https://doi.org/10.1093/nargab/lqac071

Computational identification of signals predictive for nuclear RNA exosome degradation pathway targeting

Mengjun Wu[1,2], Manfred Schmid[3], Torben Heick Jensen[3] and Albin Sandelin[1,*]

[1]The Bioinformatics Centre, Department of Biology and Biotech and Research Innovation Centre, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark, [2]SciLifeLab, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, 171 65 Solna, Sweden and [3]Department of Molecular Biology and Genetics, Aarhus University, Universitetsbyen 81, Aarhus, DK-8000, Denmark

Received April 10, 2022; Revised August 05, 2022; Editorial Decision August 22, 2022; Accepted September 01, 2022

ABSTRACT

The RNA exosome degrades transcripts in the nucleoplasm of mammalian cells. Its substrate specificity is mediated by two adaptors: the ‘nuclear exosome targeting (NEXT)’ complex and the ‘poly(A) exosome targeting (PAXT)’ connection. Previous studies have revealed some DNA/RNA elements that differ between the two pathways, but how informative these features are for distinguishing pathway targeting, or whether additional genomic features that are informative for such classifications exist, is unknown. Here, we leverage the wealth of available genomic data and develop machine learning models that predict exosome targets and subsequently rank the features the models use by their predictive power. As expected, features around transcript end sites were most predictive; specifically, the lack of canonical 3′ end processing was highly predictive of NEXT targets. Other associated features, such as promoter-proximal G/C content and 5′ splice sites, were informative, but only for distinguishing NEXT and not PAXT targets. Finally, we discovered predictive features not previously associated with exosome targeting, in particular RNA helicase DDX3X binding sites. Overall, our results demonstrate that nucleoplasmic exosome targeting is to a large degree predictable, and our approach can assess the predictive power of previously known and new features in an unbiased way.

INTRODUCTION

The mammalian genome produces a large repertoire of RNAs (1). Excessive RNAs pose a potential threat to cellular homeostasis by interfering with productive cellular processes, such as transcription and genome maintenance (2). To control nuclear RNA levels, transcripts are targeted by the RNA exosome, a highly conserved 3′–5′ exo- and endonucleolytic multisubunit complex (3–8). Nuclear RNA exosome substrates include prematurely terminated RNAs from within protein-coding loci as well as many long noncoding RNAs (lncRNAs), such as promoter upstream transcripts/upstream antisense RNAs and enhancer RNAs (eRNAs) (7,9–13).

While the RNA exosome itself is a highly efficient and processive degradation machine, it lacks substrate selectivity. In the nucleoplasm of mammalian cells, this is achieved through its ability to connect to one of two main adaptors: the ‘nuclear exosome targeting (NEXT)’ complex and the ‘poly(A) exosome targeting (PAXT)’ connection (14–16). NEXT and PAXT both connect to the RNA exosome through the common RNA helicase MTR4/Skiv2l2 (17,18). Besides MTR4, the NEXT complex consists of the Zn-knuckle protein ZCCHC8 and the RNA recognition motif-containing RBM7 protein (14), whereas the core moiety of PAXT consists of MTR4 heterodimerizing with the Zn-finger protein ZFC3H1, and making additional more transient contacts with the ZC3H3 and RBM26/RBM27 proteins as well as an RNA-dependent link with the nuclear poly(A) binding protein (PABPN1) (15,16).

NEXT primarily targets short, unspliced and nonadenylated RNAs (14,19,20), whereas PAXT mediates the exosomal degradation of polyadenylated RNAs that are often longer and include spliced RNAs with extended nuclear residence times (15,21,22). In line with this, high-resolution analyses of the 3′ ends of NEXT and PAXT targets revealed major differences in their 3′ end processing pathways (20). Besides, several sequence features have previously been associated with nuclear exosome targets: such transcripts are generally transcribed from loci with high densities of transcription start site (TSS)-proximal poly(A) sites (PASs) and low densities of 5′ splice sites (5′ SSs) (21,23–25); it is not entirely clear whether these features contribute equally to

[*]To whom correspondence should be addressed. Tel: +45 35321285; Email: albin@binf.ku.dk
Present address: Manfred Schmid, QIAGEN Aarhus A/S, Silkeborgvej 2, 8000 Aarhus, Denmark.

the substrate specificity, and whether they carry enough information to guide a specific transcript to one or both pathways.

Here, we develop a machine learning framework to assess (i) whether NEXT and PAXT pathway targeting is predictable using transcript features drawn from existing large and diverse genomic datasets, and (ii) if so, what the most informative features for respective prediction are.

To this end, we first identified NEXT and PAXT targets by performing differential expression analysis of RNA-seq data from NEXT/PAXT adaptor knockdowns, and then systematically collected relevant molecular features to serve as prediction features in random forest-based machine learning models that predict NEXT/PAXT targeting pathways of transcripts. Finally, we ranked the features by quantitatively assessing their contributions to the prediction.

Our results demonstrate that the most informative features for prediction are drawn from molecular signals previously associated with exosome sensitivity. We found that RNA processing-related features are generally more informative for NEXT/PAXT targeting than other features such as chromatin modifications. Specifically, lack of RNA 3′ end processing was most predictive for distinguishing NEXT/PAXT pathways; as expected, lack of canonical 3′ end processing by cleavage and polyadenylation (CPA) machinery was found to be most characteristic of NEXT targets. Moreover, other reported features, such as G/C content around TSSs and TSS-proximal 5′ SSs, were also highly predictive for NEXT targets but were not able to distinguish PAXT from non-NEXT/PAXT targets. Finally, we found that interaction data for several RNA binding proteins (RBPs) that were not previously associated with exosome targeting were informative for distinguishing NEXT and PAXT pathways, in particular the binding of RNA helicase DDX3X.

MATERIALS AND METHODS

Public data acquisition and processing

HeLa S3 NET-seq was described in (26) and obtained from Gene Expression Omnibus (GEO): GSE61332. The ChIP-seq datasets used in this study were described in the ENCODE project (1); HeLa S3 H3K4me1, H3k4me2, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27ac, H3K36me3, H4K20me1, H2A.Z and PolIIb were obtained from GEO: GSE29611; and DNase-seq was obtained from ENCODE: ENCSR959ZXU. The hg19 genome coordinates were converted to hg38 using the UCSC liftOver tool. For data with replicates, replicates were pooled and signals were averaged over replicates for subsequent analysis. CLIP binding sites for RBPs were obtained from POSTAR2 (http://lulab.life.tsinghua.edu.cn/postar/index.php) (27). Binding site datasets called by different peak calling methods from the same CLIP data were treated as independent datasets when extracting the features.

Transcriptome annotation

De novo HeLa transcriptome annotation from (28) was used in this study.

Classification of exosome targets by pathways

For classification, we used RNA-seq counts derived from total RNA of siEGFP-, siRBM7-, siZCCHC8- and siZFC3H1-treated HeLa cells first described in (15) (GSE84172) and similar data from siZC3H3-treated cells first described in (16) (GSE131255). Counts in exonic regions of major transcript isoforms in our in-house HeLa transcriptome annotations were collected using the featureCounts tool from Subread package (v2.0.0) (29), using parameters [-p -C -s 2 -t exon]. These counts were then subjected to differential expression analysis using DESeq2 (v1.22.2) (30) using default settings except that batch information [see (16) for details] was included in the design. Transcripts significantly upregulated [$\log_2$ fold change ($\log_2$FC) > 0 and $P_{\mathrm{adj}} < 0.1$] in siRBM7, siZCCHC8, siZFC3H1 and siZC3H3 were selected and used to define the set of NEXT targets (significantly upregulated in siRBM7 and siZCCHC8, but not significantly upregulated in siZC3H3 and siZFC3H1), PAXT targets (significantly upregulated in siZC3H3 and siZFC3H1, but not significantly upregulated in siRBM7 and siZCCHC8) and non-NEXT/PAXT targets ($\log_2\mathrm{FC} > -1$ and $\log_2\mathrm{FC} \leq 0$ in any of the four knockdowns).

Sequence analysis

Sequences were extracted from the reference genome (hg38) using getfasta from bedtools. The R package Biostrings (version 2.54.0) was used for the following sequence analysis. G/C content is defined as the percentage of DNA that is G or C, and was computed using the letterFrequencyInSlidingView function over a 10 bp window. Raw position frequency matrix of the TATA box, initiator (INR) element, 5′ SS and pA site motifs were obtained from (31), and converted to position weight matrix (PWM) using R function PWM. RBP motif PWM was obtained from the CISBP-RNA database (32). The countPWM function was used to scan for motif occurrences; a minimum score of 90% was used for counting a motif hit.

Extraction of signals from sequencing data

Signals from ChIP-seq and strand-specific signals from NET-seq over a given window were extracted using the R package rtracklayer (version 1.46.0).

Low information feature filtering

Bit entropy was calculated for assessing the empirical distribution of each feature by taking the values across three exosome target categories using the entropy function from the R package entropy (version 1.2.1). Features with entropy smaller than 0.8 bits were removed.

GC spread estimation

The GC spread metric was designed as a proxy to quantify the boundary of the G/C enriched region immediately downstream of TSS. It was calculated as the width of regions with 75% of the total G/C content in a defined region downstream of TSS. G/C content for each nucleotide is first computed as in the 'Sequence analysis' section and normalized as max(0, G/C − 0.5); the normalized G/C content is then multiplied by a dynamic scaling factor calculated using a Gaussian function as follows: $s = 2\^(-(x/a)\^2) \in (0, 1)$, where x is the width of the region immediately downstream of TSS (1 kb in this study) and a is set to half of the region width so that the scaling factor will be decreased to 0.5 at the middle point of the region. By using the scaling factor, we put more weight on nucleotides close to TSS and less weight on those more distant, thus minimizing the influence of the random G/C content fluctuation more distant from TSS. In addition, to avoid high GC spread from TSSs with generally low G/C content, we reduce the GC spread if the average of normalized and unscaled G/C content (GC'avg) in the calculated GC spread is less than 0.1, by a factor of GC'avg/0.1.

PAS strength estimation

PAS strengths were estimated as described in (20) using the deep neural network model APARENT (Python package V0.1) (33) and depicted as log odds of the prediction score.

Expression match sampling

Transcript expression levels were estimated as the nascent RNA levels measured by NET-seq and calculated as the sum of NET-seq signals around −100 to +500 bp window around TSS. To select samples with matched expression levels from two or more exosome target categories, we first obtained the largest potential expression range by taking the minimum and maximum expression levels of all exosome target categories. We then divided the expression range into 2000 bins, and for each bin we randomly selected a number of samples from each exosome target category corresponding to the number of samples in the smallest exosome target category. As shown in Supplementary Figure S1B, the sampling process yielded similar distributions of expression levels for the three exosome target categories. By performing the expression match sampling, we obtained 2277 samples in NEXT versus non-NEXT/PAXT targets, 1090 samples in PAXT versus non-NEXT/PAXT targets, 1085 samples in NEXT versus PAXT targets and 1080 samples in multiclass classification.

