PharmaSUG 2026 - Paper 166

Quantifying Expression Divergence to Identify Candidate RNA-Binding Proteins Modulating Nonsense-Mediated Decay Across Human Tissues Using t-SNE Embedding Analysis

Liangwei He, University of Southern California

ABSTRACT

Nonsense-mediated mRNA decay (NMD) is a post-transcriptional regulatory mechanism that degrades transcripts containing premature termination codons (PTCs). To prioritize candidate RNA-binding proteins (RBPs) as potential regulators of NMD, we developed a divergence-based framework using GTEx transcriptomic data. Within each tissue, samples were embedded in a t-SNE space and quantified by calculating Euclidean distances between different expression groups. We summarized these results to a global Ratio-Score and compared with a correlation baseline. Both methods capture overlapping, but different NMD-related expression patterns. For example, the Ratio-Score metric prioritized a smaller subset of canonical NMD factors, which performed better for curated NMD-associated regulators. Meanwhile, our framework showed that strong separation can sometimes reflect tissue identity rather than NMD regulation. Overall, these results support a rigorous framework for prioritizing candidate RBPs as potential NMD regulators and highlight some limitations of divergence-based rankings.

INTRODUCTION

Nonsense-mediated mRNA decay (NMD) was originally discovered as an mRNA surveillance system, whose primary role is to detect and eliminate aberrant transcripts generated by mutations or splicing errors, thereby protecting cells from producing potentially toxic proteins (Kurosaki et al. 2016). Later, studies showed that NMD is a key post-transcriptional regulatory mechanism that detects and degrades natural transcripts with premature termination codons(PTCs) following certain rules(e.g., the 50-nt rule) (Lindeboom et al. 2020; Kurosaki et al. 2016), thereby fine-tuning transcriptome (Popp and Maquat 2019) (Figure 1).

During this process, RNA-binding proteins (RBPs) are important factors because they influence splicing, transcript processing, and translation that together determine whether transcripts become susceptible to NMD mechanism (Kurosaki et al. 2016; Popp and Maquat 2019). However, identifying candidate RBPs from high-dimensional transcriptomic data can be difficult. Basic correlation analyses can capture some co-variation effects, but they often miss context-dependent structure and may not distinguish tissue-driven expression patterns from regulatory stratification related to NMD.

To address this problem, we developed a divergence-based framework that uses GTEx transcriptomic data to prioritize candidate RBPs associated with NMD-related transcriptomic structure across human tissues (Sun and Chen 2023). Within each tissue, we embedded samples using NMD-related transcript expression and quantified the separation of different expression groups(e,g. high vs. low) for each RBP. We then compared the outcome ratio with a correlation-based baseline and evaluated how different cross-tissue summary choices affected global prioritization. Rather than claiming definitive identification of NMD regulators, our study aims to provide a computational approach for ranking candidate RBPs and for examining strengths and limitations of our divergence ranking strategy.

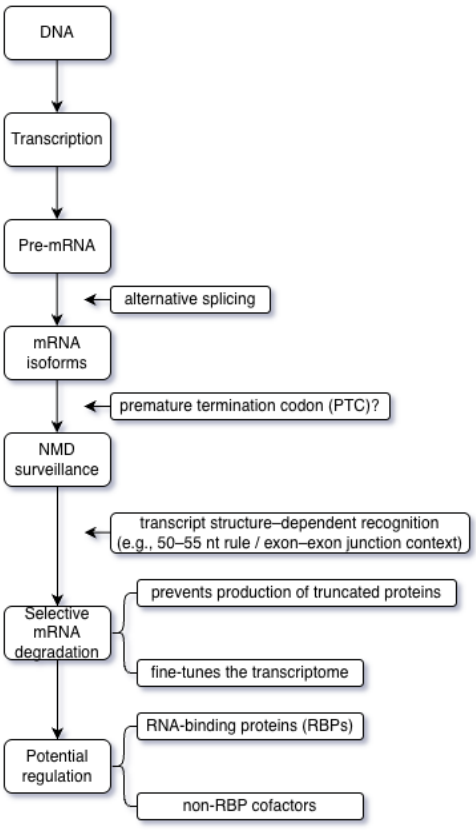

Figure 1. Overview of NMD Process

METHODS

Data Sources

We integrated several public datasets, including transcript annotation and transcriptomic resources, for the overall analysis. Human transcript annotation was based on GENCODE release v46 for GRCh38, downloaded from the GENCODE portal on October 8, 2024. This annotation was used to define NMD-related transcripts and to derive NMD target gene sets. Gene expressions values were obtained from the GTEx v10 gene TPM matrix (GTEx_Analysis_v10_RNASeQCv2.4.2_gene_tpm.gct) from the GTEx portal on December 10, 2024 (The GTEx Consortium 2020). Sample-to-tissue mappings were obtained from the GTEx v8 sample attributes table (GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt), downloaded on August 6, 2024 (The GTEx Consortium 2020). The initial reference set of 6022 RBPs was derived from the RBP2GO resource (Caudron-Herger et al. 2021).

Definition of NMD-Targeted Transcripts

NMD-targeted transcripts were defined using GENCODE v46 annotation for GRCh38. We first identified transcripts annotated as *nonsense_mediated_decay* and used this set as the primary representation of NMD targets to construct the NMD-related expression space. We then further collapsed the NMD annotations to their corresponding parent genes for downstream benchmark.

