Supplementary Information

RNAmotifs 2.0: a framework for inferring multivalent RNA motifs and their candidate regulators of splicing

Mariachiara Grieco[1,2,§], Tommaso Becchi[1,3,§], Gabriele Boscagli[3,§], Francesca Priante[2], Livia Caizzi[2,4], Uberto Pozzoli[5], Matteo Cereda[1,2,3,*]

1. Department of Oncology and Hemato-Oncology, University of Milan, 20142 Milan, Italy
2. Italian Institute for Genomic Medicine, c/o IRCCS, Str. Prov.le 142, km 3.95, 10060 Candiolo (TO), Italy
3. IFOM ETS - The AIRC Institute of Molecular Oncology, Milan, Italy
4. Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy
5. Bioinformatics Lab, Scientific Institute IRCCS E. MEDEA, Bosisio Parini, Italy

§ These authors contributed equally to this work
* Corresponding author

Correspondence to: Prof Matteo Cereda (matteo.cereda@ifom.eu)

Supplementary Notes

1. RNAmotifs 2.0: software improvements and new analytical capabilities

This section describes the software improvements and new analytical capabilities introduced in RNAmotifs 2.0, the upgraded tool presented in the main text, whose Motifs and Regulators of splicing (MaRs) module performs the multivalent RNA motif (MRM)-RNA-binding protein (RBP) association analysis. While the positional tetramer enrichment algorithm remains faithful to the original implementation [1], producing byte-identical output when run with matching parameters, the software has been rebuilt from the ground up. All computationally intensive steps (tetramer search, bootstrap FDR, tetramer selection) are implemented as self-contained C++17 binaries with OpenMP parallelisation, coordinated by a Python 3 orchestrator, and require no external C++ library dependencies. The original RNAmotifs was implemented as a Bash orchestrator (RNAmotifs.sh) that coordinated a Python 2 motif search module (m3_light), a C++ positional mapping library (GeCo++ [20]), and single-threaded R scripts for bootstrap FDR estimation and visualisation. The version 2.0 pipeline comprises six compiled C++17 binaries coordinated by a Python 3 command-line interface:

Binary	**Function**
rnamotifs_search	Tetramer scan across 512 motifs (256 ACGT + 256 IUPAC), clustering by sliding window, position mapping. OpenMP-parallel across all motifs.
rnamotifs_selection	Two-sided Fisher exact test at each position, enrichment classification by region, tetramer ranking by combined Fisher score, sequence-similarity clustering.
rnamotifs_bootstrap	Empirical p-value estimation via label permutation and bootstrap resampling. OpenMP-parallel across permutations.
rnamotifs_mars_score	Compute eCLIP peak overlap using fast hash-set lookups, then derive the Cosine Similarity (CS) and Signal Recovery Rate (SRR).
rnamotifs_core	Shared library providing coordinate mapping (genomic ↔ exon-relative), region boundary computation (R1–R3), adaptive clamping to exon limits, and BED record parsing/writing.
rnamotifs_tetramer	Tetramer index generation (256 + 256 IUPAC).

All binaries are compiled with CMake 3.10+ and a C++17-compliant compiler (GCC ≥ 7).

1.1. Key algorithmic optimisations

- *Indexed BED lookup*. The original RNAmotifs Python search used a linear scan over all BED records per exon. The version 2.0 C++ implementation sorts BED records by chromosome and position, then uses O(log N) binary search to locate the relevant records for each exon.
- *Dense arrays with prefix sums*. The original RNAmotifs sliding-window cluster detection iterated over hash-map entries within the window at each position (O(hw) per position). The version 2.0 implementation uses dense position arrays with prefix sums, enabling O(1) window queries.
- *Chromosome preloading*. Full chromosome sequences are loaded into memory at startup (one string per chromosome), replacing per-region file I/O.
- *OpenMP parallelism*. The tetramer search parallelises across all 512 motifs, and the bootstrap FDR parallelises across permutation iterations. Both achieve near-linear scaling.

1.2. Performance benchmarks

The core motif-discovery routine of RNAmotifs was reimplemented in C++17 with OpenMP parallelisation (RNAmotifs 2.0), replacing the original mixed-language pipeline (Python m3_light tetramer search, R bootstrap-FDR.R, and a Python/gMotifs/R counting and selection workflow). We benchmarked both implementations on the NOVA dataset (4,368 exons, mm9 assembly) on a single workstation (Intel Core i7-8700, 6 physical cores / 12 threads, 3.20 GHz; 64 GB RAM; Ubuntu 24.04 LTS). All values are wall-clock times, and peak resident memory remained below 4 GB throughout. The original RNAmotifs pipeline has no parallel path and runs single-threaded end to end. RNAmotifs 2.0 was timed both single-threaded and with 10 OpenMP worker threads. Single-threaded, RNAmotifs 2.0 reduced tetramer search from ~50 min to ~12 min (4.2x) and bootstrap-based FDR estimation from ~14 min to ~32 s (26x), shortening the full pipeline from ~70 min to ~13 min (5.5x). With 10 threads, search fell to ~2 min (22.9x) and bootstrap FDR to ~14 s (60x), for an end-to-end runtime of ~2.6 min (~26x). Parallel scaling of the search was sublinear beyond the six physical cores, consistent with a memory-bandwidth-bound workload. The search speed-up carried a modest memory cost, as RNAmotifs 2.0 allocates dense per-chromosome arrays (~2.6 GB) in place of the sparse Python dictionaries of the original implementation (~0.17 GB); total peak resident memory nonetheless stayed below 4 GB.

Task	RNAmotifs	RNAmotifs 2.0 (1 thread)	RNAmotifs 2.0 (10 thread)	Fold-change (1-thread)	Fold-change (10 threads)
Tetramer search	~50 min	~12 min	~2 min	4.2×	22.9×
Bootstrap FDR	~14 min	~32 sec	~14 sec	26×	60×
Full pipeline	~70 min	~13 min	~2.6 min	5.5×	~26×

1.3. Direct rMATS integration

To streamline analysis of standard differential splicing output, RNAmotifs 2.0 provides direct import of rMATS [21] results via the --from-rmats flag. Exons are automatically classified into regulatory categories according to the magnitude of their inclusion-level change and statistical significance. An exon is labelled enhanced or silenced when its absolute inclusion-level difference (|IncLevelDifference|) and false discovery rate (FDR) satisfy user-defined thresholds (--rmats-incl and --rmats-fdr, respectively), with the sign of IncLevelDifference determining the direction of regulation. Constitutive control exons are drawn from events that fall below these thresholds and are subsampled to a user-specified number (--rmats-constit, capped at --rmats-max-constit) to yield a balanced background set. Event coordinates are then converted to the internal RNAmotifs coordinate representation for downstream motif discovery. All classification and sampling parameters are user-configurable, allowing the import behaviour to be tailored to the stringency and class balance required for a given dataset.

1.4. Expanded genome assembly support

The original RNAmotifs supported human hg19 and mouse mm9 assembly only. RNAmotifs 2.0 now supports different genome assemblies across human and mouse: the human assemblies hg19 (GRCh37) and hg38 (GRCh38), and the mouse assemblies mm9 (NCBI37), mm10 (GRCm38), and mm39 (GRCm39).

1.5. Modernised orchestration

The RNAmotifs 2.0 pipeline is coordinated through a modernised Python 3 command-line interface comprising three executables. The main orchestrator, *rnamotifs*, coordinates the core pipeline from input processing through to the generation of RNA splicing maps. The *rnamotifs-mars* orchestrator implements the MaRs module, managing both the three-phase

application-mode workflow and the multi-RBP discovery mode; it handles parameter-grid orchestration, sequential per-RBP processing, and the writing of crash-safe JSON manifests. A third utility, *rnamotifs-extract*, exports the coordinates of enriched tetramers in BED and TSV formats for downstream use. To ensure robustness over long runs, both execution modes persist their state to JSON manifest files after each completed unit of work. If a run is interrupted, re-issuing the same command resumes execution automatically from the last checkpoint rather than restarting.

2. Out-of-sample parameter selection by leave-one-RBP-out cross-validation

In the main analysis, the clustering and enrichment windows for each RBP were selected in-sample, as the combination maximising that RBP's own AUROC. To assess whether this in-sample selection might inflate the reported performance, we repeated parameter selection under leave-one-RBP-out (LORO) cross-validation across both reference panels (15 RBPs in HepG2, 13 in K562; 28 RBP-cell line combinations in total). For each held-out RBP r*, parameters were chosen to maximise the mean AUROC over the remaining panel members, and r* was then evaluated at these independently selected values, so that its parameter selection was strictly out-of-sample. All other settings matched the main analysis (AUROC computed over enriched MRMs, scored only when ≥5 MRMs are enriched at the evaluated (*n, e*)), so the in-sample AUROCs coincided with the values reported in Fig. 3 (Additional file 2: Table S3 and Additional file 1: Fig. S4).

Overall, *rank order was largely preserved out-of-sample.* The Spearman correlation between in-sample and LORO AUROC was $\rho = 0.73$ ($P = 2.2 \times 10^{-3}$) in HepG2 and $\rho = 0.92$ ($P = 8.8 \times 10^{-6}$) in K562, and RBPs scoring highly in-sample also tended to perform well when held out (for example, HNRNPK, QKI, RBFOX2 and U2AF1 in HepG2; RBM15, TARDBP and U2AF2 in K562). Selected parameters were also highly consistent across folds: 14 of 15 HepG2 folds selected (*n, e*) = (10, 200) and all 13 K562 folds selected (70, 300), the single exception being the held-out-UCHL5 fold, which selected n = 70. This consistency indicates a well-defined panel-level optimum that is stable to the removal of essentially any individual RBP.

