

Uncertainty-aware benchmarking reveals ambiguous transcripts in mRNA–lncRNA classification

Daniel Garcia-Ruano

IBGC, CNRS UMR 5095, Université de Bordeaux

Mikaël Georges

IBGC, CNRS UMR 5095, Université de Bordeaux

Saswat K. Mohanty

Molecular, Cellular, and Integrative Biosciences, Huck Institutes of the Life Sciences, Penn State University

Rahma Baaziz

IBGC, CNRS UMR 5095, Université de Bordeaux

Kateryna D. Makova

Department of Biology, Penn State University

Macha Nikolski

IBGC, CNRS UMR 5095, Université de Bordeaux

Domitille Chalopin

`domitille.chalopin-fillot@u-bordeaux.fr`

ImmunoConcEpT, UMR 5164, Université de Bordeaux

Research Article

Keywords: long non-coding RNA, transcript classification, benchmarking, uncertainty analysis, transposable elements, non-B DNA motifs

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11362934/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Uncertainty-aware benchmarking reveals ambiguous transcripts in mRNA–lncRNA classification

Daniel Garcia-Ruano ¹, Mikaël Georges ^{1,2}, Saswat K. Mohanty ^{3,4}, Rahma Baaziz ¹, Kateryna D. Makova ^{4,5}, Macha Nikolski ^{1,2,&}, Domitille Chalopin ^{1,6,&}

¹ IBGC, CNRS UMR 5095, Université de Bordeaux, Bordeaux 33000, France

² Bordeaux Bioinformatics Center, University of Bordeaux, Bordeaux 33000, France

³ Molecular, Cellular, and Integrative Biosciences, Huck Institutes of the Life Sciences, Penn State University, University Park, State College, PA, 16802, USA

⁴ Department of Biology, Penn State University, University Park, State College, PA, 16802, USA

⁵ Center for Medical Genomics, Penn State University, University Park, State College, PA, 16802, USA

⁶ ImmunoConcEpT, CNRS UMR 5164, Université de Bordeaux, Bordeaux 33000, France

& corresponding authors

List of author's emails:

Garcia-Ruano Daniel dani.garciaruano@gmail.com

Georges Mikaël mikael.georges@u-bordeaux.fr

Mohanty Saswat K. skm6640@psu.edu

Baaziz Rahma rahmabaazizwork@gmail.com

Makova Kateryna D. kdm16@psu.edu

Nikolski Macha macha.nikolski@u-bordeaux.fr

Chalopin Domitille domitille.chalopin-fillot@u-bordeaux.fr

Abstract

Background. Long non-coding RNAs (lncRNAs) have gained significant attention in recent years, yet distinguishing them from protein-coding transcripts remains challenging. Indeed, many lncRNAs share mRNA-like processing and existing sequence-derived signals do not fully capture the coding/non-coding boundary. Recent GENCODE annotation efforts revealed tens of thousands of novel lncRNA sequences as well as the reclassification of some lncRNAs into the protein-coding class, highlighting the need to better characterize transcript features associated with classification uncertainty and errors.

Results. We performed uncertainty-aware benchmarking by retraining and evaluating eight transcript classifiers under a controlled protocol on a label-stable GENCODE v46–v47 subset. Beyond conventional model evaluation metrics, we quantified inter-tool agreement and entropy-based uncertainty to stratify transcripts into low and high uncertainty groups. Although aggregate performance was high, ~45% of transcripts showed inter-tool discordance, particularly among lncRNAs. To expand standard sequence and ORF-derived signals, we incorporated repeat-derived features and non-canonical secondary structure from both mature transcripts and gene bodies. Feature analyses linked low-uncertainty predictions to strong coding-like signals, whereas high-uncertainty profiles exhibited mixed signatures. Alongside classical predictors in global importance analyses, repeat-derived features appeared as main contributors.

Conclusions. By combining controlled benchmarking with transcript-level agreement and uncertainty stratification, together with extended feature profiling, we identified patterns associated with classifier disagreement and misclassification. This novel framework provides practical guidance for interpreting predictions, motivating the development of more robust coding/non-coding classifiers, while also shedding light on the sequence properties that distinguish lncRNA sequences.

Keywords

long non-coding RNA, transcript classification, benchmarking, uncertainty analysis, transposable elements, non-B DNA motifs

Background

Long non-coding RNAs (lncRNAs) constitute a large and heterogeneous class of non-coding RNAs longer than 200 nucleotides, encompassing more than 35,000 annotated human loci in the latest GENCODE update [1,2]. Recognized as regulators of gene expression, they exert their functions through interactions with DNA, RNA, and proteins, influencing chromatin organization, RNA splicing, stability, and translation [3,4]. With their ability to regulate gene expression at multiple levels, lncRNAs are central players in many biological processes, such as X-chromosome inactivation or cell differentiation (reviewed in [5,6]). However, their functional diversity, high isoform variability, and numerous similarities with mRNAs such as splicing, capping, and polyadenylation, make them difficult to distinguish from protein-coding transcripts purely from transcript sequence. Additionally, the presence of antisense transcripts and sometimes their processing from intronic or repetitive sequences further blurs the boundary between protein-coding and non-coding genes.

Distinguishing these two types of transcripts has long challenged computational genomics. Early approaches such as Fickett's TESTCODE algorithm [7] detected coding regions *ab initio*. With the recognition of the pervasive transcription of non-coding genomic regions, tools such as PhyloCSF [8], CPC [9], and CPAT [10] were developed to classify transcripts using evolutionary conservation and sequence composition. Over the past decade, machine learning methods have incorporated broader feature sets, including k-mer frequencies, RNA secondary structure, and predicted protein properties; and, more recently, deep learning models have been trained on raw transcript sequences. Although these methods can achieve high accuracy on standard within-species benchmarks, performance can degrade substantially under class imbalance, cross-species transfer, and challenging transcript profiles

(e.g., extreme length or complex isoform structure) [11]. Recently, lncRNA-BERT leveraged transformer architectures and reported strong performance on several commonly used benchmarks [12]. However, important limitations remain, both in terms of data complexity and model limitations. This includes annotation updates, dataset shifts, sequence redundancy, context-length constraints for long transcripts, cross-species model generalization and limited biological interpretability.

As these classifiers proliferated, benchmarking studies have increasingly moved beyond analyzing accuracy to exploring their biological interpretability. For example, Klapproth et al. [11] argued that most classifiers detect the absence of coding signals rather than lncRNA-specific properties. Others reported systematic biases, such as performance drops for sequences outside typical length ranges of 1-10 kb [13], and consistent misclassification of particular transcript subsets across dozens of models [14]. While these studies provided valuable insights, analyses often focused on a narrow set of transcript properties (e.g., length). Moreover, many large-scale benchmarks provide limited control over dataset composition, which can introduce data leakage via redundancy, near-duplicate isoforms, or overlapping annotations between training and evaluation sets. As a result, key questions remain: to what extent do state-of-the-art tools produce concordant predictions on the same transcripts? Do “difficult” transcripts share distinctive and interpretable properties? And can features beyond sequence-derived statistics help explain disagreement and improve classification?

In this study, we address these limitations by (1) retraining eight representative coding-potential classifiers under a controlled protocol, (2) quantifying prediction confidence together with inter-model agreement to stratify transcripts into operational “difficulty groups”, and (3) analyzing the features associated with confident, correct and discordant predictions (Fig. 1). Beyond the raw transcript sequences, secondary structure and predicted protein characteristics, we investigated additional feature sets, including non-B DNA motifs, predicted from the genomic sequence and known to modulate transcription [15–18]; repetitive element

content from gene bodies and mature transcripts, potentially influencing the functional properties of the RNA itself; and RNA folding statistics describing the structure that could drive their functions at the transcript level [19,20]. This framework supports a more robust and reproducible comparison of tools, provides practical guidance on how to interpret model disagreements, and highlights transcript types that remain challenging, not only suggesting directions for future model development but also providing insights on lncRNA complexity.

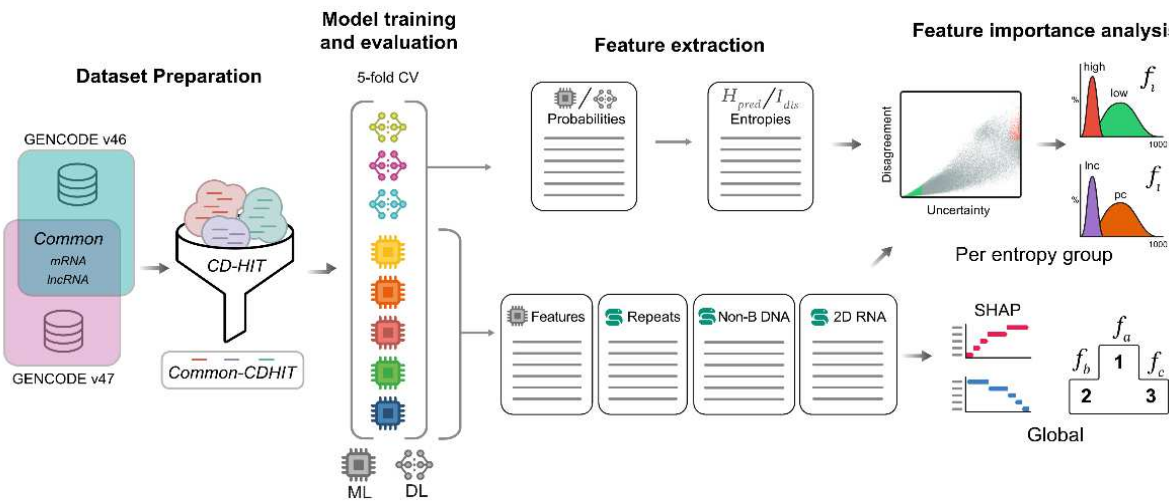


Figure 1. Overview of the benchmark approach: Identifying Classification Failures via Entropy and Atypical Structures

Results

Construction of a high-quality, non-redundant benchmarking dataset and extraction of DNA and RNA features

Fair comparison of coding/non-coding classifiers requires evaluation on identical train/test sets and explicit control over train-test leakage due to sequence redundancy (e.g., highly similar isoforms). However, most available models are trained on unpublished or outdated datasets, either making it difficult to retrain them under identical conditions and to quantify leakage and dataset shift.

To mitigate these issues in our study, we constructed a benchmarking dataset with three design principles: (1) annotation stability and curation, to reduce label noise; (2) reduced sequence redundancy, to limit near-duplicate information leakage; and (3) balanced training

classes (i.e., equal number of protein-coding and lncRNA transcripts), to avoid model learning a trivial majority class detection rule.

We used Human GENCODE v47 as the curated reference annotation for human transcripts. In a recent publication of the GENCODE consortium, Capture Long-Read Sequencing (CLS) combined with the updated TAGADA pipeline added 140,268 novel lncRNA transcripts (and 17,931 novel loci) in v47 compared to v46, roughly doubling the lncRNA catalog [1,21] (Fig. 2A, B). Because GENCODE labels are refined over time, transcripts annotated as lncRNA in one release can later be reassigned as protein-coding (or vice versa) as evidence accumulates. To reduce label instability, we defined a “common” subset comprising transcripts that (1) are present in both v46 and v47, (2) are labeled as either protein-coding or lncRNA, and (3) retain the same class label across releases (Table 1, Methods). This yields a conservative benchmark set focused on stable labels rather than maximal catalogue size.

To reduce sequence redundancy and minimize near-duplicate leakage, we clustered transcripts in the “common” set using CD-HIT [22], as is common practice in coding-potential tool development (e.g., CPC2 [23], lncDC [24], lncRNA-BERT [12]). Shortly, CD-HIT groups sequences into the same cluster based on a 90% similarity with the representative sequence. We selected the longest sequence per cluster as representative, ensuring maximal retention of exon content while avoiding redundant training examples. Our final dataset, named “common-CDHIT” contains 45,561 lncRNAs and 66,703 mRNAs, that were later stratified in five cross-validation fold subsets, in which the training set was class-balanced to support unbiased model training (36,448 transcripts per class). On the other hand, the held-out partition retained the class proportions of the non-redundant set (Table 1).