Since a single gene can produce both NMD-targeted and non-NMD protein-coding isoforms, we should interpret the results as a set of genes with at least one annotated NMD-related transcript rather than a set of genes exclusively governed by NMD. This distinction is important for interpreting downstream candidate prioritization results, especially in tissues with heterogeneous isoform usage.

Downstream Embedding and Divergence Metric

To quantify NMD-related transcriptomic structure, we analyzed each tissue separately. For a given tissue, we extracted the expression matrix of NMD-targeted transcripts across samples and used it to generate a two-dimensional t-SNE embedding. Each coordinate point corresponds to one sample, and we used a color gradient to reflect the expression magnitude (Kobak and Berens 2019; Wattenberg et al. 2016).

Based on the expression values of each RBP within each tissue, we defined the upper quartile as the high-expression group and the lower quartile group as the low-expression group. Using the t-SNE coordinates, we then measured three quantities: (1) the mean pairwise distance among samples within the high-expression group, denoted $D_{WithinHigh}$; (2) the mean pairwise distance among samples within the low-expression group, denoted $D_{WithinLow}$; and (3) the Euclidean distance between the centroids of the two groups, denoted $D_{Between}$. If we use μ_H and μ_L to represent the centroids of the high- and low-expression groups in the embedding space, we shall have:

$$D_{Between} = \|\mu_H - \mu_L\|_2 = \sqrt{(\mu_{H1} - \mu_{L1})^2 + (\mu_{H2} - \mu_{L2})^2} \quad (1)$$

We then combined these quantities into a Ratio-Score defined as:

$$R = \frac{D_{Between}}{Avg(D_{WithinHigh}, D_{WithinLow})} \quad (2)$$

This score measures the separation between two groups relative to their internal dispersion. Intuitively, we may consider that the numerator $D_{Between}$ captures how far apart the centers of the two groups are and the denominator $Avg(D_{WithinHigh}, D_{WithinLow})$ accounts for how spread out the groups are within themselves. Thus, larger R values may indicate a clearer separation between high- and low-expression samples, while smaller values indicate weaker separation or greater overlap. In this way, the Ratio-Score captures transcriptomic divergence associated with RBP expression stratification. Besides, because the

Ratio-Score was calculated separately for every RBP in each tissue, the same RBP could show various separations in different tissues. Therefore, we treated each pair of *RBP-Tissue* as the basic analytical unit and then evaluated these scores for downstream ranking.

Correlation Baseline and Ratio-Score Summary Across Tissues

For each RBP, we calculated its expression values across samples and measured its Pearson correlation with the expression of NMD-target transcripts:

$$\tilde{r} = \{r_1, r_2, r_3, \dots, r_M\} \quad (3)$$

To obtain one global summary of how strongly each RBP varies with NMD-targeted transcripts across samples, we averaged these correlations and obtained a single summary value, denoted as C_{mean}. We used it as a baseline ranking criterion for comparison with the divergence framework, as C_{mean} summarizes the average linear co-variation across the full sample set.

Meanwhile, we considered several summary measures for the Ratio-Score, including the (1) maximum observed ratio across tissues, R_{Max}; (2) the mean ratio across tissues, R_{Mean}; (3) and the median ratio across tissues, R_{Median}.

We also recorded the number of tissues in which an RBP satisfied both the statistical significance criterion and the effect-size filter; this quantity was denoted as *Count* and used to capture recurrence across tissues. Specifically, for each *RBP–tissue* pair, we performed two Welch's t-tests by comparing within-group distances from the high-expression group and the low-expression group, against the corresponding between-group distances. A pair was retained for recurrence counting if (1) either test passed the Bonferroni-adjusted threshold ($\alpha = \frac{0.05}{6022}$) with $p_{test} < \alpha$; (2) and the Ratio-Score exceeded 1.5. In this way, *Count* reflects how often an RBP shows statistically strong separation across tissues.

These summary measures served different purposes. R_{Max} was sensitive to extreme separation observed in a single tissue, whereas R_{Median} was more robust to reflect the strength of separation across tissues. *Count* provided some complementary information by highlighting RBPs whose separation signal appeared repeatedly across tissues. Therefore, in the final candidate prioritization steps, we emphasized R_{Median} as the primary global Ratio-Score and used *Count* as an additional measure of recurrence.

Benchmark and Evaluation

To assess whether our divergence framework recovered biologically significant signals, we compared ranked candidate lists against two benchmark sets. The first set contained canonical NMD factors and core exon-junction-complex-related components. The second was a curated set of NMD-associated regulators, including genes implicated in RNA surveillance, transcript processing, or NMD-related regulation. Note that these benchmark sets were used only for comparative evaluation and were not treated as complete ground truth.

For each summary measure, RBPs were ranked from highest to lowest and compared against the benchmark sets at different cutoffs. To check if a given ranking criterion selectively prioritized genes with prior biological relevance, we recorded the number of benchmark genes recovered among the top-ranked candidates and compared the observed overlap with the overlap under random selection. In addition, we constructed candidate tables for downstream interpretation by combining the separation strength, recurrence across tissues, and comparison with the correlation baseline. We inspected candidate lists with supporting information such as the tissue in which the strongest separation occurred, the number of significant tissues, and the corresponding baseline C_{Mean} score.