Absolute AUROC decreased modestly under out-of-sample selection. Mean LORO AUROC was 0.732 in HepG2 (versus 0.871 in-sample; $\Delta = -0.14$) and 0.705 in K562 (versus 0.811; $\Delta = -0.11$). These panel means are lowered by two folds in which the metric was undefined: for UCHL5 in HepG2 (from 0.857 to 0) and PUS1 in K562 (from 0.417 to 0), fewer than five

MRMs were enriched at the evaluated parameters, so the AUROC was undefined and reported as 0 (a sentinel value rather than a genuine score). Excluding these two folds, mean LORO AUROC was 0.789 in HepG2 and 0.764 in K562. Overall, 15 of 28 RBPs (8/15 HepG2, 7/13 K562) retained a LORO AUROC within 0.10 of their in-sample value. UCHL5 and PUS1 are accessory factors, a spliceosome-associated deubiquitinase and a pseudouridine synthase, respectively, rather than sequence-specific splicing regulators, consistent with their sparse and poorly localised binding signal.

Crucially, *the reductions in absolute AUROC did not compromise the core regulators*. Separating each RBP's final discrimination from the magnitude of its AUROC decrease, several factors with a larger absolute drop nonetheless remained strong classifiers: PRPF8 (from 1 to 0.79 in HepG2; from 1 to 0.83 in K562), U2AF2 (from 1 to 0.84 in HepG2), HNRNPC (from 1 to 0.86) and SRSF1 (from 0.93 to 0.81 in HepG2) all retained clearly above-chance discrimination and are more accurately described as parameter-sensitive in magnitude than as generalization failures. Every core, sequence-specific regulator retained an out-of-sample AUROC well above chance (minimum 0.66, SRSF1 in K562). The only collapses to an undefined score occurred for the two accessory factors above. The remaining low performers (RBM22, from 0.46 to 0.35; SF3A3, from 0.59 to 0.48) were already near chance in-sample; their low LORO values therefore reflect weak intrinsic signal rather than a loss under out-of-sample selection. Taken together, rank order preserved in both panels (Spearman ρ = 0.73 and 0.92 in HepG2 and K562, respectively) and above-chance discrimination retained for every core regulator, these results indicate that in-sample tuning does not substantially inflate the reported performance of the sequence-specific splicing regulators, with degradation confined to a few accessory, low-signal factors.

3. Cross-cell-line transfer: robustness to mismatched RNA-seq and eCLIP sources

To assess robustness to mismatched RNA-seq and eCLIP sources, we restricted the analysis to the seven RBPs profiled in both HepG2 and K562 (HNRNPU, PRPF8, PTBP1, SF3B4, SRSF1, U2AF1, U2AF2) and re-ran the grid search in two cross-configurations: H2K (HepG2 knockdown exon sets to K562 eCLIP) and K2H (the reverse). For each configuration, the best AUROC across the 20 parameter combinations was compared with the matched-cell-line AUROC from Fig. 4C and Additional file 2: Table S3.

Cross-cell performance approximates matched performance in aggregate. Across the 14 matched-cross pairs, the mean matched AUROC was 0.885 and the mean cross-cell AUROC 0.861 (mean Δ = -0.024). Every RBP retained an informative cross-cell AUROC (minimum 0.649, HNRNPU K2H), with no catastrophic transfer failures.

Robustness is broad and rank order is largely preserved. Eleven of the 14 pairs (79%) fell within 0.10 AUROC of their matched value, and matched performance was a good predictor of cross-cell performance (Spearman ρ = 0.73, P = 2.9×10^{-3}). U2AF2 was invariant across all configurations (AUROC = 1), and PRPF8 was near-invariant, declining only in H2K, to 0.917. SF3B4 transferred with little change in both directions (+0.03 H2K, +0.04 K2H). The deviations that did occur were modest and bidirectional. The largest single loss was SRSF1 in H2K (from 0.93 to 0.7), while the largest gains were PTBP1 in K2H (from 0.81 to 0.93) and SRSF1 in K2H (from 0.77 to 0.87, Additional file 1: Fig. S4C and Additional file 2: Table S5).

Overall, the residual cell-line bias was small and approximately symmetric across transfer directions, and no RBP was reduced to an uninformative level. This limited cross-cell bias is further supported by the PC3 validation (Fig. 5), in which HNRNPK was correctly identified in an unrelated cellular context using HepG2 eCLIP. Together, these results indicate that RBP identification transfers robustly across cell lines, remaining informative even when the RNA-seq and eCLIP sources are mismatched.

4. Ablation of the association-score components

The association score multiplies a positional-concordance term (Cosine Similarity, CS), which measures agreement between MRM enrichment and eCLIP binding, by a signal-recovery rate (SRR) term, which quantifies the reproducibility of the eCLIP map (AS = CS × SRR). To separate their contributions to the RBP rankings, we recomputed the ranking at the full-grid optimum (Additional file 1: Fig. S4) for all 28 RBP-cell-line combinations under two scoring conditions, the full score (CS × SRR) and a CS-only score with SRR fixed to 1, evaluated on the CS and SRR values stored during the grid search. Because the full score is the concordance-only score scaled by SRR, the difference between conditions isolates the effect of SRR, and the full-score column reproduces the Fig. 4 AUROC by construction. Per-RBP values are reported in Additional file 2: Table S6.

Concordance alone accounted for essentially the entire ranking. Across the 28 combinations the full and CS-only scores were statistically indistinguishable in aggregate (mean AUROC 0.84 and 0.85, the marginal difference favouring the CS-only score), and this held within each cell line (HepG2, 0.87 versus 0.89; K562, 0.81 versus 0.81). Scaling concordance by

SRR changed the AUROC by more than 0.1 for only two RBPs (NCBP2 and SF3A3) and by less than 0.02 for 21 of 28 (75%). The AUROC distributions of the two conditions did not differ (P = 0.61; Additional file 1: Fig. S4D), and, treating each combination as a paired observation within a 0.01-AUROC indifference band, SRR raised the AUROC for 5 combinations, lowered it for 6 and left 17 unchanged, indicating no directional bias (one-sided exact binomial P = 0.73). Because SRR is a single value per RBP, replacing the CS-only score with the full score multiplies the true RBP's scores by its own SRR and each competing RBP's scores by theirs. This leaves the ordering of MRMs within an RBP unchanged and moves a per-RBP AUROC only by rescaling that RBP relative to the rest of the panel.

This neutrality does not imply that the underlying eCLIP is clean. CLIP and eCLIP maps are intrinsically noisy and heterogeneous, carrying substantial background from abundant and non-specifically co-purified transcripts and differing in enrichment and between-replicate reproducibility even within standardized ENCODE data [6]; SRR was introduced precisely to down-weight the least reproducible maps. Two features of the benchmark explain the null result. First, the benchmark uses the ENCODE eCLIP maps available for these RBPs and was not constructed to span or stress-test eCLIP quality, so it does not probe the low-quality regime in which SRR is intended to act. Second, in the cases where SRR shifted the ranking it lowered the AUROC about as often as it raised it, and its two largest effects were reductions for RBPs that concordance alone already ranked well. In particular, applying SRR reduced the AUROC of NCBP2 from 0.93 to 0.79 and of SF3A3 from 0.71 to 0.59. We therefore make no claim that SRR improves ranking on noisier data, and the present benchmark cannot establish such a benefit. We nonetheless retain AS = CS × SRR as a default-on, optional safeguard for eCLIP of unknown or lower quality, exposed through --score-mode (cs-only sets SRR to 1).

5. Guided hyperparameter search: grid search versus Bayesian optimisation

To test whether a continuous optimisation strategy could improve on the discrete grid, we benchmarked grid search against Bayesian optimisation (BO) across the full reference panel (28 RBP-cell line combinations; BO returned a defined score for 27, the exception being RBM15 in K562). BO used a Gaussian-process surrogate with a Matérn-5/2 kernel that gives *n* and *e* separate length-scales (the two axes differ markedly in range and units; with only two dimensions this per-dimension length-scaling, also termed as automatic relevance

determination, is not high-dimensional feature selection) and Expected-Improvement acquisition [25, 26], implemented with scikit-learn's Gaussian-process regressor [27] and the Expected Improvement maximised by SciPy's L-BFGS-B optimiser [28], evaluating 23 parameter combinations per RBP (8 Latin-hypercube initial points [29] + 15 sequential iterations) over a continuous search space ($n \in [10, 70]$, $e \in [30, 300]$), compared with the grid's 20 discrete combinations. Both methods used identical objective functions and bootstrap seed. The full comparison is shown in Additional file 1: Fig. S5. Per-RBP grid and BO optima and AUROCs are provided in Additional file 2: Table S7.

BO did not change the ranking or improve accuracy. Across the 27 combinations for which both optimisers returned a defined score, grid and BO AUROCs were very highly correlated (Spearman $\rho = 0.844$, $P = 3.2 \times 10^{-8}$), yielding a near-identical RBP ranking (Additional file 1: Fig. S5A). BO did not improve on the grid: mean AUROC was in fact marginally lower for BO (0.78 versus 0.84), and the grid matched or exceeded BO in 22 of 27 combinations (81%; Additional file 1: Fig. S5A and Additional file 2: Table S7). While the smaller evaluation budget afforded to BO (8 + 15 iterations) argues against over-interpreting this modest difference, it nonetheless confirms that the grid entails no cost in accuracy. BO neither improved the low-scoring RBPs (PUS1, FXR1, SF3A3 and RBM22 remained low under both methods) nor diminished the strong ones (HNRNPC, HNRNPK and RBFOX2 each remained ≥ 0.98; Additional file 2: Table S7). However, it displayed greater variability. Occasional marked gains (for example PTBP1 in K562, from 0.81 to 0.93) were counterbalanced by pronounced losses on factors that the grid had resolved well (NCBP2 in HepG2, from 0.79 to 0.40; HNRNPU in K562, from 0.76 to 0.52; TARDBP, from 1.00 to 0.84 in K562; Additional file 1: Fig. S5A), whereas grid search yielded no comparable deviations. This instability provides a further argument for the deterministic search.

