA structural dynamics is the riboSNitch. The riboSNitch refers to a functional RNA structural element whose conformation is altered by a single nucleotide variant (SNV) or single nucleotide polymorphism (SNP), thereby impacting gene regulation or function. Many riboSNitches have been identified, highlighting that RNA structure mediates the regulatory effects of genetic variations beyond coding changes. For instance, a riboSNitch in the *HBB* (beta-globin) gene alters RNA structure ensembles, leading to reduced beta-globin production and contributing to β-thalassemia (Halvorsen, et al., 2010). Similarly, riboSNitches in *MAPT* RNA, which encodes the Tau protein, affect splicing patterns, leading to neuronal disease. The riboSNitches in *MAPT* RNA can generate different RNA structure ensembles to regulate splicing, leading to neuron disease (Kumar, et al., 2022). In wheat, our lab previously identified riboSNitches that influence translation (Yang, et al., 2021) in a subgenome-specific manner. These cases highlight how single-nucleotide changes can reshape local or global RNA secondary structures, resulting in downstream regulatory effects.

Tomato (*Solanum lycopersicum*), one of the world's most important horticultural crops, is valued for both economic and nutritional reasons. As a major dietary source of vitamins, minerals, antioxidants, and dietary fibre, tomato plays a crucial role in human health. It also serves as an important model plant for studying fruit development, metabolism, and gene regulations. Tomato fruit ripening involves extensive physiological, biochemical, and structural changes governed by multiple regulatory molecules. RNA structure, as a key regulator, is likely to contribute during different developmental stages, yet its role in tomato ripening remains largely unexplored. Given the dynamic changes of metabolites throughout ripening, RNA structure may function analogously to riboswitches, adopting alternative structural conformations in response to changing levels of metabolites, and thereby regulating gene expression.

Based on these insights, we aimed to investigate the functional role of RNA structures in the tomato fruit development. We focused on two representative developmental stages, immature green and mature green, and generated the first transcriptome-wide RNA structure profiling libraries in tomato using chemical probing. To comprehensively compare the RNA structurome between two stages, we developed RSDE-Tool, a new computational pipeline designed to identify RNA structural dynamic elements (RSDEs) and representative conformations across multiple reactivity profiles. Unlike previous tools, which detect only short fragments of RNAs with significant differences of chemical reactivities or deconvolute

bulk signals into multiple chemical reactivity profiles (details reviewed in 1.3.3), RSDE-Tool identified RSDEs across multiple chemical reactivity profiles for generating representative RNA structural conformations. Using this approach, we identified ~12,000 RSDEs with representative RNA structural conformations. To explore their functional roles, we generated the polysome profiling libraries for the two stages and obtained the differential translation efficiencies. We revealed that the five RSDE clusters were enriched in 5′UTR-CDS junctions or CDSs that were associated with translational regulations. Taken together, these findings revealed extensive RNA structural heterogeneity during tomato development and suggested that dynamic RNA structures in 5′UTRs and CDSs impact translation across developmental stages.

4.2 Materials and methods

4.2.1 Plant materials

Tomato (*Solanum lycopersicum*) cv. Money Maker plants were grown in the greenhouse at an average ambient temperature of 20–22 °C. Supplemental lighting was available to maintain 16 h of light per day when necessary. Fruits at stages of immature green (IMG), mature green (MG) were harvested for further analysis.

4.2.2 Construction of SHAPE-Structure-seq libraries

Tomato fruits were harvested and sliced into 1mm-thick slices. Fruit pericarps were separated and thoroughly covered in 150mM NAI at 22 °C for 15 min with shaking at 500 rpm (the control group were treated with buffer without NAI). Five times of dithiothreitol was added with vigorous vortexing until completely dissolved to quench NAI. After washing three times with sterilized water, the fruit material was immediately frozen with liquid nitrogen and ground into fine powder with a mortar and pestle. Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen) following the manufacturer's instructions. Two independent biological replicates were conducted for each treatment to isolate RNA and construct libraries.

SHAPE-Structure-seq libraries were constructed following the protocol described in Ding *et al* (2015) with modifications. Two rounds of poly(A) selection were performed using a Poly(A) Purist MAG Kit (Ambion). PolyA-selected RNA was recovered and subjected to reverse transcription using Superscript III First-Stand Synthesis System (Invitrogen), with Random hexamer fused with Illumina sequencing adapter

(5′CAGACGTGTGCTCTTCCGATCTNNNNNN3’). cDNA was ligated with an ssDNA linker (5′-PhosNNNAGATCGGAAGAGCGTCGTGTAG-/3SpC3/3′) using Circligase ssDNA Ligase (Epicentre) following the manufacturer’s instructions. Ligation products longer than 100 nt were selected by TBE-Urea 10% Gel (Invitrogen), which were then purified by QIAquick Gel Extraction Kit (Qiagen). Purified cDNA was subjected to PCR amplification using KAPA Library Amplification Kits (Roche). Three rounds of agarose gel purification were performed to purify the fragments of 200~650 bp using QIAquick Gel Extraction Kit (Qiagen). The libraries were sequenced with the Illumina HiSeq 4000 platform by BGI Genomics.

4.2.3 Construction of polysome profiling and RNA-seq libraries

Polysome-associated mRNA analysis was performed as described in established protocol with modifications (Missra and von Arnim, 2014). Briefly, a sucrose gradient (1.9 ml of 60% sucrose, 3.8 ml of 45%, 3.8 ml of 30%, and 1.9 ml of 15%) was prepared and stored at -80 °C before use. Tomato fruit was harvested at the corresponding stage, and the pericarp was isolated and ground into fine powder in liquid nitrogen with a mortar and pestle. The powder was dissolved in 500 μL pre-cooled polysome extraction buffer (200 mM Tris-HCl, pH 8.4, 50 mM KCl, 25 mM MgCl2, 1% deoxycholic acid, 2 mM DTT in 10mM sodium acetate, 50 μg/mL cycloheximide, 2% Polyoxyethylene 10 tridecyl ether, 400 U/mL recombinant Rnasin). After incubation on ice for 30 min, samples were centrifuged at 16,000 × g for 15 min at 4 °C. 50 μL of the clear supernatant (total RNA) was transferred to a new Eppendorf tube, in which 50 μL of isolation buffer, 100 μL Trizol, and 100 μL of Phenol: Chloroform: Isoamyl Alcohol (Sigma-Aldrich) was added. The mixture was kept on ice before further extraction. The supernatant (500 μL) was layered onto a sucrose gradient, and polysomes were dissociated by centrifugation at 28,000 × g for 4 h using a Beckman ST40Ti rotor. Upper 6 ml fractions were discarded, and the rest of the fractions (enriched in translating ribosomes) were selected for RNA extraction using TRIzol reagent (Ambion). Polysome-bound RNA or total RNA was subjected to RNA-seq library generation by BGI Genomics following the manufacturer’s protocol.

4.2.4 SHAPE-Structure-seq library analysis

The tomato genome and transcriptome reference were obtained and downloaded from Phytozome (tomato ITAG4.0). The SHAPE (-) libraries without SHAPE chemical treatment can be used to modify the tomato reference to the specific tomato cultivar (Money Maker). The Hisat2 tool was used to map raw reads from all the SHAPE (-) libraries to the tomato

genome reference (Kim, et al., 2019), followed by SNV calling using the FreeBayes software to obtain a tomato reference for tomato cv. Money Maker (Garrison and Marth, 2012). Due to the limitation of the original transcriptional annotation, many transcripts lack the annotations for their untranslated regions. Thus, the 'tadpole.sh' program in the BBmap package was used to extend the untranslated regions based on the SHAPE (-) libraries (Bushnell, 2014). For transcripts lacking annotated UTRs and where UTR extension was unsuccessful, the 300 nucleotides upstream and downstream of the CDS were respectively assigned as the 5′UTR and 3′UTR, respectively.

All the SHAPE-Structure-seq libraries (i.e., eight libraries in total, the SHAPE-treated (+SHAPE) and non-chemical control (-SHAPE) libraries in the two tomato developmental stages with two biological replicates) were used to calculate the SHAPE chemical reactivities. After the adapter trimming, the raw reads of all the libraries are mapped to the modified transcriptome reference using Hisat2. Next, the RT (reverse transcription)-stop count of (-) and (+) SHAPE libraries are calculated from the indexed BAM files by counting the nucleotides located one nucleotide upstream of the 5′ end of all mapped reads. The raw SHAPE reactivity was calculated using the following equation (Ding, et al., 2015; Ding, et al., 2014; Yang, et al., 2020):

$$SHAPE\ reactivity_i = \frac{log(1+P_i)}{\sum_i log(1+P_i)} - \alpha \frac{log(1+M_i)}{\sum_i log(1+M_i)}$$

where P_i is (+) SHAPE RT count and M_i is (-) SHAPE RT count at nucleotide *i*. The factor $\alpha = min\,(1,\ \sum_i log(1+P_i) / \sum_i log(1+M_i))$. Then, the box-plot normalisation was used for the normalisation of raw reactivities (Deigan, et al., 2009; Yang, et al., 2020) to obtain final reactivities. The extremely high reactivities were capped at 2, and only positions with depth higher than 20 were retained. All the procedure was described in our previous studies (Ding, et al., 2014). The RNAFramework was used with parameters 'rf-fold -md 200 -nlp -sh -dp -t 22' under the reactivity constraint to fold the RNA structures and obtain the base-pairing probability and the Shannon Entropy of all the nucleotides (Incarnato, et al., 2018).

4.2.5 Polysome profiling library analysis

The polysome profiling libraries in the two tomato developmental stages with three replicates (including polysome-associated RNA-seq and ribosome-free RNA-seq libraries) were mapped to the modified transcriptome reference by Salmon software with parameter '--validateMappings' after quality control, respectively (Patro, et al., 2017). After obtaining

the read counts of all the transcripts, RiboDiff, a software designed to calculate the differences of translation efficiency, was applied to calculate the TE differences of individual transcripts between the two developmental stages with the default parameters (Zhong, et al., 2017). The TE difference was defined by $\log_2(TE_{stage2}/TE_{stage1})$.

4.2.6 Development of the RSDE-Tool

To identify RNA structural dynamic elements (RSDEs) and their representative structural conformations, we developed a novel computational pipeline called RSDE-Tool. This pipeline integrates SHAPE-based reactivity profiles, RNA structural ensemble sampling, statistical assessments, and machine learning to systematically identify and characterize RSDEs.

Step 1: Hot spot identification and expansion.

For transcripts with the SHAPE reactivity profiles from two conditions, the diffScan R package was used to search for "hot spots" (usually <25 nt) with significant SHAPE reactivity differences (Yu, et al., 2022). Then, these "hot spots" were expanded to 200-nt "hot elements" centred with the hot spot. Only elements with at least 100 reliable SHAPE positions were retained. diffScan was designed to identify hotspots corresponding to regions with significantly different chemical reactivities between two conditions. diffScan incorporates multiple normalization steps and statistical tests to ensure that the detected hotspots are highly confident. Therefore, additional false discovery rate (FDR) correction was not applied to these hotspot elements.

Step 2: Ensemble generation and comparison.

For each hot element, two ensembles (1,000 structural conformations per ensemble) were sampled under the constraints of SHAPE reactivities of two conditions using the 'partition' and 'stochastic' program in RNAstructure package (Reuter and Mathews, 2010). Two generated ensembles will undergo two separate processes: (1) the measurement of the ensemble-wide difference using multiple statistical methods; (2) the identification of the RSDE representative structural conformations. To measure the differences between two ensembles, all the structural conformations in each structural ensembles were annotated and digitalised using the 'rnaConvert' in the Forgi package (i.e., five prime: 0, three prime: 0, stem: 1, hairpin loop: 2, multiloop:3, and interior loop:4) (Thiel, et al., 2019). Each ensemble was represented as a 1000 × 200 matrix. To test whether ensembles differed significantly, two nonparametric tests—Maximum Mean Discrepancy (MMD) and Energy Distance—were applied using the hyppo package in Python (Panda, et al., 2019). Both methods are either kernel-based or distance-based measurements that do not acquire explicit estimations

of the underlying probability density functions. Two ensembles with $P < 0.01$ in both tests were considered significantly different.

Step 3: Machine learning-based representative base-pairs identification.

All base-pairing pattens across the two ensembles (2000 conformations in total) were extracted and used to train an Extreme Gradient Boosting (XGBoost) classifier. For example, in the two structural conformations with the equal length: '.<<...>>.' and '<<.....>>', there are three distinct base-pairing patterns: '<.....>', '.<.....>.', and '..<...>..'. Among these, '<.....>' is unique to the second structural conformation. '..<...>..' is unique to the first structural conformation. '.<.....>.' is shared by both conformations. The presence or absence of these patterns in each structural ensemble can be further represented as a binary feature vector, for instance, [0,1,1] is for the first structural conformation while [1,1,0] is for the second structural conformation. Each structural ensemble is encoded in this way. All the resulting vectors are directly used for training the XGBoost classifier. If both AUROC and accuracy of the model are higher than the threshold (we set the default as 0.65), the important base-pairing patterns contributing to the model were then captured by the code '.feature_importances'. In addition to the machine learning model construction, all the base-pairing patterns are sorted by the frequency differences between two ensembles defined as $frequency_{ensemble1} - frequency_{ensemble2}$.

Step 4: Conflicting stem identification and representative structural conformation selection.

Adjacent base-pairing patterns are merged to form new stems, with a minimum stem length of three base pairs. The frequencies of individual base-pairing patterns are summed to represent the frequency weight of the newly formed stem, reflecting its enrichment in the two ensembles. Given the frequency weight and the feature importance of individual base-pairing patterns derived from the machine learning model, all the stems were divided into two groups with different enriched stems in the two ensembles, respectively. For example, the stems with the positive frequency weight containing at least one base-pairing pattern as one important feature of model were included in the ensemble1-enriched group, while the stems with the negative frequency weight with one base-pairing pattern were in the ensemble2-enriched group. Then, RSDE-Tool adopted a concept of conflicting stems from the SwitchFinder software (Khoroshkin, et al., 2024). Specifically, SwitchFinder was designed to identify RNA structure switches based on the hypothesis that the two structural conformations in such switches are often associated with conflicting stems, meaning the formation of one structure excludes the formation of the other. This mutual exclusivity ensures a noticeable conformational difference between the two structural conformations. For example, given the same RNA sequence, there are two structural conformations

‘..<<..>>’ and ‘<<..>>..’ represent the structural divergence as a conflicting stem. For each conflicting stem pair, one type of stem from the ensemble1-enriched group was favoured in ensemble 1, while another one from the ensemble2-enriched group was favoured in ensemble 2. Moreover, a total weight was assigned to each conflicting stem pair, defined as the sum of the absolute frequency weights of the two types of stems from two ensembles. All the conflicting stem pairs were then ranked in descending order based on their weights. Following this order, a recursive selection process was applied to retain only those conflicting stem pairs with the high weights. We further select those conflicting stem pair that did not conflict with any of other pairs. Through this process, two mutually exclusive sets of compatible stems were selected to represent each ensemble.

Following the SwitchFinder principle, mutually exclusive “conflicting stems” were identified, ranked by weight (absolute frequency difference), and recursively selected to construct two sets of compatible stems. Representative structures were then predicted using ‘MaxExpect’ under SHAPE constraints, maximizing expected base-pairing accuracy (Lu, et al., 2009). ‘MaxExpect’ was developed to predict RNA secondary structures with maximum expected accuracy (MEA) to search for a set of base pairs that collectively maximize the total expected accuracy based on base-pairing probabilities derived from the whole structural ensemble (Lu, et al., 2009).

Step 5: Linking RSDEs to biological functions.