RESULTS

Global Comparison with The Correlation Baseline

We first compared the divergence framework with the correlation baseline by examining the relationship between filtered R_{mean} and transcript-level C_{mean} across RBPs. As shown in Figure 2, the two measures were positively associated, with a moderate correlation ($r = 0.443$). This indicates that RBPs with stronger separation in the NMD-related embedding tended to show stronger co-variation with NMD-target transcripts as well. However, substantial dispersion remained across the ranking space, suggesting that the divergence framework and the correlation baseline capture overlapping but different aspects of NMD-related transcriptomic structure.

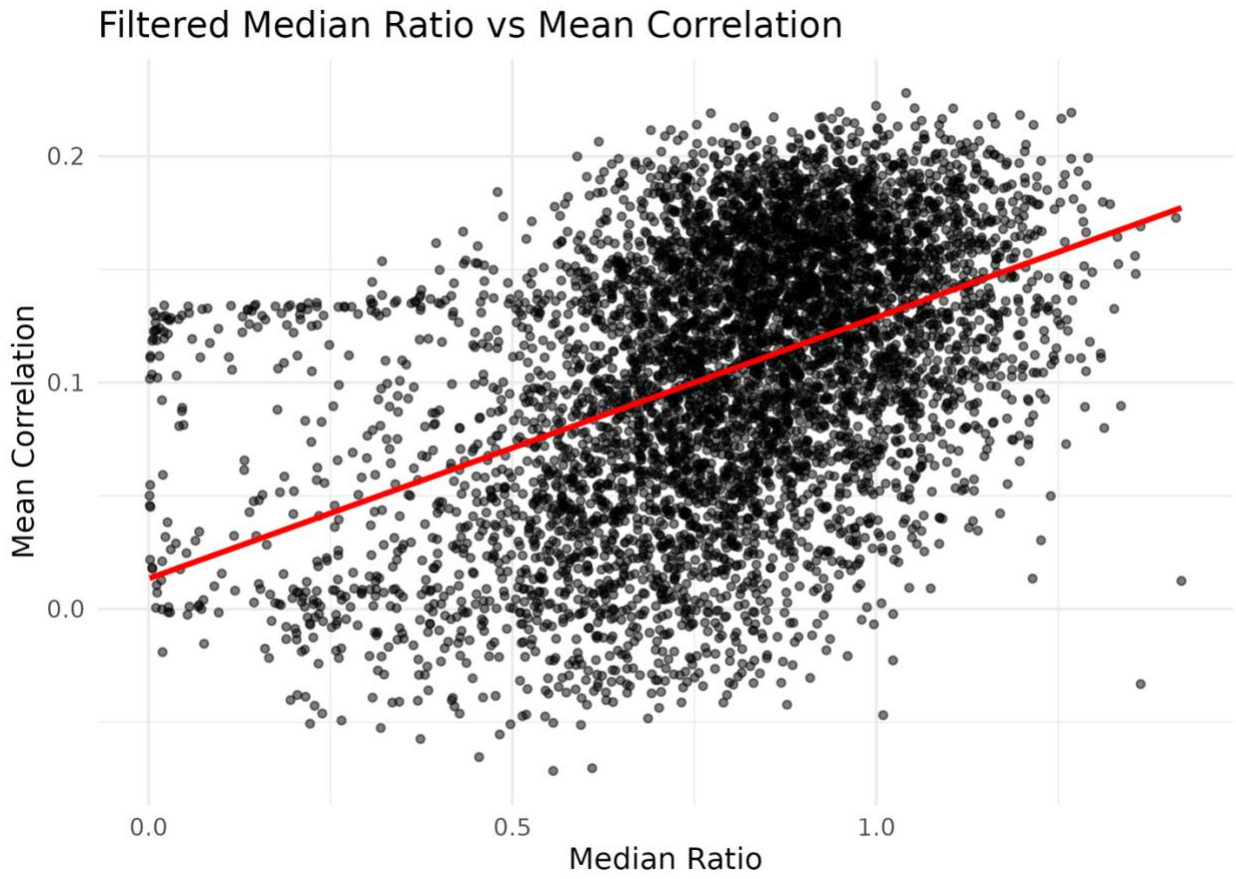

Figure 2. Filtered Median Ratio vs Mean Correlation

To get a more direct look at this overlap, we compared top-ranked candidate sets from the divergence framework with those obtained from the correlation baseline. Figure 3 shows that candidates prioritized by filtered median ratio were consistently enriched relative to random expectation, with observed-to-expected overlap ratios of 4.68, 2.34, 2.49, 2.13, and 1.70 at the top 50, 100, 200, 500, and 1000 candidates, respectively. A similar pattern was observed for *Count*, with stronger enrichment at smaller cutoffs: 6.81, 5.11, 3.12, 1.75, and 1.61 at the same thresholds. These results show that the divergence framework is not simply reproducing the correlation baseline. Instead, both approaches recover some shared candidates, while the divergence framework introduces additional structure. For this reason, we treated filtered median ratio as the main global summary of separation strength and recurrence count as a complementary measure for highlighting candidates that appeared repeatedly across tissues.