Table 1. Protein-coding and lncRNA transcript counts at each processing step					
Step	PC	lncRNA	Total	PC %	lncRNA %
All GENCODE v47	112,218	191,106	303,324	37.0	63.0

Table 1. Protein-coding and lncRNA transcript counts at each processing step					
Common (v46↔v47, same class)	111,838	59,674	171,512	65.2	34.8
Common-CDHIT (benchmarking set)	66,703	45,561	112,264	59.4	40.6
Training Set (per fold)	36,448	3,6448	72,896	50.0	50.0
Test Set (per fold)	30,255	9,113	39,368	76.9	23.1
Classified by all of the 8 tools	66,376	45,276	111,652	59.4	40.6

Transcripts from GENCODE v47 were filtered for annotation stability across versions, clustered with CD-HIT to remove redundant sequences and split into training and testing sets within a 5-fold cross validation setup. The final row reports all transcripts that were classified by all eight tools. This number differs from the total transcripts in the common-CDHIT set because some tools did not classify all transcripts, due to size thresholds or other constraints (see Methods, Table S19). PC: protein-coding transcripts; lncRNA: long non-coding RNA.

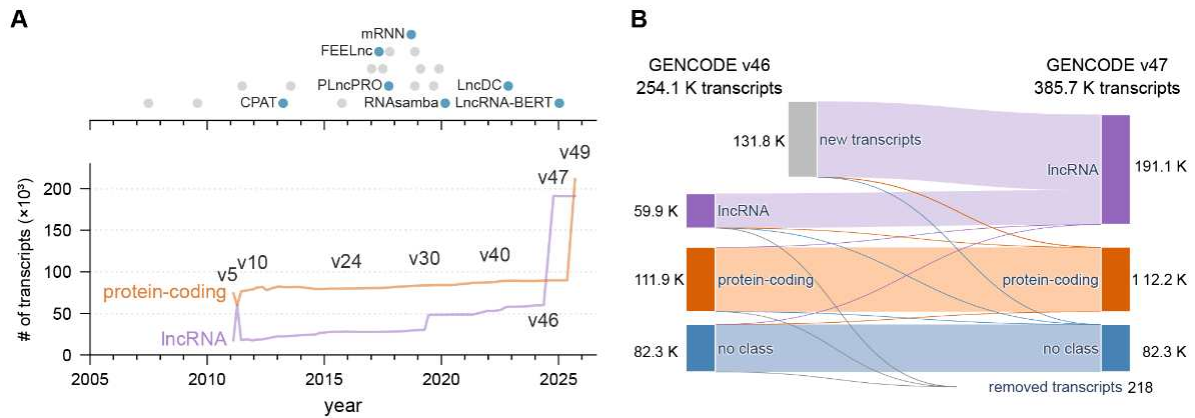


Figure 2. Evolution of transcript annotations and classification tools. (A) Timeline of transcript classification tools by publication date, with blue highlighting those benchmarked in this study (top). Number of human protein-coding and lncRNA transcripts annotated across all GENCODE database versions, with selected version numbers labelled for reference (bottom). Tools displayed in A are detailed in Table S1. (B) Sankey plot showing changes in coding class annotation (lncRNA, protein-coding, and no class) between versions v46 and v47 of the GENCODE database, including transcripts added or removed in v47. “No class” transcripts are those which do not fit in either of the protein-coding or lncRNA groups, such as short non-coding RNAs. Link widths are proportional to transcript numbers, illustrating the stability and transitions between versions. Data used to construct B are detailed in Table S2.

Beyond primary sequence features used by most classifiers, we computed additional DNA and transcript-associated features that are rarely incorporated in coding-potential tools, focusing on non-B DNA motifs, repetitive elements at the DNA and RNA level, and secondary and tertiary structure of RNA.

Repetitive elements were annotated using RepeatMasker and summarized per sequence after resolving overlaps among transposable elements (TEs), low-complexity regions, and tandem repeats (See Methods and Supplemental Material). For each transcript or gene body, we computed 170 repeat-derived features (Table S4), including counts, coverage measures and family-resolved summaries. For these repeat-derived features, we computed both absolute measures (counts/lengths) and length-normalized counterparts (counts per kb and coverage fractions) to mitigate transcript-length confounding. Likewise, non-B DNA motifs were predicted across the GRCh38 genome assembly using G4Discovery [25] and non-B gfa (https://github.com/abcsFrederick/non-B_gfa), yielding 199 motif-derived features per gene body (Table S5). Finally, RNA folding features were computed from the output of two different tools (Table S6). (1) ScanFold2 [26] was used to compute locally-stable RNA secondary structures, from which 69 features were extracted, including absolute and relative measures, together with position-relative features. (2) rg4detector [27] was used to predict the presence of RNA G-quadruplexes (G4s), which play roles in transcript regulation from transcription to translation [28–30]. Consistent with previous reports, protein-coding and lncRNA sequences in our dataset show distinct distributions of repetitive element content [31], RNA secondary structure [32], and non-B DNA motifs [33] (Table 2; Table S7), suggesting these features may carry informative signals for transcript classification.

In summary, the “common-CDHIT” dataset provides a label-stable (v46–v47 consensus) and redundancy-reduced benchmark, with standardized splits and an extended per-transcript feature compendium, enabling controlled retraining and comparative evaluation of coding-

potential classifiers. All transcripts, splits, tool versions/parameters, and derived features are released with a fully reproducible workflow (Methods; Data availability).

Table 2. Descriptive statistics of DNA and RNA features for transcripts in the common-CDHIT set

	Protein-coding		lncRNA	
	Median $\pm$ MAD		Median $\pm$ MAD	
Transcript length (nt)	1620 $\pm$ 1012		1028 $\pm$ 457	
	Counts (mean $\pm$ std)	% coverage (mean $\pm$ std)	Counts (mean $\pm$ std)	% coverage (mean $\pm$ std)
Rep (RNA, total)	1.50 $\pm$ 2.55	7.12 $\pm$ 11.57	2.27 $\pm$ 2.39	28.27 $\pm$ 24.35
Rep (DNA, total)	101.97 $\pm$ 187.69	36.57 $\pm$ 17.98	75.68 $\pm$ 140.19	41.27 $\pm$ 21.06
NBD (total)	288.02 $\pm$ 551.34	13.63 $\pm$ 16.79	215.73 $\pm$ 415.33	13.10 $\pm$ 26.96
rG4 peaks (total)	0.20 $\pm$ 0.59	NA	0.11 $\pm$ 0.40	NA

For repeats, non-B DNA motifs (NBD) and rG4 peaks, counts indicate the total number of recovered hits (i.e. number of repeat insertions extracted from RepeatMasker after nested insertion resolution, predicted NBD motifs, and rG4detector peaks), and coverage represents the percentage of the sequence that is covered by repeats or non-B DNA motifs. Mann-Whitney tests comparing protein-coding and lncRNA counts and coverage show all statistically significant p-values (Table S7). Features by repeat class and NBD motif type detailed in Table S7. Acronyms: Rep, Repeats; MAD, Median absolute deviation; NA: not applicable (Supplemental Methods).

High overall performance but substantial inter-tool disagreement

To assess coding-potential prediction performance on our leakage-controlled benchmark, we surveyed widely used and recently proposed transcript classification tools. We focused on methods with publicly available, functional code that can be retrained from scratch under a controlled protocol, enabling reproducible comparisons on identical training/test splits. We additionally prioritized tools that are actively maintained and collectively represent the major modeling families used for RNA classification.

We selected eight representative tools: lncRNA-BERT [12], CPAT [10], mRNN [34], FEELnc [35], RNAsamba [36], LncFinder [37], LncDC [24], and PLncPRO [38]. They cover transformer-based language models, neural networks, and classical machine-learning approaches (e.g., random forests and gradient-boosted trees). All tools were retrained on the

common-CDHIT benchmark using the same cross-validation splits and evaluation metrics (see Methods). We performed 5-fold cross-validation and report out-of-fold predictions, such that each transcript is evaluated exactly once in a held-out test fold, enabling transcript-level agreement analyses across tools.

Across all evaluation metrics, the eight classifiers produced broadly consistent rankings, with lncRNA-BERT, mRNN, and FEELnc forming the top tier. Most tools achieved strong overall performance on this benchmark, with all except CPAT exceeding 80% on balanced accuracy and showing comparably high precision/recall/F1 under the published thresholding rules for each tool (see Methods). Finally, lncRNA-BERT, mRNN and FEELnc consistently reached over 90% across the reported metrics (Fig. 3A; Table S8).

We next examined inter-tool agreement using the subset of transcripts that produced predictions from all eight tools (N = 111,652; Methods). More than half of transcripts (54.5%) were unanimously correct (all tools matched the GENCODE label), whereas a small fraction (0.9%) were consistently incorrect (all tools disagreed with the label) (Fig. 3B). The remaining 44.6% exhibited discordant predictions, indicating substantial tool-to-tool variability despite high aggregate scores. Stratifying by class, discordance was more prevalent among lncRNAs than protein-coding transcripts (Fig. 3B), consistent with the greater heterogeneity of lncRNA sequences and their more variable annotation status [39].

Agreement patterns further showed that many discordant cases (33.4%) corresponded to near-consensus predictions: transcripts correctly classified by most tools, with one or two tools producing the opposite label (Fig. 3C). In these near-consensus errors, CPAT accounted for a large share of lncRNA misclassifications, whereas PLncPRO and lncDC contributed disproportionately to protein-coding misclassifications (Fig. 3C).

These results highlight a key benchmarking insight: high headline metrics can coexist with substantial disagreement at the transcript level, motivating the analysis of disagreement profiles with transcript resolution. In the next section, we address this by defining transcript “difficulty” in practice, using prediction confidence and inter-tool agreement.