Following the identification of the RSDEs and their representative structures, we took advantage of both bioinformatical and statistical methods to link the RSDEs to biological functions. After annotations of representative structural conformations using ‘rnaConvert’ (i.e., five prime: 0, three prime: 0, stem: 1, hairpin loop: 2, multiloop:3, and interior loop:4) (Thiel, et al., 2019), all numeric vectors from 5′UTRs, CDSs, 3′UTRs, and junctions between genic regions are further processed with the dimensional reduction process via the t-distributed stochastic neighbour embedding (t-SNE) (Maaten and Hinton, 2008) followed by the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN is a density-based advanced clustering algorithm that does not require the number of clusters to be specified beforehand and is particularly robust to noise and outliers, making it well-suited for complex, high-dimensional datasets (McInnes, et al., 2017). Clusters containing at least 10 RSDEs were retained for downstream analysis. RSDEs grouped within the same clusters were considered as one type of RSDEs associated with the specific biological function. To assess functional associations (translational alterations), several approaches were applied to link the clusters to translational differences including Mann-Whitney *u* test, Pearson Correlation Coefficient (PCC), and the mutual information

pipeline descried in Section 3.2.2 (pyTEISER). For each RSDE cluster, transcripts with translation efficiency (TE) differences were divided into two groups based on the presence or absence of the RSDE. A Mann-Whitney *u* test was then performed to evaluate whether TE differences were significantly different between the two conditions. Next, transcripts

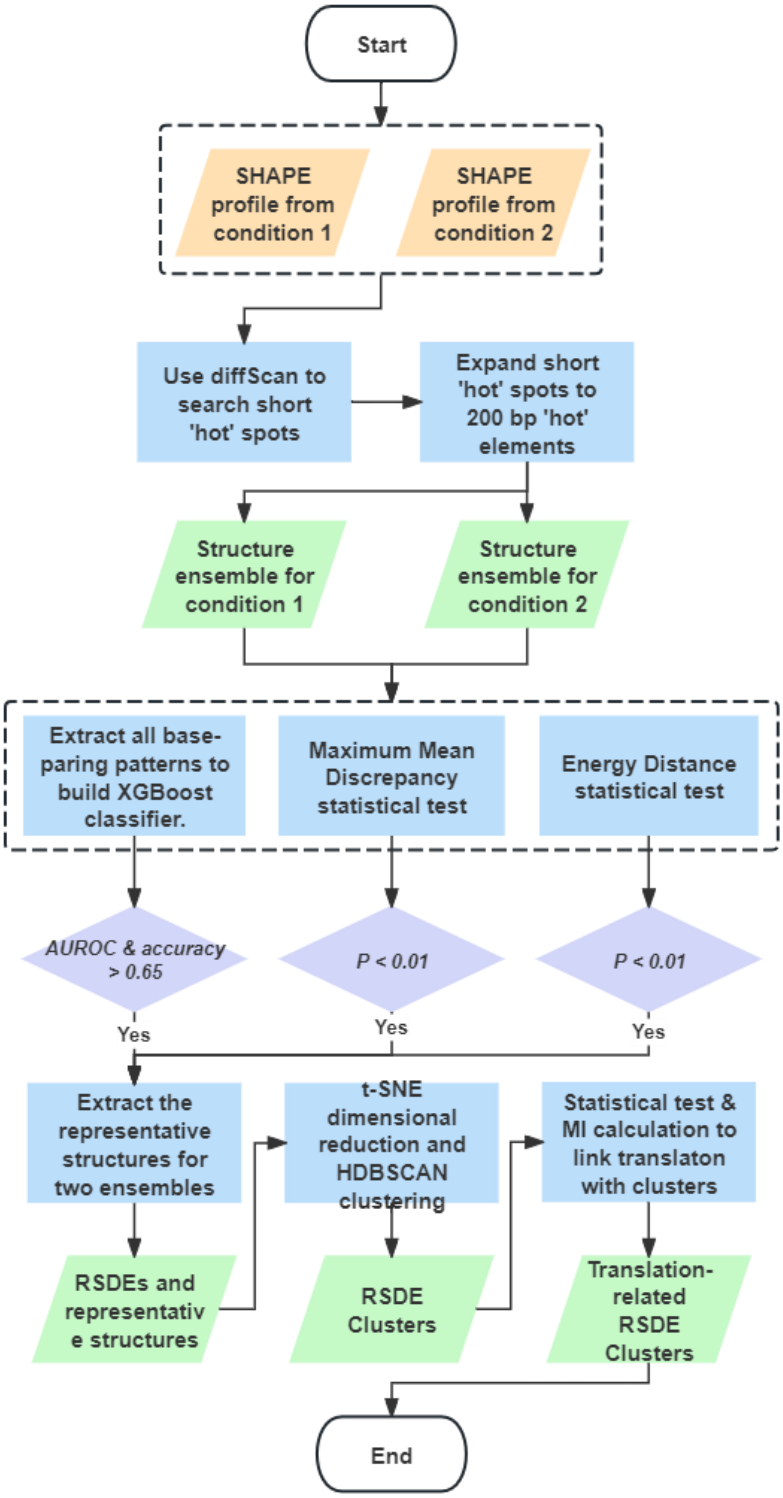


Figure 4. 1 The flow chart of RSDE-Tool.

were ranked by TE differences and binned into ten equally-size groups (bin number is 10). The mutual information, statistical significance, and Z-score were calculated following the procedure as in Section 3.2.2. In parallel, PCCs were calculated between mean TE differences and RSDE frequencies in bins. Clusters that yield P <0.05 in both the mutual information and Mann-Whitney *u* test were considered translation-associated RSDE clusters. The flow chart is shown in **Figure 4.1**.

4.2.7 Use of AI during the writing of the thesis

The Grammarly web (https://app.grammarly.com/) was used for grammar check.

4.3 Results

4.3.1 The tomato RNA structurome at two developmental stages

We used tomato (*Solanum lycopersicum*) cv Money Maker cultivar as the experimental material, selecting immature green (stage1) and mature green (stage2) fruit as representative developmental stages since numerous metabolites including sugars and organic acids, accumulate rapidly during this transition. SHAPE (**S**elective 2' **H**ydroxyl **A**cylation analysed by **P**rimer **E**xtension)-Structure-seq was applied to probe RNA structures in both stages. Tomato fruits were treated with 2-methylnicotinic acid imidazolide (NAI, one type of SHAPE chemicals) to modify single-stranded nucleotides transcriptome-wide. Two independent biological replicates with and without NAI treatment, were generated for constructing the SHAPE-Structure-seq libraries. Deep sequencing yielded 0.68 and 0.76 billion 150-nt paired-end reads for stage 1 and stage 2, with ~85% and ~84% of total reads successfully mapped to the tomato transcriptome, respectively (**Table 4.1**). Pearson correlation coefficients (PCCs) of mRNA abundances between biological replicates were very high for both SHAPE-treated (+SHAPE) and non-chemical control (−SHAPE) libraries, demonstrating strong reproducibility (**Figure 4.2**). In total, 19,331 transcripts for stage 1 and 18,302 transcripts for stage 2 with reliable SHAPE reactivity profiles were retained. To validate the accuracy of RNA structure probing, we compared SHAPE reactivities with an evolutionarily conserved structural element in tomato 18S rRNA. Both high- and low-reactivity regions showed strong agreement across the two stages (**Figure 4.3**), confirming the reliability of the structure libraries. Collectively, these results provide high-quality, transcriptome-wide RNA structure maps for tomato.

Table 4. 1 The basic statistics of SHAPE-Structure-seq libraries

Library	Total reads	Mapped reads	Mapped rate	Uniquely mapped reads	Uniquely mapped rate
Stage1 (-) rep1	155950580	136586755	87.58%	122475148	78.53%
Stage1 (-) rep2	105899246	92292128	87.15%	85695181	80.92%
Stage1 (+) rep1	128112064	100252156	78.25%	90325673	70.51%
Stage1 (+) rep2	289207060	248975160	86.09%	221229677	76.50%
Stage2 (-) rep1	192714228	170072593	88.25%	167485982	86.91%
Stage2 (-) rep2	156514640	137325993	87.74%	135318887	86.46%
Stage2 (+) rep1	194927162	155790308	79.92%	151181374	77.56%
Stage2 (+) rep2	219093784	179666488	82.00%	174126799	79.48%

To investigate the structure differences between two stages in tomato, the base-pairing probability (BPP) and Shannon entropy of every nucleotide were calculated across different genic regions via RNAFramework (Incarnato, et al., 2018). BPP represents the probability of being double-stranded of every single nucleotide from RNA structure population. Shannon entropy indicates the potential of structural alteration of each nucleotide. More transcripts tend to be more structured in stage 2, especially in 3′UTRs. For Shannon entropy, slightly more transcripts tend to have more diverse structures in stage 1 in 5′UTRs and 3′UTRs with the opposite in CDS, though the difference is not obvious (**Figure 4.4**). Additionally, the Pearson correlation coefficients (PCCs) between BPP and Shannon entropy of two stages were performed (**Figure 4.5**). Compared to 5′UTRs and 3′UTRs, PCCs of both BPP and Shannon entropy is significantly lower than those in CDS regions, indicating CDS regions are more likely to fold diversely.

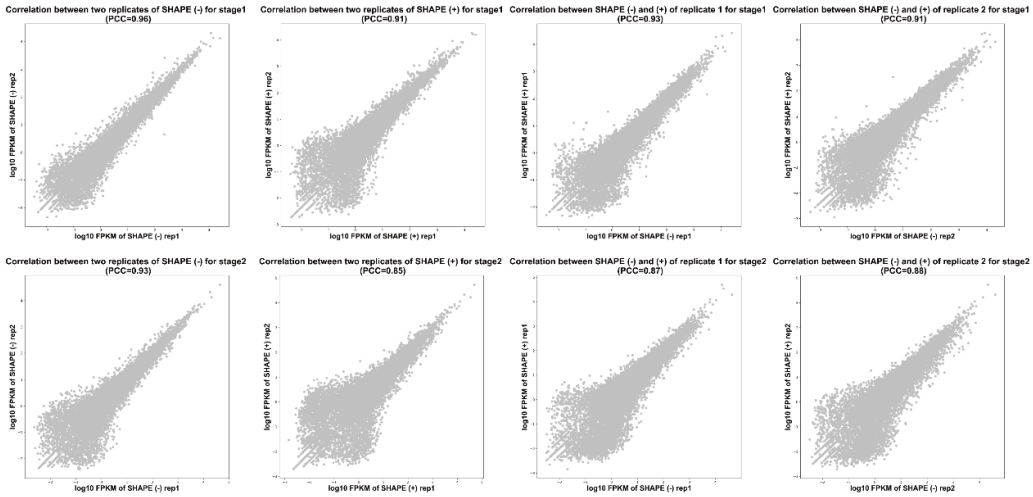
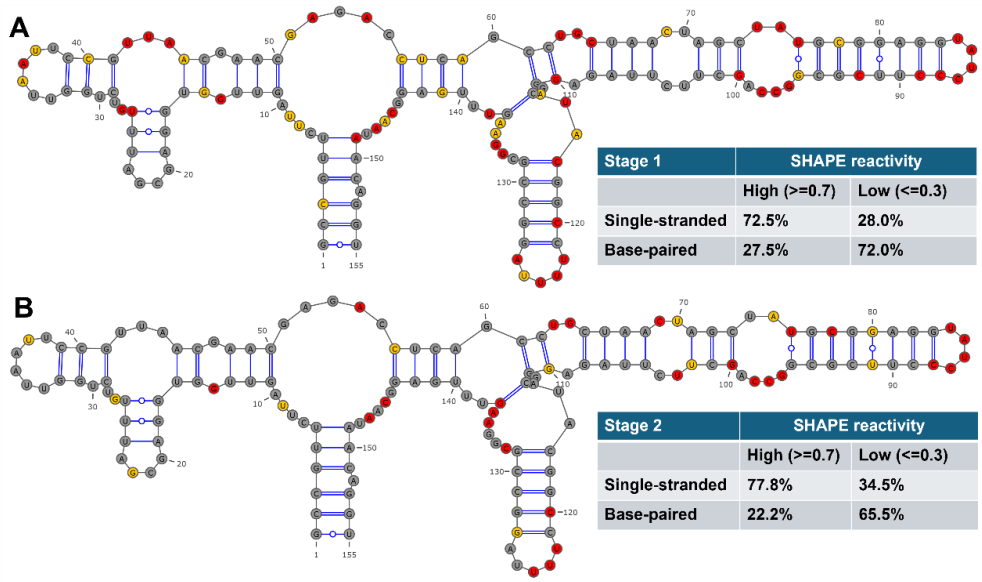


Figure 4. 2 The high reproducibility of the SHAPE-Structure-seq libraries.



Stage 1	SHAPE reactivity	
	High (>=0.7)	Low (<=0.3)
Single-stranded	72.5%	28.0%
Base-paired	27.5%	72.0%

Stage 2	SHAPE reactivity	
	High (>=0.7)	Low (<=0.3)
Single-stranded	77.8%	34.5%
Base-paired	22.2%	65.5%

Figure 4. 3 The alignment of SHAPE reactivity and known 18S rRNA.

4.3.2 RSDE detection between tomato RNA structuromes at two developmental stages

While global analyses can capture broad RNA structural trends, the biological impact of RNA structure dynamics often arises from local structural alterations. To systematically identify such regions, we developed the RSDE-Tool. Unlike earlier tools, such as diffScan, which mainly identify short (~20 nt) local differences in chemical reactivity profiles, our pipeline enables comprehensively comparison of multiple structural ensembles and generation of representative structural conformations with distinct divergence.

The RSDE-Tool integrated several strategies. First, diffScan was applied to detect small 'hot' spots of reactivity changes which were then expanded to 200 nt as the 'hot' elements. For each 'hot' element, two structural ensembles corresponding to stage 1 and stage 2 were generated. The RSDEs were called when the ensembles satisfied three criteria: two non-parametric distribution tests (Maximum Mean Discrepancy (MMD) and Energy Distance) and a machine learning classifier. MMD and Energy Distance quantified statistical

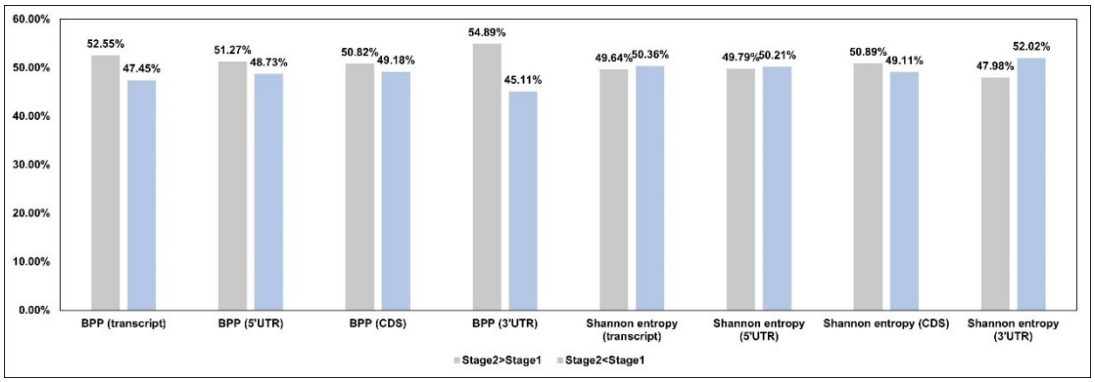


Figure 4. 4 The comparison of BPP and Shannon entropy between two stages.

differences between matrix representations derived from structure ensembles at two developmental stages. In addition to the statistical methods, an XGBoost classifier was trained on base-pairing patterns, tested separability of the two ensembles. Strong classifier performance (accuracy and AUROC > 0.65) indicated distinct base-pairing landscapes. An element was designated as an RSDE if both MMD and Energy Distance returned $P < 0.01$ and the classifier achieved good performance.

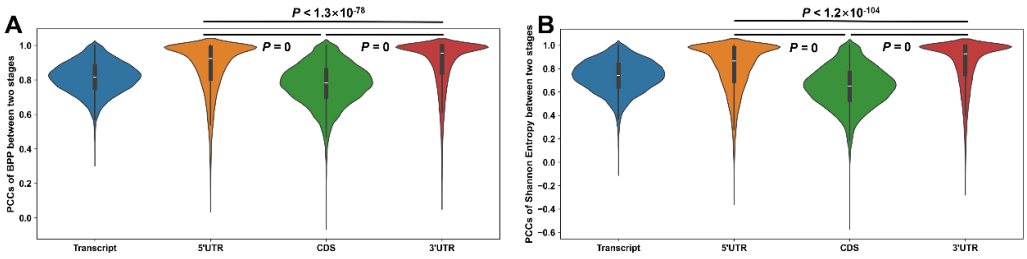


Figure 4. 5 The correlation of BPP and Shannon entropy between two stages. Mann-Whitney *u*-test was performed to compare PCCs of different regions.