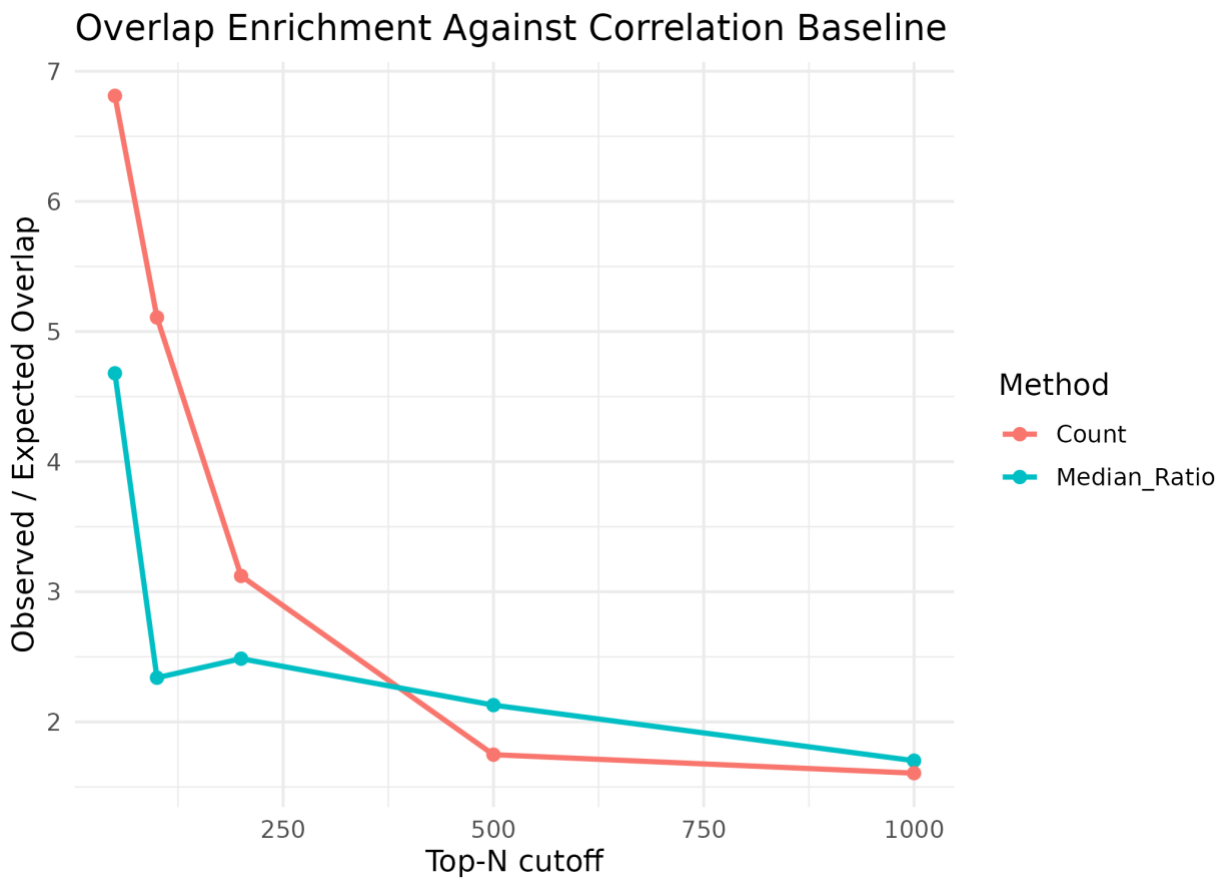

Figure 3. Overlap Enrichment vs Correlation Baseline

Candidate Prioritization and Evaluation Against Known NMD Factors

Figure 4 summarizes the highest-ranked candidates after excluding mitochondrial and ribosomal genes. Several candidates, including **RNF10, SARNP, FKBP8, MORF4L1, DNAJC8, NOL7, RTF2, USP39,** and **NKAP**, combined relatively high filtered median ratio values with recurrence across multiple tissues. This pattern suggests that the divergence framework does not simply highlight isolated outliers but also retains candidates whose separation signal appeared repeatedly across tissue contexts.