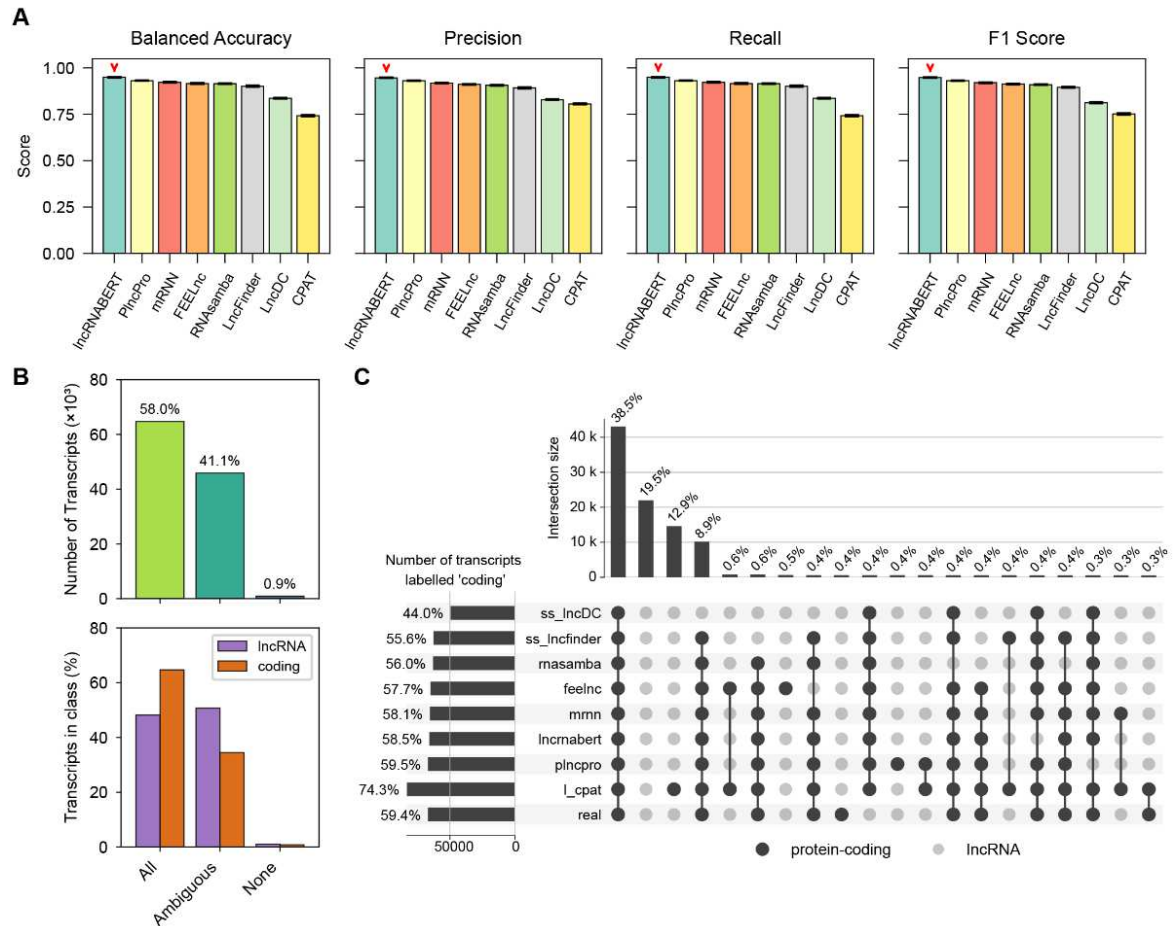


Figure 3. Aggregate performance and inter-tool label agreement. (A) Five-fold cross-validation performance (mean $\pm$ s.d.) of the eight retrained models on the common-CDHIT benchmark. Metrics: balanced accuracy, precision, recall, and F1-score (Methods). Average and per-fold values are detailed in Table S8 and Table S9, respectively. (B) Consensus analysis across tools. Transcripts are grouped as unanimously correct (all tools match the reference label), discordant (“Ambiguous”; at least one tool disagrees with the reference label), or unanimously incorrect (all tools disagree with the reference label). Left: transcript counts. Right: proportions within each reference class. (C) UpSet plot of top 20 agreement patterns. Black dots indicate the subset of tools predicting “protein-coding” for a given pattern; bars show the number of transcripts assigned “protein-coding” by that exact combination of tools (vertical bars) or by each tool (horizontal bars). Reference labels are from GENCODE.

Transcript stratification using inter-tool agreement and predictive uncertainty

Tools disagreed (produced different class labels) for about 45% of transcripts, with unbalanced results when lncRNA and protein-coding transcripts were considered separately. Moreover, simply counting how many tools disagree on how many transcripts does not provide granular, interpretable, or statistically meaningful insight into these discrepancies. To understand what drives model disagreement and what leads to full consensus on incorrect predictions, we adopted the following strategy: (1) we used an information theory-based decomposition commonly used to summarize ensemble (that is set of eight selected tools) predictions; (2) we estimated inter-tool disagreement as a disagreement score based on mutual information; and (3) we stratified transcripts into uncertainty groups using top and bottom 10th percentile threshold per coding class (Fig. 4A,B, see Methods). Our final groups contain 11,166 and 3,730 transcripts in the low- and high-entropy groups, respectively (Table 3).

Table 3. Summary of transcript counts in the low- and high-entropy groups				
	Protein-coding	lncRNA	Total	lncRNA/PC
Low entropy	6,638	4,528	11,166	0.68
High entropy	1,974	2,587	4,561	1.31
All transcripts	66,376	45,276	111,652	0.68
Low entropy transcripts correspond to those within the bottom 10th percentile of predictive entropy values computed over transcripts scored by all eight tools (N=111,652) for each coding class. High entropy transcripts were defined as those within the top 10th percentile for both predictive entropy and disagreement score.				

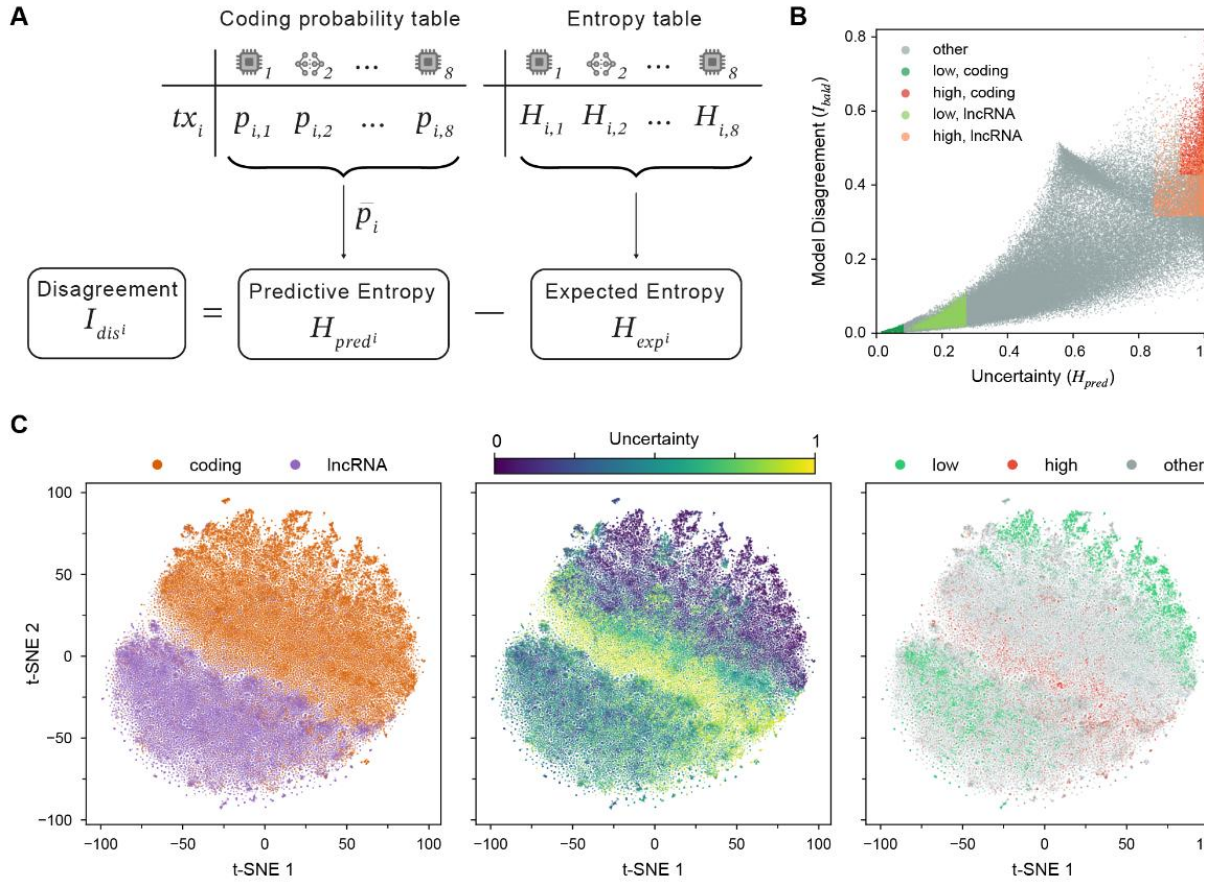


Figure 4. Transcript categorization based on confidence and agreement. (A) Process of entropy calculation. For each transcript, the predicted coding probabilities of eight models are used to calculate the mean predicted coding probability across tools, and the predictive entropy (H_{pred}) as a measure of classification uncertainty. The expected entropy is calculated as the average entropy of each model's predicted coding probability, and the disagreement score (I_{dis}) is then derived. (B) Scatter plot of transcripts by ensemble uncertainty and model disagreement. Shades of green represent transcripts with low classification uncertainty (10th percentile, coding: $H_{pred} < 0.080$; lncRNA: $H_{pred} < 0.271$ 10th percentile) and shades of red highlight transcripts with high classification uncertainty (90th percentile, coding: $H_{pred} > 0.925$; lncRNA: $H_{pred} > 0.844$ 90th percentile) and model disagreement (90th percentile, coding: $I_{dis} > 0.427$; lncRNA: $I_{dis} > 0.314$, 90th percentile). Transcripts outside these percentile ranges are represented as gray dots. (C) t-SNE visualization of transcripts after dimensionality reduction of the feature table (Table S3), comprising 124 features computed by the five machine learning models: CPAT (n=4 features) FEELnc (n=8), LncFinder (n=19), LncDC (n=28), and PlncPRO (n=65). Proximity reflects transcript similarity in the original sequence feature space. Plots are colored by coding class (left), predictive entropy (middle) and entropy groups (right).

To further investigate transcript distribution and contextualize it within their feature space, we performed t-SNE dimensionality reduction using the 124 features extracted by our five ML-based models, and projected transcript labels, entropy values, and uncertainty groups onto this embedding (Fig. 4C). In such visualization, transcript proximity reflects similarity in the original feature space. The class-colored projection revealed a broad but imperfect separation between lncRNA and protein-coding transcripts, with a fuzzy boundary and some transcripts misplaced in the opposing class neighborhood (Fig. 4C, left panel). Entropy values were distributed non-randomly, with low-entropy transcripts clustering tightly within their respective class domains, and high-entropy transcripts localizing at the lncRNA–protein-coding boundary (Fig. 4C, middle panel). This is consistent with the idea that their ambiguous feature profile is a primary driver of tool disagreement. Notably, the low entropy values cluster at the periphery of the protein-coding domain (upper part of the t-SNE) and inside the lncRNA cluster, whereas high-entropy transcripts are mostly located at the frontier between the two classes of transcripts (Fig. 4C, right panel). Closer examination of entropy value distributions revealed that protein-coding transcripts have lower values than lncRNA transcripts (Fig. S4).

The defined low- and high-entropy subsets enable downstream comparisons of feature distributions and misclassification. These subsets provide a benchmark-oriented way to localize where tools agree, disagree, and fail, using a common transcript set.

Identification of features driving confident classification and discriminating lncRNAs from protein-coding transcripts

Given the observation that high-entropy transcripts localize at the lncRNA–protein-coding classification boundary and that protein-coding transcripts tend to be less entropic than lncRNAs (Fig. S4), we next investigated which features drive this uncertainty, as well as broader lncRNA versus protein-coding discrimination. We focused on the five feature-based (ML) tools, which calculate primary sequence properties (e.g., transcript length, ORF length, k-mer frequencies) and added repeat-based (e.g., class/family diversity, coverage), non-B

DNA motif (e.g., presence of G4 motifs, etc.), and RNA structure (number/frequency of paired nucleotides, presence of RNA G4s, etc.) features previously calculated. Deep learning tools were included in agreement analyses but do not expose comparable engineered feature vectors, so they were excluded from feature-attribution analyses.

To assess how strongly each transcript feature associates with distinct entropy and classification groups, we performed univariate statistical analyses paired with effect size estimates (Fig. 5). Since the 734 features (descriptions in Tables S3-5) may share redundant information, we grouped 667 continuous features through hierarchical clustering and retained only the feature with the highest effect size per cluster, with categorical features left as is (see Methods). Transcript length was retained as the representative of its cluster, given its well-documented relevance in coding/non-coding classification. This yielded 220 clusters in total: 53 clusters from ML-based classifier features, 71 from repeat-based features (35 genomic and 36 transcript-based), 52 from non-B DNA features, 1 from rG4s, and 42 from RNA folding features (Table S17). Complete univariate test results containing non-representative features are shown in Tables S10 and S12.