To generate representative structural conformations for each RSDE, we adopted the concept of conflicting stems from SwitchFinder, identifying mutually exclusive stems that drive ensemble divergence. These were combined with the classifier-informed base-pairing patterns derived from the machine learning model to define the most representative structural conformations for each RSDE. Using this pipeline, we identified 12,320 RSDEs across the tomato transcriptome. Notably, 3,112 (25.3%) were enriched in 5′UTRs or 5′UTR-CDS junctions, indicating that RNA structures in or near 5′UTRs tend to be particularly dynamic (**Figure 4.6**). Given the known role of 5′UTR structures in translational regulation, we next focus on the relationship between RSDEs and translation.

To study the relationship between RSDEs and translation, we performed polysome profiling including both polysome-associated RNA-seq and free RNA-seq in tomato cv. Money Maker at the same developmental stages. Quality metrics confirmed high reproducibility of the datasets (**Supplementary Table 5**; **Supplementary Figure 1**). Translation efficiency

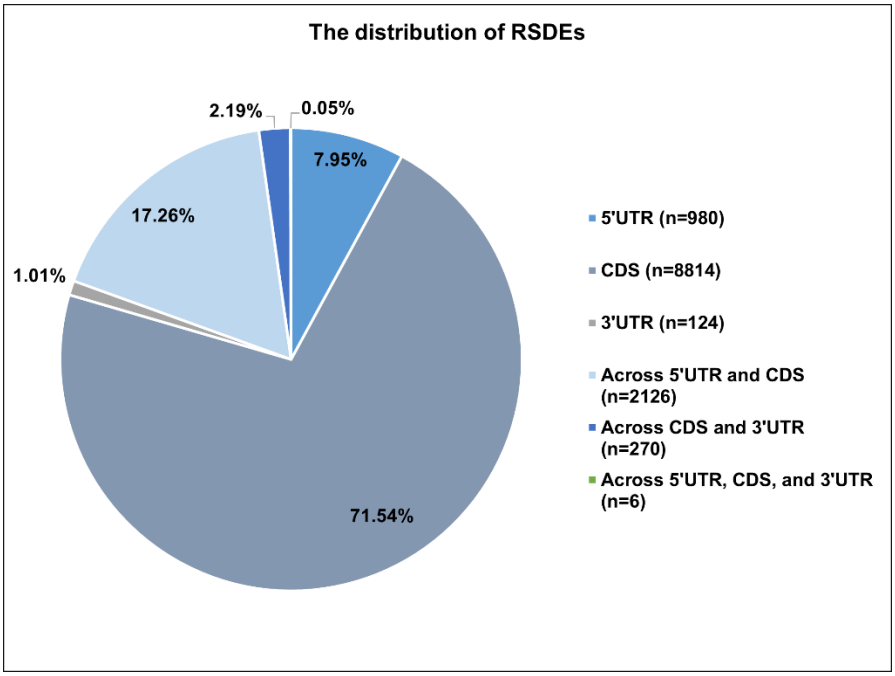


Figure 4. 6 The distribution of RSDEs across tomato transcriptome.

(TE) differences between stages were quantified with RiboDiff software (Zhong, et al., 2017). Before the functional association analysis, RSDEs were clustered by structural similarity using HDBSCAN. We then employed multiple approaches—including Mann–Whitney *u*-test, Pearson correlation coefficients (PCC), and mutual information (MI) analysis—to identify the RSDE clusters associated with translation changes. Transcripts were ranked by TE difference, partitioned into 10 bins, and RSDE frequencies per cluster were calculated. PCC and MI were derived from these frequencies, and MI significance was assessed via permutation testing. Mann–Whitney *u*-tests compared TE differences between transcripts with and without RSDEs in each cluster. Clusters were considered translation-related if both MI $P < 0.05$ and Mann–Whitney *u*-test $P < 0.05$.

Our analysis revealed five translation-associated RSDE clusters: three enriched at 5′UTR–CDS junctions and two within CDS regions. The largest, CDS_46, contained >4,000 RSDEs in CDSs and was significantly enriched in genes with higher TE at stage 1 (PCC = –0.64). A second CDS-enriched cluster, CDS_36, showed a similar trend but with weaker statistical support. By contrast, the three junction-enriched clusters were associated with higher TE in

stage 2. Together, these findings demonstrate that the RSDE-Tool effectively identifies dynamic RNA structural elements with functional links to translational regulation.

Table 4. 2 The RSDE clusters linked to translation difference between two developmental stages of tomato

The * indicates the number of RSDEs in corresponding cluster and ** indicates the number of RNAs containing RSDEs in corresponding cluster.

Cluster ID	*	**	U test *P* value	MI	MI *P* value	Z-score	PCC	PCC *P* value
CDS_46	4063	2538	8.03×10^{-18}	0.005	0	38.19	-0.64	0.05
CDS_36	19	19	0.01	0.0009	0.001	4.07	-0.32	0.36
UTR5_CDS_30	125	121	0.002	0.0006	0.04	1.98	0.70	0.03
UTR5_CDS_29	20	20	1.51×10^{-5}	0.0008	0.001	3.65	0.89	0.0005
UTR5_CDS_14	12	11	0.04	0.0007	0.008	3.10	0.57	0.09

4.4 Summary and discussion

Transcriptome-wide RNA structure profiling has been carried out in various species, revealing important regulatory roles of RNA structures in diverse biological functions. In this study, for the first time, we present the first high-quality transcriptome-wide RNA structurome of tomato fruit at two key developmental stages and systematically compared their structural landscapes. Our analyses found that RNA structures are not always consistent between stages, with particularly notable differences in coding sequence (CDS) regions (**Figure 4.5**). Such structural differences within CDSs may have implications for translational regulations in tomato. To explore these dynamics in greater depth, we developed a new computational pipeline, RSDE-Tool, designed to identify RNA structure dynamic elements (RSDE) from chemical profiling data. Unlike existing methods such as diffScan, which primarily compare reactivity profiles, the RSDE-Tool focuses on the level of the structural ensembles. RSDE-Tool integrated statistical approaches to assess divergence, which were named as RNA structure dynamic regions (RSDEs) between the structural ensembles derived from the same RNA element with different chemical profiles under different conditions. We also applied machine learning and classical RNA structure prediction models to identify the representative structural conformations for each structural ensemble. Using this pipeline, we identified over 12,000 RSDEs across the tomato transcriptome. The majority were located in 5′UTRs, 5′UTR-CDS junction, or CDSs, genic

regions well known for their regulatory impact on translation. Of particular relevance to CDS regions, changes in translational activity may alter ribosome footprinting, thereby affecting the accessibility and modification efficiency of probing reagents. In addition, condition-specific binding of RNA-binding proteins or alterations in the RNA modification landscape could further contribute to the observed shifts in reactivity. Theoretically, longer RNA molecules may be more susceptible to these additional layers of regulation compared to shorter ones, a possibility that warrants further investigation in future studies.

Given this enrichment, we also generated polysome profiling libraries of tomato fruits at the same developmental stages to quantify translation differences. By combining clustering with statistical approaches, we associated RSDE clusters with translational regulation. We revealed five translation-associated RSDE clusters: three enriched at 5′UTR-CDS junctions and two within CDS regions. These results suggest that local RNA structural dynamics, particularly in regions influencing ribosome access and initiation, may contribute to stage-specific changes in translational efficiency in tomato development.

It is important to note that while our methods are designed to prioritize RSDEs most strongly associated with translational changes, this does not exclude functional roles for other RSDEs in additional biological functions. A limitation of the present work is that the detected RSDEs and their predicted conformations have not yet been validated experimentally, and this part of the project is still ongoing. The experimental validation strategy is to introduce designed mutations into RSDEs located in 5′untranslated regions (5′UTRs). As these regions are non-coding, codon constraints do not apply which allows greater flexibility in design. Mutations will be introduced specifically within the conflicted stem regions. The design principle is that, after mutation, the resulting RNA structural ensemble resembles that observed in one of the two tomato developmental stages. This similarity can be validated using *in vitro* chemical probing-based RNA structure profiling, with particular emphasis on the abundant formation of the corresponding conflicted stems. Based on this rationale, two classes of mutant sequences will be designed: one representing the RNA structural ensemble characteristic of one developmental stage, and the other representing the alternative stage. These two mutant constructs, together with the wild-type sequence, will then be subjected to luciferase reporter assays to evaluate whether the trends in translation efficiency (TE) observed in polysome-seq data are recapitulated. If the experimental outcomes are consistent with the predicted TE changes, this would provide direct evidence supporting the existence of the RSDE.

Overall, by integrating transcriptome-wide RNA structure profiling with translation analysis, we provide the first comprehensive view of dynamic RNA structural elements in tomato fruit development. Our findings support the idea that RNA structural dynamics are not merely passive features but may actively shape translational regulation and contribute to developmental transitions in tomato.

4.5 Other contributors to the work described in this chapter

I acknowledge all the internal and external collaborators in this chapter. The project was in the collaboration with Prof. Huakun Zhang's group [Northeast Normal University, China] and Prof. Jie Li [John Innes Centre, UK, now Shanghai Jiao Tong University, China]. All the main collaborators who have contributed to work in this chapter are list below:

- Prof. Jie Li [John Innes Centre, UK, now Shanghai Jiao Tong University, China] contributed to generate the SHAPE-Structure-seq libraries and Polysome profiling libraries (partly).
- Dr. Xiaofei Yang [John Innes Centre, UK, now Chinese Academy of Sciences, China] contributed to generate the SHAPE-seq libraries (partly).
- Dr. Zongyun Yan [John Innes Centre] contributed to construct the Polysome profiling libraries (partly).

Chapter 5 Concluding discussion

5.1 Summary of the thesis

This thesis presents three independent projects that apply computational biology approaches to investigate RNA structures in diverse functional contexts across multiple species.

The first project focuses on i-motifs (iM), a specific cytosine-rich structure motifs with distinctive features. To address the lack of iM-specific prediction frameworks, we developed iM-Seeker, a computational tool trained on iM-specific experimental datasets. iM-Seeker can identify putative iM-forming sequences in both DNA and RNA, predict DNA iM folding status, and estimate folding strength, which also serve as a reference for RNA iMs. We released both command-line tools and a webserver. The webserver integrated automated machine learning to allow users to customized iM-prediction machine learning models with their own datasets. Despite limitations from small training datasets, iM-Seeker demonstrates the power of integrating multiple data types with tailored algorithms to study a sequence-defined structure motif. Using iM-Seeker, we systematically identified iMs across transcriptomes over 400 diverse species. For RNA iMs, we observed a strong enrichment of iMs in 5′UTRs across all plants, with monocots showing particularly high enrichment, especially in 5′UTRs. The enrichment correlated positively with temperature, suggesting that plants in warmer environments tend to adopt more 5′UTR iMs. Translatome analysis further indicated that 5′UTR iMs can suppress translation although their stability does not appear to directly determine translational repression.

The second project extends beyond the sequence-defined structure motifs to identify general and functional RNA structure motifs associated with RNA stability in wheat. Here, we explored a wide range of potential factors to RNA stability (length, nucleotide composition, codon usage, and RNA secondary structure). We found that RNA structure is the main factor contributing to RNA stability in wheat. We then optimised pyTEISER, one of the few available functional RNA structure discovery pipelines, by expanding the motif seed library and improving computational efficiency. This allowed us to enrich for RNA structure motifs associated with RNA stability in wheat. These RNA stability-associated RNA structure motifs also displayed subgenome-specific preferences, highlighting a functional role for RNA structure-mediated stability in polyploid plant genomes.

The third project addresses the functional role of RNA structural dynamics. We generated high-quality RNA structuromes of tomato fruit at two developmental stages and conducted a systematic comparison. To identify dynamic structural elements, we developed RSDE-Tool, which integrates ensemble-based RNA structure analysis, statistical models, and machine learning to identify RNA Structural Dynamic Elements (RSDEs) and their representative structural conformations under different conditions. Using RSDE-Tool, we identified more than 12,000 RSDEs, with the ~25% RSDEs enriched in 5′UTRs or 5′UTR–CDS junctions. Further analyses revealed RSDEs in these regions as well as within CDSs that were significantly associated with the alterations of translational efficiency, demonstrating both the utility of RSDE-Tool and the regulatory importance of dynamic RNA structures in tomato fruit development.

Together, these three projects illustrate how computational methods can be leveraged to investigate RNA structure functionalities from different angles: (i) sequence-defined structural motifs, (ii) functional structure motifs associated with RNA stability, and (iii) RNA structural dynamic elements associated with translational changes.

5.2 Future directions

RNA is a central molecule in living systems, and its structural features play crucial roles in regulating gene expression across diverse biological pathways. However, RNA structure is highly complex and heterogeneous: even a single RNA can adopt multiple structural conformations. This structural plasticity complicates efforts to identify functional RNA structures. The complexity increases further when considering six additional issues: (i) current RNA structure prediction accuracy remains limited; (ii) RNA structures can change dynamically across the RNA life cycle; (iii) structures vary between subcellular compartments (e.g., nucleus vs. cytoplasm); (iv) structural heterogeneity exists across different cell types and tissues; (v) external stimuli such as temperature fluctuations or pathogen infection can globally reshape RNA structuromes; and (vi) natural genomic variation introduces RNA structural diversity across individual natural variants, raising questions about the role of such heterogeneity during the evolution. These interconnected complexities make RNA structure research particularly challenging.

Over the past decade, genome-wide RNA structure probing techniques have rapidly advanced, producing increasingly high-quality RNA structurome datasets. These resources

provide unprecedented opportunities for computational biology to extract insights into RNA function. The three projects presented in this thesis showcase how diverse computational approaches—including statistical modeling, supervised and unsupervised machine learning, optimization algorithm, and classical RNA structure prediction tools—can be applied to uncover functional RNA structure regions. Looking ahead, several key directions emerge:

- **Improving RNA structure prediction**: AI has already enhanced RNA secondary and tertiary structure prediction, but further development on experimental approaches for generating extensive bench markers is needed to transform current RNA structure prediction software, RNAfold, that was derived from the biophysical measurements over ~400-500 short RNAs in 2004. Compared to the training dataset of tens of thousands of proteins structure available before AlphaFold2, the current scale of RNA structural data obtained through biophysical methods is extremely limited, amounting to only around 2,000 structures, with merely a few hundred derived from cryo-electron microscopy experiments. Given this small dataset and the more dynamic nature of RNA structures compared to proteins, directly training models on biophysical experimental data using a strategy analogous to AlphaFold2 is impractical. Consequently, foundation models are likely to play a crucial role in the task of RNA structure prediction. Foundation models pre-trained on vast amounts of cross-species RNA sequences can be considered to have already keenly captured the patterns and rules embedded within RNA sequences, which naturally may include those related to RNA structure. After specific fine-tuning, foundation models, including RNA-FM (Chen, et al., 2022) and PlantRNA-FM (Yu, et al., 2024), have demonstrated excellent performance in RNA secondary structure prediction, highlighting the advantage of large models in this task. Furthermore, to address the shortage of RNA structural data generated by biophysical methods for AI model training, a substantial number of chemical probing-based structurome datasets can be directly utilized for training. Currently, there are ample data reserves for at least a dozen important organisms, including humans, mice, *Arabidopsis thaliana*, rice, and so forth. These datasets help models learn the preference for individual nucleotide sites to form single-stranded regions, greatly enriching the information incorporated into the models. If the field of RNA structure were to initiate a project similar to the Human Cell Atlas in single-cell biology, generating large-scale, high-quality data under unified standards in the future, such standardized data would

provide immense room for enhancing the performance of AI-driven RNA structure prediction.