Random forest model and prediction performance evaluation

The random forest model was built using the R package caret (version 6.0.86) by choosing the 'parRF' method. We first performed expression match sampling to obtain balanced and expression matched data for different exosome target categories. We then used the resulting balanced dataset for classification, where 70% of the dataset was used for training and the remaining part for testing. When training the model, we chose the default 500 trees as we consider it an ample amount given the number of features we have and used 5-fold cross-validation with five repeats in order to tune the parameter 'number of randomly selected features at each tree split'; the optimal model with the largest average accuracy value was selected to evaluate the performance on the test data. To measure the prediction performance, we used 'area under the receiver operator characteristic' (AUC) and F_1 score, i.e. the harmonic mean of precision and recall.

To ensure the downsampling process accurately reflects the larger exosome target category and to minimize the biases of random splitting training and testing data, we repeated the above processes 10 times, and calculated the mean and standard deviation of AUC and F_1 score.

Feature importance score calculation

The feature importance score is calculated using the varImp function from the caret package. The importance score of a feature was measured by calculating the average values of the difference in prediction error rate with and without permuting the values of the feature on the out-of-bag portion of data over all trees, normalized by the standard deviation of the difference yielding a z-score. We averaged the importance score of each feature from the 10 independent random forest models by repeating random splitting of training and test data 10 times. In addition, we compared the significant features identified by our iterative selection with the significant features identified by the Boruta method (34), which overall supports our conclusion.

Data visualization

We used R and the ggplot2 R package (35) unless otherwise noted for visualizations.

RESULTS

NEXT and PAXT target classification, feature selection and machine learning framework

To identify determinants for PAXT- and NEXT-mediated RNA decay pathways, we used the analysis outline summarized in Figure 1. We first used the *de novo* HeLa transcriptome annotation from (28) as a transcription unit (TU) framework. To this set of TUs, we mapped previously published RNA-seq libraries (15,16) derived from HeLa cells depleted for RBM7, ZCCHC8, ZFC3H1, ZC3H3 or EGFP (control) and used differential expression analysis to define TUs that were targeted by PAXT or NEXT or were unaffected (Figure 2A; see the 'Materials and Methods' section). Briefly, RNAs from 2575 TUs that compared to their controls were upregulated [DESeq2 $\log_2$FC > 0 and false discovery rate (FDR) < 0.1] in siRBM7 and siZCCHC8 samples and unaffected in siZC3H3 and siZFC3H1 samples were defined as NEXT targets, RNAs from 1116 TUs that were upregulated in siZC3H3 and siZFC3H1 samples but unaffected in siRBM7 and siZCCHC8 samples were defined as PAXT targets and RNAs from 3664 TUs were defined as non-NEXT/PAXT targets ($\log_2$FC > −1 and $\log_2$FC ≤ 0 in siRBM7, siZCCHC8, siZC3H3 and siZFC3H1 versus control) (Supplementary Table S1).

Consistent with previous observations (15,21), non-NEXT/PAXT targets mostly consisted of full-length multi-exonic protein-coding RNAs, whereas NEXT targets were generally short and mono-exonic, while PAXT targets mainly consisted of mono-exonic protein-coding RNAs and lncRNAs with lengths between those of non-NEXT/PAXT

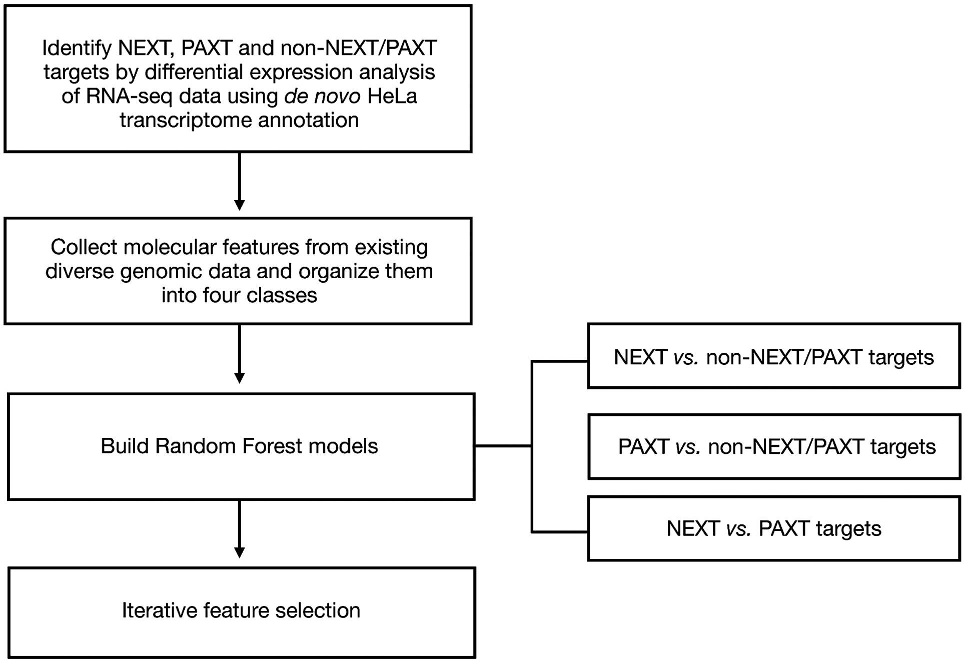
Identify NEXT, PAXT and non-NEXT/PAXT targets by differential expression analysis of RNA-seq data using *de novo* HeLa transcriptome annotation
Collect molecular features from existing diverse genomic data and organize them into four classes
Build Random Forest models
NEXT *vs.* non-NEXT/PAXT targets
PAXT *vs.* non-NEXT/PAXT targets
NEXT *vs.* PAXT targets
Iterative feature selection

Figure 1. Overview of the computational framework for identifying molecular features for NEXT/PAXT targeting.

and NEXT targets (Figure 2B). In order to find biological information with the potential to classify pathways, we considered four classes of features based on their location with respect to TUs, summarized below, and in Figure 2C, and described in greater detail in Table 1.

Class 1 contained chromatin and sequence features around TSSs. This was motivated by our prior results of chromatin and sequence differences between TSSs producing exosome-sensitive versus exosome-insensitive transcripts (13,21,24). For example, when compared to TSSs of exosome-insensitive mRNAs, TSSs of exosome substrates, like eRNAs, harbored higher levels of H3K4me1 and lower levels of H3K4me3 (13). In terms of sequence content, high-density G/C regions were often found to extend further downstream of TSSs of exosome-insensitive transcripts compared to those of exosome-sensitive ones (21,24). Besides the mentioned features, we also included less studied ones for exploratory reasons, which might potentially lead to new hypotheses. Specifically, we included HeLa cell chromatin data from ENCODE (ChIP-seq for H3K4me1, H3k4me2, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27ac, H3K36me3, H4K20me1 and H2A.Z; DNase-seq), and DNA sequence features, including the presence of TATA and INR patterns, G/C content and GC spread, which is a designed proxy of the width of regions with high G/C content (above 50%) immediately downstream of TSSs (Supplementary Figure S1A; see the 'Materials and Methods' section). Finally, as exosome targets were previously found to generally have lower levels of transcription than non-NEXT/PAXT targets (13), we also included levels of transcription by including NET-seq and Pol II ChIP-seq data.

Classes 2 and 3 consisted of features related to RNA processing proximal to TSSs and transcript end sites (TESs), respectively, including RNA sequence features and RBP binding sites. Previous studies suggested a relevance for TSS- and TES-proximal RNA processing to exosome targeting (20,24,25). In TSS-proximal regions, low density of U1 snRNP recognition sites (5′ SSs) and high density of PASs were previously observed in exosome-sensitive transcripts (24,25). With respect to TES-proximal RNA processing, differences in 3′ end processing mechanisms were observed between PAXT and NEXT targets, i.e. whether these RNAs are processed by the CPA machinery or not (20). Therefore, in Class 2, we included 5′ SSs and PASs, and in Class 3 cleavage-related PAS features (PAS strength and cleavage PAS motif match) as well as 5′ SS and PAS motif match – the same as Class 2 – as a control. Additionally, for exploratory reasons and given the importance of RBPs in RNA processing, we included in both classes RBP motif match and CLIP-based RBP binding data from the POSTAR2 database (27), which includes CLIP-seq from both the ENCODE consortium and other recent publications. We collected 401 sets of CLIP-seq peaks called from 171 RBPs across different cell lines.

Finally, Class 4 TES features included the same chromatin data as in Class 1 for exploratory reasons. While chromatin features around TES are known to play important roles in transcription termination, their importance in RNA exosome targeting is unclear.

An important consideration was that exosome targets are generally less expressed than non-NEXT/PAXT targets (Supplementary Figure S1B), and many of the features included above, e.g. histone modifications and RBP binding strengths, correlate with gene expression levels (13,36,37). Hence, to avoid that our classification was confounded by differing expression levels, we performed expression match sampling between the three exosome target categories (see the 'Materials and Methods' section and Supplementary Figure S1B). To assess the diversity of the generated fea-