Relative performance tracked eCLIP signal strength. The concordance of the two search strategies on the same low-scoring RBPs indicates that performance is constrained by the biological signal in the eCLIP data rather than by the optimisation procedure. Strong, position-specific binders reached AUROC >0.9 under both methods (*i.e.,* HNRNPC, PTBP1 and RBFOX2), whereas factors with weak or non-discriminating binding signals scored low under both (*i.e.,* PUS1, FXR1, SF3A3, RBM22; Additional file 1: Fig. S5A-B and Additional file 2: Table S7).

The two optimisers selected different parameters, not different performance. The grid- and BO-selected optima were only weakly and non-significantly correlated across RBPs (Spearman $\rho = 0.06$, $P = 0.78$ for n; $\rho = 0.16$, $P = 0.42$ for e), even though the resulting AUROCs were highly correlated ($\rho = 0.844$, $P = 3.2 \times 10^{-8}$; Additional file 2: Table S7). This

degeneracy is directly visible in the parameter fold-changes (Additional file 1: Fig. S5C). The log2 fold-change between BO- and grid-selected values is frequently large for both *n* and *e*, while the corresponding change in AUROC is minimal (Additional file 1: Fig. S5C, predominantly pale points). Many (*n, e*) settings are thus near-optimal for a given RBP, and the two searches settle at different points on a broad near-optimal plateau while achieving equivalent discrimination. This degeneracy of the parameter landscape is a further reason to prefer the grid, whose discrete optima correspond to interpretable biological clustering scales, over the continuous values returned by BO.

Rationale for retaining the grid as the default. In the absence of any accuracy advantage for BO, the choice between the two strategies rests on practical considerations that favour the grid: (*i*) reproducibility (BO is contingent on GP initialisation and the stochasticity of the acquisition step, whereas the grid is deterministic), (*ii*) interpretability (its discrete optima correspond to biologically meaningful clustering scales, $n \in \{10, 30, 50, 70\}$), and (*iii*) computational cost (the 20 grid evaluations are fully parallelisable, whereas the acquisition steps of BO are inherently sequential). On these data, the grid therefore dominates BO across every axis that matters (*i.e.* equal discrimination, full reproducibility, biologically interpretable optima, and trivial parallelisation) with none of the stochastic variability that a Gaussian-process optimiser introduces. The grid is therefore retained as the principled default, validated by direct comparison rather than assumed, with BO exposed as an optional backend.

6. Motif width comparison

The use of tetramers (k = 4) was inherited from RNAmotifs [1] and rests on two considerations. First, most RBP recognition elements are short and degenerate, with information-rich cores of roughly 3-5 nts, and multivalent regulation arises from the clustering of such short elements rather than from long contiguous motifs [50, 51]. Position-dependent action of these clustered motifs is the basis of the RNA-map framework on which the method builds [39]. Tetramers are therefore expected to capture the recognition core, with the surrounding multivalency recovered by clustering and by the positional enrichment window rather than by a longer k. Second, k = 4 keeps the combinatorial search space tractable ($2 \times 4^4 = 512$ candidate motifs). Together these predict that widening k should not improve, and may degrade, RBP identification, which we tested directly.

We repeated motif enrichment and MaRs association scoring at k = 4, 5 and 6 for four RBPs that are well identified at k = 4 (HNRNPC, QKI and U2AF2 at AUROC, and PTBP1; Fig. 3B)

and differ in recognition-motif composition. High-performing RBPs were chosen so that any effect of wider motifs would be clearly interpretable. For each RBP the clustering and enrichment window were held at that RBP's k = 4 optimum, leaving motif width as the only variable, and AUROC was computed against the HepG2 eCLIP panel. The search space grows as 2×4^k (512, 2,048 and 8,192 candidate motifs at k = 4, 5 and 6), making k = 6 roughly 16x more expensive to enumerate and bootstrap than k = 4.

No RBP improved at k > 4 (Additional file 1: Fig. S38A). HNRNPC remained essentially saturated (AUROC 1.00, 0.98 and 0.99 at k = 4, 5 and 6), whereas the other three lost discrimination as k increased. At k = 5, the largest losses were observed for QKI (0.67) and U2AF2 (0.80). PTBP1 declined gradually at k = 5 (0.91) and then sharply at k = 6 (0.76). The mean ΔAUROC relative to k = 4 was -0.14 at k = 5 and -0.18 at k = 6. The number of significantly enriched MRMs did not track AUROC (Additional file 1: Fig. S38B), rising substantially for PTBP1 (to 45 at k = 5) and QKI (to 48 at k = 6) even as their AUROC declined.

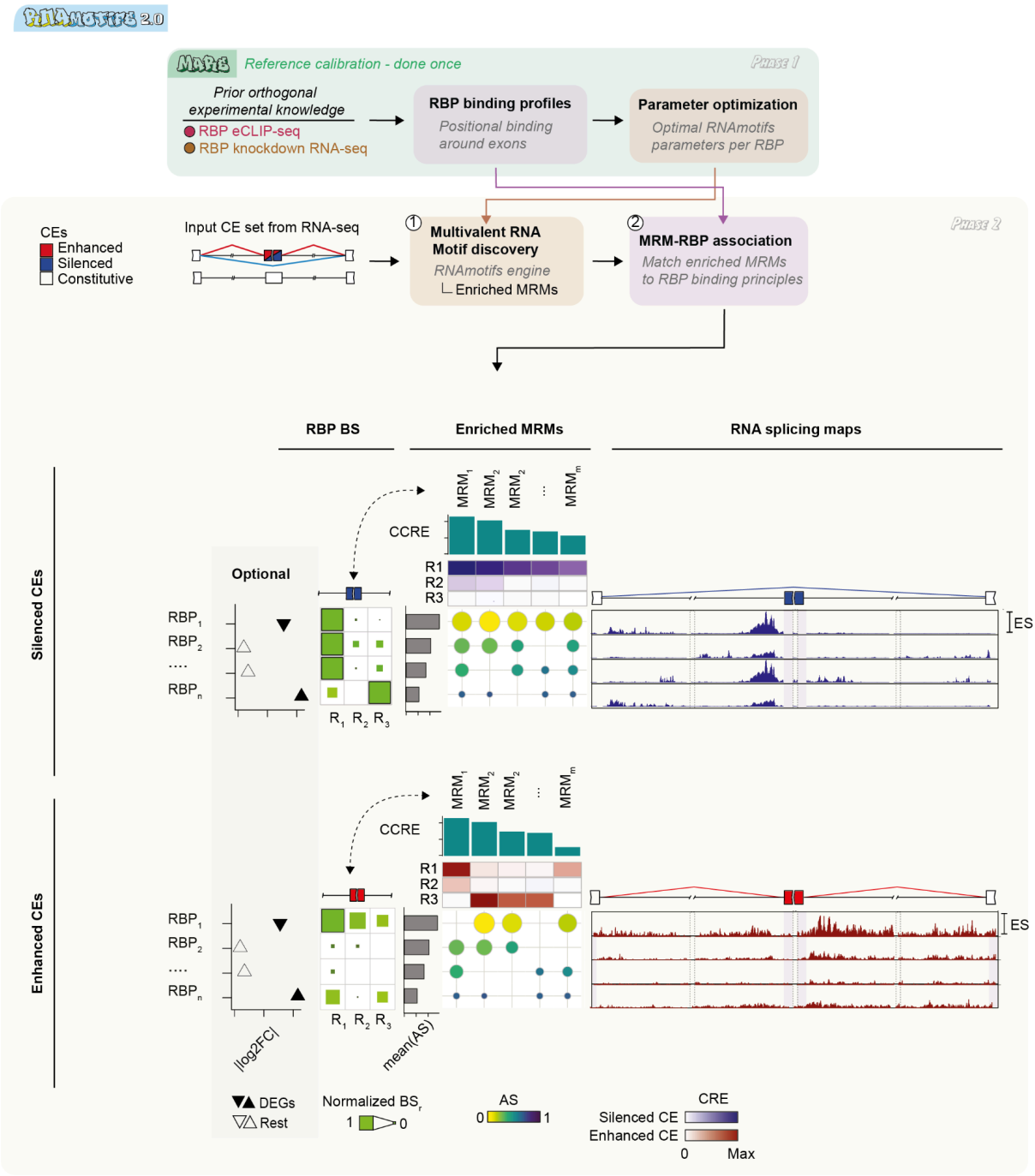
Consistent with the short-recognition-element rationale above, these additional MRMs represent lower-count fragments of the same binding signal, such that distributing a finite amount of signal over an exponentially larger and sparser motif space dilutes the positional-enrichment statistic rather than sharpening it. U2AF2 illustrates the same effect independently of MRM proliferation. Its enriched-MRM count remained essentially flat across k (16, 14 and 16), yet its AUROC fell sharply at k ≥ 5. This behaviour matches U2AF2's bipartite recognition mode, in which each RNA recognition motifs contacts a short (~4-nt) pyrimidine half-site and productive binding requires two such half-sites rather than one longer contiguous element [52, 53]. A k = 4 motif corresponds to a single half-site, whereas a wider k matches no natural recognition unit and instead spreads the recognition signal across wider, less-specific MRMs.

Computational cost constitutes a second, independent argument in favour of k = 4. The candidate motif space grows deterministically as 2×4^k, a 16-fold increase from k = 4 to k = 6, and this search dominates the pipeline, with total per-RBP wall-clock rising from seconds at k = 4 to ~42 min at k = 6 (Additional file 2: Table S13). This increase yielded no corresponding gain in AUROC and, for three of the four RBPs, a substantial loss. Because all three values of k were evaluated within a single benchmark harness at a fixed core count, confirming that the comparison isolates the effect of motif width. Taken together, the empirical results and the recognition-element architecture indicates that for the short, degenerate motifs characteristic of the RBPs examined here, a motif width of four nucleotides captures the relevant recognition signal, and wider motifs neither improve

identification nor justify their markedly greater computational cost. We therefore retain k = 4 as the validated default while exposing motif width as a user-configurable parameter (-k), so that the setting can be revisited for RBP families whose recognition elements are known to exceed this width.