- **Advancing functional RNA motif discovery**: Tools such as pyTEISER and PlantRNA-FM demonstrate potential, but more efficient and robust approaches are required for multi-functional motif discovery.
- **Expanding comparative analyses**: Rather than focusing solely on individual structure conformation predicted by RNA structure prediction software, future efforts will shift toward analysing the entire structural ensemble. Such methods could reveal the functional roles of structural heterogeneity. RSDE-Tool demonstrated the utility of ensemble-based structural comparisons. However, computational tools for analysing transcriptome-wide structural dynamics across species, natural variants, or diverse conditions remain limited. For instance, the current RSDE-Tool is limited to comparing structural ensembles sampled based on chemical probing reactivity between two conditions. When the number of conditions increases—such as in the presence of structural variations or multiple experimental states—the machine learning model within RSDE-Tool could be extended into a multi-class classification model. This would allow for the identification of condition-specific structural stems, thereby enabling comparisons across multiple structural ensembles. Deep learning methods are also well-suited to address this challenge. For example, graph neural networks could be employed to learn the topological features of all conformations within an ensemble, followed by comparing similarities between different ensembles for ensemble-level comparative analysis. Such a model could further be applied for genome-wide searches of specific structural dynamic regions.
- **Developing integrated analysis platforms**: While many pipelines exist for specific tasks (e.g., RNA structure prediction, probing data processing, motif discovery), unified platforms that integrates diverse analyses—short/long-read data processing, ensemble modeling (e.g., DRACO, DaVinci), motif detection, etc.—would greatly benefit the field, especially for labs without extensive bioinformatics expertise.
- **Harnessing AI for RNA design**: AI has already revolutionized protein design, particularly through diffusion models and other generative AI approaches. Applying similar methods to RNA design could unlock novel opportunities for synthetic biology and therapeutic development.

Future advances will be driven by the synergy between next-generation experimental technologies and AI-driven computational modelling. For instance, combining RNA

chemical probing with long-read sequencing can reveal isoform-specific structures at higher resolution, while single-cell RNA structuromics could uncover cell-type–specific structural features across tissues (Wang, et al., 2024). Although current costs limit genome-wide applications, technological improvements will make such approaches increasingly feasible.

Equally transformative is the rise of foundation models in biology, which integrate multimodal data to capture hidden relationships across complex systems. For example, Evo2—a genomic foundation model trained on collected genomes from all domains of life—demonstrates both strong discriminative and generative capacities (Brixi, et al., 2025). Similarly, single-cell foundation models such as scGPT (Cui, et al., 2024), scBERT (Yang, et al., 2022), and xTrimoGene (Gong, et al., 2023) have advanced our understanding of cell fate, communication, and tissue-specific functions. These breakthroughs suggest enormous potential for building virtual cells that integrate RNA structural information across developmental, spatial, and environmental contexts.

In the short term, AI-driven methods will help address individual complexities of RNA structural heterogeneity. In the long term, integrating structural, transcriptomic, and functional data into unified AI-powered frameworks may enable the creation of digital, multimodal models of cellular function, reshaping our understanding of RNA biology in development, stress response, and evolution.

Chapter 6 Appendix

6.1 Supplementary figures and tables

Supplementary Table 1 Dataset of putative i-motif sequences and transitional pH from literatures

The sequences not highlighted in gray were used as input data of Putative-iM-Searcher followed by iM-Seeker.

Sequence Name	Sequence 5' - 3'	pH_T	Ref.
hTeloC	TAACCCCCCCCCCC	6.5	(Wright, et al., 2017)
C2T3	CCTTTCCTTTCCTTTCC	6.1	
C3T3	CCCTTTCCCTTTCCCTTTCCC	6.7	
C4T3	CCCCTTTCCCCTTTCCCCTTTCCCC	7.1	
C5T3	CCCCCTTTCCCCCCCTTTCCCCCCCTTTC CCCCCCTTTCCCCCCCCTTTCCCCCCC	7.2	
C6T3	CCCCCCTTTCCCCCCTTTCCCCCCTTTCC CCCC	6.8	
C7T3	CCCCCCCTTTCCCCCCCTTTCCCCCCCTT TCCCCCCC	7.4	
C8T3	CCCCCCCCTTTCCCCCCCCTTTCCCCCCC CTTTCCCCCCCC	7.1	
C9T3	CCCCCCCCCTTTCCCCCCCCCTTTCCCCC CCCCTTTCCCCCCCCC	7.3	
C10T3	CCCCCCCCCCTTTCCCCCCCCCCTTTCCC CCCCCCCTTTCCCCCCCCCC	7.3	
C5T1	CCCCCTCCCCCTCCCCCTCCCCC	6.9	
C5T2	CCCCCTTCCCCCTTCCCCCTTCCCCC	7.1	
C5T4	CCCCCTTTTCCCCCTTTTCCCCCTTTTCC CCC	6.7	
AC017019.1	CCCCCCTCCCCCCCTCCCCCCTCCCCCC	7.1	
AC018878.3	CCCCCACCCCCAGCCCCCTTTCCCCC	7.1	
ATXN2L	CCCCCCCCCCCCCCCCCCCCCCCC	7	

CAMK2G	CCCCCAGGCCCCGCCAGTCCCCCCCCCCGCCCGGCCCCCGGCCCGCCCCC	6.9
DAP	CCCCCGCCCCCGCCCCCGCCCCCGCCCCC	7
DRP2	CCCCCTCTTCCCCTCTCCCCCTCTCCCCCTCTCTCCCTCTTCCCCCTCTCCTTGTCTCCTTCTCTCCCCC	6
DUX4L22	CCCCCGAAACGCGCCCCCCTCCCCCCTCCCCCCTCTCCCCC	7.1
GH2	CCCCCACCCCCACCCCCATCCCCACGCCCCGCCCCCGCCCCC	7.1
HIC2	CCCCCGGGACAGGGACCCTGGCCCCCCCCGACAGGCTGACGCCCACCCCCTCAAACTCTGGTGGACTTACCCCC	6.4
HOXC10	CCCCCACCCCCACCCCCACCCCCC	7.1
HOXD10	CCCCCCCCCCCTCCCCCGCGGCCCCC	7.1
JAZF1	CCCCCCCCGCCCCCGCCCCCGCCCTCCCCCC	7.1
MSMO1	CCCCCGCCCCCGCCCCCGCCCCC	6.7
NFATC1	CCCCCGTTTCCCCCGCCAGCCCCAGCGCCCCCCTGCCCGGCCCCC	7.1
PIM1	CCCCCGACGCGCCCCCCAACACACAAACCCCCAGAATCCGCCCCC	7
PLCB2	CCCCCGCCTCTTCTGGAGGCCCCCGCCCCCACCCCC	7
QSOX1	CCCCCGCCCCCGAGCCCCCGCCCCC	7.1
RAE1	CCCCCCGCCCCCCCCCGCCCCCCCGCGCCGCCCCCCCCCCGCCCCCCCGCCCCCGTCCCCCCGCCCCCCCCCGCCCCCCCCCGCCCCCCGTCCCCCCCGCCCCCCCCCGCCCCCCCCGTCCCCCCC	6.8
RUNX1–1	CCCCCCCCGCACCCCTTCCCCCGGCCCCCCC	6.7

RUNX1–2	CCCCCCTCCCCCTGCCTCTCCCTCCCCCCTTTCCCC	6.5	
RUNX1–3	CCCCCCTTTCCCCTGCCCCCCCTGCCTCCCCC	6.7	
SHANK1b	CCCCCCTCCCCCCACCCCCCACCCCCCC	7.1	
SHANK3	CCCCCGCCTCCGGCGCAGCCCCCTCGCCACCCCCGCTTCCCTCCCGTCTCAGGCCCCCTCCCCCCCGCCGCCCCCGCCCCC	6.6	
SHANK3b	CCCCCCGCACCGAGGCCTAGGACTCCCCCCCCCAACCCCGTCACAGCCCCCCAGACCCCCGCCCCGTGGCTCGGCCCCC	6.5	
SNORD112	CCCCCCCCCCGCCCCCCACCCCCCCACCCCCCCCCC	7.2	
SOX1	CCCCCTGCAGGCCCCCCTGCGCCTCCCCCCCCCCGCCACTGGCGCCTGGCTTCCCCC	6.9	
STX17	CCCCCGCCCCCGCCCCCGCCCCGCAGGGCCCCC	7	
Tandem Repeat (LA16c-OS12.2)	CCCCCCGTGTCGCTGTTCCCCCCGTGTCGCTGTTCCCCCCGTGTCGCTGTTCCCCCC	6.6	
TRABD	CCCCCGCCCCCCCCCCCCCCCCC	6.9	
WNT7A	CCCCCGCCCCTCCCTCCTTTCCCCCGTCCCTCCCCCCGCCCCCTCCCCC	7.1	
ZBTB7B	CCCCCCATCCCTCCCCTCCCTCCCCCCGCCCCTGCCACCCCCCAAACTCCCCCCCCCC	7.1	
ZFP41	CCCCCAGCCCCCGCCGACCCCCAGCTCCCGCCTCCGCCGACCCCCAGCCCCC	7	
ZNF480	CCCCCGCCCCCGCCCCCGCCCCC	6.7	
RET20	CCCCGCCCCGCCCCGCCCCA	7.1	(Bielecka, et al., 2019)
c-MYC	CCCCACCTTCCCCACCCTCCCCACCC	6.6	

BCL-2	CCCGCTCCCGCCCCCTTCCTCCCGCGCCCGCCCC	6.6	(Brooks, et al., 2010)
VEGF	CCCGCCCCCGGCCCGCCCC	5.8	
RET	CCCGCCCCGCCCCGCCC	6.4	
Rb	CCGCCCAAAACCCCCC	5.9	
h-TELO	TAACCCTAACCCTAACCCTAACCC	6.5	(Debnath, et al., 2019)
c-MYC	TCCCCACCTTCCCCACCCTCCCCACCCTCCCCA	6.6	
PDGFR-β	GCGTCCACCCTCCCTGCCCCGCCGCCCCCCCTTCTCCCAGC	6.6	
BCL-2	CAGCCCCGCTCCCGCCCCCTTCCTCCCGCGCCCGCCCCT	6.6	
KRAS	GCCCGGCCCCCGCTCCTCCCCCGCCGGCCCGGCCCGGCCCCCTCCTTCTCCCCG	6.9	
VEGF	GACCCCGCCCCCGGCCCGCCCCGG	6	
Rb	GCCGCCCAAAACCCCCCG	5.9	
RET	CCGCCCCCGCCCCGCCCCGCCCCTA	6.4	
c-KI-RAS	GCTCCCTCCCTCCCTCCTTCCCTCCCTCCC	6.6	
c-KIT	CCCTCCTCCCAGCGCCCACCCT	6.8	
HIF-1α	CGCGCTCCCGCCCCCTCTCCCCTCCCCGCGC	7.2	
DAP	CCCCCGCCCCCGCCCCCGCCCCCGCCCCC	7	
JAZF1	CCCCCCCCCGCCCCCGCCCCCGCCCTCCCCCC	7.1	
n-MYC	ACCCCCTGCATCTGCATGCCCCCTCCCACCCCCCT	6.5	
4CT	CTTCTCCCCACCTTCCCCACCCTCCCCACCCTCCCC	6.9	(Sutherland, et al., 2016)
5CT	CTTCTCCCCACCTTCCCCACCCTCCCCACCCTCCCCATAAGCGCCCCTCCCG	6.5	
WT PDGFR-	GCGTCCACCCTCCCTGCCCCGCCGCCCCCCCTTCTCCCAGC	6.6	(Brown, et al., 2017)

beta NHE Py41			
R1 T-to-C mutant Py41	GCGCCCACCCTCCCTGCCCCGCCGCCCCCCCTTCTCCCAGC	6.6	
R5 C-to-A mutant Py41	GCGTCCACCCTCCCTGCCCCGAAGCCCCCCCTTCTCCCAGC	6.4	
R6 C-to-A mutant Py41	GCGTCCACCCTCCCTGCCCCGCCGCCACACCTTCTCCCAGC	6.2	
C3T333	CCCTTTCCCTTTCCCTTTCCCT	6.6	(Gurung, et al., 2015)
C3T444	CCCTTTTCCCTTTTCCCTTTTCCCT	6.4	
C3T555	CCCTTTTTCCCTTTTTCCCTTTTTCCCT	6.2	
C3T666	CCCTTTTTTCCCTTTTTTCCCTTTTTTCCCT	5.8	
C3T777	CCCTTTTTTTCCCTTTTTTTCCCTTTTTTTCCCT	5.6	
C3T888	CCCTTTTTTTTCCCTTTTTTTTCCCTTTTTTTTCCCT	5.4	
C3T388	CCCTTTCCCTTTTTTTTCCCTTTTTTTTCCC	6.1	
C3T338	CCCTTTCCCTTTCCCTTTTTTTTCCC	6.5	
C3T383	CCCTTTCCCTTTTTTTTCCCTTTCCC	6.6	
C3T833	CCCTTTTTTTTCCCTTTCCCTTTCCC	6.5	
C3T883	CCCTTTTTTTTCCCTTTTTTTTCCCTTTCCC	6.2	
C3T838	CCCTTTTTTTTCCCTTTCCCTTTTTTTTCCC	6.1	
APE1-4 track	TACCCACCCCCACCCTGCCCTG	6.1	(Rogers, et al., 2018)
APE1-5 track	AACCCCCAGGGCTACCCACCCCCACCCTGCCCTG	6.2	

FEN1	GTCCCCACTCCACCCACACCAGGTCCCCGCAGGCCCCTGCTCCCTC	6.4
MGMT	CCGCCCCAGCTCCGCCCCCGCGCGCCCCGGCCCCGCCCCCGC	6.7
NEIL1	CGCCCCTCCCTGCGCCCCCCTCCCCCCAC	6.5
NEIL2-4 track	GGCCCGGGGCCCGCCCTCCCTT	5.7
NEIL2-5 track	GGCCCGGGGCCCGCCCTCCCTTCCTGTCCCCTC	6.1
NEIL2-6 track	GGCCCGGGGCCCGCCCTCCCTTCCTGTCCCCTCCCGA	6.3
NEIL3-4 track	GGCCCCGCCCAGGCCCCGCCCAA	5.1
NEIL3-5 track	GGCCCCGCCCAGGCCCCGCCCAAACAGCACCCTA	5.8
NTHL1-4 track	GTCCCGGGCCCTCACCCGCGCCCAC	5.3
NTHL1-5 track	GTCCCGGGCCCTCACCCGCGCCCACTGCAACCCGA	5.7
PCNA: sequence 1	TTCCCTAGCCCCGACCCGAGAGCTCCCTCTCCCGG	5.9
PCNA: sequence 2	CGCCCCGCCCCGCCCCCGTCGCCCTGCCTCCCTG	6.6
POLβ	CGCCCCTCTAGCCCCGCCCCGCCCCGCCCAG	6.5
Polη	GTCCCGACACCCTCTCCCAGCCCCAG	6.2
RAD17-Sequence 1	CGCCCCCAGCCTGCCCCAGCCCAGTCCCTCCCGG	6.2
RAD17-Sequence 2	CCACCCCCCCCCCGCCCCCCCCCGGA	6.9
RAD21	TTCCCCACCCCCTCCCCCGACCCTTTTCCCCTCCCCGG	6.5