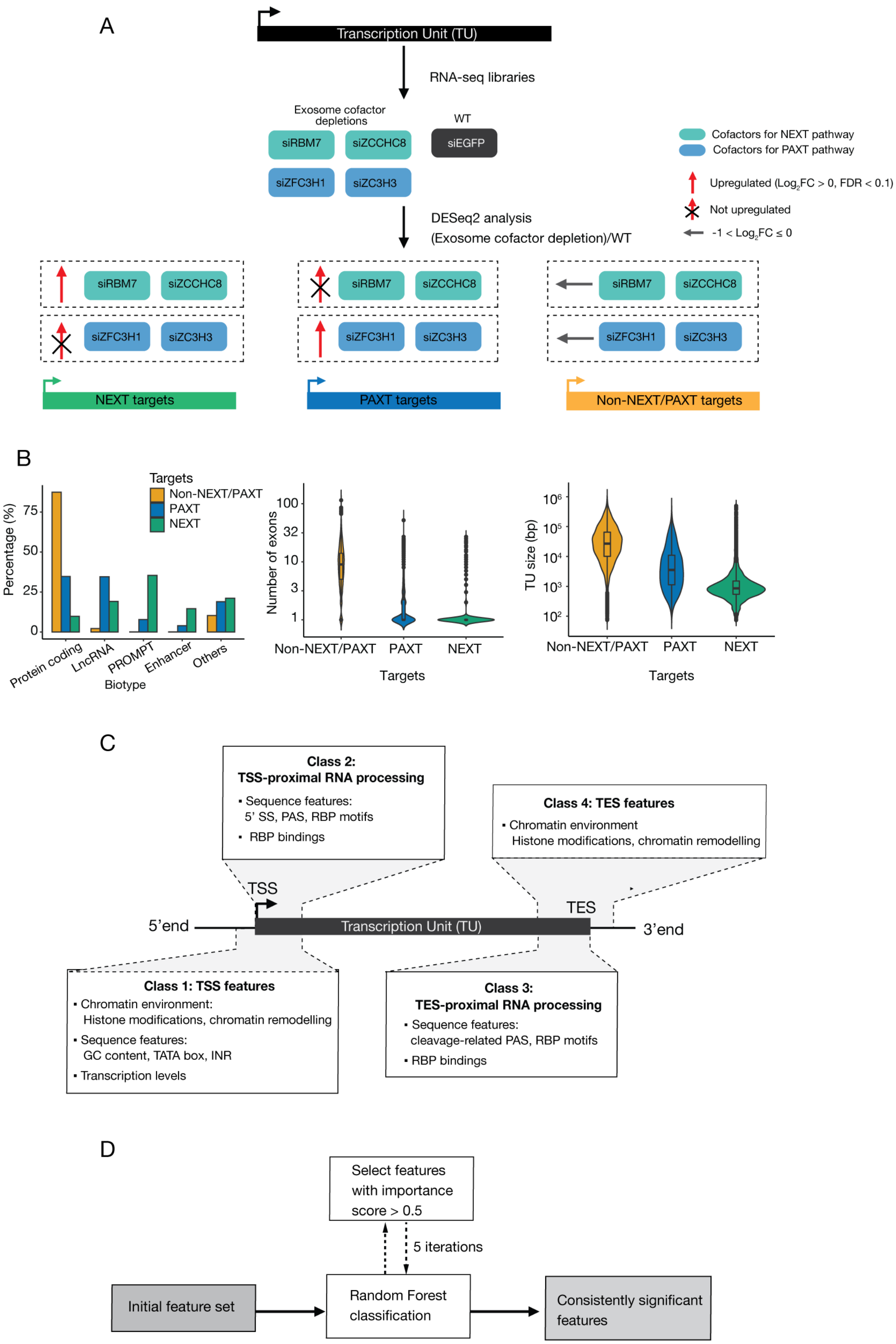

Figure 2. Exosome target dataset, feature design and machine learning framework. (**A**) Schematic overview of nucleoplasmic exosome target definition based on RNA-seq libraries of siRNA-depleted cofactors of NEXT and PAXT pathways versus wild-type (WT) cells. (**B**) Characterization of nucleoplasmic exosome targets. Bar plots on the left show the percentage (%) (*Y*-axis) of specific RNA biotypes for each exosome target category [*X*-axis; the 'Others' biotype consists of 17 types: unprocessed_pseudogene, transcribed_unitary_pseudogene, polymorphic_pseudogene, processed_pseudogene, rRNA_pseudogene, misc_RNA, TEC, snRNA, snoRNA, histone_coding, miRNA, NAT, intragenic, intergenic, TtT3, overlapping, nTtT, ambiguous; for details, see (28)]. Combined violin box plots in the middle show the distribution of the number of exons (*Y*-axis) for each exosome target category; combined violin box plots on the right show the distribution of TU length (*Y*-axis) for each exosome target category. (**C**) Schematic representation of the feature classes. For detailed feature descriptions, see Table 1 and the 'Materials and Methods' section. (**D**) Schematic overview of the machine learning framework.

Table 1. Molecular features used in this study

	Feature definition	Feature extraction	No. before filtering	No. after filtering
Class 1 TSS-localized transcription, chromatin and DNA sequence features				
Chromatin environment	Levels of histone modification and chromatin remodeling (ChIP-seq for H3K4me1, H3k4me2, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27ac, H3K36me3, H4K20me1, H2A.Z; DNase-seq)	Sum of signals in ±500 bp window around TSS	11	11
Transcription levels	Nascent RNA levels by NET-seq	Sum of signals in −100 to +500 bp window around TSS	1	1
	Polymerase II loading (Pol II ChIP-seq)	Sum of signals in ±500 bp window around TSS	1	1
Sequence features	G/C content	G or C content in ±500 bp window around TSS	1	1
	GC spread	Calculated (see the 'Materials and Methods' section) in +1 to +500 bp window to TSS	1	1
	Presence of TATA box	Frequency of motif hit in −50 to +1 bp window to TSS	1	0
	Presence of INR element	Frequency of motif hit in −15 to +10 bp window to TSS	1	0
Class 2 TSS-proximal RNA processing-related features				
Sequence features	Presence of 5′ SS motif	Frequency of motif hit in +1 to +500 bp window to TSS	1	1
	Presence of PAS motif	Same as above	1	1
RBP motifs and binding sites	Presence of RBP binding motifs	Same as above	193	77
	Presence of RBP binding sites by CLIP-seq	Frequency of binding sites in +1 to +500 bp window to TSS	401	26
Class 3 TES-proximal RNA processing-related features				
Sequence features	PAS strength	PAS score calculated by APARENT (see the 'Materials and Methods' section)	1	1
	Cleaved PAS	Frequency of motif hit in −50 to −1 bp window upstream of TES	1	1
	Presence of 5′ SS motif	Frequency of motif hit in −500 to +1 bp window to TES	1	1
	Presence of PAS motif	Same as above	1	1
RBP motifs and binding sites	Presence of RBP binding motifs	Same as above	193	96
	Presence of RBP binding sites by CLIP-seq	Frequency of binding sites in −500 to +1 bp window to TES	401	58
Class 4 TES-localized chromatin features				
Chromatin environment	Levels of histone modification and chromatin remodeling (ChIP-seq for H3K4me1, H3k4me2, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27ac, H3K36me3, H4K20me1, H2A.Z; DNase-seq)	Sum of signals in ±500 bp window around TES	11	11

The features are divided into four classes by their locations with respect to TUs. To avoid misleading high z-score from features of low information content, features are further filtered by entropy; only those with entropy > 0.8 across three exosome target categories are retained. The columns 'No. before filtering' and 'No. after filtering' correspond to the number of features before and after entropy filtering.

ture space, we calculated the pairwise Pearson correlation coefficients between different features by taking values in all three exosome target categories after expression sampling matching (Supplementary Figure S1C). This showed that only few features correlate strongly, while the whole feature space does not, which confirmed the high dimensionality of the data in the three exosome target categories.

Feature filtering and iterative feature selection

We next used random forests (38) as the machine learning framework to predict NEXT/PAXT targets. We chose to make three initial binary classifications independently: NEXT versus non-NEXT/PAXT, PAXT versus non-NEXT/PAXT and NEXT versus PAXT. These were chosen instead of multiclass classifications as it allowed us to identify specific features distinguishing the individual categories from each of the other two categories. A primary aim of our study was to identify molecular features distinguishing two exosomal decay pathways. Hence, a key endpoint was to be able to compare the relative contribution of the input features and select the most discriminative features contributing to learning a classification problem, rather than focusing on the predictive performance alone. In the random forest method, feature ranking and selection can be facilitated by using the individual importance z-scores, which are based on 'out-of-bag' errors in the decision tree process employed by the algorithm. While useful, such z-scores have two weaknesses.

First, feature importance assessment can overemphasize features that have consistent but small effects on performance, resulting in high z-scores (see the 'Materials and Methods' section). As such features have low information content and would not be considered to be generally discriminative, we only retained features with an entropy > 0.8 across the three exosome target categories (Table 1, Supplementary Figure S1D and Supplementary Table S2). This filtering removed TATA box and INR features. For CLIP-seq RBP binding sites, 26 sets of peaks of 24 RBPs for Class 2 and 58 sets of peaks of 37 RBPs for Class 3 were retained after filtering. To investigate whether the retained RBP features share certain functions, we performed Gene Ontology over-representation analysis of RBPs from both CLIP-seq and motif features using 1542 RBPs from (39) as background. We found that Class 2 and 3 input RBP features were enriched for processes related to RNA splicing and regulation of cellular metabolic processes (Supplementary Figure S2).

Second, ranking of features by z-score can be arbitrary if it is only based on a single prediction outcome and when there are many redundant features, in which case the importance score of each feature will be diluted. To obtain consistent results, we designed an iterative feature selection strategy. We opted to choose a relatively low importance score of 0.5 as a threshold and performed iterative feature selection by first training a model, then removing features below the threshold and subsequently retraining the model on the remaining features (Figure 2D). Training and feature selection was repeated until the feature set became stable. A stable feature set of consistently significant features was obtained after five iterations in each feature class (data not shown). Compared to other iterative feature selection strategies, such as recursive feature elimination, in which one feature with lowest rank is removed in each iterative round, our method needed fewer iterative rounds and was more efficient given that we have four feature classes. In addition, our method was based directly on importance score instead of ranking, yielding a clear cutoff during the feature selection process. The iterative feature selection resulted in only minor changes of prediction performance and the importance scores of the selected features of the final models were used for determining the most discriminative features. To confirm our feature selection results, we compared the selected features from our method with results from another feature selection method, Boruta (34), which estimates the importance threshold using 'shadow' features obtained by shuffling the values of original features and selects significant features by iteratively comparing the importance score of original feature with estimated importance threshold. As shown in Supplementary Figures S3–S5, the two methods gave largely consistent sets of features, where the top ranked features were consistent across the methods and both methods identified the same features with largest effect size.

Classification of NEXT versus non-NEXT/PAXT targets

We first attempted to distinguish NEXT targets from non-NEXT/PAXT targets. The iterative feature selection yielded a reduction of the number of features in Classes 2 and 3, while the number of features in Classes 1 and 4 remained unchanged (Supplementary Figure S1D and Figure 3A). However, overall, this feature selection resulted in little, if any, changes in classification performance for all classes (Figure 3B). To explore the classification potential of each class of features, we ran the classification algorithm using each feature class separately, and subsequently using every possible combination of feature classes. All single-class and multiclass models showed moderate to very good performance (F_1 score ranging from 0.69 to 0.90 and AUC from 0.78 to 0.94; Figure 3B). Notably, classification using Class 1 of TSS features was substantially less predictive (F_1 score 0.69, AUC 0.78) than the remaining feature classes (lowest F_1 score ~ 0.870 and AUC ~ 0.92, discussed further below). Interestingly, combinations of different feature classes did not affect the performance substantially, indicating that there were little or no synergistic effects of features from different classes in terms of prediction.

Next, we assessed the most discriminative features in each feature class (z-score > 1; top features are shown in Figure 3C–F). For Class 1, the GC spread, a proxy for the width of high G/C content regions downstream of TSSs, was among the top discriminative features and more discriminative than the accumulated G/C content around the TSS (Figure 3C, left panel). Other top ranked features included histone modifications H3K4me1, H3K4me3, H3K4me2 and H3K27ac, the combination of which is commonly associated with enhancer/promoter activities (40). To identify how differences of individual features between the two target categories may affect classification, we plotted the distributions for selected top ranked features (Figure 3C, right panel). This showed that the GC spread was on average smaller in NEXT targets compared to the non-NEXT/PAXT targets, and had higher variance, consistent with previous observations (21,24). For histone modifications, NEXT target loci had higher average levels of H3K4me1 and H3K27ac, and similar levels of H3K4me3 compared to non-NEXT/PAXT targets.