Supplementary Figures

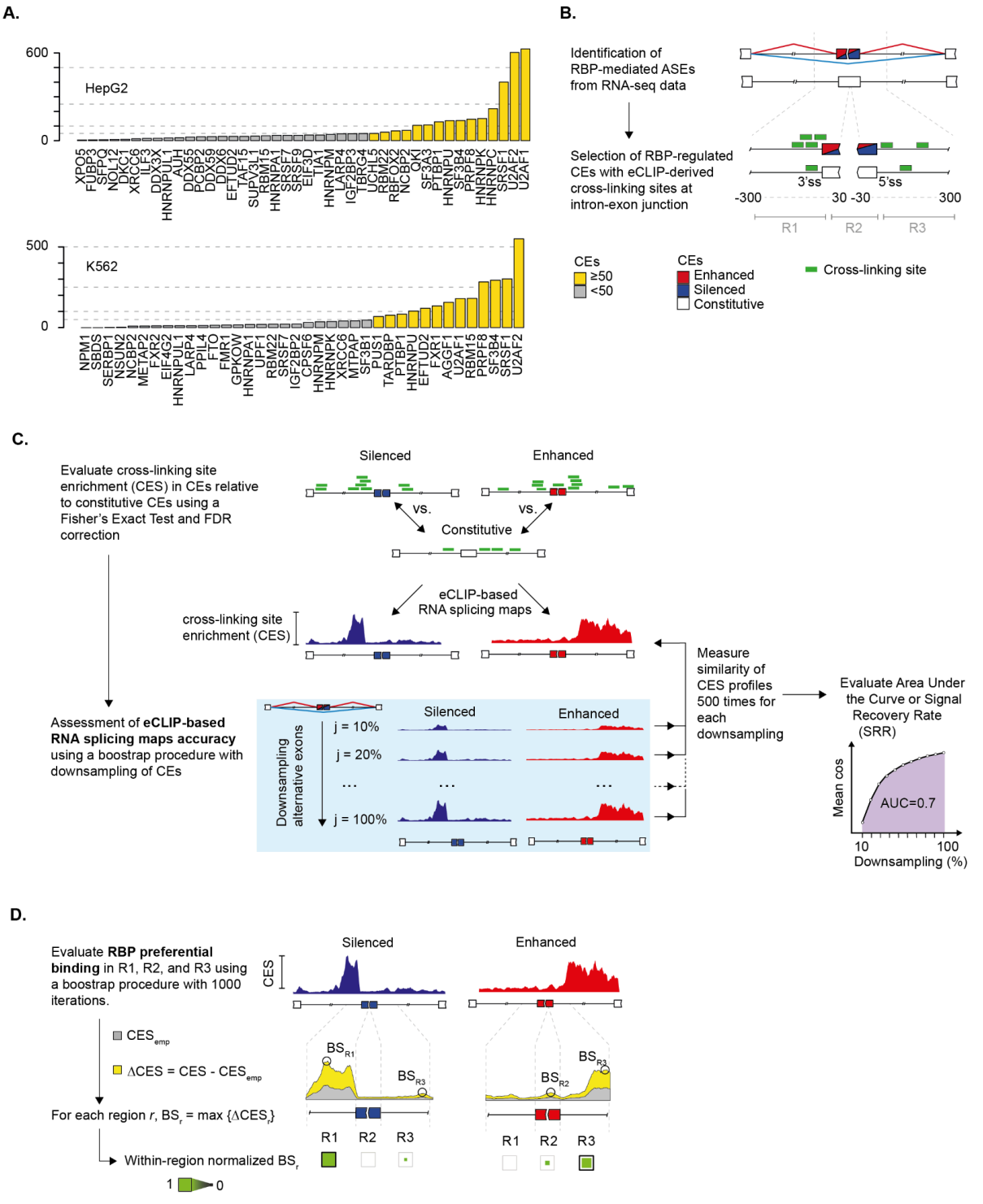
Figure S1. Detailed overview of the RNAmotifs 2.0 framework.



RNAmotifs 2.0 runs in two phases. In Phase 1 (MaRs, reference calibration, performed once), the MaRs module converts prior orthogonal experimental knowledge from ENCODE eCLIP-seq binding profiles (physical association) and RNA-seq following RNA-binding

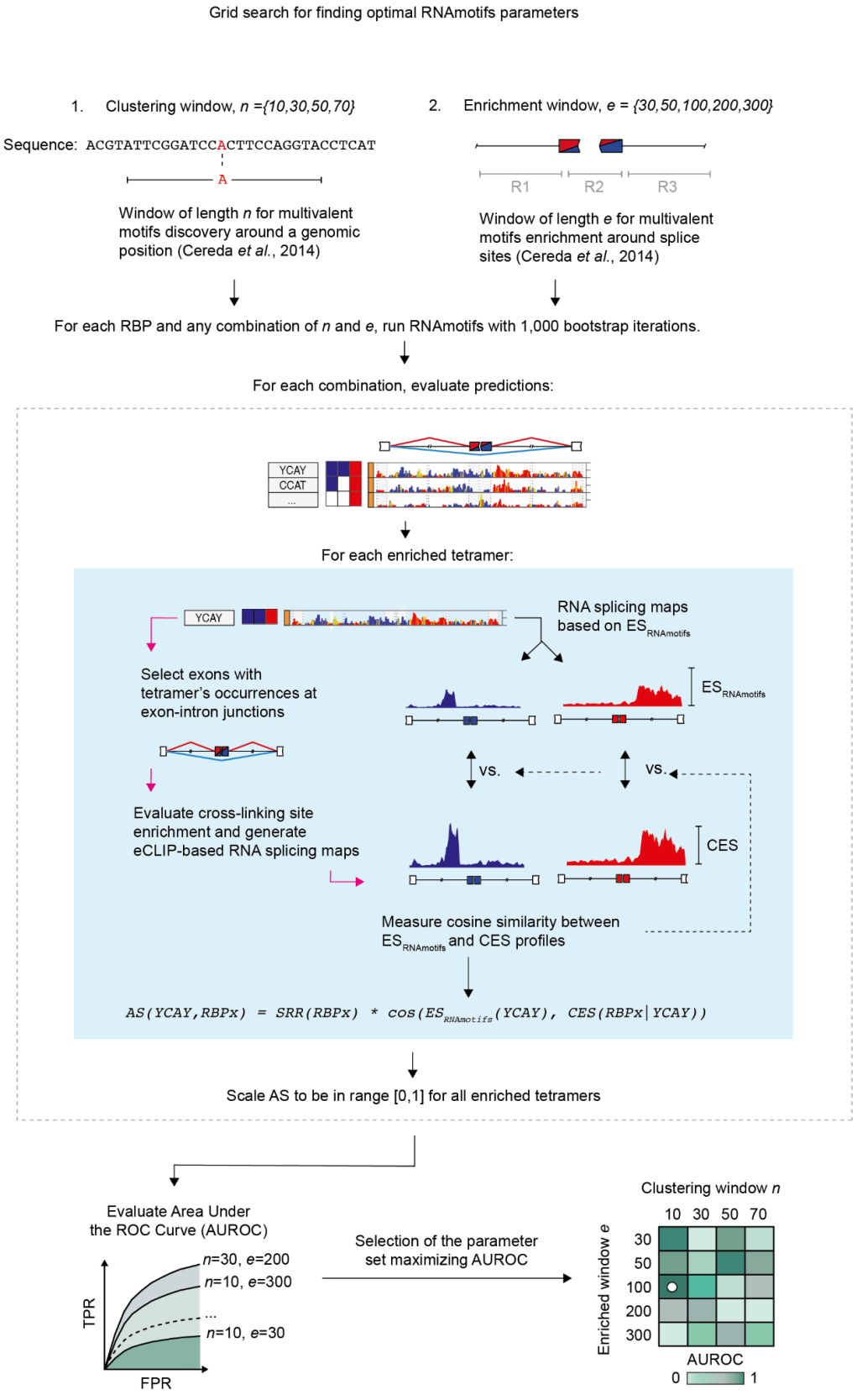
protein (RBP) knockdown (functional evidence) into (i) region-resolved RBP binding profiles describing positional binding around exons and (ii) the optimal RNAmotifs parameters for each RBP. Because this step depends only on reference data, it is computed once and reused across analyses: the optimized parameters feed multivalent RNA motif discovery, whereas the RBP binding profiles feed multivalent RNA motif (MRM)-RBP association in Phase 2. In Phase 2, RNAmotifs 2.0 takes as input a set of cassette exons (CEs) derived from RNA-seq and annotated as enhanced, silenced, or constitutive based on differential exon inclusion levels. (1) The RNAmotifs engine identifies enriched MRMs from the input CEs. (2) RNAmotifs 2.0 then links MRMs to their cognate RBPs by comparing MRM-specific and RBP-specific RNA splicing maps, producing MRM-RBP association scores (AS) alongside region-specific RBP binding scores (BS). Outputs, shown separately for silenced and enhanced CEs, include ranked candidate RBPs, enriched MRMs, and RNA splicing maps that summarize positional binding and regulatory signatures. For each MRM, combined regional enrichment (CRE) values across regions R1-R3 are displayed as a top heatmap annotation, cumulative CRE (CCRE) as a top bar plot, and MRM-RBP AS as a bubble plot; per-region RBP binding is shown as normalized BSr and mean AS per RBP as a bar plot. Differential gene expression (DEG) results for candidate RBPs (|log2FC|; DEGs vs. Rest) are optionally included.