Low- and high-entropy transcripts showed distinct feature associations (Fig. 5A,B). Continuous features such as transcript length, ORF molecular weights and the peak frame signal were significantly higher in low-entropy transcripts (Mann-Whitney U, FDR-corrected $p < 0.01$, centered VDA > 0.15 ; panel A). The number of structured nucleotides (z-score < -1) and maximum average z-score were also associated with low-entropy transcripts. Moreover, presence of repetitive elements (low complexity regions, simple repeats, transposons, LINEs, LTRs, and SINEs) and non-B DNA motifs (Z-DNA, triplex DNA, A-phased, direct, and mirror repeats, etc.) was also predictive of low-entropy transcripts, with higher motif counts associating with lower entropy. Notably, the maximum gap between either repetitive elements in transcript sequences was also associated with low entropy transcripts (Fig. 5A,

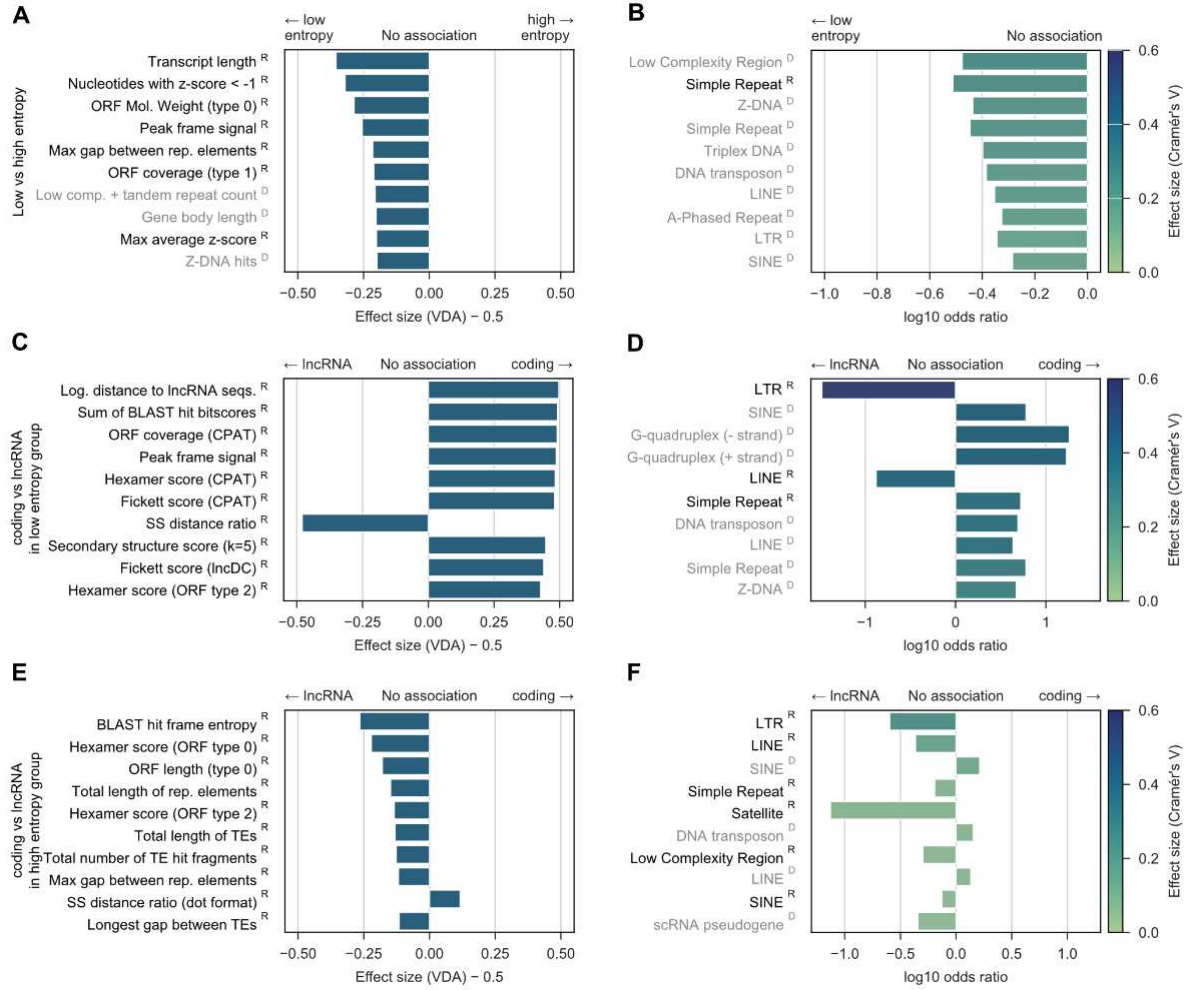


Figure 5. Differential feature associations with entropy groups and transcript class. Feature enrichment testing and prevalence summaries are shown for low- vs high-entropy transcripts (A–B) and for lncRNA vs protein-coding transcripts within the low-entropy (C–D) and high-entropy (E–F) subsets. Statistical tests were selected based on feature type, and all p-values were corrected for multiple testing (FDR; significance threshold $p < 0.01$). R superscripts indicate that features were computed at the transcript/RNA level, while D superscripts refer to gene bodies/DNA level. (A,C,E) Top 10 continuous features ranked by effect size (Mann–Whitney U). Effect size of Mann–Whitney U is reported as centered Vargha–Delaney A (cVDA = VDA - 0.5), where 0 indicates no effect and ± 0.5 indicates complete separation; the sign indicates association with the left vs right group. Statistically significant features with $|cVDA| > 0.1$ were considered to have a non-negligible association. Only top feature per feature cluster is reported, except transcript length, which was selected as representative for its cluster. (B, D, F) Top 10 categorical features ranked by effect size (chi-squared). Effect size of chi-squared is reported as Cramér's V (color scale truncated at 0.6 for visualization; true values in Tables S10–15). Odds ratios indicate direction and magnitude of enrichment. Statistically significant features with $V > 0.1$ were considered to have a non-negligible association.

B). Interestingly, none of the top 10 features showed association with the high-entropy transcript group.

To further investigate transcripts associated with low and high model uncertainty, we evaluated the differences between lncRNA and protein-coding transcripts within each entropy group (Fig. 5C-F).

In the low-entropy subset, protein-coding transcripts exhibited longer ORFs with higher coverage than lncRNAs (Fig. 5C). Sequence composition metrics, including the Fickett and hexamer scores had larger values in protein-coding transcripts. Interestingly, features derived from non-B DNA motifs and repetitive elements also emerged as discriminative (although not all of them passed the 0.1 significance threshold): presence of long terminal repeats (LTR) and long interspersed nuclear elements (LINE) was characteristic of low-entropy lncRNAs, while simple repeats, low-complexity regions and several non-B DNA motifs types were enriched in genes encoding protein-coding transcripts (Fig. 5D). Interestingly, LINEs were enriched in spliced lncRNAs but were mainly associated with protein-coding genes at the gene body level. Presence of G4s on either the plus or minus strands was also indicative of coding transcripts. Finally, a higher secondary-structure-derived frequency ratio (SS distance ratio) distinguished low-entropy lncRNAs from protein-coding transcripts, although its biological interpretation remains less straightforward, given that this ratio is derived from logarithmic rather than Euclidean distances.

On the other hand, feature associations shifted in high-entropy transcripts: lncRNAs displayed longer ORFs and higher hexamer scores, characteristics typically associated with coding potential (Fig. 5E). Moreover, lncRNAs also showed presence of simple repeats and low-complexity regions, previously identified as related to low-entropy coding transcripts (Fig. 5F). This suggests that high-entropy lncRNAs evade confident classification by mimicking coding sequence properties.

Comparing feature associations across entropy groups revealed both consistent and context-dependent patterns. Enrichment of LTR and LINE sequences was associated with spliced lncRNAs in both the low- and high-entropy groups, suggesting that the presence of such elements provides a broadly informative signal for lncRNA classification. In contrast, sequence composition and ORF-related features tend to have opposite associations in low- versus high-entropy transcripts. This divergence is consistent with the interpretation of high entropy transcripts occupying an ambiguous feature space between the two classes.

Overall, statistical enrichment tests showed that: (1) low-entropy transcript features align closely with well-established predictors, validating entropy-based stratification as a reliable post-hoc strategy for curating high-confidence datasets; (2) within the high-entropy group, discriminative signals are substantially weaker and sometimes inverted, reflecting the inherent ambiguity of these sequences; and (3) genomic features underused for transcript classification show concrete and interpretable associations, with LTR presence emerging as a consistent lncRNA signal across entropy regimes. These group-stratified analyses shed light on local drivers of classification uncertainty, but do not assess the global predictive power of individual features. Therefore, we next quantified the importance of these features for classification using multivariate models.

Identification of global informative features

To identify which features in our new dataset capture global discriminatory signals in a machine learning context, we assessed overall feature importance by examining the learning patterns of Random Forest (RF) models trained on a combined feature dataset. We chose RF because it (1) offers a strong performance–interpretability trade-off for attributing feature contributions and (2) is the same model family used by both PlncPro [38] and FEELnc [35], the best-performing feature-based tools in our benchmark (Fig. 3; Table S8). Similarly to the univariate analyses, we picked one representative feature per cluster, although this time

arbitrarily. We trained one RF model per fold (as defined previously in our cross-validation framework), computed SHAP values for each fold, and averaged them across folds to obtain stable estimates of feature contributions (Fig. 6; Table S16). This approach enabled us to explain global learning patterns by quantifying how each feature influenced transcript classification across the full dataset.

Feature importance analysis based on mean absolute SHAP values indicated that features with stronger contributions to classification were those capturing classical coding-potential information at the RNA level (Fig. 6A). Indeed, the top-ranked feature was the sum of all BLAST hit bitscores, reflecting the cumulative sequence similarity to known proteins as a dominant signal for protein-coding transcripts. Next, sequence composition features (k-mer and Fickett scores), together with ORF-related features (coverage, molecular weight) were in the top five.

Interestingly, eight repetitive element-derived RNA features were ranked among informative features, with mean absolute SHAP values between 0.005 and 0.018. While their individual contribution was modest, their collective presence in the top 20 suggests that the repeat content of transcripts carries discriminative information. In this regard, LTR presence stands out, as it was previously identified as a significant predictor in both low- and high-entropy transcript groups. This consistency across low- and high-entropy groups and model-level importance suggests that LTR presence reflects a distinct trait of lncRNAs. Together, these repeat-based features may provide discriminative value beyond classical coding-potential metrics.

To complement the global importance results and provide a feature-based interpretation of misclassification across our entropy-based groups, we examined representative transcripts characterized by distinct confidence and inter-tool agreement. We visualized SHAP values for well-annotated transcripts with available literature and experimental validation, each arbitrarily

selected as representative of one group (Fig. 6B-E): (1) a protein-coding transcript from the low-entropy group (EGF); (2) a protein-coding transcript from the high-entropy group (SWI5); (3) a lncRNA from the low-entropy group (MALAT1); and (4) a lncRNA from the high-entropy group (MEG3). We reported the top ten features, and one aggregated term summarizing the contributions of 228 features.