RAD54L	GGCCCCGCCCCTCCCCCGCCACCCCCGCCCCCGCCCCGCCCCTC	6.5	
UDG-4 track	TTCCCAGCCCCCTCCCCCCGACCCCAC	6.7	
UDG-5 track	CACCCCTAAGGGGCAGGAACTTTTCTTCCCAGCCCCCTCCCCCGACCCCAC	6.3	
XRCC2	CGCCCACCGGCGGCCTTGTTCCCATCTCCCTCACTCCCAACCCGG	6	
XRCC3	GACCCGCCCCGCCGCCCCGGCCCGGCCCCGC	5.8	
XRCC5	TACCCACCCATCCCATCCCTCTTCTCCCTC	6.4	
hTeloCT	TCCCTAACCCTAACCCTAACCCAA	6.3	(Wright, et al., 2020)
ODN2	CCCGTTGCCCTTTCCCGTTGCCC	7.6	(Fujii and Sugimoto, 2015)
ODN3	CCCGTTGCCCTTTCCCTTTTCCC	7.4	
ODN6	CCCGTTGCCCTTTCCCATTACCC	7.1	
ODN7	CCCCGTTGCCCTTTCCCCTTTTCCC	7.9	
dC19	CCCCCCCCCCCCCCCCCCCCC	7.5	(Fleming, et al., 2017)
dC18-4132413	CCCCTCCCTTCCCCTCCC	6.6	
dC19-4133413	CCCCTCCCTTTCCCCTCCC	6.4	
dC19-4141414	CCCCTCCCCTCCCCTCCCC	6.9	
dC20-4142414	CCCCTCCCCTTCCCCTCCCC	6.2	
LL3	TCGTTCCGTTTCGTTCCGT	7.8	(Mir, et al., 2017)
LL4	TCGTTCCGTTTTCGTTCCGT	7.9	
LL3rep	TCGTTCCGTTTTTCGTTCCGTTTTTCGTTCCGTTTTTCGTTCCGT	7.6	
LL3long	TCGTTCCGTTTTTCGTTCCGTTTTTTTTCGTTCCGTTTTTCGTTCCGT	7.5	
TT	TTCCCTTTCCCTTTCCCTTTCCCTT	6.5	

AA	TTCCCTATCCCTTTCCCTATCCCTT	6.2	(Benabou, et al., 2016)
CC	TTCCCTCTCCCTTTCCCTCTCCCTT	6.4	
GG	TTCCCTGTCCCTTTCCCTGTCCCTT	6.2	
GC	TTCCCTGTCCCTTTCCCTCTCCCTT	6.3	
TA	TTCCCTTTCCCTTTCCCTATCCCTT	6.3	
CA	TTCCCTCTCCCTTTCCCTATCCCTT	6.3	
TG	TTCCCTTTCCCTTTCCCTGTCCCTT	6.4	
1C	TGTCCCCACACCCCTGTCCCCACACCCCTGT	6.5	(Guneri, et al., 2023)
2C	TGTGCCCACACCCCTGTGCCCACACCCCTGT	5.2	
3C	TGTCCTCACACCCCTGTCCTCACACCCCTGT	6.0	
4C	TATCCCCACACCCCTATCCCCACACCCCTAT	6.7	
5C	TATCCACACACCCCTATCCACACACCCCTAT	6.8/ 5.9	
6C	TGTCCCCAGACCCCTGTCCCCAGACCCCTGT	6.2	
7C	TGTCCTCAGACCCCTGTCCTCAGACCCCTGT	5.4	
8C	TGTCCCCGGACCCCTGTCCCCGGACCCCTGT	5.1	
9C	TGTCCCCAGGACCCCTGTCCCCAGGACCCCTGT	5.5	
10C	TGTCCCCAGGACCCTGTCCCCAGGACCCTGT	4.7	
11C	TGTCCCCGGGACCCCTGTCCCCGGGACCCCTGT	4.7	
AC017019.1	CCCCCCTCCCCCCCTCCCCCCTCCCCCC	7.0	(Williams, et al., 2023)
DAP	CCCCCGCCCCCGCCCCCGCCCCCGCCCCC	5.9/ 7.1	

PIM1	CCCCCGACGCGCCCCCCAACACACAAA CCCCCAGAATCCGCCCCC	6.6
ZBTB7B	CCCCCCATCCCTCCCCTCCCTCCCCCCGC CCCTGCCACCCCCCAAACTCCCCCCCCC C	6.7
DUX4L22	CCCCCGAAACGCGCCCCCCTCCCCCCTC CCCCCTCTCCCCC	6.9
DUX4L22 MUT	CCTCCGAAACGCGCCTTCCTCCTTCCTC CTTCCTCTCCTCC	6.3
SNORD112	CCCCCCCCCGCCCCCCACCCCCCCACCC CCCCCCCC	7.1
SNORD112 MUT	CCTCCTTCCGCCTTCCACCTTCTCACCTC CTCCTCC	5.5/ 6.6
MYCPu22r ev comp	TTACCCACCCTACCCACCCTCA	5.9
GGGT rev comp	ACCCACCCACCCACCC	5.8

Supplementary Table 2 In-house biophysical dataset of putative i-motif sequences and transitional pH

Biophysical analysis of c-rich sequences with indicated location of nucleotide replacement or deletion (Δ) of the predominant ILPR (ACA) sequence. Variations of hTeloC and some known iM forming sequences are included in the experimental set up. The sequences not highlighted in grey were used as input data of Putative-iM-Searcher followed by iM-Seeker. Thermodynamics data are presented as Mean ± SD (n=3).

Name	**Sequence 5' - 3'**	**Thermodynamics UV spectroscopy (295 nm)**			**CD Spectroscopy**	**TDS**
		T_M (°C)	T_A (°C)	ΔH (°C)	pH_T	Structure

ACA	TGTCCCCAC ACCCCTGTCC CCACACCCC TGT	55 ± 1.7	52 ± 0.9	3 ± 1.0	6.6	iM
ACA(C1/9=T)	TGT**T**CCCACA CCCCTGT**T**CC CACACCCCTG T	50 ± 0.3	48 ± 0.0	2 ± 0.2	6.4	iM
ACA (C2/10=T)	TGTC**T**CCACA CCCCTGTC**T**C CACACCCCTG T	40 ± 0.5	38 ± 0.6	2 ± 0.4	5.7 / 6.5	iM
ACA(C3/11=T)	TGTCC**T**CAC ACCCCTGTCC **T**CACACCCC TGT	40 ± 0.3	37 ± 1.0	3 ± 1.2	6.0	iM
ACA (C4/12=T)	TGTCCC**T**ACA CCCCTGTCCC **T**ACACCCCTG T	48 ± 0.3	45 ± 0.4	3 ± 0.7	5.7 / 6.5	iM
ACA (C5/13=T)	TGTCCCCACA **T**CCCTGTCCC CACA**T**CCCTG T	53 ± 0.6	50 ± 0.6	3 ± 0.6	5.4 / 6.7	iM
ACA (C6/14=T)	TGTCCCCACA C**T**CCTGTCCC CACAC**T**CCTG T	44 ± 0.0	41 ± 0.5	2 ± 0.5	5.9 / 6.7	iM
ACA (C7/15=T)	TGTCCCCACA CC**T**CTGTCCC CACACC**T**CTG T	37 ± 0.6	35 ± 0.6	2 ± 0.0	5.9	iM
ACA (C8/16=T)	TGTCCCCACA CCC**T**TGTCCC	44 ± 0.0	43 ± 0.6	1 ± 0.6	6.0	iM

	CACACCC**T**TG T					
ACA (C1/5/9/13=T)	TGT**T**CCCACA **T**CCCTGT**T**CC CACA**T**CCCTG T	49 ± 0.6	45 ± 0.1	4 ± 0.6	6.3	iM
ACA (C2/6/10/14=T)	TGTC**T**CCACA C**T**CCTGTC**T**C CACAC**T**CCTG T	32 ± 0.4	28 ± 0.5	4 ± 0.3	6.0	iM
ACA (C3/7/11/15=T)	TGTCC**T**CACA CC**T**CTGTCC**T** CACACC**T**CTG T	29 ± 0.3	27 ± 0.5	2 ± 0.4	5.6	iM
ACA (C4/8/12/16=T)	TGTCCC**T**ACA CCC**T**TGTCCC **T**ACACCC**T**TG T	44 ± 0.4	41 ± 0.2	3 ± 0.3	6.1	iM
ACA (C1-odd)	TGT**T**C**T**CACA **T**C**T**CTGT**T**C**T** CACA**T**C**T**CTG T	ND	ND	ND	ND	B-DNA
ACA (C2-even)	TGTC**T**C**T**ACA C**T**C**T**TGTC**T**C **T**ACAC**T**C**T**TG T	ND	ND	ND	ND	B-DNA
ACA (C1/3/7/9=T)	TGT**T**C**T**CACA CCCCTGT**T**C**T** CACACCCCTG T	31 ± 0.0	29 ± 0.2	2 ± 0.2	5.9	iM
ACA (C6/8/14/16=T)	TGTCCCCACA C**T**C**T**TGTCCC CACAC**T**C**T**TG T	30 ± 0.5	29 ± 0.0	2 ± 0.5	5.7	iM

ACA (C2/6/10/14 =T)	TGTC**T**CCACA C**T**CCTGTC**T**C CACAC**T**CCTG T	32 ± 0.4	28 ± 0.5	4 ± 0.3	6.0	iM
ACA (G=T)	TTTCCCCACA CCCCTTTCCCC ACACCCCTTT	55 ± 0.0	54 ± 0.6	1 ± 0.6	6.7	iM
ACA (A=T)	TGTCCCCTCTC CCCTGTCCCCT CTCCCCTGT	61 ± 0.0	56 ± 0.0	2 ± 0.0	6.9	iM
ACA=TTT	TGTCCCCTTTC CCCTGTCCCCT TTCCCCTGT	60 ± 0.3	58 ± 0.0	2 ± 0.3	6.8	iM
ACA (C1/2/9/10=d)	TGT**ΔΔ**CCACA CCCCTGT**ΔΔ**CC ACACCCCTGT	36 ± 0.3	34 ± 0.1	2 ± 0.3	5.7	iM
ACA (C5/6/9/10=d)	TGTCCCCACA **ΔΔ**CCTGT**ΔΔ**CC ACACCCCTGT	ND	ND	ND	5.5	iM
ACA (C_3)	TGTCCC**Δ**ACA CCC**Δ**TGTCCC**Δ** ACACCC**Δ**TGT	40 ± 0.4	38 ± 0.1	2 ± 0.4	6.0	iM
ACA (C2)	TGTCC**ΔΔ**ACA CC**ΔΔ**TGTCC**ΔΔ** ACACC**ΔΔ**TGT	36 ± 0.2	34 ± 0.1	2 ± 0.3	ND	Z-DNA
ACA (C1/9=d)	TGTCCC**Δ**ACA CCCCTGT**Δ**CCC ACACCCCTGT	25 ± 1.0 49 ± 0.6	18 ± 0.6 47 ± 0.6	7 ± 0.6 2 ± 0.6	6.9 6.0	iM
ACA (C5/13=d)	TGTCCCCACA **Δ**CCCTGTCCCC ACA**Δ**CCCTGT	46 ± 0.0	44 ± 0.0	2 ± 0.0	6.3	iM

ACA (G=d)	T**Δ**TCCCCACA CCCCT**Δ**TCCCC ACACCCCT**Δ**T	54 ± 0.2	52 ± 0.0	2 ± 0.2	6.7	iM
ACA (T1/3/5=d)	**Δ**GTCCCCACA CCCC**Δ**GTCCC CACACCCC**Δ**G T	58 ± 0.1	55 ± 0.5	2 ± 0.6	6.9	iM
ACA (T2/4/6=d)	TG**Δ**CCCCACA CCCCTG**Δ**CCC CACACCCCTG **Δ**	53 ± 0.0	51 ± 0.3	2 ± 0.3	6.5	iM
ACA (T=d)	**Δ**G**Δ**CCCCACA CCCCGCCCCA CACCCC**Δ**G**Δ**	54 ± 0.2	52 ± 0.0	2 ± 0.2	6.7	iM
ACA (A1/3=d)	TGTCCCC**Δ**CA CCCCTGTCCC C**Δ**CACCCCTG T	58 ± 0.6	55 ± 0.0	2 ± 0.6	6.7	iM
ACA (A2/4=d)	TGTCCCCAC**Δ** CCCCTGTCCC CAC**Δ**CCCCTG T	59 ± 0.0	57 ± 0.0	2 ± 0.0	6.7	iM
ACA (A=d)	TGTCCCC**Δ**C**Δ**C CCCTGTCCCC**Δ** C**Δ**CCCCTGT	58 ± 0.1	56 ± 0.0	2 ± 0.2	5.0 / 6.9	Mixed
ACA (L-C=d)	TGTCCCCA**Δ**A CCCCTGTCCC CA**Δ**ACCCCTG T	53 ± 0.1	51 ± 0.0	2 ± 0.1	6.6	iM
ACA(C9=G)	TGTCCCCACA CCCCTGT**G**CC CACACCCCTG T	49 ± 0.0	47 ± 0.0	2 ± 0.0	6.0	Mixed

ACA(C1=G)	TGT**G**CCCACACCCCTGTCCCCACACCCCTGT	49 ± 0.6	47 ± 0.9	2 ± 1.1	6.0	iM
AGG	TGTCCCCAGGCCCCTGTCCCCAGGCCCCTGT	53 ± 0.3	50 ± 0.6	3 ± 0.5	5.8	iM
AGG (C5=A)	TGTCCCCAGGACCCTGTCCCCAGGCCCCTGT	35 ± 1.3 41 ± 0.2 62 ± 2.1	26 ± 5.4 38 ± 0.6 45 ± 1.1	10 ± 1.3 15 ± 3.3 15 ± 0.2	5.0 / 5.8	Mixed
AGG (C13=A)	TGTCCCCAGGCCCCTGTCCCCAGGACCCTGT	36 ± 2.7 56 ± 3.2	23 ± 3.7 44 ± 3.5	12 ± 3.2 12 ± 2.7	5.5	Mixed
ACA(C1/9=A)	TGTACCCACACCCCTGTACCCACACCCCTGT	48 ± 0.0	46 ± 0.8	3 ± 0.8	6.0	iM
(TAT)ACA(C3/11=A)	TATCCACACACCCCTATCCACACACCCCTAT	40 ± 0.6	38 ± 0.8	2 ± 1.3	5.8 / 6.7	Mixed
(TAT)ACA(C11=A)	TATCCCCACACCCCTATCCACACACCCCTAT	43 ± 0.7	40 ± 0.0	2 ± 0.6	5.8 / 6.2	iM
(TAT)ACA(C3=A)	TATCCACACACCCCTATCCCCACACCCCTAT	42 ± 0.6	40 ± 0.0	2 ± 0.6	6.0	iM

(TAT)ACA(C3/11=G)	TATCCGCACACCCCTATCCGCACACCCCTAT	42 ± 0.7	40 ± 0.0	3 ± 0.9	5.6 / 6.2	iM
(TAT)ACA(C3/11=T)	TATCCTCACACCCCTATCCTCACACCCCTAT	42 ± 0.0	40 ± 0.6	2 ± 0.6	6.0	iM
postILPR1	CCCCTGCCGCCTGGCCC	ND	ND	ND	5.1	Mixed
postILPR2	CCCCCCCACCCCAGGCCC	51 ± 0.2	44 ± 0.5	8 ± 0.6	6.25	iM
preILPRC	CCCCTCCCTCACTCCCACTCTCCCACCCCCACCACC	50 ± 0.4	48 ± 0.2	2 ± 0.4	6.3	iM
CHairpin 2:2	CCTACCTACCTACCTTTTCCTACCTACCTACC	31 ± 0.3	29 ± 0.0	2 ± 0.3	6.0	iM
CHairpinhTeloC TTA/AAT	AATCCCAATCCCATTCCCATTCCC	46 ± 0.4	43 ± 0.4	3 ± 0.3	6.5	iM
hTeloC	TAACCCTAACCCTAACCCTAACCC	47 ± 0.2	43 ± 0.2	4 ± 0.3	6.5	iM
hTeloCTGG	TGGCCCTGGCCCTGGCCCTGGCCC	ND	ND	ND	5.8	Hairpin
NRF2C	CCCTCCCGCCCTTGCTCCCTTCCC	46 ± 0.2	43 ± 0.4	2 ± 0.3	6.7	iM

Supplementary Table 3 RNA structure motifs in 3'UTR suppressing mRNA decay in wheat