Figure S2. Robustness and positional specificity of eCLIP-derived RBP binding profiles



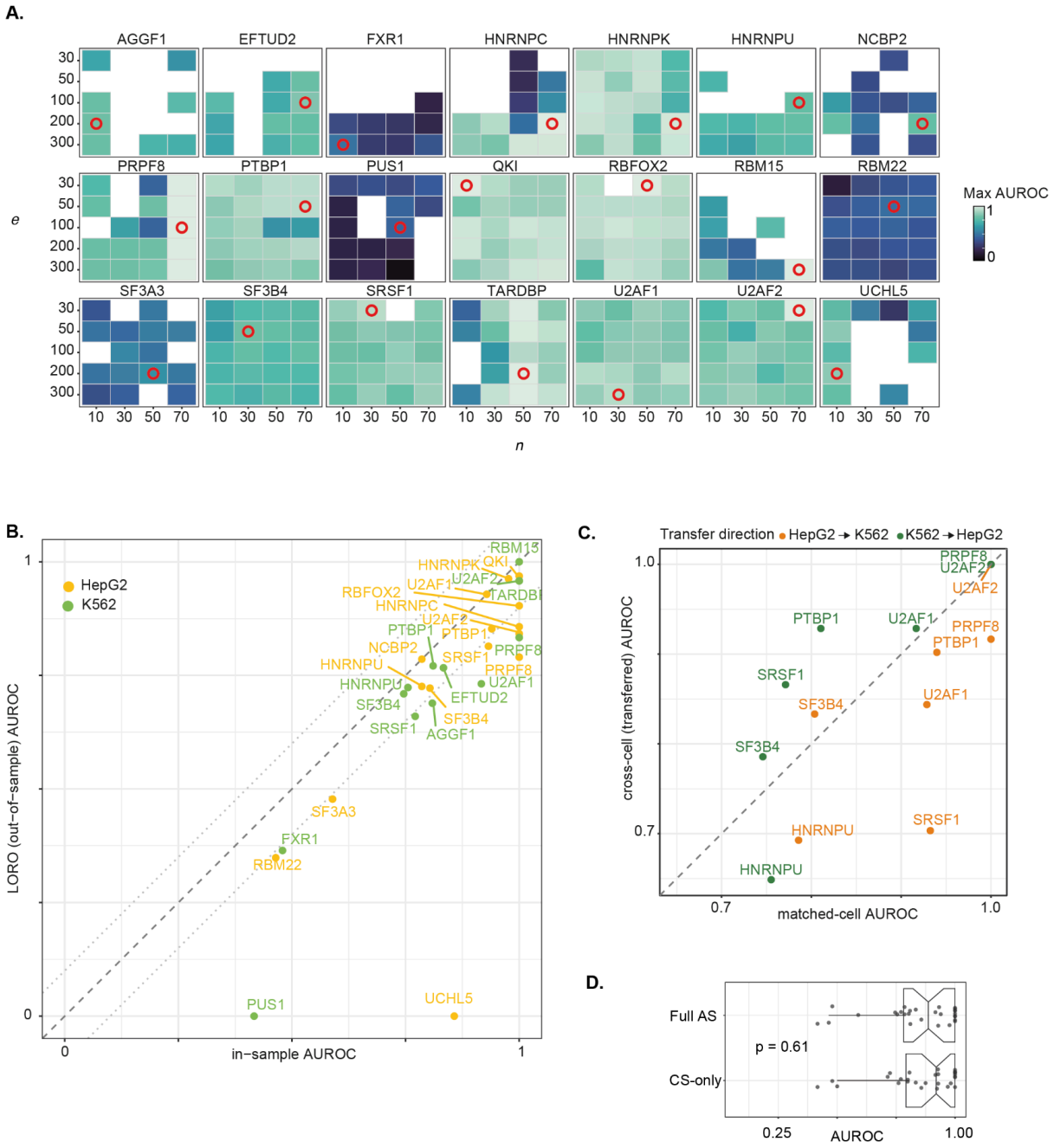
A. Number of alternative splicing events (ASEs) per RBP in HepG2 and K562 cells with cross-linking sites at intron/exon junctions. Yellow bars represent RNA-binding proteins (RBPs) with at least 50 regulated cassette exons (CEs). **B.** Overview of the framework for the selection of CEs regulated by RBP binding. **C.** Schematic of the computation of CES profiles (top) and evaluation of their robustness through downsampling for SRR calculation (bottom). **D.** Overview of the framework for evaluating RBP preferential binding and quantifying the Binding Score (BS) through comparison with a background obtained via bootstrapping.

Figure S3. Overview of the framework for the ROC-guided selection of optimal RNAmotifs parameters.



For each RNA-binding protein (RBP), RNAmotifs results are generated using 20 different combinations of clustering (*n*) and enrichment (*e*) windows. Based on RNAmotifs results, the association score (AS) for each multivalent RNA motif (MRM) is computed by comparing the enrichment score (ES) and cross-linking site enrichment (CES) profiles restricted to exons containing the MRM at exon/intron junctions. AS values are scaled to the [0,1] range and used to compute combination-specific AUROC values.

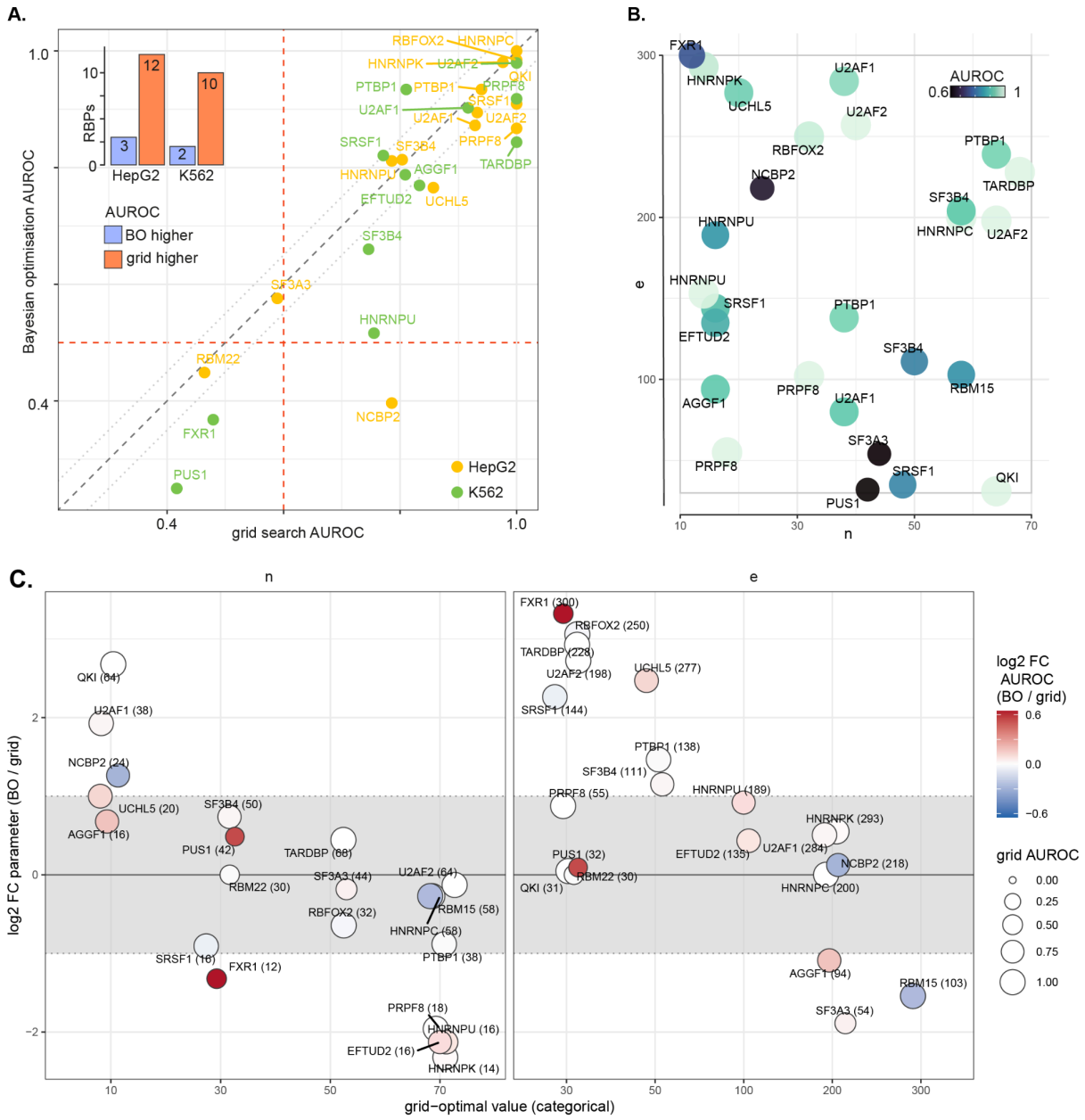
Figure S4. Optimization of RNAmotifs parameters and evaluation of MRM-RBP association scores.



A. Heatmaps summarizing AUROC values obtained for each combination of clustering (*n*) and enrichment (*e*) window for each RNA-binding protein (RBP). For each combination, the maximum AUROC across regulation classes and cell lines is reported. Only values derived from RNAmotifs results with at least five enriched multivalent RNA motifs (MRMs) are included. For each RBP, the maximum AUROC across all combinations is highlighted with a red circle. **B.** Out-of-sample generalization assessed by leave-one-RBP-out (LORO)

cross-validation. For each RBP, the in-sample AUROC (obtained on the RBP's own data) is plotted against the LORO out-of-sample AUROC (obtained when the RBP is held out), stratified by cell line (HepG2, K562). The dashed line marks the identity, points near it indicate parameters that generalize beyond the RBP used to select them. **C.** Cross-cell-line transferability of optimal parameters. For each shared RBP, the matched-cell AUROC (parameters applied within the cell line in which they were calibrated) is plotted against the cross-cell (transferred) AUROC (parameters applied to the other cell line). Colour denotes transfer direction. The dashed line marks the identity. **D.** Comparison of the full association score (Full AS) against the profile-similarity component alone (CS-only), showing the distribution of per-RBP AUROC values under each scoring scheme. The p-value is from a paired Wilcoxon comparison across conditions.