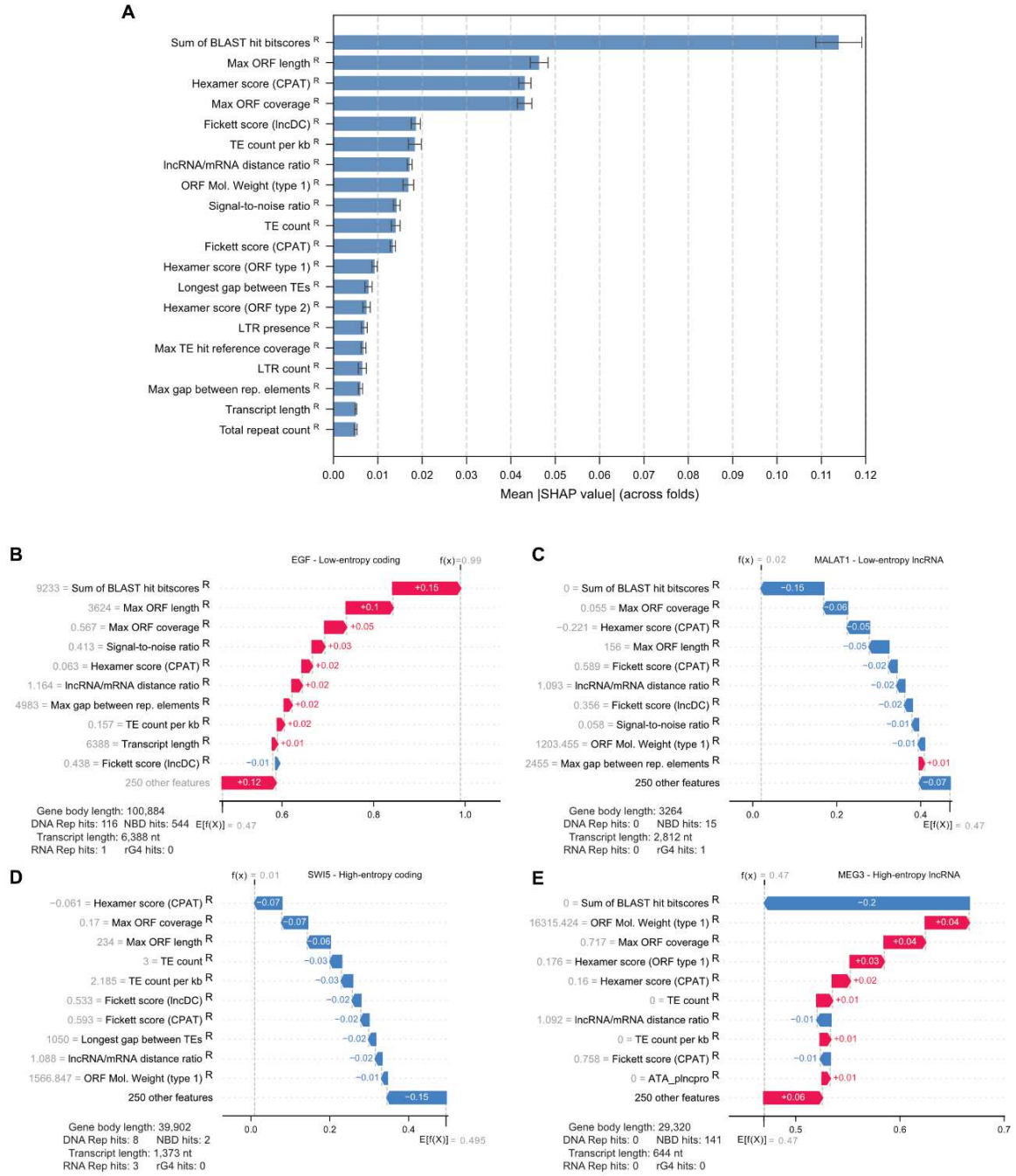


Figure 6. Global feature importance and transcript-level SHAP explanations. (A) Global importance of 25 selected features in the Random Forest model, summarized as mean absolute SHAP value ($\pm$ s.d.) across cross-validation folds (computed on held-out predictions; Methods). (B-E). SHAP waterfall plots for representative transcripts showing how individual features shift the model output from the baseline expectation $E[f(x)]$ to the final prediction $f(x)$ (coding-class output; reported as probability, Methods). Red bars indicate features increasing the coding prediction (positive SHAP values); blue bars indicate features decreasing the coding prediction (negative SHAP values, toward a lncRNA prediction). General descriptive statistics are indicated below the top 10 features for each transcript and feature values are reported on the left of the feature name.

The RF-predicted coding probability of EGF was 0.99, which classifies it correctly and with high confidence as a protein-coding transcript (Fig. 6B). The primary drivers of “protein-coding” classification of EGF were BLAST hit scores (+0.15) and ORF-related (total contribution of +0.15). Interestingly, repeat-related features contribute an additional 4 points. Finally, the 250 residual features accounted for +0.12 of the coding probability, suggesting that multiple features contribute with minimal but additive coding signal. Similar interpretation of the SHAP results for low entropy lncRNA MALAT1 confirmed its correct classification (0.02 coding probability, Fig. 6C), where the absence of BLAST hits (Sum of BLAST hit bitscores = 0; -0.15), short predicted ORFs (cumulative -0.12) and sequence composition features (Hexamer, K-mer, and Fickett scores, as well as signal-to-noise ratio; -0.11) were the main drivers of non-coding classification. Interestingly, the residual features pushed coding probability of MALAT1 toward 0 with a much smaller contribution to that observed in EGF SHAP values.

In contrast, SHAP values of high-entropy, high-disagreement transcripts revealed why these transcripts are misclassified. While the lack of BLAST hits contributed significantly (-0.2) to the non-coding classification of MEG3 (a lncRNA), eight of the top ten features pushed the classification towards a protein-coding transcript (cumulative +0.16). The residual features reinforced this decision, with an additional 6 points. On the other hand, the lack of features contributing towards the protein-coding class led to the wrong labelling of transcript SWI5. (Fig. 6D-E).

Beyond the representative cases examined here, the local interpretability of SHAP values of the RF models makes this approach applicable to any transcript of interest, to understand which sequence features drive its classification. Collectively, these analyses revealed the feature profiles underlying inter-tool disagreement. They highlighted discriminative signals exploited by the RF model alongside conflicting features that reduce prediction confidence.

Discussion

In this study, we aimed to go beyond aggregate performance metrics and characterize where and why tools disagree in protein-coding vs lncRNA classification. By benchmarking eight transcript classifiers with a robust dataset and a controlled protocol, we were able to (1) define transcript groups based on predictive entropy and inter-tool disagreement, (2) investigate feature enrichment across these groups; and (3) assess feature global contributions by training and interpreting RF models. Hypothesizing that limitations of tool's performances stem, in part, from the difficulty in capturing features beyond standard sequence and ORF-derived signals, we included repetitive elements, RNA structure, and non-B DNA motifs, and demonstrated their contribution to transcript classification.

Model performance and uncertainty

The eight tools we benchmarked showed good overall performance on our dataset, with lncRNA-BERT, mRNN and FEELnc scoring over 90% across balanced accuracy, precision, recall, and F1-score. More than half of the transcripts (54.5%) were correctly classified by all tools, but about 45% showed disagreement, especially for lncRNAs where only half were consistently identified correctly compared to 60% for protein-coding transcripts. A small fraction (0.9%) was misclassified by all tools despite high confidence. To dig into this, we used entropy from ensemble predictions to measure uncertainty per transcript, with high entropy means confusion. High-entropy transcripts clustered at the lncRNA/protein-coding boundary in t-SNE embedding, showing they have mixed features that confuse classifiers. Protein-coding transcripts had lower entropy overall than lncRNAs, suggesting lncRNAs are inherently harder to classify due to their diverse sequences and weaker coding signals. This uncertainty global analysis reveals why tools disagree: it is not just errors, but real biological ambiguity in many transcripts. Finally, transcripts with high confidence but wrong classification suggest they mimic features of the opposite class [40–42]. Together, these cases justify investigating

less common genomic features such as repeats and non-B DNA motifs to better capture these differences.

Feature drivers of uncertainty and (mis)classification

Our entropy-based stratification aimed to clarify why some transcripts are confidently classified while others remain ambiguous. Low-entropy sequences largely conformed to classical protein-coding and lncRNA signatures, with distinctive ORF lengths, k-mer and Fickett scores, as well as characteristic TE and non-B DNA profiles aligning with their reference labels. In contrast, high-entropy transcripts combined discordant features, such as coding-like hexamer composition and ORF metrics in lncRNAs. Thus, these high-entropic transcripts had poor signal dominance, consistent with their position at the lncRNA–protein-coding boundary. As opposed to repeat- and non-B-DNA-based features, the RNA folding features we studied showed weak correlation with entropy and transcript class. We hypothesize that this reflects a granularity gap: most of these features capture per-nucleotide or transcript-wide information, while previously used features (such as the dot-bracket k-mer composition or the SS ratios) capture structure at an intermediate scale.

SHAP analyses of the RF models reinforced this view and highlighted where our framework added biological interpretability. As expected, BLAST, hexamer and ORF-related features remained the strongest global contributors, but TE-derived metrics, although modest, collectively participated in decision confidence. The four representative transcripts illustrated how both the amount and organization of TE information might modulate classification: EGF and MALAT1 presented opposite extremes of transcript-related size information, supporting clear protein-coding versus lncRNA profiles. The protein-coding transcript SWI5 showed a clear lncRNA prediction: substantial TE content combined with low coding-potential scores. By contrast, MEG3 displayed a more ambiguous profile, with mixed feature-level predictions, pulling predictions toward the decision boundary. More broadly, the consistent importance of LTR signals across entropy groups suggests that TEs are a meaningful axis of lncRNA–

protein-coding discrimination in the “ambiguous zone” where current tools disagree, potentially reflecting underlying biological signals.

Implications for lncRNA biology

Repetitive elements show different patterns between lncRNAs and protein-coding transcripts. Several studies show that lncRNAs associate closely with TEs: Kapusta et al. [43] reported 75% of lncRNAs (GENCODE v13) harbor TE fragments, and Carlevaro-Fita et al. [44] found ~83% contain TEs. Our RepeatMasker results indicate that 70% of lncRNAs contain at least one repeat, compared to 30% of protein-coding transcripts. In our benchmark, repeat-derived features emerged as informative contributors alongside classical coding signals, supporting a predictive relevance of TE-associated sequence patterns. Consistent with the RIDL hypothesis, which proposes that TEs can contribute functional domains to lncRNAs through RNA-, DNA-, or protein-binding activities [27,42], the presence of these repetitive elements at the RNA level may be informative of biological function beyond simple classification performance. For example, deleting the Alu-derived SINEB1 domain in MALAT1 disrupts nuclear localization and triggers apoptosis [45]. In this context, some TE-derived exonic fragments could have been retained through co-option because they confer functional RNA domains, potentially contributing to the heterogeneous and ambiguous sequence profiles.

Our results further suggest that repeat-derived features contribute to lncRNA–mRNA discrimination in a context-dependent manner. LTRs and LINEs were consistently enriched in lncRNAs at the transcript level, whereas SINEs, DNA transposons, and LINEs were enriched in protein-coding gene bodies. Thus, repeat-related features may capture both the composition of mature transcripts and the genomic architecture of their corresponding loci. The additional enrichment of SINEs, simple repeats, and satellite sequences in high-entropy lncRNAs suggests that these elements may contribute to the ambiguous feature profiles observed near the classification boundary. SINEs may therefore help distinguish or, in ambiguous cases, blur the distinction between lncRNAs and protein-coding transcripts.

521
522 Non-B DNA motifs leading to non-canonical secondary structures encode regulatory
523 information in genes beyond primary sequence. Non-B DNA structures in the body of genes,
524 including genes transcribing lncRNAs, affect replication [18,46], transcription initiation and
525 progression. RNA G4s in lncRNAs also suggest regulatory roles [29]. In our analyses, non-B
526 DNA motif features (G4, Z-DNA, and triplex motifs) were overrepresented in the gene bodies
527 of low-entropy transcripts, and particularly in protein-coding genes. This is consistent with a
528 model in which non-B DNA motifs are subject to differential selective pressures in protein-
529 coding versus non-coding sequences, reflecting both the compositional constraint of GC
530 content on motif formation and the selection against destabilizing structures in coding exons
531 through synonymous codon choice [47,48]. For instance, G4s are overrepresented at UTRs
532 but are underrepresented at the non-transcribed strand of protein-coding exons [25]. Beyond
533 G4s, other non-B DNA structures depend on underlying sequence repeat motifs, and whether
534 these are subject to comparable coding/non-coding asymmetries remains an open question.
535
536 High-entropy lncRNAs may represent a heterogeneous group with weaker or conflicting
537 sequence signals, resulting in greater disagreement among classification tools. Although
538 previous studies have incorporated conservation-related features into lncRNA classifiers or
539 examined lncRNA age and conservation separately [49,50], the relationship between
540 evolutionary age and classifier performance has not been systematically assessed. One
541 possible explanation is that they are evolutionarily younger or less conserved than low-entropy
542 lncRNAs and therefore lack features consistently recognized by current classifiers. However,
543 entropy is not itself a direct measure of age or evolutionary constraint. This hypothesis could
544 be tested by comparing conservation scores using phastCons [51], syntenic conservation,
545 and, for transcripts with putative ORFs, PhyloCSF [8] scores between high- and low-entropy
546 groups. In this context, the enrichment of repeat-derived features in high-entropy lncRNAs
547 may reflect the contribution of TEs to lncRNA emergence and diversification as mentioned
548 earlier.

Limitations and future directions

The model ensemble included eight classifiers; however, feature importance analyses were restricted to the feature-based models due to the inherent interpretability challenges of DL models. Further work integrating attention weights from the DL models would help bridge this gap. Accordingly, our modular pipeline architecture allows the inclusion of additional tools, extending the ensemble with future methods as the state of the art evolves.