ID	Motif Sequence	Structure	MI	*Z*-Score	PCC
sRSM1	WKNVSBNBHKKHH	((.((....))))	0.0031	12.50	-0.45
sRSM2	BAGVBNNBHMTBBK	(((.......).))	0.0028	5.98	-0.24
sRSM3	DNDDDKYSVNVTBNH	(((((....)).)))	0.0023	3.61	-0.41
sRSM4	KHBDDHNVBDBBNDVHBNV	((..((.......))..))	0.0023	4.30	-0.37
sRSM5	GGAGNUNNNGNNYUYY	((((........))))	0.0023	5.79	-0.47
sRSM6	DNBDVNBVNVSKRKDD	(((((.....))).))	0.0020	4.08	-0.58
sRSM7	BNBNDNBDNRVHNMNDVBW	(((...((...))...)))	0.0019	4.04	-0.36
sRSM8	KNDKWRNDNDNBDNKDVDN	(((.((.....))...)))	0.0018	3.64	-0.57
sRSM9	UUAADHUNNDNNUURR	((((........))))	0.0016	3.37	-0.72
sRSM10	GCASDUNVNNNBUGY	((((.......))))	0.0015	3.61	-0.28
sRSM11	GBHDVDDBWDNBMYVDDN	((..((......))..))	0.0031	5.43	-0.20
sRSM12	HSBHMKHMHRNYBBD	(((.........)))	0.0030	7.80	-0.22
sRSM13	VKYBNSWNVWDRNMW	(((........).))	0.0030	6.65	-0.27
sRSM14	DMVNVNDDMVYNDV	((.((....)).))	0.0030	6.38	-0.35
sRSM15	NVWNHWYVSBYNVN	((((......))))	0.0029	6.24	-0.21
sRSM16	KAKSCHCDNBNUN	((((.....))))	0.0028	8.06	-0.27
sRSM17	KNDWNNBNDHRDVYV	(.((((...)))).)	0.0027	6.24	-0.35
sRSM18	VDDRKBKYSMBBN	(.(((...))).)	0.0027	5.91	-0.34
sRSM19	KDKBKGNWHHVVWDB	((..((...))..))	0.0027	5.04	-0.35
sRSM20	KDVHMYVNBDNNHHDN	(((..((...)).)))	0.0025	5.67	-0.35
sRSM21	SDKNKVNYDSDVNRB	(((((.....)))))	0.0025	6.06	-0.35
sRSM22	VWBNDSTDVBWNDN	(.((((...)))))	0.0025	4.15	-0.29
sRSM23	RCVBBDNNDVRNN	((((....))).)	0.0025	3.55	-0.23
sRSM24	VDNHVNYVVDHWVBHYN	(((((......)).)))	0.0025	5.48	-0.32
sRSM25	VDDHDNDRHNDNBVBDWVY	(((..((.......)))))	0.0024	4.03	-0.37
sRSM26	DHBSSDYBYWRNB	((((.....))))	0.0024	4.58	-0.26
sRSM27	BVNDNVHVRDWHKNYKB	((...((...))...))	0.0024	5.28	-0.40
sRSM28	MAAWDNNNNNGUDUUK	((((........))))	0.0024	8.63	-0.23
sRSM29	DBNYKSBNANVBVD	(((((...)))).)	0.0024	4.06	-0.37
sRSM30	NSHVNDBKDSYDSNBN	((...((....)).))	0.0024	5.03	-0.24

sRSM31	HBVNWVVDHNNNVKNHDM	(((.(.....).)))	0.0023	4.37	-0.35
sRSM32	BSDMDBWNDDHBN	((((....).)))	0.0023	4.13	-0.18
sRSM33	WGAWANNNNUKDNDUDUY	.((((.....))))	0.0023	5.39	-0.54
sRSM34	SVBVHDDDHWVDHBN	((((.....).)))	0.0022	3.98	-0.28
sRSM35	HDRVVBHHVDWNBBNBNNN	(((((.....)))))	0.0022	3.43	-0.26
sRSM36	NVNSVVWMDDKYDNNNH	((...((....))..))	0.0022	5.26	-0.31
sRSM37	NNKVBNBNDBKVVNHNNRVH	((....(((....)))..))	0.0022	4.21	-0.30
sRSM38	RVYDNBSBVYHYBVDK	((((.....))))	0.0022	4.86	-0.23
sRSM39	KNKMNKMBVDHBHHBDBN	(.(((.....))))	0.0022	3.37	-0.29
sRSM40	NNYBNDVNHBBDDNNYVVD	(.((((.....))))))	0.0022	4.27	-0.22
sRSM41	WNDKSDYBKDBVHNVD	((((.....))).)	0.0022	4.14	-0.24
sRSM42	NKBKNHDNNHKKKRW	(((..((...)))))	0.0022	4.06	-0.21
sRSM43	HDNHNVKVHDYKGBNRVD	((...((.....))..))	0.0021	3.98	-0.28
sRSM44	HKBBRWNWHKDKVHK	((.(((....)))))	0.0021	4.85	-0.40
sRSM45	BBYKMSNMBHDRKBH	((((.....))).)	0.0021	3.73	-0.26
sRSM46	NWBDBHBMHNHNNSHNNN	((((((....))..))))	0.0021	3.45	-0.24
sRSM47	DDYDDKBBVVNBKNMD	((.(((....))).))	0.0021	3.98	-0.41
sRSM48	YMDRVHBNHDDHBDR	(.((((....))))	0.0021	3.57	-0.38
sRSM49	VNDNNVDMHNDDYNYD	(.((((....)))).)	0.0021	4.12	-0.35
sRSM50	NCDVDDWHVBVHDNBD	((((.....).)))	0.0021	3.42	-0.45
sRSM51	NDDSDNKDBBNDDRDBV	(((((.....)))).)	0.0021	4.56	-0.22
sRSM52	NBDHBNKBWNNBYNYYYH	((((.....)).))	0.0020	4.52	-0.22
sRSM53	BNNBHVDYNHHDBKKWNN	((.(((.....)))))	0.0020	5.65	-0.22
sRSM54	NMVNVHHNVVBDNKNBBH	(((.((.....)).)))	0.0020	3.49	-0.39
sRSM55	NUCUNNNGUNNNURGRN	((((.....))))	0.0018	4.63	-0.17
sRSM56	YRUARNNYNNYNNYURY	.((((.....))))	0.0018	5.11	-0.05
sRSM57	NNDNNKBNVBWNBVDDDDR	((((..((...))..))))	0.0018	3.32	-0.25
sRSM58	BSGUANNNNKNHNUURYB	.((((.....))))	0.0018	5.24	-0.21
sRSM59	UAGUNVUDGNDNRYUR	((((.....))))	0.0017	3.79	-0.14
sRSM60	GAAGNNGYNYUUY	((((.....))))	0.0017	3.35	-0.06
sRSM61	RDBHNBBNBBBNNDVDDDD	((...((((....))))))	0.0017	3.39	-0.22
sRSM62	NURAGNNCKHYUYR	.((((.....))))	0.0016	3.54	-0.36
sRSM63	BGYGANVNDDNNUBUYRY	.((((.....))))	0.0016	3.33	-0.38
sRSM64	NGUAGNNANNNNAURYN	((((.....))))	0.0016	3.34	-0.01

Supplementary Table 4 RNA structure motifs in 3'UTR promoting mRNA decay in wheat

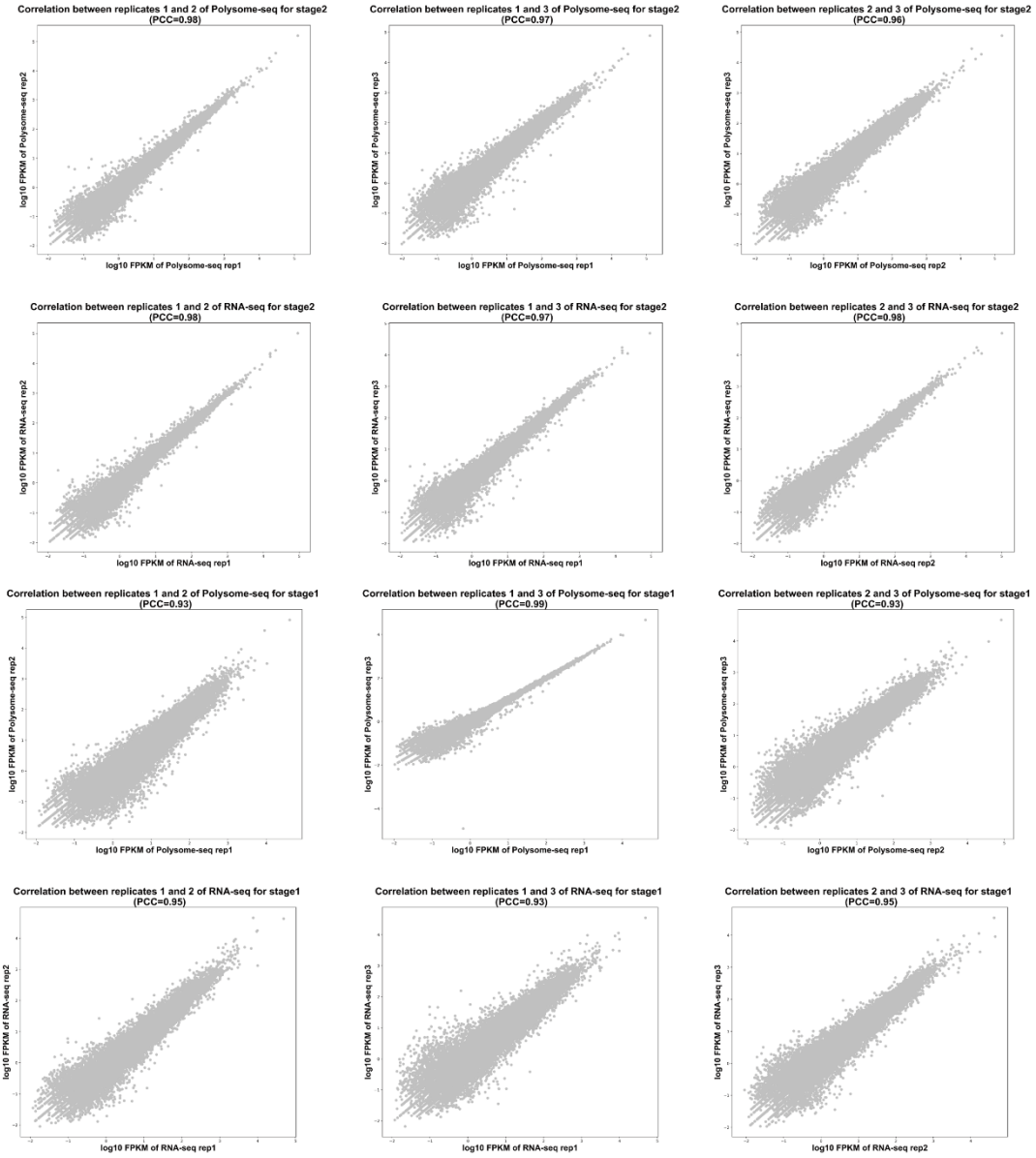
ID	Motif Sequence	Structure	MI	*Z*-Score	PCC
uRSM1	HBBNBVRNWDVNVSNV	((.((.....)).))	0.0029	5.68	0.63
uRSM2	RBNRVVNHBDYNHKH	(((.((...)).)))	0.0026	4.78	0.63
uRSM3	VMBVKBDNNNNBNBNDBN	(.((((((.....))))))	0.0025	8.02	0.68
uRSM4	NDNNBHVBKCVRY	((.((...)).))	0.0025	6.76	0.76
uRSM5	NNKDHBNHRVHBKNBYH	((..((....))...))	0.0023	6.47	0.62
uRSM6	KBDDWBNHBHBRNSH	((..((....)).))	0.0022	4.07	0.79
uRSM7	DWBWBBVNNHDVVBDV	(((.((....)).)))	0.0021	4.02	0.74
uRSM8	BVHDNHHBYNNNDNHDHB	(((.(((.....)))))))	0.0021	3.38	0.53
uRSM9	KVNNBVVKANNBNBHBNN	(((..((....))..)))	0.0020	3.43	0.71
uRSM10	HRGUGGCNNNNYRYY	.((((......))))	0.0019	5.04	0.84
uRSM11	DHNRBVBMVVWVVDH	((((......)).))	0.0030	7.11	0.59
uRSM12	BNHDVHKYDKNHNGB	((((......).)))	0.0029	6.41	0.62
uRSM13	VNNBNBDKDNDKDVHSDVND	(((..((.....))...)))	0.0028	6.27	0.58
uRSM14	MRNYDRBWRDBNN	((((....))).)	0.0027	5.67	0.51
uRSM15	UKAUHNNUNKADNRUNR	((((.........))))	0.0027	7.60	0.61
uRSM16	RHVHHBDWVMDVBH	((.((....)).))	0.0026	6.18	0.55
uRSM17	DDKVBWWNNBBVWD	(((((...)))).)	0.0026	5.99	0.54
uRSM18	UYUGUNNGNRYRRR	(((((....)))))	0.0026	9.66	0.39
uRSM19	YBDBNBVVMNVNDDDB	((..((....))..))	0.0025	5.85	0.53
uRSM20	NBVRDNHDVVNKHVBYNDBH	((..((........))..))	0.0025	4.10	0.56
uRSM21	BAVNDNYHVNNHNRBDBDB	((...((......))..))	0.0024	5.36	0.51
uRSM22	RVHNVKDNDYDKDBRNW	((...((....))..))	0.0023	3.97	0.59
uRSM23	VVNVNBVKDNWNDBT	(((((....).))))	0.0023	4.14	0.63
uRSM24	BDRNNRBNBBVNDYNYS	((..(((.....)))))	0.0023	3.81	0.61
uRSM25	BBKHBNDVNVVVDBKR	(((.((.....)))))	0.0023	4.16	0.54
uRSM26	VVBBBDSVVDKWRDDNB	(((.(........))))	0.0023	5.30	0.63
uRSM27	DKNNBKNSNHNHNRYBRWB	((....((....))...))	0.0023	5.89	0.65
uRSM28	NHNHNKNNVNVNVGRB	(((.((...))..)))	0.0023	5.54	0.63

uRSM29	BDDDWNVDDBWMVDBS	((.((....))...))	0.0022	3.87	0.71
uRSM30	NBNKWHDBSNDBHHRN	(((((.....).))))	0.0022	4.66	0.59
uRSM31	YHBKNVDNNVNDVWHKVDV	((...(((.....))).))	0.0022	4.04	0.51
uRSM32	CANSGNNVNNANBNUG	((((.....))))	0.0022	5.56	0.45
uRSM33	NWVDDYDBNHNBVKHBB	(((.((....)).).))	0.0022	4.30	0.54
uRSM34	CBSGANNNUNNNNYBNG	((((.....))))	0.0021	5.53	0.23
uRSM35	YSWAWNNHNDCDUDBR	(((((.....)))))	0.0021	5.78	0.39
uRSM36	KNDHVKNNBVBNRVDDVD	((...((.....)).))	0.0021	4.09	0.54
uRSM37	HDDBHNDMNHKBWVDV	(((((.....))).))	0.0020	3.86	0.67
uRSM38	GNGCNNGNANNNGYNY	((((.....))))	0.0020	5.35	0.47
uRSM39	SUCGNNNSDGYGRB	((((.....))))	0.0019	5.89	0.54
uRSM40	GNGUGNGUNYRYNY	(((((....)))))	0.0019	6.46	0.40
uRSM41	UNGGUNNNNCUNYYNR	((((.....))))	0.0019	5.02	0.66
uRSM42	UUUANNKKNNBGURRR	((((.....))))	0.0019	5.91	0.43
uRSM43	YGKUANNHUDNURNYR	(((((.....)))))	0.0018	4.08	0.32
uRSM44	YRKAMADBGHHNUNYR	((((.....))))	0.0018	4.16	0.58
uRSM45	NAUUUNNNNANUNRRUN	((((.....))))	0.0018	4.27	0.64
uRSM46	NGUUGRDVNNNUVYRRY	.((((.....))))	0.0018	3.93	0.26
uRSM47	NUDUUKNNNNYNNRRNR	.(((((.....)))))	0.0017	4.84	0.25
uRSM48	WAMSMNNCNNNNNBKUD	((((.....))))	0.0017	3.88	0.33
uRSM49	NUGCNNNNUNUNNGYRN	((((.....))))	0.0017	5.09	0.70
uRSM50	UCNCNUNNNUNGNGR	((((.....))))	0.0017	4.38	0.19
uRSM51	UUNUUNNANGRRNRR	(((((.....)))))	0.0016	3.32	0.46
uRSM52	USYUGNCNKYNRRBR	((((.....))))	0.0016	3.41	0.36
uRSM53	AUNUNNGNCNUNRNRU	((((.....))))	0.0016	3.43	0.25
uRSM54	YGSUNCABNNRBYR	((((.....))))	0.0015	3.50	0.22

Supplementary Table 5 The basic statistics of RNA-seq and polysome-seq libraries

Library	Mapped rate	Library	Mapped rate
RNA-seq stage1 rep1	80.14%	Polysome-seq stage1 rep1	84.44%
RNA-seq stage1 rep2	81.12%	Polysome-seq stage1 rep2	78.66%
RNA-seq stage1 rep3	86.97%	Polysome-seq stage1 rep3	86.78%
RNA-seq stage2 rep1	84.52%	Polysome-seq stage2 rep1	85.20%

RNA-seq stage2 rep2	85.97%	Polysome-seq stage2 rep2	84.33%
RNA-seq stage2 rep3	84.40%	Polysome-seq stage2 rep3	85.37%



Supplementary Figure 1 The reproducibility of libraries for RNA-seq and polysome-seq in tomato.