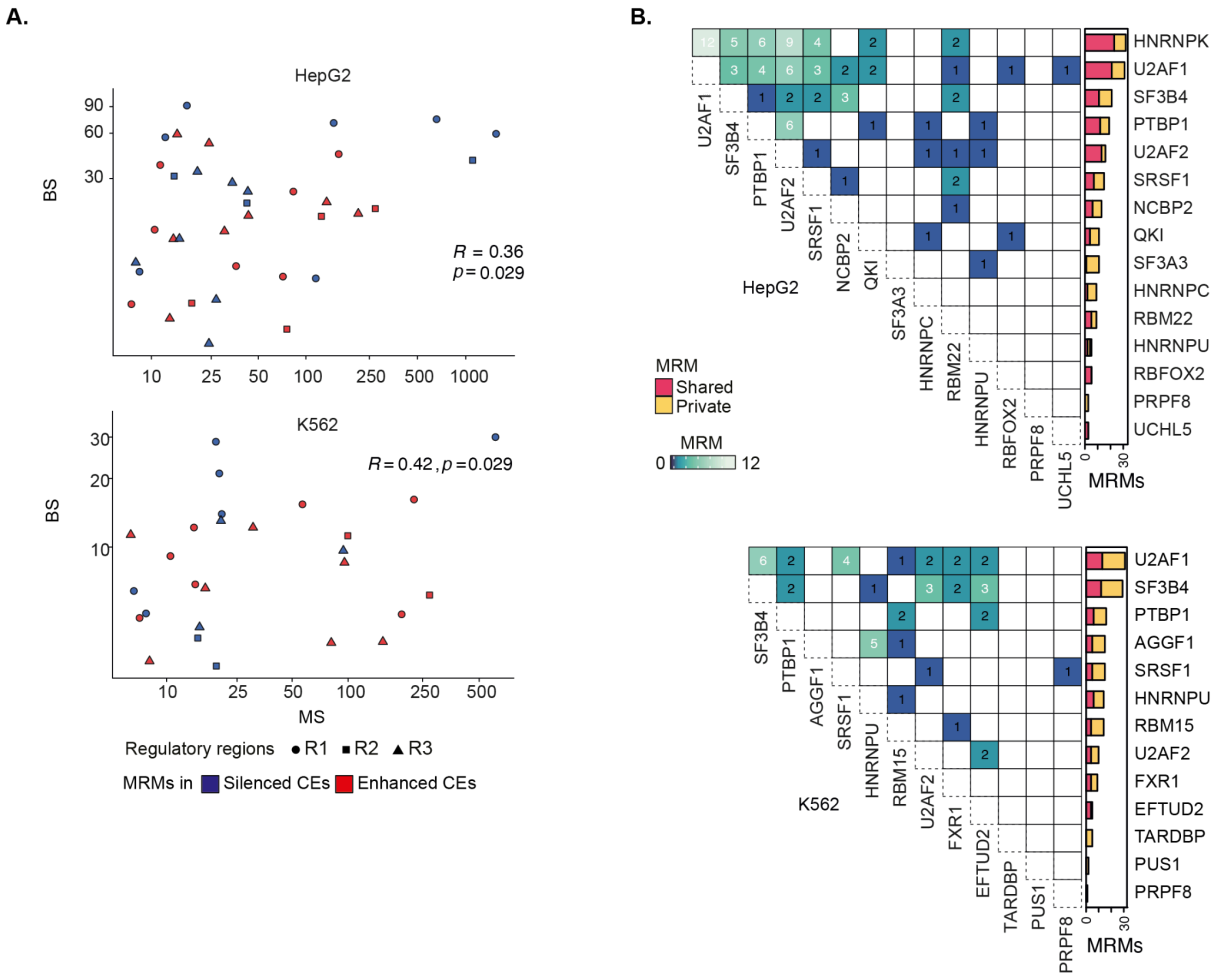
Figure S5. Bayesian optimization of RNAmotifs 2.0 parameters versus grid search.



A. Per-RNA-binding protein (RBP) AUROC obtained by Bayesian optimisation (BO) plotted against grid search, coloured by cell line (HepG2, K562); the black dashed line marks the identity and the dotted lines a tolerance band. Red dashed lines depict AUROC of 0.5. Inset: number of RBPs per cell line for which each strategy achieved the higher AUROC. **B.** BO-selected optima for each RBP in the continuous clustering-window (*n*) × enrichment-window (*e*) space, coloured by the AUROC achieved at the selected setting and stratified by cell line. Only RBPs with AUROC >0.5 are shown. **C.** Relationship between the

parameter change and the performance change induced by BO, shown separately for *n* (left) and *e* (right). For each RBP, the log2 fold-change of the selected parameter (BO/grid) is plotted against its grid-optimal value (given in parentheses); point colour encodes the log2 fold-change in AUROC (BO/grid) and point size the grid-search AUROC. The shaded band marks a two-fold parameter window.

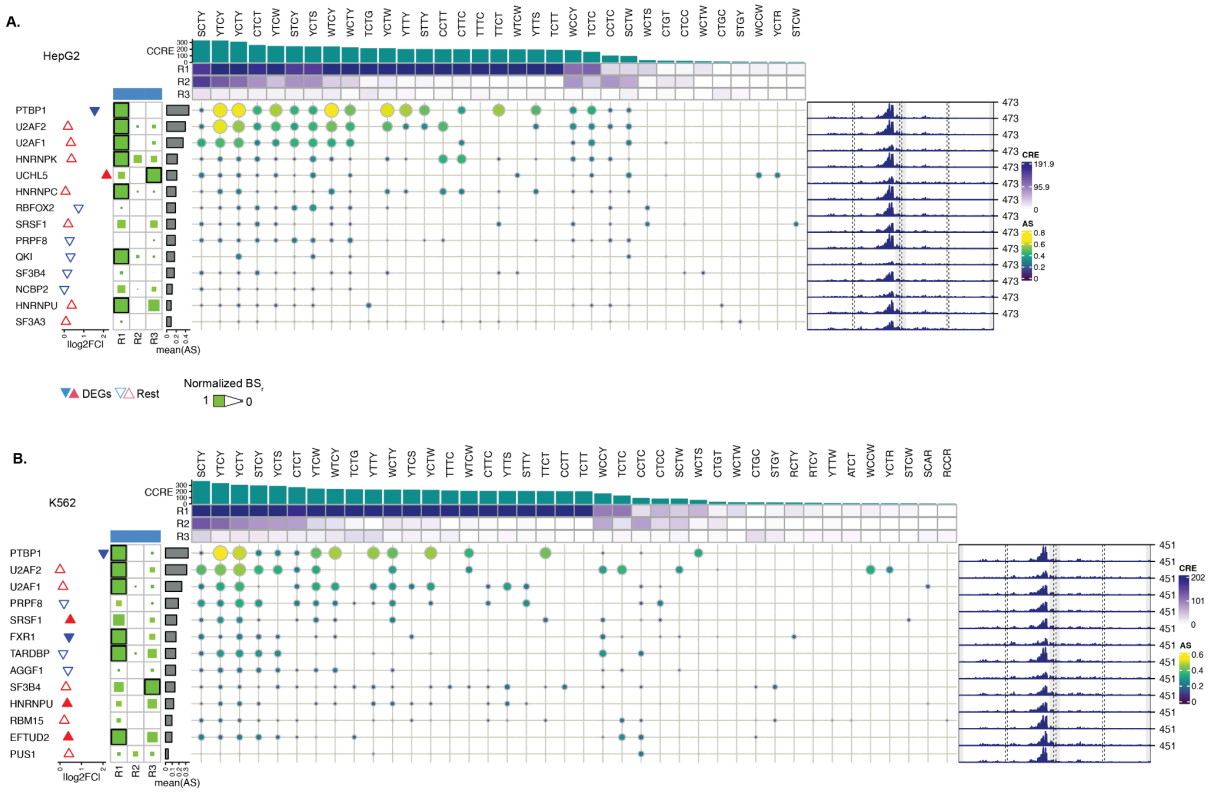
Figure S6. Binding-motif score correlation and cross-RBP MRM overlap.



A. Relationship between Binding Score (BS) and multivalent RNA motif (MRM) score (MS) values for each RNA-binding protein (RBP) across R1-R3 regions and regulation classes, stratified by cell line. Point shape indicates the regulatory region, while colour represents the regulation class. Only cases in which both scores are greater than 0 are included. For each cell line, the overall Pearson correlation coefficient is reported. **B.** Overlap of enriched MRMs across RBPs in each cell line, distinguishing shared and RBP-private motifs. Heatmap cells