A related concern is the stability of the entropy group assignments upon changes in ensemble composition. Defining these groups by using thresholds on entropy values instead of vote-count agreement limits the impact of ensemble variation. Future analyses could treat entropy as a continuous response variable in a regression framework, which would allow for a more granular characterization of features predictive of classification uncertainty.

Our framework relies on quantifying prediction disagreement across ensemble members based on sequence-derived features, pinpointing transcripts for which the sequence information alone is insufficient to confidently assign a class. Importantly, this could be useful to flag transcripts for experimental validation or reannotation. However, model uncertainty should not be interpreted as biological ambiguity. Coding potential is inherently context-dependent: lncRNAs frequently show tissue-specific expression, some undergo condition-specific translation through tightly regulated ribosome recruitment, and dual-function transcripts with both coding and non-coding roles blur the boundary between classes. As a result, such transcripts may be biologically ambiguous yet confidently classified by our ensemble, or vice versa.

While the current set of features has proven useful for high-performing models, it exhibits high multicollinearity, with many features correlated with transcript length. This can affect interpretability: when globally important features are highly correlated, feature-importance rankings may distribute importance across redundant predictors and can obscure

characteristics that matter for specific outlier transcripts or borderline cases. Moreover, SHAP-based interpretation depends on the choice of explanatory model (here a Random Forest), which can itself be influenced by collinearity in the feature space. Exploring alternative interpretability frameworks and model classes less sensitive to correlated predictors may help reveal complementary signals.

Finally, our analyses incorporated repeat-derived features at both the mature-transcript and gene-body levels. Transcript-level features capture repeat sequences retained in mature RNAs, whereas gene-body features provide a broader locus-level context that includes intronic sequences. These intronic features may also influence transcript biology from transcriptional regulation to RNA maturation [52,53]. Although intronic features were not analysed as a separate class, their contribution is partly reflected in the gene-body measurements. Notably, the most informative repeat-derived features were transcript-level, supporting the relevance of repeat sequences retained in mature RNAs for coding/non-coding classification. As recent work has demonstrated the value of alternative genomic-context features in coding/non-coding classification [54], future exploration could include positional information (intergenic, antisense or overlapping regions), exon-intron structure, chromatin accessibility, and phylogenetic conservation metrics to better capture evolutionary and regulatory constraints on transcripts structure, regulation and function.

Conclusions

Our benchmarking strategy helps elucidate transcript profiles that drive misclassification and inter-tool disagreement in protein-coding versus lncRNA prediction. It serves two main purposes: (1) highlight transcript properties associated with coding versus non-coding types, and (2) identify feature families that can improve the robustness and interpretability of future classifiers.

Our study further shows that benchmarking can be informative beyond aggregate performance metrics, by revealing transcript-level agreement patterns and systematic characteristics of “easy” versus “difficult” cases. Refining this framework will support more interpretable and biologically informed models, and provide a clearer view of the heterogeneous sequence profiles that underlie lncRNA annotation.

Altogether, our results motivate uncertainty-aware, feature-rich classifiers and a community practice of propagating prediction confidence into downstream annotation and validation. Beyond improving computational workflows, our study contributes to refining the biological definition of lncRNAs.

Methods

This study consists of four main components: (1) constructing a label-stable, redundancy-reduced benchmark dataset with extended genomic and structural features (common-CDHIT); (2) benchmarking eight lncRNA/mRNA classifiers under a standardised 5-fold cross-validation framework, producing out-of-fold coding probabilities and binary labels; (3) characterising transcript-level classification uncertainty and inter-tool disagreement using entropy-based metrics, and stratifying transcripts into distinct uncertainty regimes; and (4) performing feature importance analyses across complementary feature sets to identify the sequence features driving classification behaviour. Each component is described in the sections below. Additionally, the Fig. S1 summarizes the general workflow of the study, including dataset pre-processing, model training and evaluation as well as downstream analyses.

Construction of the benchmarking dataset

Dataset construction involved three steps: (1) selecting transcripts with stable protein-coding or lncRNA biotype assignments across GENCODE releases v46 and v47; (2) applying sequence-level redundancy reduction using CDHIT to prevent data leakage; and (3) computing an extended set of genomic features (repetitive elements and non-B DNA structural

motifs) to complement sequence-intrinsic features of the benchmarked tools. The resulting dataset is referred to as the common-CDHIT set.

1. Label-stable GENCODE dataset

We downloaded human transcript sequences and annotations from the GENCODE project (releases v46 and v47) based on the GRCh38 genome assembly. Transcripts were classified into two main categories: protein-coding transcripts and lncRNA, based on the GENCODE FASTA files (e.g., gencode.v46.pc_transcripts.fa and gencode.v46.lncRNA_transcripts.fa) and the respective files for each release. Transcripts outside these categories were discarded for our analyses.

We then generated a dataset containing only transcripts present in both releases v46 and v47 with no change in class label. We refer to this subset as the “common dataset”. This cross-release filtering focuses model training and evaluation on transcripts with stable annotations, improving the reliability of downstream analyses.

2. Redundancy reduction of transcripts

To minimize sequence similarity within and between training and testing sets and prevent overfitting, we surveyed the strategies used by the original publications of the tools used in the study and followed the most used approach. Clustering was performed on the “common dataset” using CD-HIT with default parameters (90% sequence identity threshold, 90% coverage, word size of 8 nucleotides) [22]. For each cluster, we retained only the representative sequence, creating a non-redundant dataset “common-CDHIT” for downstream analyses.

3. Computation of DNA and RNA features

Repetitive element analyses

Transcripts in the common-CDHIT dataset were scanned for repetitive elements using RepeatMasker v4.2.1 [55] in sensitive mode with the default human Dfam library (v3.9). When

overlaps occurred, overlapping annotations were resolved by retaining high-confidence hits based on alignment scores and annotation quality (see Supplemental Methods). Retained hits were separated in three major classes: (1) transposable elements, (2) pseudogenes, and (3) simple repeats and low complexity regions. Then, each class was processed using a custom hierarchical feature extraction pipeline (https://github.com/cbib/rep_extraction_pipeline), which computed 170 manually curated features at three levels: (1) individual hits (length, divergence, indels, alignment scores), (2) aggregated elements accounting for hit fragmentation (total length, fragment count, class and family assignment), and (3) transcript-level summaries (total hit counts, total coverage, hit counts per family, presence/absence flags, and inter-hit statistics).

In addition, unspliced versions of the transcripts (DNA) were extracted from GENCODE annotations (see Supplemental methods), and a second set of 170 repetitive element-related features were computed following the same pipeline.

Non-B DNA prediction

Non-B DNA motifs were predicted on the GRCh38 reference assembly using the *non-B_gfa* (gfa) pipeline (https://github.com/abcsFrederick/non-B_gfa), generating genomic intervals for A-phased repeats (APR), direct repeats (DR), inverted repeats (IR), mirror repeats (MR), short tandem repeats (STR), and Z-DNA. Triplex-forming motifs (TRI) were defined as the subset of MR predictions flagged as triplex-prone by the gfa output. G-quadruplexes (G4) were predicted on GRCh38 using G4Discovery [25] with default parameters, reporting both pqsfinder [56] and G4Hunter [57] scores. All non-B DNA motif calls were intersected with GENCODE v47 gene intervals using BEDtools to assign motif occurrences to genes encoding transcripts in the common-CDHIT dataset (Supplemental Information).

For each transcript, we computed 178 non-B DNA features, including per-motif-type presence/absence, hit counts, length and gap statistics, as well as global summary features

(e.g., diversity, coverage) calculated on the full gene bodies(exons+introns). Because a locus can support multiple non-B DNA structural annotations, overlapping motif calls were retained (no overlap resolution) when computing features. Feature extraction was performed using a custom pipeline (https://github.com/cbib/nbd_extraction_pipeline).

RNA secondary structure predictions

Two sets of secondary structure features were computed: (1) RNA G-quadruplex (rG4) predictions and (2) transcript secondary structure metrics.

rG4 prediction was performed by rg4detector, which uses a ML model trained on real rG4-seq sequencing data [27]. This makes rg4detector more adapted to predicting RNA G4-forming sequences than tools such as G4Discovery, and shows a different distribution of predictions along transcript sequences (Fig. S5). This tool outputs a per-nucleotide score that indicates how biased the sequencing reads are under rG4-prone conditions. We defined rG4 elements as score peaks above a default threshold (1.56), and computed a presence/absence flag, peak count, density per kilobase and average relative position (see Supplemental methods).

Secondary structure features were derived from ScanFold2 [26] predictions, which estimate local RNA folding within sliding windows, and aggregate these results into per-nucleotide metrics and overall structural profiles. ScanFold2 was run using default parameters (window size of 120 nt; dinucleotide shuffling mode). From these outputs, we computed 4 groups of transcript features: (1) summary statistics (min, mean, median, max) for each structural metric (z-score, minimum free energy [MFE], and ensemble diversity [ED]); (2) the relative position of the minimum value for each metric; (3) counts and proportions of nucleotides that are paired, involved in competing base-pair interactions, or exhibit low z-scores; and (4) positional enrichment of nucleotides of group 3, calculated as the proportion that fall within each of ten equally-sized transcript segments. Feature extraction was performed using a custom pipeline (https://github.com/cbib/secondrna_pipeline).

Finally, all features were aggregated together and dimensionality was reduced through a hierarchical clustering approach described above.

Benchmarking protocol

The benchmarking workflow consisted of: (1) selecting retrainable tools with public, runnable code; (2) retraining each tool under the same 5-fold cross-validation splits; (3) producing coding scores and binary labels for out-of-fold transcripts; and (4) evaluating global classification performance with macro-averaged metrics. These steps are detailed below.

1. Selection of retrainable tools

Selection of tools for the benchmark was guided by five criteria:

- public source code availability,
- compatibility with custom retraining,
- capacity to yield soft predictions,
- algorithmic diversity,
- support for feature-based interpretability.

These criteria are further detailed in Supplemental Methods. The final selection of tools is detailed below in Table 4.

Table 4: Tools selected for benchmarking				
Tool	Release date	Category	Core Algorithm(s)	References
CPAT	2013	Feature-based	Logistic Regression (LR)	[10]
FEELnc	2017	Feature-based	Random Forest (RF)	[34]
PlncPro	2017	Feature-based	Random Forest	[37]
mRNN	2018	DL-based	Recurrent neural network (RNN)	[33]
LncFinder	2019	Feature-based	SVM/RF/LR/ELM	[36]

Table 4: Tools selected for benchmarking				
RNAse	2020	DL-based	IGLOO (RNN)	[35]
LncDC	2022	Feature-based	XGBoost	[23]
lncRNA-BERT	2025	DL-based	Transformer (BERT-medium) + CSE	[12]

Technical details about the selected models are available in Supplemental Methods. For tools that proposed several training methods, we selected the one displaying better classification performance. Finally, we ensured tool predictions were diverse enough to be able to analyze classification differences (Figure S1). All tools were used for comparison analyses (see uncertainty and agreement analyses). However, note that feature analyses were only performed using the five feature-based tools (i.e., CPAT, FEELnc, PlncPro, LncFinder and LncDC).

2. Cross-validation setup and model training

All tools were trained following the same benchmarking setup. The common-CDHIT dataset was split in 5 cross-validation folds, where 20% transcripts from each class were held-out as a fixed test set per fold. Given the class imbalance in the original dataset, all transcripts belonging to the minority class (i.e. all lncRNAs) were retained in the training set. To avoid training bias due to the higher prevalence of protein-coding sequences, a random sample of equal size was drawn from the remaining protein-coding transcripts. Additionally, for tools incorporating a validation step during model training, a subset of training transcripts was held-out for this purpose, following each tool's published recommendations (Supp. Methods)

The eight models were trained using the original protocols, using either (1) scripts based on the tool libraries (e.g., LncFinder), (2) original source code with minimal modifications (e.g., LncDC, mRNN) or (3) directly executable training scripts provided by the authors. Exact parameters for each tool are detailed in Supplemental Methods.