The top discriminative features in Class 2 were CLIP-seq RBP binding and RBP binding motif matches, and much of the predictive power derived from a single feature — the CLIP binding data of the DDX3X protein, which belongs to the DEAD-box RNA family of helicases (Figure 3D, left panel). DDX3X functions in both the nucleus and the cytoplasm, and plays diverse roles in regulating transcription, mRNA maturation, export and translation (41). Aside from DDX3X, several RBP motifs related to RNA splicing were among the most discriminative features, including top ranked 5′ SS, MBNL3, SRSF1 and SRSF4 motifs (Figure 3D, left panel). Plotting distributions of the top features showed a clear depletion of all these features in NEXT compared to non-NEXT/PAXT targets (Figure 3D, right panel). This implied that NEXT targets are to a lesser degree subject to RNA processing, such as splicing, compared to non-NEXT/PAXT targets, which may be due to the absence of relevant sequence motifs and/or the

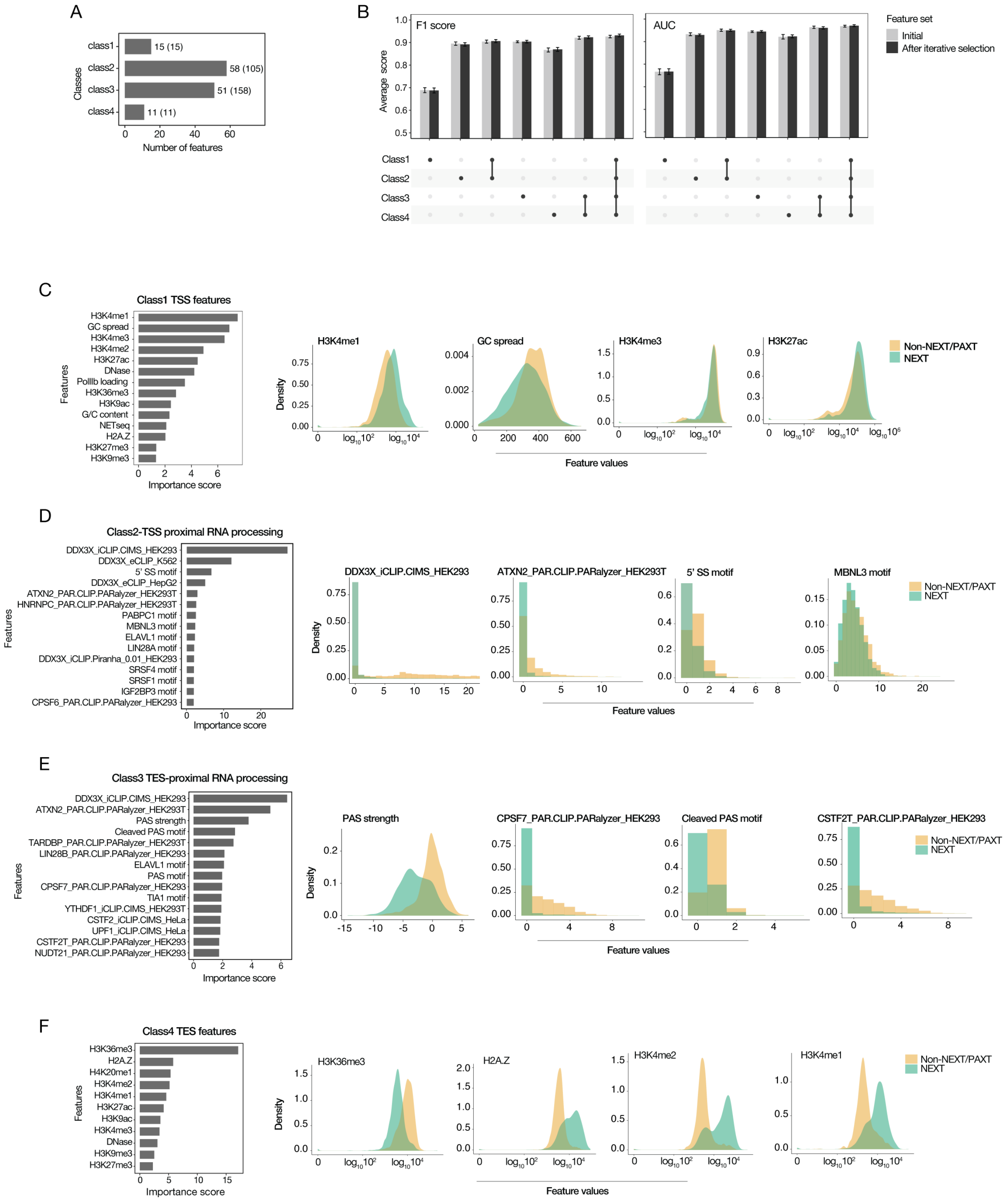

Figure 3. Predictive model of NEXT versus non-NEXT/PAXT targets. (**A**) Number of features retained after iterative selection, per feature class. Bar plot shows the number of features (*X*-axis) for each feature class (*Y*-axis) after iterative selection. Numbers in the parentheses show the number of initial features in each feature class. (**B**) Classification performance by random forest using single feature classes or combinations thereof. Bar plots in the upper panel show the average performance (F_1 score on the left, AUC on the right) over 10 repetitions for initial feature set and consistent feature set after iterative selection, as indicated by bar color. Error bars show the standard deviation of the performance over 10 repetitions. The lower panel shows which feature class (dots) or combinations thereof (dots connected by black lines) were used for classification. (**C**) Feature importance of Class 1 features. Bar plot (left panel) shows the feature importance score (*X*-axis) of top ranked features (*Y*-axis) ordered by importance score. The distributions of feature values of selected features, split by NEXT and non-NEXT/PAXT targets, are shown as density plots to the right. (**D–F**) Feature importance of Class 2, 3 and 4 features, organized as in panel (C). Distributions to the right are shown as either density plots or histograms, depending on data type.

lack of actual RBP binding. While the enrichment of TSS-proximal PAS motifs was previously observed in exosome targets (21,24,25,42), it was not ranked as an important feature for distinguishing NEXT and non-NEXT/PAXT targets, likely owing to the specificity of NEXT toward unadenylated transcripts (20).

The top two features of Class 3 were also among the top discriminative features in Class 2, but DDX3X was no longer the single standout feature (Figure 3E). Instead, it had more similar importance score as the rest of the top discriminative features in Class 3. Examining these top features, together with their distributions, we found that most of them are relevant to canonical 3′ end RNA processing by the CPA machinery, and as expected NEXT targets were depleted of these features. For example, compared to non-NEXT/PAXT targets, NEXT targets lack a well-positioned PAS motif upstream of the transcript TES, weaker PAS strength and lack of binding of RBPs involved in the 3′ end cleavage process, such as CSTF2, CSTF2T, CPSF7 and NUDT21 (CPSF5). Again, this was consistent with previous studies that NEXT targets generally do not undergo canonical CPA-dependent 3′ end processing (20,28).

The performance of Class 4 features, like those of Class 2, relied mostly on a single standout feature: H3K36me3 levels (Figure 3F, left panel). Lower average levels of H3K36me3 around TESs were observed for NEXT versus non-NEXT/PAXT targets (Figure 3F, right panel). H3K36me3 is known to be enriched in the gene body during active transcription and associated with efficient transcription elongation (43) and it has also been reported to be deposited downstream of the first intron (44); thus, lower levels of H3K36me3 around NEXT TESs agreed with the notion that NEXT targets are usually short and unspliced, and derive from inefficient transcription elongation (15,28). In addition, and even though Class 4 features were centered around TESs, other top ranked features were associated with transcription initiation, including H2A.Z, H3K4me2 and H3K4me1 levels, and showed higher average values around the TESs of NEXT than non-NEXT/PAXT targets (Figure 3F, right panel). We reasoned that because many NEXT targets are short (Figure 2B), regions around the TESs of such targets may, at least partially, share the same larger chromatin environments with the cognate TSSs, which might explain our observations. Specifically, in our analysis, ±500 bp regions around TESs — in which chromatin modification signals were collected (Table 1) — might overlap with those around TSSs of transcripts $<$ 1 kb. We found that the analyzed TSS and TES regions had a median overlap of 200 bp, which was substantially higher than the corresponding overlap in PAXT and non-NEXT/PAXT targets (median 0 bp) (Supplementary Figure S6A, upper panel). However, when plotting the distribution of the levels of these chromatin modification data in NEXT and non-NEXT/PAXT targets that were long enough to not have overlaps between the TSS and TES windows, we observed similar levels of H3K36me3 as in Figure 3F (Supplementary Figure S6B). The differences are less prominent between NEXT and non-NEXT/PAXT targets for H2A.Z, H3K4me2 and H3K4me1 (Supplementary Figure S6B), but the levels of these histone features are consistently higher in NEXT compared to non-NEXT/PAXT targets as in Figure 3F. Thus, overlap between NEXT TSS and TES regions cannot fully account for the observed enrichment of TSS-associated chromatin modifications. Taken together, our observations support the idea that transcription of NEXT targets is likely to fail transition into efficient elongation (28).

To further validate the significance of the most important features of the single-class models, we assessed the predictive power of the features in the full model with features of all four classes (Supplementary Figure S7). Consistent with the prediction power (Figure 3B), the top ranked features were mostly from Classes 2–4, while the Class 1 features had lower ranks. In accordance with the individual models, the two top ranked features, H3K36me3 levels and DDX3X binding, corresponded to the two standout features of Classes 4 and 2, followed by top ranked features of Class 3 ATXN2.

NEXT targets are heterogeneous in terms of biotypes, so we wondered whether the identified discriminative features might only reflect the predominant biotypes. In fact, the length filtering performed in Supplementary Figure S6B increased the proportion of protein-coding biotype RNAs in NEXT targets (Supplementary Figure S6A, lower panel), which might account for changed distribution of certain histone features. Thus, we further investigated whether the identified discriminative features are dependent on biotypes by only assessing protein-coding RNA (252 NEXT and 3204 non-NEXT/PAXT targets). The prediction performance of the four feature classes was generally worse for protein-coding NEXT versus non-NEXT/PAXT targets compared to the original analysis, but the trend of prediction power between different classes remained similar (Supplementary Figure S8A and Figure 3B). Class 1 TSS-proximal features were least predictive, performing only slightly better than random with F_1 score 0.54 and AUC 0.58. Compared to Class 1, Class 4 TES-proximal chromatin features showed better prediction performance and were moderately predictive with F_1 score 0.64 and AUC 0.71. Class 2 and 3 RNA processing features were more predictive compared to Classes 1 and 4, and Class 3 TES-proximal features have the best prediction performance (F_1 score 0.69, AUC 0.77). As Class 1 features were not very predictive, we only performed feature selection on Classes 2–4. The top features of all the three classes were highly consistent with those in the original analysis (Supplementary Figure S8B–D and Figure 3D–F) and we did not identify new significant features. These results suggest that Class 1 TSS features are mostly informative for predicting noncoding RNAs but not protein-coding NEXT targets, while discriminative features identified in Classes 2–4 are informative for NEXT targeting irrespective of biotypes.