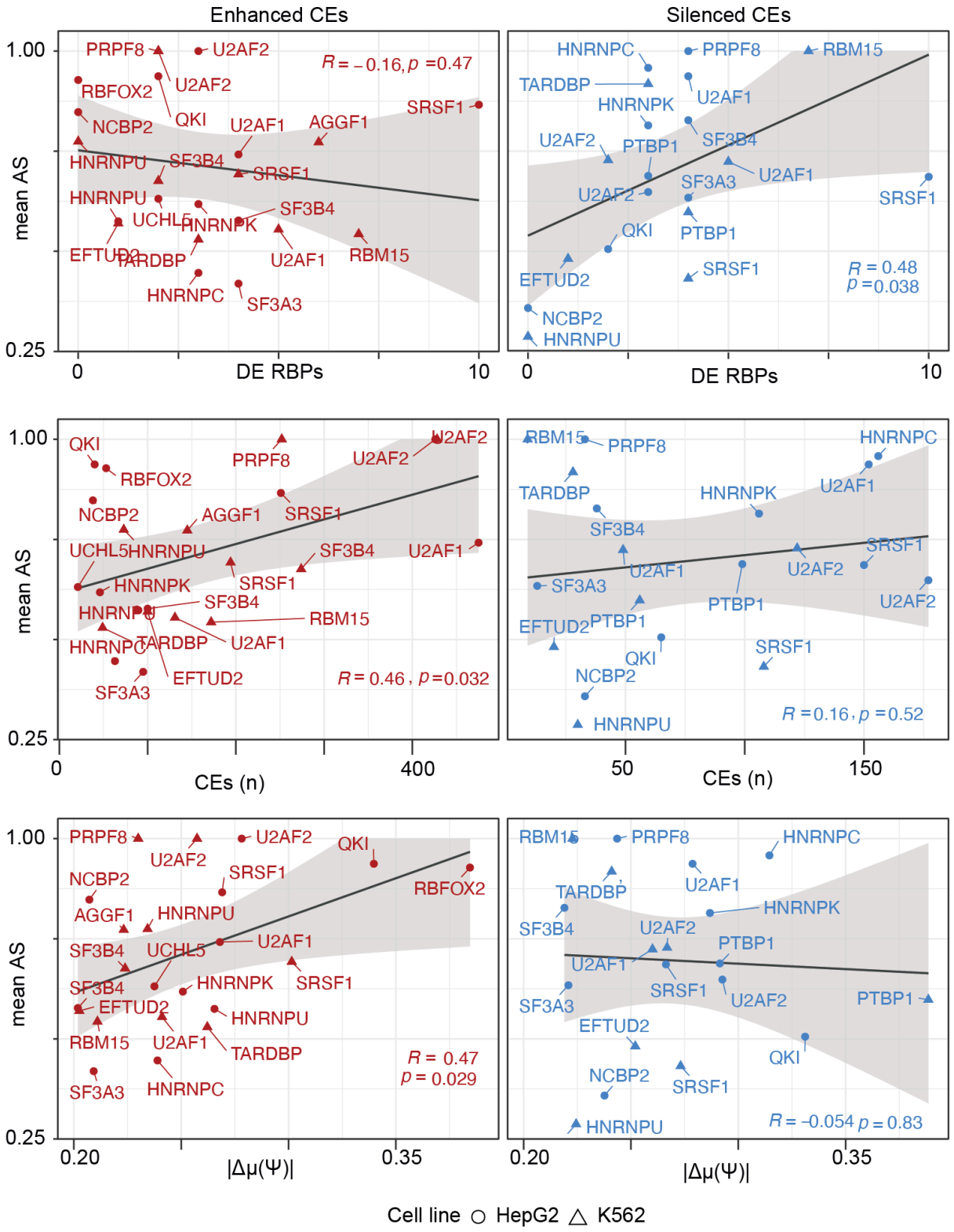
report the number of MRMs shared between each pair of RBPs. The stacked bar plot shows the total MRM count per RBP, partitioned into shared and private motifs.

Figure S7. RNAmotifs 2.0 results on PTBP1-silenced exons in HepG2 and K562 cells.



Complete RNAmotifs 2.0 outputs for PTBP1-regulated silenced exons in HepG2 (**A**) and K562 (**B**). The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

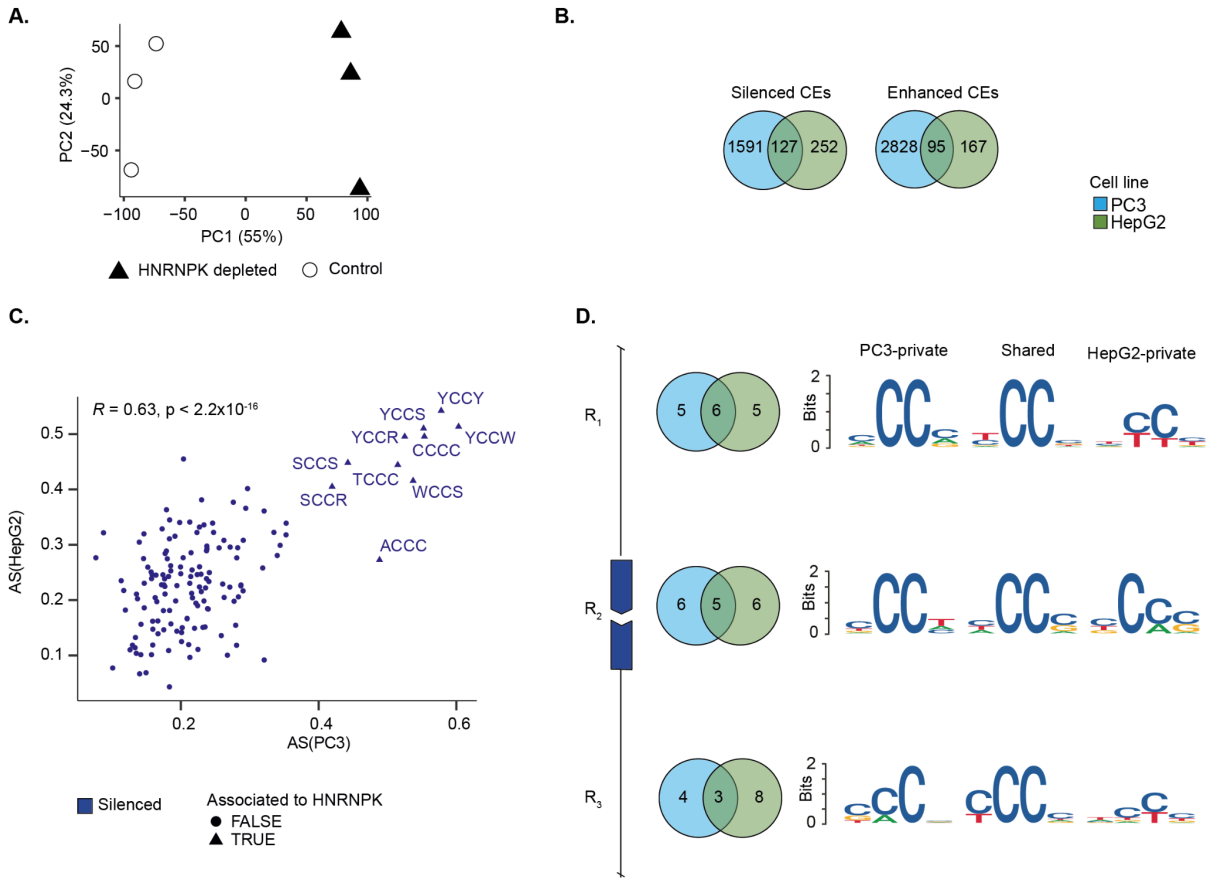
Figure S8. Regulation-class-specific determinants of mean AS



Per-RBP-binding protein (RBP) mean association score (mean AS) plotted against three features, shown separately for enhanced (left, red) and silenced (right, blue) cassette exons (CEs). Rows correspond to: (top) the number of differentially expressed RBPs (DE RBPs);

(middle) the number of input cassette exons per RBP (CEs, n); and (bottom) the mean absolute change in inclusion level (|Δμ(Ψ)|). Point shape denotes cell line (circle, HepG2; triangle, K562). In each panel, the solid line is a linear fit with its 95% confidence interval (shaded band), and the Pearson correlation coefficient (R) with its associated p-value is reported.

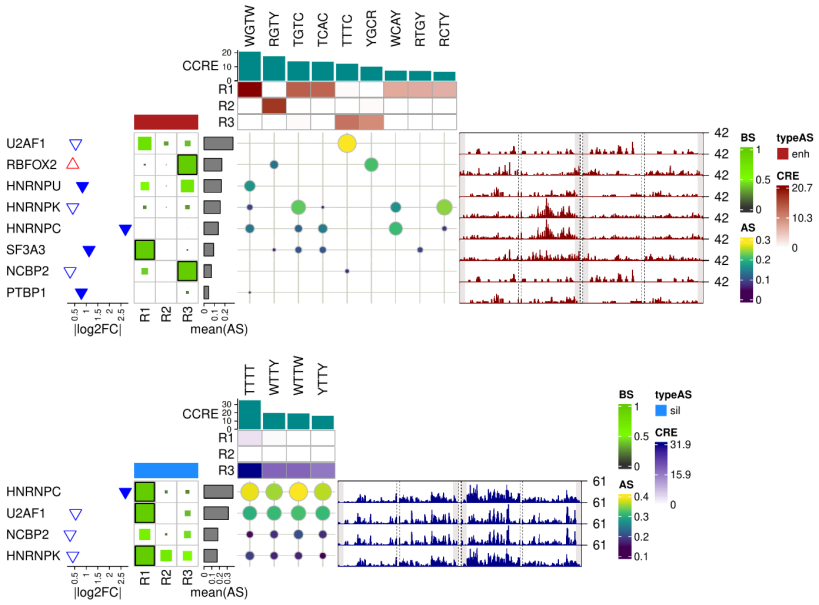
Figure S9. RNAmotifs 2.0 identifies HNRNPK as a splicing regulator in HNRNPK-depleted PC3 cells



A. Scatter plot of the first two components of principal component analysis (PCA) for PC3 dataset (n=3 for both control and HNRNPK-depleted conditions). Percentage of variance explained by each component is reported on each axis. **B.** Venn diagram showing the overlap of silenced and enhanced differentially spliced cassette exons (CEs) between PC3 and HepG2 cell lines. **C.** Relationship between association scores (AS) of each tetramer in PC3 and HepG2 cells for silenced exons. Tetramers identified as significantly associated to HNRNPK in both datasets are indicated with triangles. The overall Pearson correlation coefficient is reported. **D.** Venn diagram showing the overlap of multivalent RNA motifs

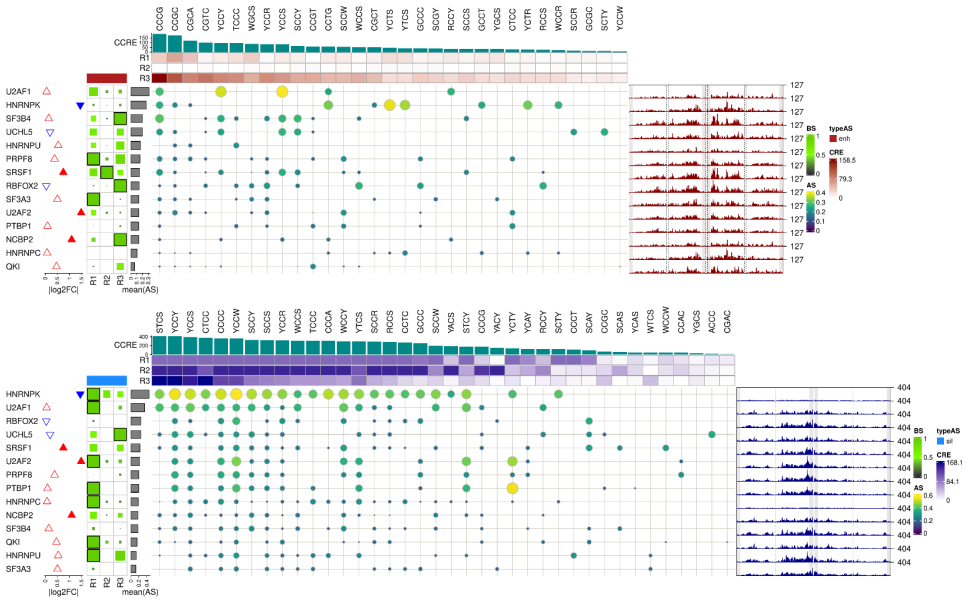
(MRMs) enriched in silenced exons between PC3 and HepG2 cell lines, along with the corresponding sequence logos summarizing MRMs shared and cell line-specific, stratified by splice-proximal regions (R1-R3).