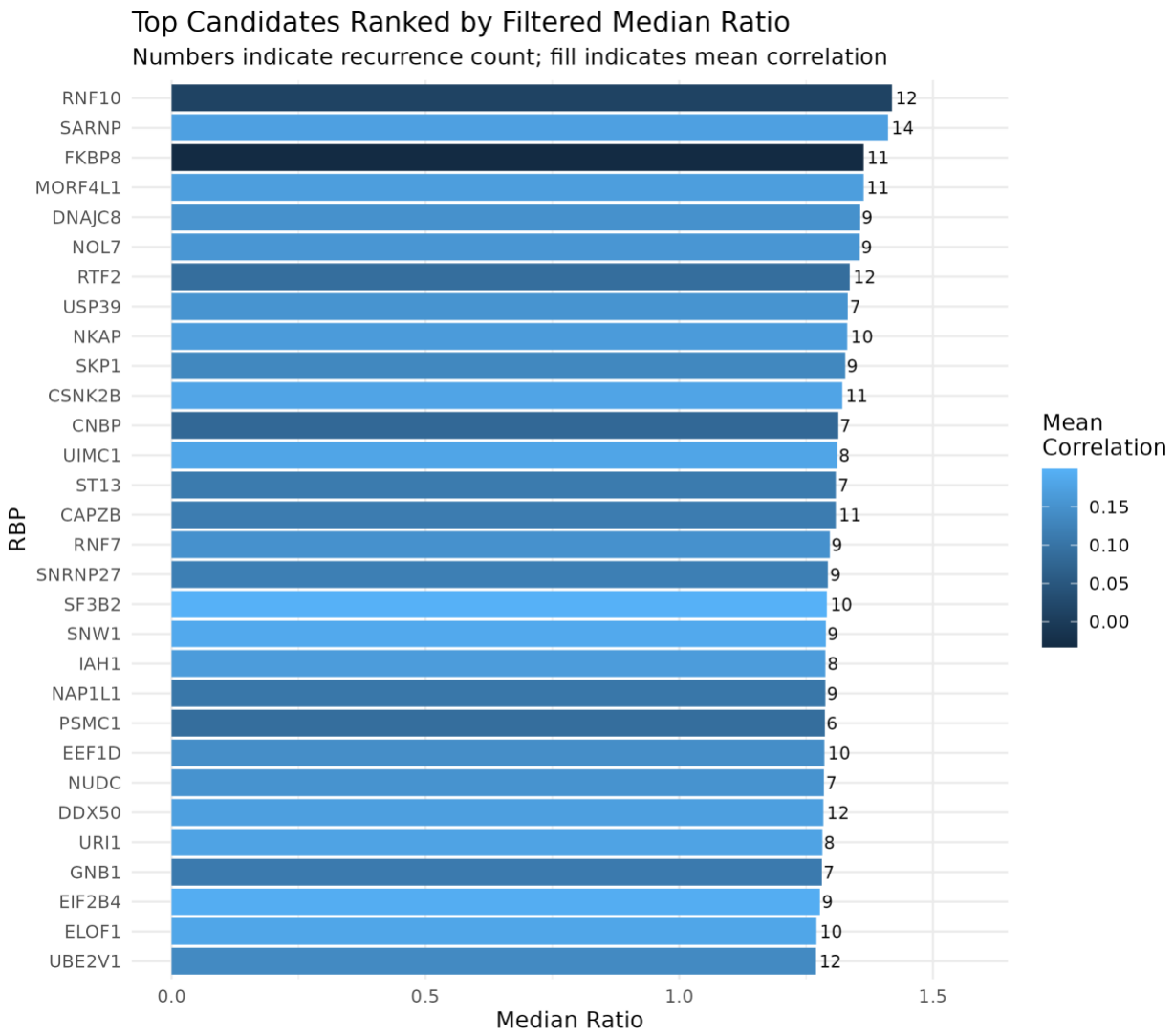

Figure 4. Top Candidate RBPs Ranked by Filtered Median Ratio

To evaluate biological plausibility, we compared ranked candidate lists against two benchmark sets: (1) a strict core set of canonical NMD factors and exon-junction-complex-related components, (2) and a broader curated set of NMD-associated regulators. For example in Table 1, filtered median ratio recovered no strict core factors within the top 200 candidates and only one (UPF2) within the top 500 candidates, whereas recurrence count recovered UPF2 within the top 100 and added RBM8Aand SMG6 by the top 500. In contrast, we found stronger recovery in the broader associated set. Filtered median ratio recovered STAU1 and RNPS1 within the top 100 and added NBAS by the top 200, while recurrence count recovered the same three genes by the top 200.

Table 1. Recovery of Known NMD-Related Genes Among Top-Ranked Candidates

Metric	Benchmark Set	Top 100	Top 200	Top 500
Filtered R_{Median}	Strict Core	0	0	1 (UPF2)
Count	Strict Core	1 (UPF2)	2 (UPF2, RBM8A)	3 (UPF2, RBM8A, SMG6)
Filtered R_{Median}	Associated	2 (STAU1, RNPS1)	3 (STAU1, RNPS1, NBAS)	3 (STAU1, RNPS1, NBAS)
Count	Associated	1 (STAU1)	3 (STAU1, NBAS, RNPS1)	3 (STAU1, NBAS, RNPS1)

These benchmark results indicate that the prioritization framework captures genes associated with broader RNA surveillance and post-transcriptional regulation than a small set of strict canonical NMD factors. This pattern is also consistent with the foundation of our divergence framework. We initially intend to rank RBPs according to transcriptomic separation structure rather than direct relevance in the core NMD machinery. At the same time, the benchmark comparison also showed that recurrence count contributed useful complementary information. In the overlap analysis against the correlation baseline, recurrence count showed stronger enrichment than filtered median ratio at smaller cutoffs.

Taken together, the highest-ranked candidates therefore represent a mixture of broadly recurrent RBPs, genes with moderate correspondence to the correlation baseline, and a smaller subset of previously implicated RNA surveillance factors.

Recurrence Across Tissues

We next examined how often significant RBPs recurred across tissues to avoid bias from potential outliers. In our framework, recurrence was summarized by *Count*, defined as the number of tissues in which an RBP satisfied both the statistical significance criterion and the effect-size filter. This quantity was used to distinguish recurring candidates from RBPs whose strong separation arose only in a single tissue set.

To visualize this recurrence pattern, we summarized significant RBPs across tissues using an UpSet plot as shown in Figure 5. This representation shows how significant RBPs are shared across tissues and how many are unique to individual tissues. The resulting pattern indicates that many significant RBPs were not restricted to a particular tissue but instead appeared repeatedly in overlapping tissue sets.