In summary, all four feature classes contained substantial power to discriminate NEXT from non-NEXT/PAXT targets, suggesting that substantial differences exist in terms of chromatin configuration, sequence content and RNA processing capacity. Comparing the prediction performance of the four feature classes, RNA processing features at both transcript 5′ and 3′ ends possessed the highest discriminatory power. The most informative features for TSS-proximal processing consisted of RBP binding and sequence features mainly related to 5′ splicing, while the most informative TES-proximal features are mostly involved in canonical 3′ end processing by the CPA machinery, including both cleavage-related RBP binding and sequence features, such as well-positioned PAS motifs. TES histone modifications associated with efficient transcription elongation could also separate NEXT from non-NEXT/PAXT targets, but to a lesser extent. Interestingly, although many TSS-related features were observed to be different between exosome and non-exosome targets, such as H3K4 methylation status and G/C content, they were substantially less predictive for NEXT targeting. These features could only partially distinguish NEXT from non-NEXT/PAXT targets and lacked predictive power to distinguish protein-coding NEXT versus non-NEXT/PAXT targets.

Determinants for classifying PAXT from non-NEXT/PAXT targets

We next applied the model to classify PAXT versus non-NEXT/PAXT targets. Iterative feature selection resulted in fewer retained features than the NEXT versus non-NEXT/PAXT classification for all classes, and this was particularly true for Class 2 and even more so for Class 3 (Figures 3A and 4A). This classification was not affected by feature selection (Figure 4B). However, it was in general worse for all four classes compared to NEXT versus non-NEXT/PAXT (Figures 3B and 4B): the F_1 score and AUC for the best-performing class (Class 3) were ∼0.78 and 0.83; the worst-performing class had values around 0.55 and 0.57 (Class 1), which were only slightly better than random classification. Indeed, the distributions of the three top features of Class 1 were all largely similar between the two target categories (Figure 4C). Combinations of feature classes only showed minor synergistic effects, if any (Figure 4B).

Class 2 features yielded a moderately good prediction performance with an F_1 score around 0.7 and AUC around 0.8. The top ranked features were all related to RBP binding, and, similarly to the NEXT versus non-NEXT/PAXT target classification, DDX3X binding was the most informative feature. Other top features consisted of binding of splicing-related RBPs, such as PRPF8, U2AF2 and YTHDC1 (Figure 4D, left panel). Examination of the distribution of these features showed that, as for NEXT targets, the PAXT targets generally showed a lack of RBP binding events (Figure 4D, right panel). Unlike NEXT versus non-NEXT/PAXT targets, the TSS-proximal 5′ SS motif and other RBP motifs were not highly ranked features. Thus, classification relied on differences from the RBP binding instead of sequence features.

Class 3 features showed the best classification performance (F_1 score 0.78 and AUC 0.83). Similar to Class 2, all the top ranked features were related to RBP binding and with similar degrees of importance (Figure 4E). While there was some overlap in terms of top class RBP features between the PAXT and NEXT versus non-NEXT/PAXT target classification, e.g. DDX3X, ATXN2, YTHDF1 and LIN28B, 3′ end CPA-related features were not among the most discriminative for classifying PAXT versus non-NEXT/PAXT targets. This indicates that while PAXT might differ from non-NEXT/PAXT targets in their TES-proximal RNA processing, they may share similar canonical 3′ end processing, i.e. CPA.

Class 4 features were moderately predictive (F_1 score 0.65 and AUC 0.7) but to a lesser degree than Class 2 and 3 features. The top ranked discriminative features showed a clear similarity to those of the NEXT versus non-NEXT/PAXT target classification, where H3K36me3 levels were a single standout significant feature, with on average lower levels around TESs of PAXT targets (Figure 4F). As with NEXT classification features, transcription initiation-associated chromatin features, such as H3K4me3, H3Kme2 and H2A.Z, were also informative, and more enriched in PAXT targets than non-NEXT/PAXT targets.

Next, we assessed the feature importance of the four classes in the full model with features of all four classes. As with NEXT versus non-NEXT/PAXT classification, the result was largely consistent with the prediction power and the feature rank observed for the individual classes (Supplementary Figure S9).

As in the NEXT analysis, we also trained and evaluated on protein-coding biotype alone, comparing 388 protein-coding PAXT with 3204 protein-coding non-NEXT/PAXT targets. Compared to the original analysis, the prediction performance of the four feature classes was worse but with similar trends: as above, Class 1 features were the least predictive and Class 3 the best with F_1 score 0.66 and AUC 0.70 (Supplementary Figure S10A). The top features were to a large extent consistent with the original analysis (Supplementary Figure S10B–D and Figure 4D–F). These results indicate that the discriminative features, especially TES RNA processing features identified in original analysis, are generally informative for PAXT targeting.

In summary, the prediction performance showed that PAXT targets were more challenging to distinguish from non-NEXT/PAXT targets as compared to NEXT targets. The poor prediction performance of Class 1 features suggests that PAXT and non-NEXT/PAXT targets are to a large extent similar with respect to their promoter properties, such as histone modifications, G/C content and spread. The most informative features for distinguishing PAXT and non-NEXT/PAXT targets were RBP binding features, especially those proximal to TESs. Notably, CPA-related features were not highly discriminative, suggesting that the CPA machinery is shared across the two classes. In addition, sequence features were not important for classifying PAXT and non-NEXT/PAXT targets, which included features previously known for exosome targeting such as promoter G/C content, 5′ SS, PAS motifs and RBP binding motifs.

Determinants for classifying NEXT from PAXT targets

Finally, we sought features distinguishing NEXT from PAXT targets using the same framework as above. After iterative feature selection, fewer features of all classes were retained, particularly for Classes 2 and 3, compared to the other two classifications above (Figure 5A). Feature selection only had a small effect on classification performance (Figure 5A), which was in general worse than the NEXT versus non-NEXT/PAXT classification and slightly better than the PAXT versus non-NEXT/PAXT classification,

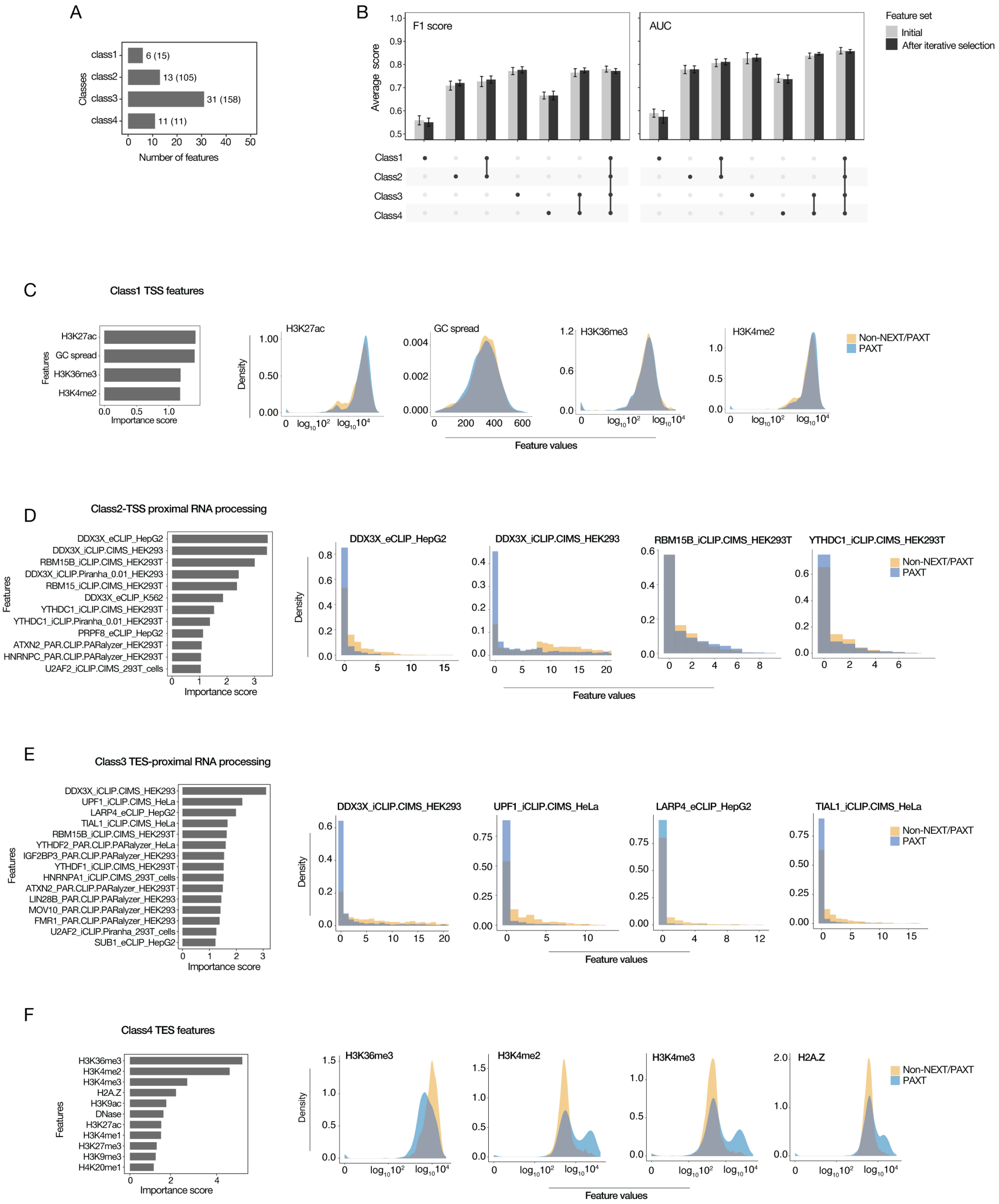

Figure 4. Predictive model of PAXT versus non-NEXT/PAXT targets. Panels (**A**)–(**F**) are organized as those in Figure 3A–F, but based on PAXT versus non-NEXT/PAXT targets.