3. Coding scores and label assignment

Transcript classification was performed independently for each tool and fold. Each model produced both a coding probability score (ranging from 0 to 1) and a binary classification label (coding / non-coding). Decision thresholds were established according to published procedures, with most tools applying a default 0.5 probability cutoff. For tools using custom thresholds, we recalculated them accordingly per fold before determining the classification label. For tools that define lncRNAs as the positive class, coding scores were derived by inverting the non-coding scores. A summary of model probabilities and entropies, reflecting how confident model predictions are, is provided in Fig. S4.

Classification results aggregation

Due to constraints imposed by certain benchmarked tools (e.g., minimum or maximum transcript length, absence of identified ORF, presence of non-canonical nucleotides), 612 transcripts could not be assigned a label by all classifiers. These transcripts were excluded from downstream analyses requiring classification scores from all tools, resulting in a final dataset of 111,652 sequences with associated features, coding probabilities and classification labels. This set of classified transcripts was used for both global performance evaluation of models, entropy analyses and feature importance.

4. Classification performance evaluation

Global model performance was evaluated using standard metrics: balanced accuracy, precision, recall and F1-score. Given the class imbalance, macro-averaged scores were computed. Detailed formulas for the scores are further detailed in Supplemental Methods.

Uncertainty and agreement analyses

Our 8-tool ensemble was used to investigate transcript-level correctness, agreement, and uncertainty using entropy-based summary statistics of the predicted coding probabilities. Specifically, we (1) computed predictive entropy, which quantifies uncertainty in the ensemble mean prediction; (2) produced an inter-tool disagreement score, which captures variability

across tools; and (3) stratified transcripts into different uncertainty groups for downstream comparisons.

1. Uncertainty

Transcript classification uncertainty was quantified using Shannon entropy of the mean predicted coding probability across tools. For each transcript, coding probabilities were averaged across tools to compute the mean posterior probability for the coding class (p_{coding}).

Shannon entropy was calculated as:

$$H_{pred} = -p_{coding} \times \log_2(p_{coding}) + 1 - p_{coding} \times \log_2(1 - p_{coding})$$

This metric produces values from 0 to 1, where $H = 0$ corresponds to extreme mean probabilities (near 0 or 1) indicating complete certainty (all tools agree with high confidence) and $H = 1$ indicates maximum uncertainty of the ensemble mean $p_{coding} \approx 0.5$ (average model prediction is uncertain). Classification entropy distributions per model and ensemble predictive entropy are shown in Figure S3B-C.

2. Disagreement

Inter-tool disagreement was summarized using an entropy-gap score. We first computed the expected entropy of each transcript as the average entropy of individual model predictions:

$$H_{exp} = \frac{1}{M} \sum_{m=1}^M H_m$$

where H_m represents Shannon entropy of the coding probability of a transcript given by a model m . Model disagreement was then defined as the difference between predictive and expected entropy.

$$I_{dis} = H_{pred} - H_{exp}$$

Low I_{dis} values indicate high ensemble consensus, while high I_{dis} values indicate substantial model disagreement.

3. Uncertainty-based transcript stratification

We categorized transcripts into uncertainty groups using a percentile-based approach. We split all predicted transcripts (111,652) in coding and non-coding and computed for each group the 10th and 90th percentiles of our uncertainty metric (H_{pred}) and the 90th percentile of the disagreement metric (I_{dis}). Transcripts with $H_{pred} < 10$ th percentile were assigned to the “low entropy” group, while sequences with H_{pred} and $I_{dis} >$ their respective 90th percentiles were assigned to the “high entropy” group. This grouping strategy enabled us to identify transcripts with distinct prediction difficulty profiles and assess systematic differences in biological features across uncertainty groups. These differences were evaluated using univariate statistical tests described below.

Feature-based analyses

From the eight models benchmarked here, five are machine learning tools based on engineered transcript features. We generated the corresponding feature tables and assembled a unified feature matrix (ML set). Additionally, we included repeat-derived features (Rep set) and Non-B DNA motif features (NBD set) features computed as described above (see Supplemental Material for additional details).

1. Feature processing

We applied the following preprocessing steps (1) removal of constant features (i.e., those with the same value across all transcripts); (2) separation of categorical and continuous variables; (3) estimation of continuous feature redundancy by calculation of a Spearman correlation matrix; (4) clustering of correlated features; (5) selection of a random representative feature per cluster; and (6) combination of representative continuous features with categorical features encoded as binary indicator variables via one-hot encoding). The number of features after each processing step is detailed in Table S17.

Clustering and representative feature selection

Specifically, continuous features were grouped using hierarchical clustering (Ward linkage) on a Spearman correlation-derived distance matrix. The number of clusters was chosen by maximizing the silhouette score over a grid of cut thresholds (Fig. S3), and one feature was randomly selected per cluster as representative

Dimensionality reduction for feature space exploration

To visualize the feature space of transcripts classified by the benchmarked models, t-SNE embeddings were computed using the scikit-learn Python library with 1000 iterations and a perplexity of 30. The input feature space comprised all 128 numerical, non-constant engineered features from the 5 feature-based models benchmarked in this study: CPAT (N=4 features) FEELnc (N=8), LncFinder (N=19), LncDC (N=28) and PlncPRO (N=69).

Univariate statistical testing of transcript features

Differences in continuous features between transcript groups (coding class or entropy groups) were assessed using Mann-Whitney U test, with effect size quantification by the Vargha-Delaney A (VDA) statistic. For visualization, we report a centered effect size (VDA – 0.5), where 0 indicates no effect and ± 0.5 indicates complete separation between groups. Differences in categorical feature prevalence were assessed using the chi-squared test of independence, with effect size quantified by Cramér's V statistic and the directional magnitude expressed in odds ratios (OR). To account for multiple comparisons, p-values were corrected using the Benjamini-Hochberg method. Results with adjusted p-value lower than 0.01 were considered statistically significant.

2. Global feature importance

A Random Forest model was then trained for each of the 5-fold train sets on the selected features, comprising 220 representative continuous features and 40 categorical features. SHAP values were then computed for the held-out set of each fold to obtain transcript-level

feature explanations (waterfall plots). Importantly, SHAP values are expressed directly in probability space, such that for each transcript, the sum of all feature SHAP values and the baseline (mean predicted probability assigned by the RF classifier) equals the coding probability assigned by the RF classifier. Finally, SHAP values from all held-out sets were combined to obtain global feature importance summaries (mean absolute SHAP value across all transcripts). The classification performance metrics of the trained RF classifiers are detailed in Table S18.

Four transcripts were chosen arbitrarily to showcase the categories of protein-coding and lncRNA with either low or high entropy.

Development and computation tools

Dataset creation, model training and data analyses were implemented as a Snakemake pipeline (v9.13.2) [58], using Python and R. Separate conda environments were created for each model to ensure reproducibility and portability. A dedicated analysis environment was used for data processing and visualisation, including Biopython and gffutils (biological data handling); numpy and pandas (data processing); scikit-learn, scipy and shap (statistics and dimensionality reduction), and matplotlib/seaborn/plotly/upsetplot (visualization). Environment configuration is available through environment files in the GitHub repository (https://github.com/cbib/uncertainty_aware_Inclassifier).

All computations were performed on the University of Bordeaux high-performance computing (HPC) infrastructure operated by the Center of Bioinformatics of Bordeaux (CBiB). GPU-accelerated tasks (RNAsamba and lncRNA-BERT training/inference) were run on NVIDIA H100 GPUs.

Declarations

Ethics approval and consent to participate

“Not applicable”

Consent for publication

“Not applicable”

Availability of data and materials

Data produced during the analyses is deposited in Zenodo

<https://doi.org/10.5281/zenodo.19551649>

Additional materials are provided together with the manuscript:

Additional files	
Supplemental_information.pdf	Additional information, figures and methods complementing the main manuscript.
Supplemental_tables.xlsx	Tabular file with supplemental tables mentioned in the manuscript (Tables S1 to S

Software

Additional files	
Project name	Uncertainty-aware benchmarking reveals ambiguous transcripts in mRNA–lncRNA classification
Project home page	https://github.com/cbib/uncertainty_aware_lncclassifier/
Archived version	https://doi.org/10.5281/zenodo.23043213
Operating systems	Linux
Programming languages	Snakemake, Python and R
Other requirements	Conda; Snakemake >= 9; Git NVIDIA GPU
License	MIT License

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by funds from the INCA ITMO Aviesan MCMP2022 [N°251534] grant number 22CE052-00 PREDICAPA to MN and DC and in part by the NIH grant R35GM151945 to KDM.

Authors' contributions

DGR, DC and MN designed the benchmarking protocol. DGR, MG, RB and SKM prepared the benchmarking datasets. DGR implemented the reproducible Snakemake pipeline and performed model retraining and evaluation. DGR and MG analyzed the data. DGR, MG, MN and DC interpreted the results. DGR drafted the manuscript; DGR, SKM, KDM, MN and DC reviewed and edited the manuscript. MN, DC and KDM supervised the study and acquired funding. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge support from the CBiB (Centre de Bioinformatique de Bordeaux) of the University of Bordeaux for the computing resources. We are thankful to Jacob Sieg for constructive discussion regarding RNA folding. We are also thankful to all the members of the CB&B team for their scientific discussions.

References

1. Kaur G, Perteghella T, Carbonell-Sala S, Gonzalez-Martinez J, Hunt T, Mądry T, et al. GENCODE: massively expanding the lncRNA catalog through capture long-read RNA sequencing [Internet]. bioRxiv; 2024 [cited 2025 July 17]. p. 2024.10.29.620654. <https://doi.org/10.1101/2024.10.29.620654>
2. Perteghella T, Kaur G, Carbonell-Sala S, Gonzalez-Martinez J, Hunt T, Mądry T, et al. The GENCODE CLS project: massively expanding the lncRNA catalog through capture long-read RNA sequencing [Internet]. bioRxiv; 2026 [cited 2026 Mar 12]. p. 2024.10.29.620654. <https://doi.org/10.1101/2024.10.29.620654>
3. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023;24:430–47. <https://doi.org/10.1038/s41580-022-00566-8>
4. Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021;22:96–118. <https://doi.org/10.1038/s41580-020-00315-9>

- 944 5. Froberg JE, Yang L, Lee JT. Guided by RNAs: X-inactivation as a model for lncRNA
945 function. *J Mol Biol.* 2013;425:3698–706. <https://doi.org/10.1016/j.jmb.2013.06.031>
- 946 6. Flynn RA, Chang HY. Long Noncoding RNAs in Cell-Fate Programming and
947 Reprogramming. *Cell Stem Cell.* Elsevier; 2014;14:752–61.
948 <https://doi.org/10.1016/j.stem.2014.05.014>
- 949 7. Fickett JW. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*
950 1982;10:5303–18. <https://doi.org/10.1093/nar/10.17.5303>
- 951 8. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish
952 protein coding and non-coding regions. *Bioinformatics.* 2011;27:i275–82.
953 <https://doi.org/10.1093/bioinformatics/btr209>
- 954 9. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, et al. CPC: assess the protein-coding
955 potential of transcripts using sequence features and support vector machine. *Nucleic Acids*
956 *Res.* 2007;35:W345. <https://doi.org/10.1093/nar/gkm391>
- 957 10. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential
958 Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*
959 2013;41:e74. <https://doi.org/10.1093/nar/gkt006>
- 960 11. Amunét T, Wang N, Khan S, Elo LL. Deep learning tools are top performers in long non-
961 coding RNA prediction. *Brief Funct Genomics.* 2022;21:230–41.
962 <https://doi.org/10.1093/bfpg/elab045>
- 963 12. Romeijn L, Cats D, Wolstencroft K, Mei H. LncRNA-BERT: an RNA language model for
964 classifying coding and long non-coding RNA. *Connect Sci.* Taylor & Francis;
965 2026;38:2624131. <https://doi.org/10.1080/09540091.2026.2624131>
- 966 13. Zheng H, Talukder A, Li X, Hu H. A systematic evaluation of the computational tools for
967 lncRNA identification. *Brief Bioinform.* 2021;22:bbab285. <https://doi.org/10.1093/bib/bbab285>
- 968 14. Singh D, Roy J. A large-scale benchmark study of tools for the classification of protein-
969 coding and non-coding RNAs. *Nucleic Acids Res.* 2022;50:12094–111.
970 <https://doi.org/10.1093/nar/gkac1092>
- 971 15. Smestad JA, Maher LJ. Relationships between putative G-quadruplex-forming sequences,
972 RecQ helicases, and transcription. *BMC Med Genet.* 2015;16:91.
973 <https://doi.org/10.1186/s12881-015-0236-4>
- 974 16. Lee C-Y, McNerney C, Ma K, Zhao W, Wang A, Myong S. R-loop induced G-quadruplex
975 in non-template promotes transcription by successive R-loop formation. *Nat Commun.* Nature
976 Publishing Group; 2020;11:3392. <https://doi.org/10.1038/s41467-020-17176-7>
- 977 17. Broxson C, Beckett J, Tornaletti S. Transcription Arrest by a G Quadruplex Forming-
978 Trinucleotide Repeat Sequence from the Human c-myc Gene. *Biochemistry.* American
979 Chemical Society; 2011;50:4162–72. <https://doi.org/10.1021/bi2002136>
- 980 18. Spiegel J, Cuesta SM, Adhikari S, Hänsel-Hertsch R, Tannahill D, Balasubramanian S. G-
981 quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*
982 2021;22:117. <https://doi.org/10.1186/s13059-021-02324-z>