6.2 Code repositories

Putative-iM-Searcher is available at https://github.com/YANGB1/Putative-iM-Searcher.

iM-Seeker is available at https://github.com/YANGB1/iM-Seeker.

Web server version iM-Seeker is available at https://im-seeker.org/.

RSDE-Tool is available at https://github.com/YANGB1/RSDE-Tool.

6.3 Published work in PhD period

Co-first authors are indicated by *.

1. Dong, Q., **Yang, B.**, Sun, W., Ding, Y. and Zhang, H. (2025) The post-transcriptional landscapes of long non-coding RNAs in plants. *Plant Communications*, under revision.
2. Yan, Z., **Yang, B.**, and Ding, Y. (2025) G-Quadruplexes and i-Motifs: Emerging Regulatory Elements in Plant Genomes. *Current Opinion in Plant Biology*, under review.
3. **Yang, B.**, Ding, Y. and Zhang, Y. (2025) Profiling RNA G-quadruplexes *in vivo*. *Current Protocols*, 5, e70209.
4. Olazagoitia-Garmendia, A., Rojas-Márquez, H., Trobisch, T., Moreno-Castro, C., Etxebarria, A.R., Mentxaka, J., Tripathi, A., **Yang, B.**, Ruiz, I.M., Anguita, J., Meana, J.J., Ding, Y., Dutta, R., Schirmer, L., Igoillo-Esteve, M., Santin, I. and Castellanos-Rubio, A. (2025) An Inflammation Associated lncRNA Induces Neuronal Damage via Mitochondrial Dysfunction. *Molecular Therapy Nucleic Acids*, 36, 102533.
5. Wu, H.*, Yu, H.*, Zhang, Y.*, **Yang, B.***, Sun, W., Ren, L., Li, Y., Li, Q., Liu, B., Ding, Y. and Zhang, H. (2024) Unveiling RNA structure-mediated regulations of RNA stability in wheat. *Nature Communications*, 15, 10042.
6. Yu, H.*, Li, F.*, **Yang, B.***, Qi, Y., Guneri, D., Chen, W., Waller, Z.A., Li, K. and Ding, Y. (2024) iM-Seeker: a webserver for DNA i-motifs prediction and scoring via automated machine learning. *Nucleic Acids Research*, 52, W19-W28.
7. **Yang, B.***, Guneri, D.*, Yu, H.*, Wright, E.P., Chen, W., Waller, Z.A. and Ding, Y. (2024) Prediction of DNA i-motifs via machine learning. *Nucleic Acids Research*, 52, 2188-2197.
8. Yu, H.*, Qi, Y.*, **Yang, B.**, Yang, X. and Ding, Y. (2023) G4Atlas: a comprehensive transcriptome-wide G-quadruplex database. *Nucleic Acids Research*, 51, D126-D134.

The iM prediction in Chapter 2 (2.3.1-2.3.4) is based on two publications 6 and 7 where Bibo Yang is a co-first author. The whole Chapter 3 is based on publication 5 where Bibo Yang is a co-first author.

Bibliography

Abou Assi, H., *et al.* i-Motif DNA: structural features and significance to cell biology. *Nucleic acids research* 2018;46(16):8038-8056.

Achar, A. and Sætrom, P. RNA motif discovery: a computational overview. *Biology direct* 2015;10(1):61.

Allan, M.F., *et al.* Discovery and Quantification of Long-Range RNA Base Pairs in Coronavirus Genomes with SEARCH-MaP and SEISMIC-RNA. *bioRxiv* 2024:2024.2004.2029.591762.

Andrews, R.J., Roche, J. and Moss, W.N. ScanFold: an approach for genome-wide discovery of local RNA structural elements—applications to Zika virus and HIV. *PeerJ* 2018;6:e6136.

Aviran, S. and Incarnato, D. Computational Approaches for RNA Structure Ensemble Deconvolution from Structure Probing Data. *Journal of Molecular Biology* 2022;434(18):167635.

Aw, J.G.A., *et al.* Determination of isoform-specific RNA structure with nanopore long reads. *Nature biotechnology* 2021;39(3):336-346.

Aw, J.G.A., *et al.* In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Molecular cell* 2016;62(4):603-617.

Bedrat, A., Lacroix, L. and Mergny, J.-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic acids research* 2016;44(4):1746-1759.

Begik, O., *et al.* Long-read transcriptome-wide RNA structure maps using DMS-FIRST-seq. *bioRxiv* 2025:2025.2004. 2022.649934.

Belmonte-Reche, E. and Morales, J.C. G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. *NAR genomics and bioinformatics* 2020;2(1):lqz005.

Benabou, S., *et al.* Understanding the effect of the nature of the nucleobase in the loops on the stability of the i-motif structure. *Physical Chemistry Chemical Physics* 2016;18(11):7997-8004.

Bielecka, P., Dembska, A. and Juskowiak, B. Monitoring of pH using an i-motif-forming sequence containing a fluorescent cytosine analogue, tC. *Molecules* 2019;24(5):952.

Bizuayehu, T.T., *et al.* Long-read single-molecule RNA structure sequencing using nanopore. *Nucleic acids research* 2022;50(20):e120-e120.

Bohn, P., *et al.* Nano-DMS-MaP allows isoform-specific RNA structure determination. *Nature Methods* 2023;20(6):849-859.

Borovská, I., *et al.* Identification of conserved RNA regulatory switches in living cells using

RNA secondary structure ensemble mapping and covariation analysis. *Nature Biotechnology* 2025:1-13.

Brixi, G.*, et al.* Genome modeling and design across all domains of life with Evo 2. *BioRxiv* 2025:2025.2002. 2018.638918.

Brooks, T.A., Kendrick, S. and Hurley, L. Making sense of G‐quadruplex and i‐motif functions in oncogene promoters. *The FEBS journal* 2010;277(17):3459-3469.

Brown, R.V.*, et al.* The consequences of overlapping G-quadruplexes and i-motifs in the platelet-derived growth factor receptor β core promoter nuclease hypersensitive element can explain the unexpected effects of mutations and provide opportunities for selective targeting of both structures by small molecules to downregulate gene expression. *Journal of the American Chemical Society* 2017;139(22):7456-7475.

Brownlee, J. XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn. Machine Learning Mastery; 2016.

Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. 2014.

Cagirici, H.B., Budak, H. and Sen, T.Z. G4Boost: a machine learning-based tool for quadruplex identification and stability prediction. *BMC bioinformatics* 2022;23(1):1-18.

Campos Carrillo De Preciado, Z.: University of East Anglia; 2018. Investigating stability of RNA i-motif.

Cao, X.*, et al.* Identification of RNA structures and their roles in RNA functions. *Nature Reviews Molecular Cell Biology* 2024;25(10):784-801.

Carbone, A., Zinovyev, A. and Képes, F. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 2003;19(16):2005-2015.

Chen, C.-Y.A. and Shyu, A.-B. AU-rich elements: characterization and importance in mRNA degradation. *Trends in biochemical sciences* 1995;20(11):465-470.

Chen, J.*, et al.* Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300* 2022.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 2016:785-794.

Chlebowski, A.*, et al.* RNA decay machines: the exosome. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 2013;1829(6-7):552-560.

Cho, H.*, et al.* Translational control of phloem development by RNA G-quadruplex–JULGI determines plant sink strength. *Nature plants* 2018;4(6):376-390.

Choudhary, K.*, et al.* dStruct: identifying differentially reactive regions from RNA structurome profiling data. *Genome Biology* 2019;20(1):40.

Collin, D. and Gehring, K. Stability of chimeric DNA/RNA cytosine tetrads: implications for i-motif formation by RNA. *Journal of the American Chemical Society* 1998;120(17):4069-4072.

Cordero, P. and Das, R. Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis. *PLOS Computational Biology* 2015;11(11):e1004473.

Cui, H.*, et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods* 2024;21(8):1470-1480.

Danaee, P.*, et al.* bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic acids research* 2018;46(11):5381-5394.

Day, H.A., Huguin, C. and Waller, Z.A. Silver cations fold i-motif at neutral pH. *Chemical Communications* 2013;49(70):7696-7698.

Debnath, M., Fatma, K. and Dash, J. Chemical regulation of DNA i‐motifs for nanobiotechnology and therapeutics. *Angewandte Chemie* 2019;131(10):2968-2983.

Decker, C.J.*, et al.* dsRNA-Seq: Identification of viral infection by purifying and sequencing dsRNA. *Viruses* 2019;11(10):943.

Deep, A.*, et al.* i-Motifs as regulatory switches: Mechanisms and implications for gene expression. *Molecular Therapy Nucleic Acids* 2025;36(1).

Deigan, K.E.*, et al.* Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences* 2009;106(1):97-102.

Delli Ponti, R.*, et al.* Rnavigator: a pipeline to identify candidates for functional RNA structure elements. *Frontiers in Virology* 2022;2:878679.

Deng, H.*, et al.* Rice in vivo RNA structurome reveals RNA secondary structure conservation and divergence in plants. *Molecular plant* 2018;11(4):607-622.

Dhapola, P. and Chowdhury, S. QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic acids research* 2016;44(W1):W277-W283.

Ding, Y.*, et al.* Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nature protocols* 2015;10(7):1050-1066.

Ding, Y.*, et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 2014;505(7485):696-700.

Do, C.B., Woods, D.A. and Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 2006;22(14):e90-e98.

Eddy, J. and Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic acids research* 2006;34(14):3887-3896.

Elemento, O., Slonim, N. and Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Molecular cell* 2007;28(2):337-350.

Elimelech-Zohar, K. and Orenstein, Y. An overview on nucleic-acid G-quadruplex prediction: from rule-based methods to deep neural networks. *Briefings in Bioinformatics* 2023;24(4):bbad252.

Feng, L. and Niu, D.-K. Relationship between mRNA stability and length: an old question with a new twist. *Biochemical genetics* 2007;45(1):131-137.

Fick, S.E. and Hijmans, R.J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 2017;37(12):4302-4315.

Fish, L.*, et al.* A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements. *Science* 2021;372(6543):eabc7531.

Fleming, A.M.*, et al.* 4 n－1 is a “sweet spot” in DNA i-motif folding of 2′-deoxycytidine homopolymers. *Journal of the American Chemical Society* 2017;139(13):4682-4689.

Fojtík, P. and Vorlícková, M. The fragile X chromosome (GCC) repeat folds into a DNA tetraplex at neutral pH. *Nucleic Acids Research* 2001;29(22):4684-4690.

Fu, L.*, et al.* UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic acids research* 2022;50(3):e14-e14.

Fujii, T. and Sugimoto, N. Loop nucleotides impact the stability of intrastrand i-motif structures at neutral pH. *Physical Chemistry Chemical Physics* 2015;17(26):16719-16722.

Garrison, E. and Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* 2012.

Gautheret, D., Konings, D. and Gutell, R.R. GU base pairing motifs in ribosomal RNA. *Rna* 1995;1(8):807-814.

Geisberg, J.V.*, et al.* Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* 2014;156(4):812-824.

Gong, J.*, et al.* xTrimoGene: an efficient and scalable representation learner for single-cell RNA-seq data. *Advances in Neural Information Processing Systems* 2023;36:69391-69403.

Goodarzi, H.*, et al.* Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* 2012;485(7397):264-268.

Guneri, D.*, et al.* Structural Insights into Regulation of Insulin Expression Involving i-Motif DNA Structures in the Insulin-Linked Polymorphic Region. *bioRxiv* 2023:2023.2006. 2001.543149.

Gurung, S.P.*, et al.* The importance of loop length on the stability of i-motif structures. *Chemical Communications* 2015;51(26):5630-5632.

Halvorsen, M.*, et al.* Disease-associated mutations that alter the RNA structural ensemble. *PLoS genetics* 2010;6(8):e1001074.

Hamada, M., *et al.* Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics* 2006;22(20):2480-2487.

Hanson, G. and Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nature reviews Molecular cell biology* 2018;19(1):20-30.

Harmanci, A.O., Sharma, G. and Mathews, D.H. Stochastic sampling of the RNA structural alignment space. *Nucleic Acids Research* 2009;37(12):4063-4075.

He, P.C. and He, C. m6A RNA methylation: from mechanisms to therapeutic potential. *The EMBO journal* 2021;40(3):e105977.

He, S., *et al.* Ribonanza: deep learning of RNA structure through dual crowdsourcing. *bioRxiv* 2024.

Hiller, M., *et al.* Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research* 2006;34(17):e117-e117.

Hochsmann, M., *et al.* Local similarity in RNA secondary structures. In, *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*. IEEE; 2003. p. 159-168.

Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic acids research* 2003;31(13):3429-3431.

Homan, P.J., *et al.* Single-molecule correlated chemical probing of RNA. *Proceedings of the National Academy of Sciences* 2014;111(38):13858-13863.

Hon, J., *et al.* pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* 2017;33(21):3373-3379.

Houseley, J., LaCava, J. and Tollervey, D. RNA-quality control by the exosome. *Nature reviews Molecular cell biology* 2006;7(7):529-539.

Hu, Y.-J. GPRM: a genetic programming approach to finding common RNA secondary structure elements. *Nucleic Acids Research* 2003;31(13):3446-3449.

Huppert, J.L. and Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic acids research* 2005;33(9):2908-2916.

Iaccarino, N., *et al.* Assessing the influence of pH and cationic strength on i-motif DNA structure. *Analytical and bioanalytical chemistry* 2019;411:7473-7479.

Incarnato, D., *et al.* In vivo probing of nascent RNA structures reveals principles of cotranscriptional folding. *Nucleic acids research* 2017;45(16):9716-9725.

Incarnato, D., *et al.* RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Research* 2018;46(16):e97-e97.

Jia, J., *et al.* Homology-mediated inter-chromosomal interactions in hexaploid wheat lead to specific subgenome territories following polyploidization and introgression. *Genome*

Biology 2021;22(1):26.

Jiao, Y., Riechmann, J.L. and Meyerowitz, E.M. Transcriptome-wide analysis of uncapped mRNAs in Arabidopsis reveals regulation of mRNA degradation. *The Plant Cell* 2008;20(10):2571-2585.

Kang, H.-J.*, et al.* The transcriptional complex between the BCL2 i-motif and hnRNP LL is a molecular switch for control of gene expression that can be modulated by small molecules. *Journal of the American Chemical Society* 2014;136(11):4172-4185.

Kastner, B.*, et al.* Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harbor perspectives in biology* 2019;11(11):a032417.

Kavita, K. and Breaker, R.R. Discovering riboswitches: the past and the future. *Trends in biochemical sciences* 2023;48(2):119-141.

Kertesz, M.*, et al.* The role of site accessibility in microRNA target recognition. *Nature genetics* 2007;39(10):1278-1284.

Kertesz, M.*, et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* 2010;467(7311):103-107.

Kharel, P.*, et al.* Stress promotes RNA G-quadruplex folding in human cells. *Nature communications* 2023;14(1):205.