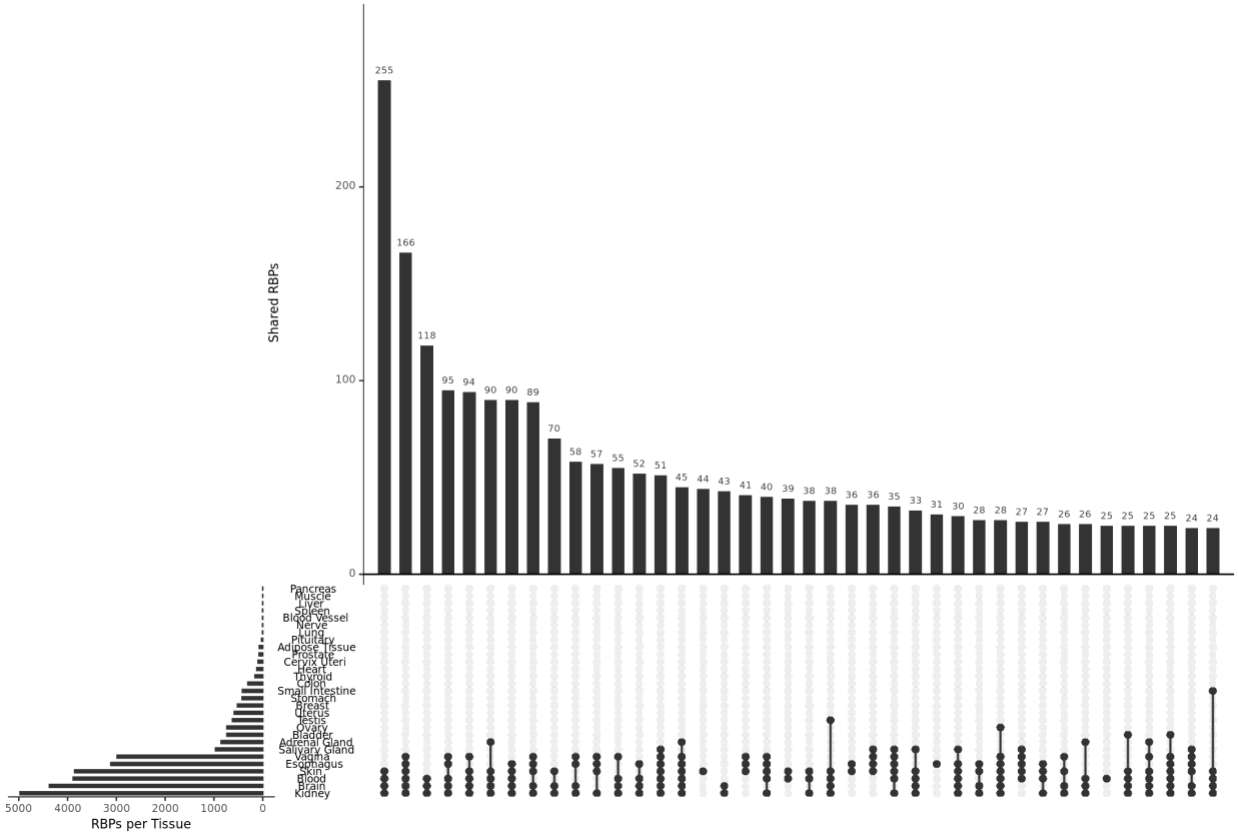

Figure 5. Shared Significant RBPs Across Tissues

Case Studies from Brain and Skin

To further illustrate how the divergence framework can capture distinct separation patterns, we examined two representative tissue-specific examples. We selected these cases primarily as visual examples showing how high ratio scores can arise under different biological interpretations.

In brain, CWC27 showed a clear separation between high- and low-expression groups in the t-SNE embedding (Figure 6). The corresponding ratio score was 3.26, with both group-to-between comparisons reaching the significance threshold. In the global candidate ranking, CWC27 also showed a relatively high filtered median ratio and a positive correlation baseline score. Taken together, these features make CWC27 a plausible example of a candidate whose separation pattern is consistent with broader transcriptomic structure.