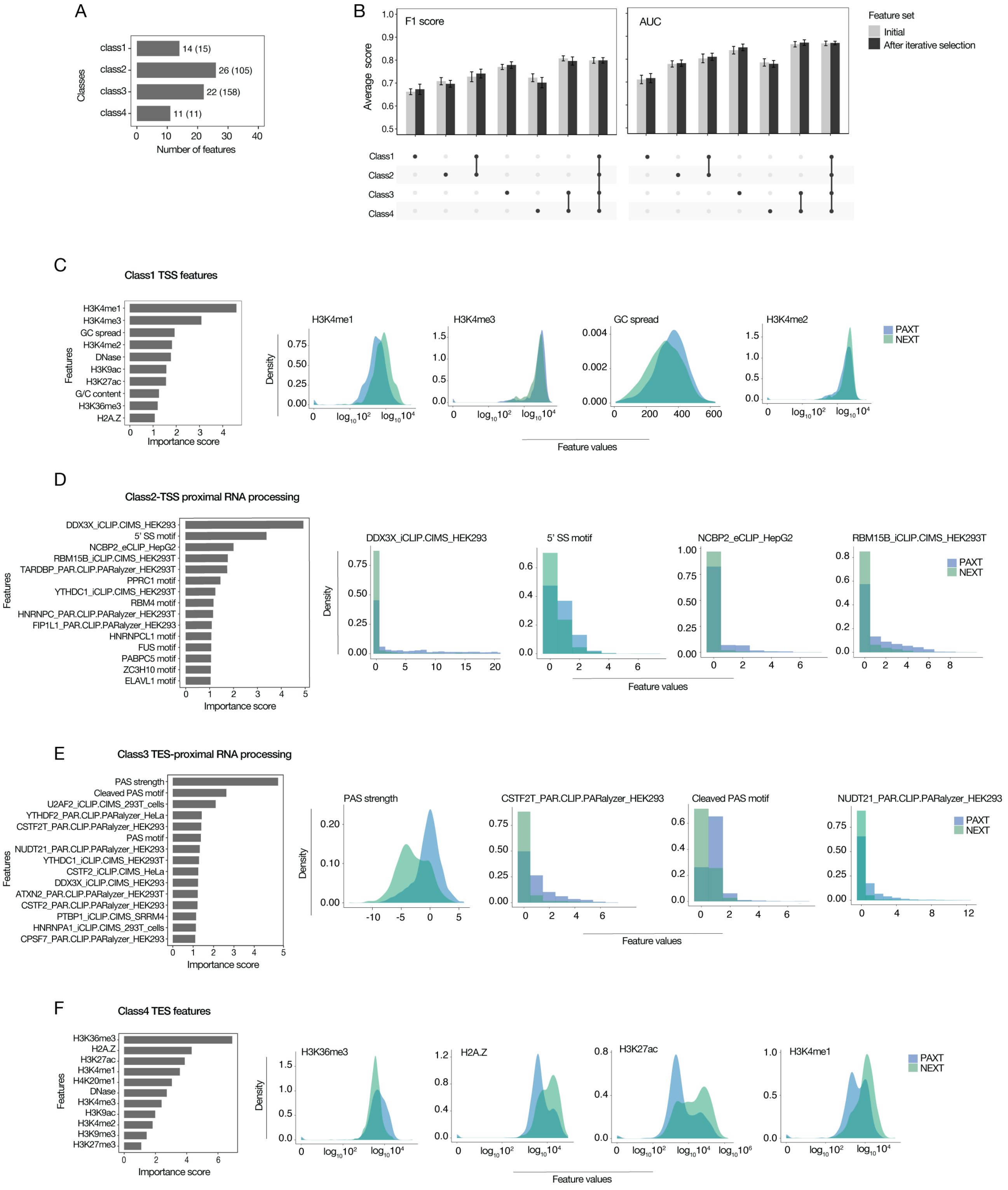

Figure 5. Predictive model of PAXT versus NEXT targets. Panels (**A**)–(**F**) are organized as those in Figure 3A–F, but based on NEXT and PAXT targets.

with F_1 score and AUC ranging from 0.67 to 0.78 and 0.72 to 0.85, respectively. Class 3 features of TES-proximal RNA processing were comparable to PAXT versus non-NEXT/PAXT targets and gave the best prediction performance (Figure 5B). Only minor synergistic effects were observed when combining feature classes (Figure 5B).

Although Class 1 features were not highly discriminative, H3K4me1, H3K4me2 and H3K4me3 levels and GC spread were the top ranked features, similarly to the NEXT versus non-NEXT/PAXT target analyses (Figures 3C and 5C). The shared top Class 1 features between NEXT versus PAXT and NEXT versus non-NEXT/PAXT targets indicated that PAXT and non-NEXT/PAXT targets are somewhat similar for Class 1 features, which was also consistent with the poor prediction performance of Class 1 features in PAXT versus non-NEXT/PAXT targets (Figure 4B and C). Similar to NEXT versus non-NEXT/PAXT targets, we found that NEXT targeted loci had on average higher levels of H3K4me1 and H3K4me2 and lower levels of H3K4me3, despite matched expression sampling. The GC spread distribution indicated that the width of the TSS-downstream G/C-rich region is on average smaller for NEXT than PAXT targets, which displayed a more well-defined width as evidenced by the smaller variance.

Class 2 features had slightly more classification power than Class 1 features (Figure 5B). The top features consisted of both sequence and RBP binding features. Examining the top ranked features and their distribution, binding of DDX3X and the 5′ SS motif were the two most outstanding top ranked features (Figure 5D). Distribution plots showed that NEXT targets in general lack RNA processing features compared to PAXT targets (Figure 5D, right).

Class 3 features yielded the best classification performance (Figure 5E, left panel). PAS strength around the TES stood out as the most discriminative feature followed by well-positioned PAS motif upstream of the TES, and both had higher values for PAXT than for NEXT targets (Figure 5E, right panel). Most of the other top features were RBP binding, which had less signal in NEXT compared to PAXT targets. Examining these features, many were RBP binding relevant to 3′ end processing by the CPA machinery, including CSTF2, CPSF7 and NUDT21. The difference in RBP binding, PAS strength and well-positioned PAS motif features between classes confirmed a major difference of TES-proximal processing between PAXT and NEXT targets with PAXT targets being dependent on the CPA machinery (20).

Class 4 features had slightly better prediction performance than Class 1 but worse than Class 2 and 3 features (Figure 5B). Top ranked features were similar to those observed in NEXT versus non-NEXT/PAXT and PAXT versus non-NEXT/PAXT targets (Figure 5F, left panel). Here, H3K36me3 was still the most important feature for the classification performance, which on average had higher levels around TES of PAXT than NEXT targets, followed by features correlated to active transcription initiation, such as H3K27ac and H2A.Z with higher levels around TES of NEXT than PAXT targets (Figure 5F, right panel). This suggests that NEXT and PAXT targets both undergo less efficient elongation compared to non-NEXT/PAXT targets, but to a different extent. This fits with the exosome's role as a quality control pathway to remove transcripts that would arise from misconfigured RNAPII/nonproductive elongation (28).

Assessing the feature importance of four classes in the full model with features of all four classes showed consistent results that agree with prediction power and feature rank observed in the individual classes (Supplementary Figure S11).

As with NEXT and PAXT analyses above, we also performed prediction using protein-coding biotype RNAs alone (252 NEXT with 388 PAXT targets). Because enough data were available, we repeated this analysis also in the lncRNA biotype RNAs independently (492 NEXT with 386 PAXT targets). Classes 1 and 4 were only predictive for lncRNA with limited power, and the top features for lncRNA biotype were largely consistent with the original analysis (Supplementary Figure S12). Both Class 2 and 3 RNA processing features showed similar predictive power for protein-coding RNA and lncRNA (Supplementary Figure S12A). Class 3 TES RNA processing features had the best prediction performance. Examining the top discriminative features, we found Class 2 TSS-proximal 5′ SS was only important for distinguishing NEXT and PAXT in the lncRNA but not protein-coding biotype. While a few new significant features were identified in Class 2 and 3 features, many top features overlapped between biotype stratified analysis and original analysis (Supplementary Figure S12D), suggesting in spite of biotypes, the differences of TES RNA processing are generally most informative for distinguishing NEXT and PAXT targets.

In summary, NEXT could be distinguished from PAXT targets by both sequence and other types of features, especially RBP binding. While both TSS- and TES-proximal RNA processing features had good prediction performance, the latter can best distinguish NEXT from PAXT targets, reiterating that a prominent difference between NEXT and PAXT targeting is dictated by 3′ end processing, where the top ranked features, such as PAS motifs and CPA-related RBPs, suggest that the difference is determined by the CPA machinery.

Identification of NEXT and PAXT pathway-specific features

A general pattern derived from the three binary classifications discussed above was that Class 2 and 3 features overall had good prediction performances with Class 3 yielding the best results. Therefore, we investigated whether NEXT- and PAXT-specific features could be identified from these two classes. For NEXT specificity, we considered features retained after iterative selection and that occurred in both NEXT versus PAXT and NEXT versus non-NEXT/PAXT target classifications. Similarly, features that occurred in both NEXT versus PAXT and PAXT versus non-NEXT/PAXT target classifications were considered PAXT specific. Finally, features that occurred in both NEXT versus non-NEXT/PAXT and PAXT versus non-NEXT/PAXT target classifications were considered as 'non-NEXT/PAXT'-specific features and features shared in all comparisons were considered discriminative for all three target categories (Figure 6, top panel).