Khoroshkin, M.*, et al.* A systematic search for RNA structural switches across the human transcriptome. *Nature Methods* 2024;21(9):1634-1645.

Kikin, O., D'Antonio, L. and Bagga, P.S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic acids research* 2006;34(suppl_2):W676-W682.

Kim, D.*, et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* 2019;37(8):907-915.

Kim, M.Y.*, et al.* Staufen1-mediated mRNA decay induces Requiem mRNA decay through binding of Staufen1 to the Requiem 3′ UTR. *Nucleic acids research* 2014;42(11):6999-7011.

Klein, R.J. and Eddy, S.R. RSEARCH: finding homologs of single structured RNA sequences. *BMC bioinformatics* 2003;4(1):44.

Knudsen, B. and Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research* 2003;31(13):3423-3428.

Komatsu, K.R.*, et al.* RNA structure-wide discovery of functional interactions with multiplexed RNA motif library. *Nature communications* 2020;11(1):6275.

Krueger, F. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data.

Babraham Institute 2015.

Kudlicki, A.S. G-quadruplexes involving both strands of genomic DNA are highly abundant and colocalize with functional sites in the human genome. *PLoS One* 2016;11(1):e0146174.

Kumar, J.*, et al.* Quantitative prediction of variant effects on alternative splicing in MAPT using endogenous pre-messenger RNA structure probing. *elife* 2022;11:e73888.

Kurosaki, T. and Maquat, L.E. Nonsense-mediated mRNA decay in humans at a glance. *Journal of cell science* 2016;129(3):461-467.

Kwok, C.K.*, et al.* A stable RNA G-quadruplex within the 5′-UTR of Arabidopsis thaliana ATR mRNA inhibits translation. *Biochemical Journal* 2015;467(1):91-102.

Kwok, C.K.*, et al.* rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nature methods* 2016;13(10):841-844.

Lee, B.D. Python implementation of codon adaptation index. *Journal of Open Source Software* 2018;3(30):905.

Leebens-Mack, J.H.*, et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 2019;574(7780):679-685.

Lemaître, G., Nogueira, F. and Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research* 2017;18(1):559-563.

Leppek, K., Das, R. and Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature reviews Molecular cell biology* 2018;19(3):158-174.

Levy, A.A. and Feldman, M. Evolution and origin of bread wheat. *The Plant Cell* 2022;34(7):2549-2567.

Liu, Z.*, et al.* In vivo nuclear RNA structurome reveals RNA-structure regulation of mRNA processing in plants. *Genome Biology* 2021;22(1):11.

Loke, J.C.*, et al.* Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant physiology* 2005;138(3):1457-1468.

Low, J.T. and Weeks, K.M. SHAPE-directed RNA secondary structure prediction. *Methods* 2010;52(2):150-158.

Lowe, T.M. and Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* 1997;25(5):955-964.

Lu, Z.*, et al.* RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* 2016;165(5):1267-1279.

Lu, Z.J., Gloor, J.W. and Mathews, D.H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *Rna* 2009;15(10):1805-1813.

Ma, X.*, et al.* Genome-wide characterization of i-motifs and their potential roles in the stability and evolution of transposable elements in rice. *Nucleic Acids Research* 2022;50(6):3226-3238.

Maaten, L.v.d. and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* 2008;9(Nov):2579-2605.

Man, O., Sussman, J.L. and Pilpel, Y. Examination of the tRNA adaptation index as a predictor of protein expression levels. In, *RECOMB Workshop on Systems Biology*. Springer; 2005. p. 107-118.

Manfredonia, I. and Incarnato, D. Structure and regulation of coronavirus genomes: state-of-the-art and novel insights from SARS-CoV-2 studies. *Biochemical Society Transactions* 2021;49(1):341-352.

Manfredonia, I.*, et al.* Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic acids research* 2020;48(22):12436-12452.

Marangio, P.*, et al.* diffBUM-HMM: a robust statistical modeling approach for detecting RNA flexibility changes in high-throughput structure probing data. *Genome Biology* 2021;22(1):165.

Marchand, B.*, et al.* Median and small parsimony problems on RNA trees. *Bioinformatics* 2024;40(Supplement_1):i237-i246.

Marinus, T.*, et al.* A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic acids research* 2021;49(6):e34-e34.

Mauger, D.M.*, et al.* mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences* 2019;116(48):24075-24083.

McInnes, L., Healy, J. and Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2017;2(11):205.

Meers, M.P., Tenenbaum, D. and Henikoff, S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics & chromatin* 2019;12:1-11.

Mergny, J.-L.*, et al.* Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic acids research* 2005;33(16):e138-e138.

Mir, B.*, et al.* Prevalent sequences in the human genome can form mini i-motif structures at physiological pH. *Journal of the American Chemical Society* 2017;139(40):13985-13988.

Mishima, Y. and Tomari, Y. Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish. *Molecular cell* 2016;61(6):874-885.

Missra, A. and von Arnim, A.G. Analysis of mRNA translation states in Arabidopsis over the diurnal cycle by polysome microarray. In, *Plant Circadian Networks: Methods and Protocols*. Springer; 2014. p. 157-174.

Mitchell III, D.*, et al.* Mutation signature filtering enables high-fidelity RNA structure probing at all four nucleobases with DMS. *Nucleic Acids Research* 2023;51(16):8744-8757.

Morandi, E.*, et al.* Genome-scale deconvolution of RNA structure ensembles. *Nature Methods* 2021;18(3):249-252.

Morandi, E., van Hemert, M.J. and Incarnato, D. SHAPE-guided RNA structure homology search and motif discovery. *Nature Communications* 2022;13(1):1722.

Mortimer, S.A., Kidwell, M.A. and Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* 2014;15(7):469-479.

Mullen, M.A.*, et al.* RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles. *Nucleic acids research* 2010;38(22):8149-8163.

Narsai, R.*, et al.* Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *The Plant Cell* 2007;19(11):3418-3436.

Nawrocki, E.P. and Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29(22):2933-2935.

Niu, K.*, et al.* BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in Bombyx mori. *Nucleic Acids Research* 2018;46(4):1710-1723.

Olson, S.W.*, et al.* Discovery of a large-scale, cell-state-responsive allosteric switch in the 7SK RNA using DANCE-MaP. *Molecular Cell* 2022;82(9):1708-1723.e1710.

Pan, X.*, et al.* Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC genomics* 2018;19(1):511.

Panda, S.*, et al.* hyppo: A multivariate hypothesis testing Python package. *arXiv preprint arXiv:1907.02088* 2019.

Parisien, M. and Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 2008;452(7183):51-55.

Parkhomchuk, D.*, et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic acids research* 2009;37(18):e123-e123.

Patro, R.*, et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* 2017;14(4):417-419.

Pavlova, N., Kaloudas, D. and Penchovsky, R. Riboswitch distribution, structure, and function in bacteria. *Gene* 2019;708:38-48.

Pedregosa, F.*, et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011;12:2825-2830.

Presnyak, V., *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* 2015;160(6):1111-1124.

Puig Lombardi, E. and Londoño-Vallejo, A. A guide to computational methods for G-quadruplex prediction. *Nucleic acids research* 2020;48(1):1-15.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841-842.

Ramírez-González, R., *et al.* The transcriptional landscape of polyploid wheat. *Science* 2018;361(6403):eaar6089.

Reuter, J.S. and Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics* 2010;11(1):129.

Rivas, E. RNA structure prediction using positive and negative evolutionary information. *PLoS computational biology* 2020;16(10):e1008387.

Rivas, E., Clements, J. and Eddy, S.R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods* 2017;14(1):45-48.

Rocher, V., *et al.* DeepG4: a deep learning approach to predict cell-type specific active G-quadruplex regions. *PLOS Computational Biology* 2021;17(8):e1009308.

Rogers, R.A., Fleming, A.M. and Burrows, C.J. Rapid screen of potential i-motif forming sequences in DNA repair gene promoters. *ACS omega* 2018;3(8):9630-9635.

Rouskin, S., *et al.* Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 2014;505(7485):701-705.

Roy, B. and Jacobson, A. The intimate relationships of mRNA decay and translation. *Trends in Genetics* 2013;29(12):691-699.

Saberi, A., *et al.* A long-context RNA foundation model for predicting transcriptome architecture. *BioRxiv* 2024:2024.2008. 2026.609813.

Sabi, R. and Tuller, T. Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias. *DNA Research* 2014;21(5):511-526.

Sabi, R., Volvovitch Daniel, R. and Tuller, T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* 2017;33(4):589-591.

Sahakyan, A.B., *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Scientific reports* 2017;7(1):14535.

Saldi, T., *et al.* Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Molecular cell* 2021;81(8):1789-1801. e1785.

Sato, K., Akiyama, M. and Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature communications* 2021;12(1):941.

Schuller, A.P. and Green, R. Roadblocks and resolutions in eukaryotic translation. *Nature*

Reviews Molecular Cell Biology 2018;19(8):526-541.

Sherpa, C.*, et al.* The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. *Nucleic acids research* 2015;43(9):4676-4686.

Sidorenko, L.V.*, et al.* GC-rich coding sequences reduce transposon-like, small RNA-mediated transgene silencing. *Nature plants* 2017;3(11):875-884.

Siegfried, N.A.*, et al.* RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods* 2014;11(9):959-965.

Singh, J.*, et al.* RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications* 2019;10(1):5407.

Snoussi, K., Nonin-Lecomte, S. and Leroy, J.-L. The RNA i-motif. *Journal of molecular biology* 2001;309(1):139-153.

Spasic, A.*, et al.* Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic acids research* 2018;46(1):314-323.

Spiegelman, Z., Golan, G. and Wolf, S. Don't kill the messenger: long-distance trafficking of mRNA molecules. *Plant Science* 2013;213:1-8.

Spitale, R.C.*, et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* 2015;519(7544):486-490.

Spitale, R.C. and Incarnato, D. Probing the dynamic RNA structurome and its functions. *Nature Reviews Genetics* 2023;24(3):178-196.

Su, Z.*, et al.* Genome-wide RNA structurome reprogramming by acute heat shock globally regulates mRNA abundance. *Proceedings of the National Academy of Sciences* 2018;115(48):12170-12175.

Sun, L.*, et al.* RNA structure maps across mammalian cellular compartments. *Nature structural & molecular biology* 2019;26(4):322-330.

Sun, L.*, et al.* Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell research* 2021;31(5):495-516.

Sutherland, C.*, et al.* A mechanosensor mechanism controls the G-quadruplex/i-motif molecular switch in the MYC promoter NHE III1. *Journal of the American Chemical Society* 2016;138(42):14138-14151.

Tao, S.*, et al.* i-Motif DNA: identification, formation, and cellular functions. *Trends in Genetics* 2024.

Thiel, B.C.*, et al.* 3D based on 2D: Calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements. *F1000Research* 2019;8:ISCB Comm J-287.

Thieme, C.J., *et al.* Endogenous Arabidopsis messenger RNAs transported to distant tissues. *Nature Plants* 2015;1(4):1-9.

Todd, A.K., Johnston, M. and Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic acids research* 2005;33(9):2901-2907.

Tomezsko, P.J., *et al.* Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature* 2020;582(7812):438-442.

Uhl, M., *et al.* GraphProt2: A graph neural network-based method for predicting binding sites of RNA-binding proteins. *BioRxiv* 2019:850024.

Underwood, J.G., *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods* 2010;7(12):995-1001.

Wachter, A., *et al.* Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing of mRNAs. *The Plant Cell* 2007;19(11):3437-3450.

Wan, R., *et al.* How is precursor messenger RNA spliced by the spliceosome? *Annual review of biochemistry* 2020;89(1):333-358.

Wang, J., *et al.* RNA structure profiling at single-cell resolution reveals new determinants of cell identity. *Nature Methods* 2024;21(3):411-422.

Wang, T., *et al.* RNA motifs and modification involve in RNA long-distance transport in plants. *Frontiers in Cell and Developmental Biology* 2021;9:651278.

Wang, Y., *et al.* Stable stem enabled Shannon entropies distinguish non-coding RNAs from random backgrounds. In, *2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*. IEEE; 2011. p. 184-189.

Washietl, S. Prediction of structural noncoding RNAs with RNAz. In, *Comparative Genomics*. Springer; 2007. p. 503-525.

Williams, S.L., *et al.* Replication-induced DNA secondary structures drive fork uncoupling and breakage. *The EMBO Journal* 2023;42(22):e114334.

Wright, E.P., *et al.* Epigenetic modification of cytosines fine tunes the stability of i-motif DNA. *Nucleic Acids Research* 2020;48(1):55-62.

Wright, E.P., Huppert, J.L. and Waller, Z.A. Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic acids research* 2017;45(6):2951-2959.

Wu, X. and Bartel, D.P. Widespread influence of 3′-end structures on mammalian mRNA processing and stability. *Cell* 2017;169(5):905-917. e911.

Xu, Y., *et al.* PrismNet: predicting protein–RNA interaction using in vivo RNA structural information. *Nucleic Acids Research* 2023;51(W1):W468-W477.

Yang, F., *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* 2022;4(10):852-866.

Yang, M., *et al.* Intact RNA structurome reveals mRNA structure-mediated regulation of miRNA cleavage in vivo. *Nucleic acids research* 2020;48(15):8767-8781.

Yang, M., *et al.* In vivo single-molecule analysis reveals COOLAIR RNA structural diversity. *Nature* 2022;609(7926):394-399.

Yang, X., *et al.* RNA G-quadruplex structures exist and function in vivo in plants. *Genome biology* 2020;21(1):226.

Yang, X., *et al.* RNA G-quadruplex structure contributes to cold adaptation in plants. *Nature communications* 2022;13(1):6224.

Yang, X., *et al.* Wheat in vivo RNA structure landscape reveals a prevalent role of RNA structure in modulating translational subgenome expression asymmetry. *Genome Biology* 2021;22(1):326.

Yao, Z., Weinberg, Z. and Ruzzo, W.L. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006;22(4):445-452.

Yu, A.M., *et al.* Computationally reconstructing cotranscriptional RNA folding from experimental data reveals rearrangement of non-native folding intermediates. *Molecular Cell* 2021;81(4):870-883.e810.

Yu, B., *et al.* Differential analysis of RNA structure probing experiments at nucleotide resolution: uncovering regulatory functions of RNA structure. *Nature Communications* 2022;13(1):4227.

Yu, G., *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* 2012;16(5):284-287.

Yu, H., *et al.* An interpretable RNA foundation model for exploring functional RNA motifs in plants. *Nature Machine Intelligence* 2024;6(12):1616-1625.

Yu, Y., Jia, T. and Chen, X. The ‘how’and ‘where’of plant micro RNA s. *New Phytologist* 2017;216(4):1002-1017.

Zanin, I., *et al.* Genome-wide mapping of i-motifs reveals their association with transcription regulation in live human cells. *Nucleic Acids Research* 2023;51(16):8309-8321.

Zeraati, M., *et al.* I-motif DNA structures are formed in the nuclei of human cells. *Nature chemistry* 2018;10(6):631-637.

Zhang, H. and Ding, Y. RNA Structure: Function and Application in Plant Biology. *Annual Review of Plant Biology* 2025;76.

Zhang, T., *et al.* Structured 3′ UTRs destabilize mRNAs in plants. *Genome Biology* 2024;25(1):54.

Zhang, W.*, et al.* tRNA-related sequences trigger systemic mRNA transport in plants. *The Plant Cell* 2016;28(6):1237-1249.

Zhang, Y.*, et al.* G-quadruplex structures trigger RNA phase separation. *Nucleic acids research* 2019;47(22):11746-11754.

Zhong, Y.*, et al.* RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* 2017;33(1):139-141.

Zhou, J.*, et al.* IRIS: A method for predicting in vivo RNA secondary structures using PARIS data. *Quantitative Biology* 2020;8(4):369-381.

Zhu, Q., Petingi, L. and Schlick, T. RNA-As-Graphs motif atlas—dual graph library of RNA modules and viral frameshifting-element applications. *International journal of molecular sciences* 2022;23(16):9249.

Zuker, M. On finding all suboptimal foldings of an RNA molecule. *Science* 1989;244(4900):48-52.