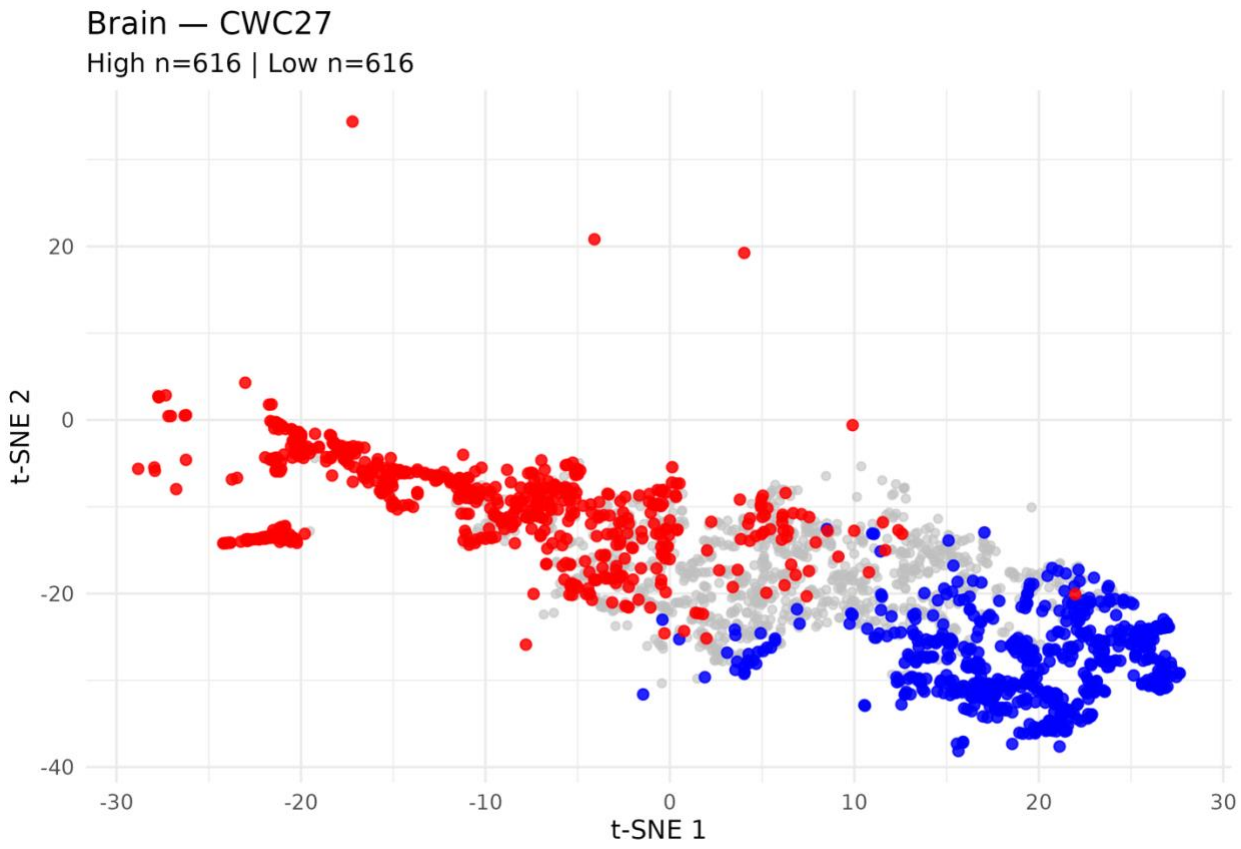

Figure 6. Separation of High- and Low- Expression Samples for CWC27 in Brain

In contrast, KRT10 in skin showed extremely strong separation in the embedding space with Raio-Score = 11.85(Figure 7). The corresponding global filtered median ratio was lower, recurrence across tissues was limited, and its mean correlation with NMD-target transcripts was nearly zero. This example suggests that very high separation within a single tissue can sometimes reflect broader tissue influence rather than direct relevance to NMD regulation.

Together, these examples show that the divergence framework is useful not only for ranking candidates, but also for distinguishing between biologically plausible signals and potentially tissue-restricted patterns. This distinction helps explain why filtered median ratio and recurrence count were retained as the main global prioritization measures, whereas maximum ratio alone was not used as the primary ranking criterion.