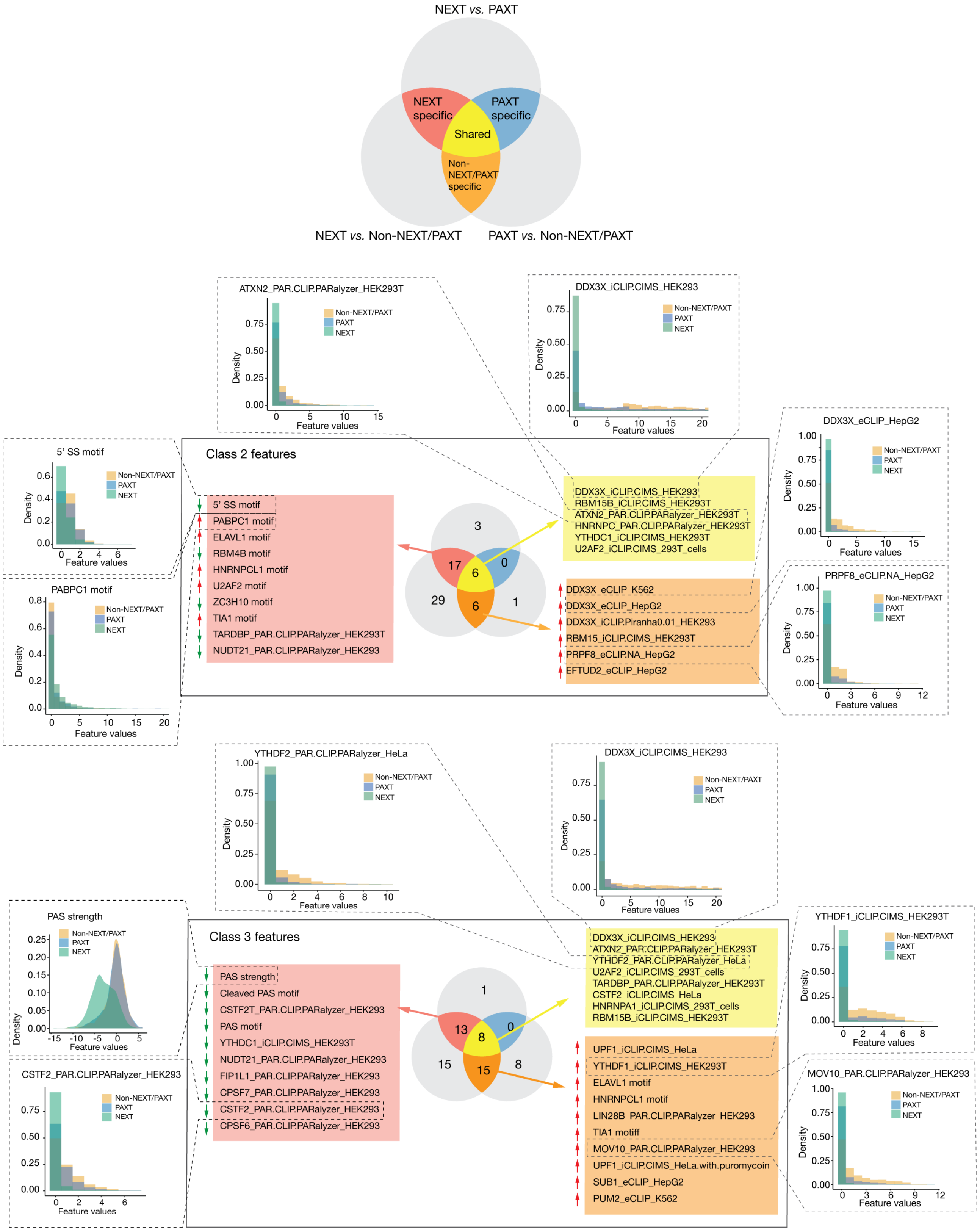

Figure 6. Exosome target pathway-specific features. Schematic Venn diagram in the top panel shows the definition of pathway-specific features. Lower panels show the corresponding Venn diagram of Class 2 (upper box) and Class 3 (lower box); the pathway-specific and shared features are listed in colored boxes. Arrows in front of NEXT- and non-NEXT/PAXT-specific features indicate whether the feature is depleted (arrow downward) or enriched (arrow upward) in NEXT or non-NEXT/PAXT targets compared to the other two targets.

Notably, we found no PAXT-specific features (as defined above) in either Class 2 or 3. The NEXT-specific features from Class 2 were mostly sequence features in which the 5′ SS motif was the most informative and with an importance score that was roughly 3.2-fold higher than the second highest ranked feature. While some of these NEXT-specific sequence features (indicated by red upward arrow in Figure 6) were more enriched in NEXT compared to non-NEXT/PAXT targets, these patterns could most likely be explained by the distinct TSS-proximal nucleotide compositions of NEXT targets that generally have higher A/U than G/C contents (20). Consistently, the enriched RBP motifs in NEXT targets contained high AU content, while depleted ones were more G/C rich (Supplementary Figure S13). In Class 3, NEXT-specific features consisted of both sequence and RBP binding features, all of which were depleted in NEXT targets, and which are mostly relevant to RNA 3′ end cleavage, e.g. PAS strength, cleaved PAS motif, and RBP binding of CSTF2T, NUDT21, CPSF7 and CSTF2. Most shared and non-NEXT/PAXT-specific features of Classes 2 and 3 were RBP binding signals, which were observed to be generally enriched in non-NEXT/PAXT targets compared to PAXT and NEXT targets (Figure 6).

As mentioned in the ‘Materials and Methods’ section, we compared our results with pathway-specific features identified by the Boruta feature selection method. In general, there was high agreement between the features selected by either method: in both, most Class 2 NEXT-specific features were RBP motifs, Class 3 NEXT-specific features were mostly relevant to RNA 3′ end cleavage and processing, and most shared and non-NEXT/PAXT-specific features were RBP binding signals (Supplementary Figure S14). Although the Class 2 5′ SS motif was considered a shared feature by the Boruta method instead of a NEXT-specific feature, this feature ranked much higher in NEXT versus non-NEXT/PAXT (ranked 5 of 64) and NEXT versus PAXT (ranked 2 of 50), compared to PAXT versus non-NEXT/PAXT (ranked 31 of 34). Thus, it seems reasonable to consider the 5′ SS motif more informative for NEXT prediction.

We then asked how well features of the four classes were able to distinguish the respective exosome degradation pathways in a multiclass classification model. To this end, we used both the initial feature sets and the consistently retained feature sets of the four classes established across all three pairwise comparisons above, and then trained a multiclass random forest model. The feature selection had only a very minor influence on the prediction performance (Figure 7). Consistent with the individual binary models, Class 1 TSS features had a poor prediction performance with an average accuracy of around 0.5, albeit with a better than random expectation (0.33). Class 2 features of TSS-proximal RNA processing and Class 4 features of TES chromatin environment both had better prediction performances with accuracies around 0.6, while Class 3 features of TES-proximal RNA processing had the best accuracy of around 0.7. We observed minor positive synergistic effects of Class 1 and 2 features. Examining the confusion matrix of the multiclass classification model using the combined consistently significant features across all three comparisons, we found that for all four feature classes, NEXT targets were generally correctly classified with a higher confidence compared to PAXT targets, which could be best classified only by Class 3 features, consistent with the binary classification results (Supplementary Figure S15).

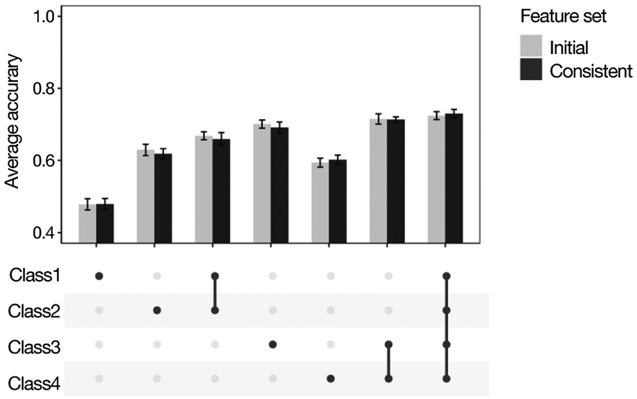

Figure 7. Multiclass classification. Bar plots (upper panel) showing the average accuracy over 10 repetitions for all features of the four classes and combined consistent significant features in four feature classes across all three comparisons. An error bar shows the standard deviation of the performance over 10 repetitions. The lower panel shows the feature class (dark dots) or combination of feature classes (dark dots connected by black solid line) used for classification.

DISCUSSION

The NEXT and PAXT exosome adaptors define two main nucleoplasmic targeting pathways for RNA decay via the nuclear RNA exosome. Previous observations have shown that RNPs targeted by these pathways harbor different characteristics with respect to features of the transcribed loci as well as RNA processing elements, such as splicing and 3′ end processing patterns (15,20,23–25). Here, leveraging the wealth of available genomic datasets, we utilize predictive models for unbiased and comprehensive exploration of molecular features associated with NEXT and PAXT targeting. Our goal was to find out to what degree NEXT/PAXT targets are predictable by either already established features or additional biological features previously not associated with NEXT/PAXT sensitivity.

To rank how informative these established and newly hypothesized biological features would be for predicting targets of these two pathways, we categorized features into four different classes, stratifying by location (i.e. TSS or TES), RNA processing or TSS/TES-related configurations. Based on the different feature classes, we trained random forest models to assess their abilities to distinguish targets and to identify relevant features within each class that drive the prediction performance. In contrast to many machine learning models that often display a ‘black box’ character, the importance of individual features can, with random forest models, be quantitatively assessed directly. The degrees of feature importances were therefore used to eliminate less relevant features in an iterative process, which led to a final model containing a consistent set of high-priority and discriminative features.

Given most exosome targets are noncoding RNAs while non-exosome targets are primarily protein-coding RNAs, it is interesting to evaluate how informative molecular features previously known to differ between coding and noncoding RNAs are for exosome targeting, for example features like G/C content and histone modifications. In our analysis, these features are less informative compared to RNA processing-related features and they are only distinctive for NEXT targets and unable to distinguish PAXT from non-NEXT/PAXT targets. This may be due to the fact that we are controlling for expression levels in our analysis, which otherwise would be expected to be higher on average for protein-coding genes. While NEXT targets could be most accurately classified from the other two target categories by all four feature classes based on both binary and multiclassification models, NEXT specificity seems more likely to be driven by features relevant to the RNA processing, especially 3′ end processing, than chromatin state. Moreover, these NEXT-specific features showed that one of the major differences contributing to distinguishing NEXT from the other two categories is related to sequence: many top ranked features are RBP binding motifs. Previous analysis of the nucleotide composition showed that NEXT targets are enriched for A and U and depleted of G and C. Thus, the RBP binding motifs in NEXT targets likely result from this distinct nucleotide composition, where many enriched motifs are AU rich while depleted ones are GC rich.

In terms of TSS-proximal RNA processing, it is well established that some exosome targets are depleted of 5′ SSs and enriched with PAS motifs (21,24,25,42). We found that a lack of 5′ SSs is only a characteristic for NEXT targets, while enrichment of PASs is not a discriminative feature. Previous studies have suggested that 5′ SSs may serve as a regulatory step to keep Pol II in active transcription via the binding of U1 snRNP, which subsequently suppresses 3′ end processing and premature transcription termination (42,45,46). However, it makes sense that this is not relevant for NEXT targets since their transcription termination was found to be mostly mediated by integrator rather than CPA complex (20,28). Consistently, among TES-proximal RNA processing features, well-positioned PAS motifs, PAS strength and RBP binding related to the RNA cleavage process in canonical 3′ end processing were found to be most discriminative and yielded the best prediction performance to distinguish NEXT targets from the other two categories.

Compared to NEXT targets, PAXT targets were more difficult to distinguish from non-NEXT/PAXT targets by the designed features in both binary and multiclass models. From prediction performance, PAXT targets were more similar to non-NEXT/PAXT targets in terms of TSS-relevant features, including histone modifications and G/C content. Prediction from histone modifications around the TES showed that PAXT targets share some similarities with non-NEXT/PAXT targets and some with NEXT targets. RNA processing-related features gave an overall better prediction performance for classifying PAXT from non-NEXT/PAXT targets. However, unlike NEXT targets, the most discriminative features were not sequence motifs but RBP binding sites. This suggests that the major differences between PAXT and non-NEXT/PAXT targets derive from differential RBP binding patterns instead of sequence. In line with this, we found that NEXT versus PAXT and NEXT versus non-NEXT/PAXT targets share many discriminative sequence features.