983 19. Liu F, Somarowthu S, Pyle AM. Visualizing the secondary and tertiary architectural
984 domains of lncRNA RepA. *Nat Chem Biol.* 2017;13:282–9.
985 <https://doi.org/10.1038/nchembio.2272>

986 20. Somarowthu S, Legiewicz M, Chillón I, Marcia M, Liu F, Pyle AM. HOTAIR forms an
987 intricate and modular secondary structure. *Mol Cell.* 2015;58:353–61.
988 <https://doi.org/10.1016/j.molcel.2015.03.006>

989 21. Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al.
990 GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.*
991 2025;53:D966–75. <https://doi.org/10.1093/nar/gkae1078>

992 22. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
993 sequencing data. *Bioinformatics.* 2012;28:3150–2.
994 <https://doi.org/10.1093/bioinformatics/bts565>

995 23. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate
996 coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*
997 2017;45:W12–6. <https://doi.org/10.1093/nar/gkx428>

998 24. Li M, Liang C. LncDC: a machine learning-based tool for long non-coding RNA detection
999 from RNA-Seq data. *Sci Rep. Nature Publishing Group;* 2022;12:19083.
1000 <https://doi.org/10.1038/s41598-022-22082-7>

1001 25. Mohanty SK, Chiaromonte F, Makova KD. Evolutionary dynamics of predicted G-
1002 quadruplexes in human and other great apes. *Genome Biol.* 2025;26:161.
1003 <https://doi.org/10.1186/s13059-025-03635-1>

1004 26. Andrews RJ, Rouse WB, O’Leary CA, Booher NJ, Moss WN. ScanFold 2.0: a rapid
1005 approach for identifying potential structured RNA targets in genomes and transcriptomes.
1006 *PeerJ.* 2022;10:e14361. <https://doi.org/10.7717/peerj.14361>

1007 27. Turner M, Danino YM, Barshai M, Yacovzada NS, Cohen Y, Olender T, et al. rG4detector,
1008 a novel RNA G-quadruplex predictor, uncovers their impact on stress granule formation.
1009 *Nucleic Acids Res.* 2022;50:11426–41. <https://doi.org/10.1093/nar/gkac950>

1010 28. Varshney D, Cuesta SM, Herdy B, Abdullah UB, Tannahill D, Balasubramanian S. RNA
1011 G-quadruplex structures control ribosomal protein production. *Sci Rep. Nature Publishing*
1012 *Group;* 2021;11:22735. <https://doi.org/10.1038/s41598-021-01847-6>

1013 29. Sahayasheela VJ, Sugiyama H. RNA G-quadruplex in functional regulation of noncoding
1014 RNA: Challenges and emerging opportunities. *Cell Chem Biol. Elsevier;* 2024;31:53–70.
1015 <https://doi.org/10.1016/j.chembiol.2023.08.010>

1016 30. Dumas L, Herviou P, Dassi E, Cammas A, Millevoi S. G-Quadruplexes in RNA Biology:
1017 Recent Advances and Future Directions. *Trends Biochem Sci. Elsevier;* 2021;46:270–83.
1018 <https://doi.org/10.1016/j.tibs.2020.11.001>

1019 31. Johnson R, Guigó R. The RIDL hypothesis: transposable elements as functional domains
1020 of long noncoding RNAs. *RNA.* 2014;20:959–76. <https://doi.org/10.1261/rna.044560.114>

1021 32. Yang J-R, Zhang J. Human long noncoding RNAs are substantially less folded than
1022 messenger RNAs. *Mol Biol Evol.* 2015;32:970–7. <https://doi.org/10.1093/molbev/msu402>

1023 33. Ghafouri-Fard S, Abak A, Baniahmad A, Hussen BM, Taheri M, Jamali E, et al. Interaction
1024 between non-coding RNAs, mRNAs and G-quadruplexes. *Cancer Cell Int.* 2022;22:171.
1025 <https://doi.org/10.1186/s12935-022-02601-2>

1026 34. Hill ST, Kuintzle R, Teegarden A, Merrill E, Danaee P, Hendrix DA. A deep recurrent neural
1027 network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic
1028 Acids Res.* 2018;46:8105–13. <https://doi.org/10.1093/nar/gky567>

1029 35. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long
1030 non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.*
1031 2017;45:e57. <https://doi.org/10.1093/nar/gkw1306>

1032 36. Camargo AP, Sourkov V, Pereira GAG, Carazzolle MF. RNAsamba: neural network-based
1033 assessment of the protein-coding potential of RNA sequences. *NAR Genomics Bioinforma.*
1034 2020;2:lqz024. <https://doi.org/10.1093/nargab/lqz024>

1035 37. Han S, Liang Y, Ma Q, Xu Y, Zhang Y, Du W, et al. LncFinder: an integrated platform for
1036 long non-coding RNA identification utilizing sequence intrinsic composition, structural
1037 information and physicochemical property. *Brief Bioinform.* 2019;20:2009–27.
1038 <https://doi.org/10.1093/bib/bby065>

1039 38. Singh U, Khemka N, Rajkumar MS, Garg R, Jain M. PLncPRO for prediction of long non-
1040 coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive
1041 lncRNAs in rice and chickpea. *Nucleic Acids Res.* 2017;45:e183.
1042 <https://doi.org/10.1093/nar/gkx866>

1043 39. Palma A, Buonaiuto G, Ballarino M, Laneve P. Genome biology of long non-coding RNAs
1044 in humans: A virtual karyotype. *Comput Struct Biotechnol J.* 2025;27:575–84.
1045 <https://doi.org/10.1016/j.csbj.2025.01.026>

1046 40. Fiannaca A, La Rosa M, La Paglia L, Rizzo R, Urso A. nRC: non-coding RNA Classifier
1047 based on structural features. *BioData Min.* 2017;10:27. <https://doi.org/10.1186/s13040-017-0148-2>

1049 41. Bai H, Wang J, Jiang X, Guo Z, Yang W, Yang Z, et al. TetraRNA, a tetra-class machine
1050 learning model for deciphering the coding potential derivation of RNA world. *Comput Struct
1051 Biotechnol J. American Association for the Advancement of Science;* 2025;27:1305–17.
1052 <https://doi.org/10.1016/j.csbj.2025.03.039>

1053 42. Liu J, Gough J, Rost B. Distinguishing Protein-Coding from Non-Coding RNAs through
1054 Support Vector Machines. *PLOS Genet. Public Library of Science;* 2006;2:e29.
1055 <https://doi.org/10.1371/journal.pgen.0020029>

1056 43. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable
1057 Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate
1058 Long Noncoding RNAs. *PLoS Genet.* 2013;9:e1003470.
1059 <https://doi.org/10.1371/journal.pgen.1003470>

1060 44. Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R. Ancient exapted
1061 transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome
1062 Res.* 2019;29:208–22. <https://doi.org/10.1101/gr.229922.117>

1063 45. Nguyen TM, Kabotyanski EB, Reineke LC, Shao J, Xiong F, Lee J-H, et al. The SINEB1
1064 element in the long non-coding RNA Malat1 is necessary for TDP-43 proteostasis. *Nucleic
1065 Acids Res.* 2020;48:2621–42. <https://doi.org/10.1093/nar/gkz1176>

1066 46. Stein M, Hile SE, Weissensteiner MH, Lee M, Zhang S, Kejnovský E, et al. Variation in G-
1067 quadruplex sequence and topology differentially impacts human DNA polymerase fidelity.
1068 DNA Repair. 2022;119:103402. <https://doi.org/10.1016/j.dnarep.2022.103402>

1069 47. Makova KD, Weissensteiner MH. Noncanonical DNA structures are drivers of genome
1070 evolution. Trends Genet. 2023;39:109–24. <https://doi.org/10.1016/j.tig.2022.11.005>

1071 48. Mirihana Arachchilage G, Hetti Arachchilage M, Venkataraman A, Piontkivska H, Basu S.
1072 Stable G-quadruplex enabling sequences are selected against by the context-dependent
1073 codon bias. Gene. 2019;696:149–61. <https://doi.org/10.1016/j.gene.2019.02.006>

1074 49. Sun L, Liu H, Zhang L, Meng J. IncRScan-SVM: A Tool for Predicting Long Non-Coding
1075 RNAs Using Support Vector Machine. PLOS ONE. Public Library of Science;
1076 2015;10:e0139654. <https://doi.org/10.1371/journal.pone.0139654>

1077 50. Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. Developmental dynamics of
1078 lncRNAs across mammalian organs and species. Nature. 2019;571:510–4.
1079 <https://doi.org/10.1038/s41586-019-1341-x>

1080 51. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al.
1081 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome
1082 Res. 2005;15:1034–50. <https://doi.org/10.1101/gr.3715005>

1083 52. Georgakopoulos-Soares I, Parada GE, Wong HY, Medhi R, Furlan G, Munita R, et al.
1084 Alternative splicing modulation by G-quadruplexes. Nat Commun. 2022;13:2404.
1085 <https://doi.org/10.1038/s41467-022-30071-7>

1086 53. Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the
1087 regulation of mammalian transcription. Nat Rev Mol Cell Biol. 2022;23:481–97.
1088 <https://doi.org/10.1038/s41580-022-00457-y>

1089 54. dos Santos RF, Baptista T, Marques GS, Homem CCF. FLYNC: a machine-learning-
1090 driven framework for discovering long noncoding RNAs in Drosophila melanogaster. NAR
1091 Genomics Bioinforma. 2026;8:lqaf216. <https://doi.org/10.1093/nargab/lqaf216>

1092 55. [repeatmasker.org/webrepeatmaskerhelp.html](https://www.repeatmasker.org/webrepeatmaskerhelp.html) [Internet]. [cited 2026 Mar 13].
1093 <https://www.repeatmasker.org/webrepeatmaskerhelp.html>. Accessed 13 Mar 2026

1094 56. Hon J, Martínek T, Zendulka J, Lexa M. pqsfinder: an exhaustive and imperfection-tolerant
1095 search tool for potential quadruplex-forming sequences in R. Bioinformatics. 2017;33:3373–
1096 9. <https://doi.org/10.1093/bioinformatics/btx413>

1097 57. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter.
1098 Nucleic Acids Res. 2016;44:1746–59. <https://doi.org/10.1093/nar/gkw006>

1099 58. Mölder F, Jablonski KP, Letcher B, Hall MB, Dyken PC van, Tomkins-Tinch CH, et al.
1100 Sustainable data analysis with Snakemake [Internet]. F1000Research; 2025 [cited 2026 Mar
1101 13]. <https://doi.org/10.12688/f1000research.29032.3>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GarciaRuanoSupplementalInformation.pdf](#)
- [GarciaRuanoSupplementalTables.xlsx](#)