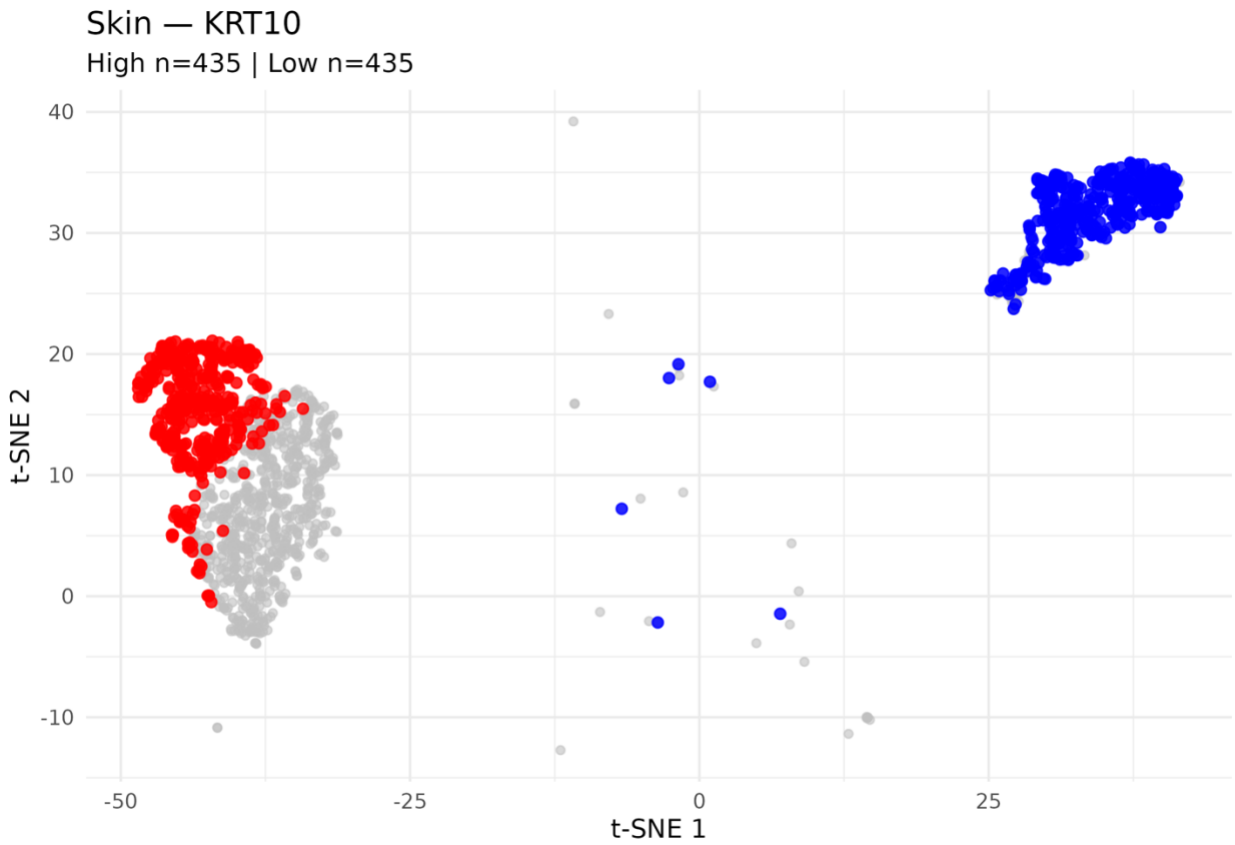

Figure 7. Separation of High- and Low- Expression Samples for KRT10 in Skin

DISCUSSION

In this study, we developed a divergence framework to prioritize candidate RBPs associated with NMD-related transcriptomic structure across human tissues. Rather than relying only on average linear correlation, our framework tries to answer whether samples with high and low expression of a given RBP occupy different regions in an NMD embedding space. Filtered median ratio remained moderately aligned with mean correlation, while top-ranked overlap was consistently enriched above random expectation. Together, these results suggest that divergence and correlation capture partly shared but still distinct features of NMD-related transcriptomic organization.

One of the main findings is that the way we summarize the Ratio-Scores may largely affect biological interpretation. For example, maximum ratio was very sensitive to extreme tissue-specific separation and often prioritized genes with little limited relevance to NMD regulation. On the other hand, filtered median ratio provided a more stable summary of separation across tissues. The recurrence count also added useful information about how often an RBP showed statistically strong separation across tissues. One counter example of relying on a single extreme score is the contrast between the brain CWC27 embedding and the skin KRT10. We should acknowledge that strong separation may not necessarily be relevant to NMD regulation.

Our next benchmark comparison further clarified what the functionality of our framework. Recovery of a strict set of canonical NMD factors was limited but stronger for a broader curated set of NMD-associated regulators. This pattern suggests that the divergence framework didn't perform well for recovering the small core NMD machinery. Instead, it appears more sensitive to RBPs associated with broader transcriptomic regulation, RNA surveillance, or post-transcriptional processes. This distinction is still important as a high-ranking candidate in this framework should not automatically be interpreted as a direct NMD factor. Again, our ranking system should be viewed as a prioritization method for future study.

Several limitations are present as well. First, the framework depends on a two-dimensional embedding space and is therefore influenced by the geometry imposed by t-SNE (Kobak and Berens 2019; Wattenberg et al. 2016). For example, two sample groups that appear well separated in one embedding may appear less distinct if the local arrangement of points changes under different hyperparameters or other embedding configurations. As a result, the Ratio-Score should be interpreted as a summary of separation in the chosen embedding space rather than direct evidence to quantify NMD regulation. Second, the high- and low-expression grouping strategy is based on quartile thresholds, which may oversimplify how NMD transcriptomic structure changes across the full range of RBP expression. Finally, current benchmark sets for NMD regulation remain incomplete (Popp and Maquat 2019).

CONCLUSION

We presented a divergence framework for prioritizing candidate RNA-binding proteins associated with NMD-related transcriptomic structure across human tissues. Rather than identifying direct NMD regulators, our framework provides a practical way to organize candidates according to transcriptomic separation and recurrence across tissues. Hence, we should view this method as a screening strategy that can support future transcript-level and exploratory follow-up studies.

REFERENCES

Caudron-Herger, M., R. E. Jansen, E. Wassmer, and S. Diederichs. 2021. “RBP2GO: A Comprehensive Pan-Species Database on RNA-Binding Proteins, Their Interactions and Functions.” *Nucleic Acids Research* 49(D1): D425–D436.

Frankish, A., M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47(D1): D766–D773.

Kobak, D., and P. Berens. 2019. “The Art of Using t-SNE for Single-Cell Transcriptomics.” *Nature Communications* 10(1): 5416.

Kurosaki, T., M. W. Popp, and L. E. Maquat. 2016. “Mechanism and Regulation of the Nonsense-Mediated Decay Pathway.” *Nucleic Acids Research* 44(4): 1483–1495.

Lindeboom, R. G. H., F. Supek, and B. Lehner. 2020. “To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases.” *Trends in Genetics* 36(10): 755–767.

Popp, M. W., and L. E. Maquat. 2019. “Quality and Quantity Control of Gene Expression by Nonsense-Mediated mRNA Decay.” *Nature Reviews Molecular Cell Biology* 20: 406–420.

Sun, B., and L. Chen. 2023. “Mapping Genetic Variants for Nonsense-Mediated mRNA Decay Regulation Across Human Tissues.” *Genome Biology* 24(1): 164.

The GTEx Consortium. 2020. “The GTEx Consortium Atlas of Genetic Regulatory Effects Across Human Tissues.” *Science* 369(6509): 1318–1330.

Wattenberg, M., F. Viégas, and I. Johnson. 2016. “How to Use t-SNE Effectively.” *Distill* 1(10): e2.