In addition to known features, our study also found that binding of yet unexplored RBPs is informative for distinguishing NEXT and PAXT pathways, especially RNA binding of the DDX3X helicase. While this may lead to new hypotheses, some technical biases need to be taken into consideration. We note that many of these discriminative RBP binding features were in general least abundant (had less average signal) in NEXT targets compared to the other two categories and less abundant in PAXT than non-NEXT/PAXT targets. This may indicate that both NEXT and PAXT targets generally lack RBP binding, which agrees with models suggesting that nuclear decay is considered as a default fate for all transcripts that lack specific protective features (47). However, technical biases introduced by CLIP-seq cannot be ignored. The instability of exosome targets in normal conditions with a fully functional exosome makes RBP binding difficult to capture by CLIP-seq and may only be revealed upon inactivation/perturbation of exosome-mediated decay pathways.

There are three important limitations to our approach. First, discrimination power in any classification is correlative and may at best give hypotheses for causal relationships. Second, as mentioned above, there may be additional features that are highly discriminative but that are not included in the model: this is highly likely for the classification of PAXT targets since the classification power is limited as well as protein-coding biotypes where the prediction performance is generally worse, and such features may include RNA modifications and exact isoform usage, including choice and efficiency of splice sites and 3′ end processing signals. Lastly, not all the NEXT and PAXT targets are included in the *de novo* HeLa transcriptome annotation that we use. For example, prematurely terminated transcripts originating from protein-coding gene TSS, which are often PAXT and/or NEXT targets (20–22), are often not included as they are uncharacterized transcripts that are typically not sequenced from 5′ to 3′ end. Because these transcripts are often lowly expressed and overlapping with highly expressed protein-coding transcript isoforms, it is difficult to accurately infer the exact isoform and quantify the abundance based on RNA-seq only. It would be interesting to include this class of targets in the future using quantitative full-length sequencing approaches.

In summary, our study systematically evaluated the ability and contribution of different molecular features to specifying the RNA exosome decay pathways, and directly compared 5′ end and 3′ end features. Our results show that NEXT, and to some degree PAXT sensitivity is feasible to computationally predict with current features, and our analysis of feature importance in predictions qualitatively validates previously identified differences between NEXT and PAXT targeting.

DATA AVAILABILITY

Code for all the analyses is available at GitHub (https://github.com/MengjunWu/ClassifyExosomePathways_ML).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Natsuda Navamajiti for helping with CLIP-seq data; William Garland for comments on the manuscript; Robin Andersson and the members of the Sandelin and Jensen laboratories for helpful discussions; and three anonymous reviewers for helpful comments for improving the manuscript.

Author contributions: M.W. and A.S. conceived the project. M.W. designed the computational framework. M.W. and M.S. analyzed the data. T.H.J. and A.S. supervised the project. M.W. and A.S. wrote the manuscript with input from all authors.

FUNDING

Lundbeck Foundation [to A.S. and T.H.J.]; Danish Cancer Society [to A.S.]; Novo Nordisk Foundation [to A.S. and T.H.J.]; Independent Research Fund Denmark [to A.S. and T.H.J.]; Carlsberg Foundation [to A.S.].

Conflict of interest statement. None declared.

REFERENCES

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Pefanis,E., Wang,J., Rothschild,G., Lim,J., Chao,J., Rabadan,R., Economides,A.N. and Basu,U. (2014) Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. *Nature*, **514**, 389–393.
3. Houseley,J. and Tollervey,D. (2009) The many pathways of RNA degradation. *Cell*, **136**, 763–776.
4. Jensen,T.H., Jacquier,A. and Libri,D. (2013) Dealing with pervasive transcription. *Mol. Cell*, **52**, 473–484.
5. Schmid,M. and Jensen,T.H. (2008) The exosome: a multipurpose RNA-decay machine. *Trends Biochem. Sci.*, **33**, 501–510.
6. Kilchert,C., Wittmann,S. and Vasiljeva,L. (2016) The regulation and functions of the nuclear RNA exosome complex. *Nat. Rev. Mol. Cell Biol.*, **17**, 227–239.
7. Schmid,M. and Jensen,T.H. (2018) Controlling nuclear RNA levels. *Nat. Rev. Genet.*, **19**, 518–529.
8. Mitchell,P., Petfalski,E., Shevchenko,A., Mann,M. and Tollervey,D. (1997) The exosome: a conserved eukaryotic RNA processing complex containing multiple $3' \rightarrow 5'$ exoribonucleases. *Cell*, **91**, 457–466.
9. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
10. Seila,A.C., Calabrese,J.M., Levine,S.S., Yeo,G.W., Rahl,P.B., Flynn,R.A., Young,R.A. and Sharp,P.A. (2008) Divergent transcription from active promoters. *Science*, **322**, 1849–1851.
11. Preker,P., Nielsen,J., Kammler,S., Lykke-Andersen,S., Christensen,M.S., Mapendano,C.K., Schierup,M.H. and Jensen,T.H. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, **322**, 1851–1854.
12. Flynn,R.A., Almada,A.E., Zamudio,J.R. and Sharp,P.A. (2011) Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 10460–10465.
13. Andersson,R., Refsing Andersen,P., Valen,E., Core,L.J., Bornholdt,J., Boyd,M., Heick Jensen,T. and Sandelin,A. (2014) Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.*, **5**, 5336.
14. Lubas,M., Christensen,M.S., Kristiansen,M.S., Domanski,M., Falkenby,L.G., Lykke-Andersen,S., Andersen,J.S., Dziembowski,A. and Jensen,T.H. (2011) Interaction profiling identifies the human nuclear exosome targeting complex. *Mol. Cell*, **43**, 624–637.
15. Meola,N., Domanski,M., Karadoulama,E., Chen,Y., Gentil,C., Pultz,D., Vitting-Seerup,K., Lykke-Andersen,S., Andersen,J.S., Sandelin,A. *et al.* (2016) Identification of a nuclear exosome decay pathway for processed transcripts. *Mol. Cell*, **64**, 520–533.
16. Silla,T., Schmid,M., Dou,Y., Garland,W., Milek,M., Imami,K., Johnsen,D., Polak,P., Andersen,J.S., Selbach,M. *et al.* (2020) The human ZC3H3 and RBM26/27 proteins are critical for PAXT-mediated nuclear RNA decay. *Nucleic Acids Res.*, **48**, 2518–2530.
17. Schuch,B., Feigenbutz,M., Makino,D.L., Falk,S., Basquin,C., Mitchell,P. and Conti,E. (2014) The exosome-binding factors Rrp6 and Rrp47 form a composite surface for recruiting the Mtr4 helicase. *EMBO J.*, **33**, 2829–2846.
18. Schneider,C. and Tollervey,D. (2013) Threading the barrel of the RNA exosome. *Trends Biochem. Sci.*, **38**, 485–493.
19. Lubas,M., Andersen,P.R., Schein,A., Dziembowski,A., Kudla,G. and Jensen,T.H. (2015) The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Rep.*, **10**, 178–192.
20. Wu,G., Schmid,M., Rib,L., Polak,P., Meola,N., Sandelin,A. and Jensen,T.H. (2020) A two-layered targeting mechanism underlies nuclear RNA sorting by the human exosome. *Cell Rep.*, **30**, 2387–2401.
21. Wu,M., Karadoulama,E., Lloret-Llinares,M., Rouviere,J.O., Vaagensø,C.S., Moravec,M., Li,B., Wang,J., Wu,G., Gockert,M. *et al.* (2020) The RNA exosome shapes the expression of key protein-coding genes. *Nucleic Acids Res.*, **48**, 8509–8528.
22. Ogami,K., Richard,P., Chen,Y., Hoque,M., Li,W., Moresco,J.J., Yates,J.R., Tian,B. and Manley,J.L. (2017) An Mtr4/ZFC3H1 complex facilitates turnover of unstable nuclear RNAs to prevent their cytoplasmic transport and global translational repression. *Genes Dev.*, **31**, 1257–1271.
23. Almada,A.E., Wu,X., Kriz,A.J., Burge,C.B. and Sharp,P.A. (2013) Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature*, **499**, 360–363.
24. Chen,Y., Pai,A.A., Herudek,J., Lubas,M., Meola,N., Järvelin,A.I., Andersson,R., Pelechano,V., Steinmetz,L.M., Jensen,T.H. *et al.* (2016) Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.*, **48**, 984–994.
25. Ntini,E., Järvelin,A.I., Bornholdt,J., Chen,Y., Boyd,M., Jørgensen,M., Andersson,R., Hoof,I., Schein,A., Andersen,P.R. *et al.* (2013) Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.*, **20**, 923–928.
26. Mayer,A., di Iulio,J., Maleri,S., Eser,U., Vierstra,J., Reynolds,A., Sandstrom,R., Stamatoyannopoulos,J.A. and Churchman,L.S. (2015) Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, **161**, 541–554.
27. Zhu,Y., Xu,G., Yang,Y.T., Xu,Z., Chen,X., Shi,B., Xie,D., Lu,Z.J. and Wang,P. (2019) POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.*, **47**, D203–D211.
28. Lykke-Andersen,S., Žumer,K., Molska,E.Š., Rouvière,J.O., Wu,G., Demel,C., Schwalb,B., Schmid,M., Cramer,P. and Jensen,T.H. (2021) Integrator is a genome-wide attenuator of non-productive transcription. *Mol. Cell*, **81**, 514–529.
29. Liao,Y., Smyth,G.K. and Shi,W. (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
30. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
31. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
32. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.
33. Bogard,N., Linder,J., Rosenberg,A.B. and Seelig,G. (2019) A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, **178**, 91–106.
34. Kursa,M.B. and Rudnicki,W.R. (2010) Feature selection with the Boruta package. *J. Stat. Softw.*, **36**, 1–13.
35. Wickham,H. (2009) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer, Berlin.

36. Karlić,R., Chung,H.-R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 2926–2931.
37. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
38. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
39. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
40. Andersson,R. and Sandelin,A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, **21**, 71–87.
41. Bol,G.M., Xie,M. and Raman,V. (2015) DDX3, a potential target for cancer treatment. *Mol. Cancer*, **14**, 188.
42. Chiu,A.C., Suzuki,H.I., Wu,X., Mahat,D.B., Kriz,A.J. and Sharp,P.A. (2018) Transcriptional pause sites delineate stable nucleosome-associated premature polyadenylation suppressed by U1 snRNP. *Mol. Cell*, **69**, 648–663.
43. Wagner,E.J. and Carpenter,P.B. (2012) Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.*, **13**, 115–126.
44. Huff,J.T., Plocik,A.M., Guthrie,C. and Yamamoto,K.R. (2010) Reciprocal intronic and exonic histone modification regions in humans. *Nat. Struct. Mol. Biol.*, **17**, 1495–1499.
45. Zhang,S., Aibara,S., Vos,S.M., Agafonov,D.E., Lührmann,R. and Cramer,P. (2021) Structure of a transcribing RNA polymerase II–U1 snRNP complex. *Science*, **371**, 305–309.
46. Kaida,D., Berg,M.G., Younis,I., Kasim,M., Singh,L.N., Wan,L. and Dreyfuss,G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664–668.
47. Bresson,S. and Tollervey,D. (2018) Surveillance-ready transcription: nuclear RNA decay as a default fate. *Open Biol.*, **8**, 170270.