Figure S10. RNAmotifs 2.0 results for HNRNPC-regulated exons in HepG2 cells.



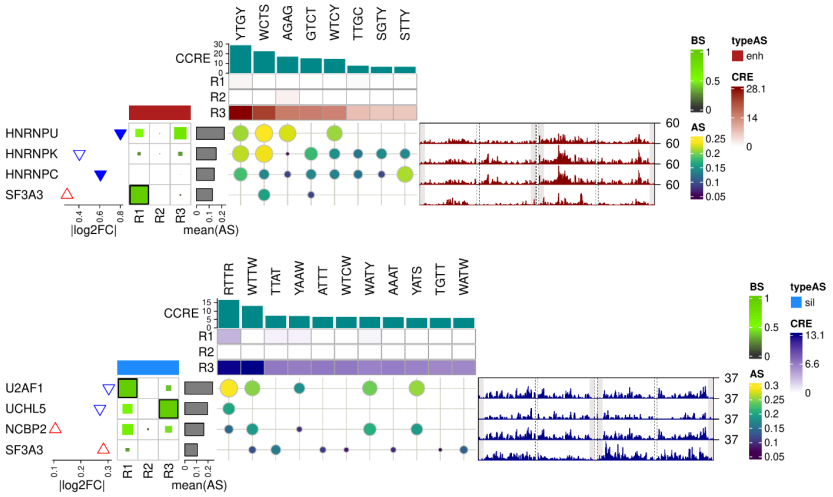
Complete RNAmotifs 2.0 outputs for HNRNPC-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S11. RNAmotifs 2.0 results for HNRNPK-regulated exons in HepG2 cells.



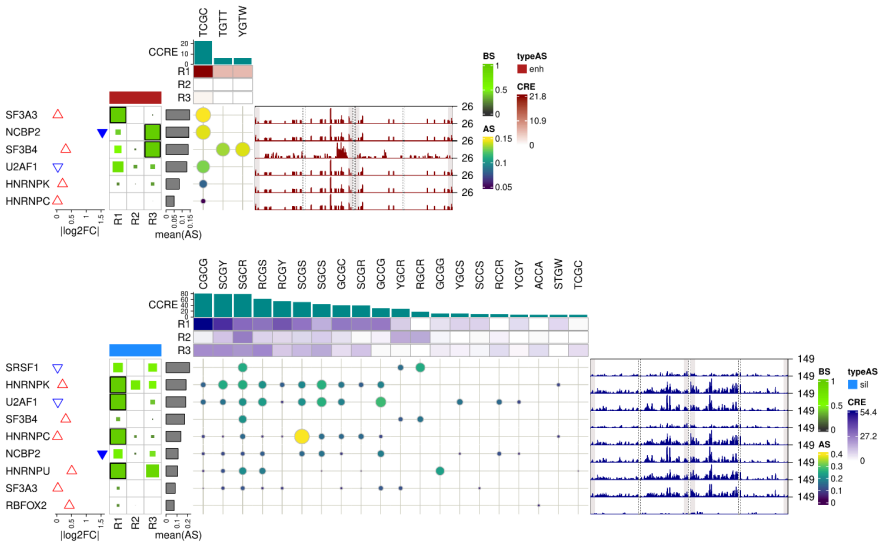
Complete RNAmotifs 2.0 outputs for HNRNPK-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S12. RNAmotifs 2.0 results for HNRNPU-regulated exons in HepG2 cells.



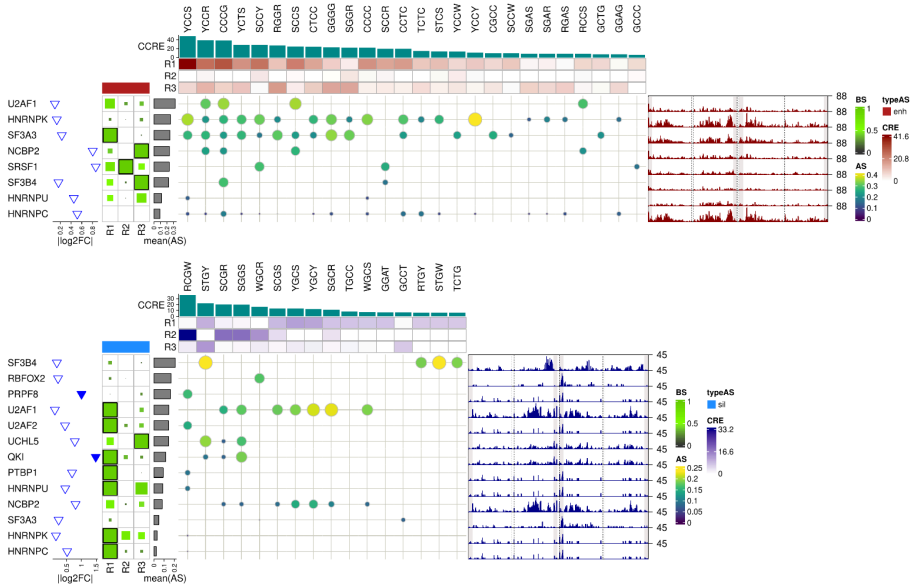
Complete RNAmotifs 2.0 outputs for HNRNPU-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S13. RNAmotifs 2.0 results for NCBP2-regulated exons in HepG2 cells.



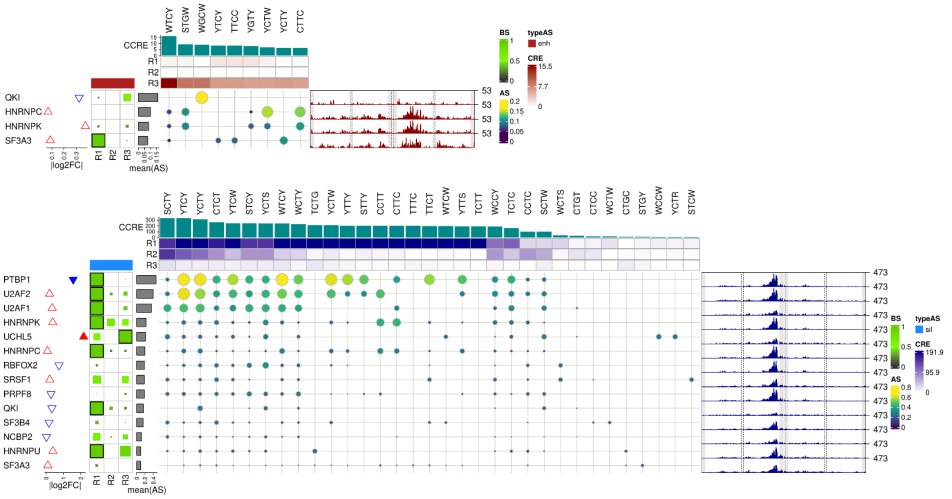
Complete RNAmotifs 2.0 outputs for NCBP2-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S14. RNAmotifs 2.0 results for PRPF8-regulated exons in HepG2 cells.



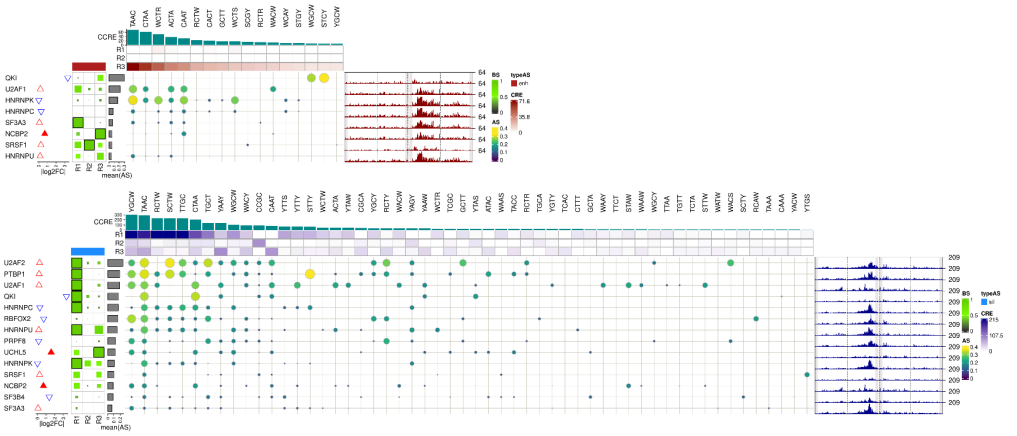
Complete RNAmotifs 2.0 outputs for PRPF8-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S15. RNAmotifs 2.0 results for PTPB1-regulated exons in HepG2 cells.



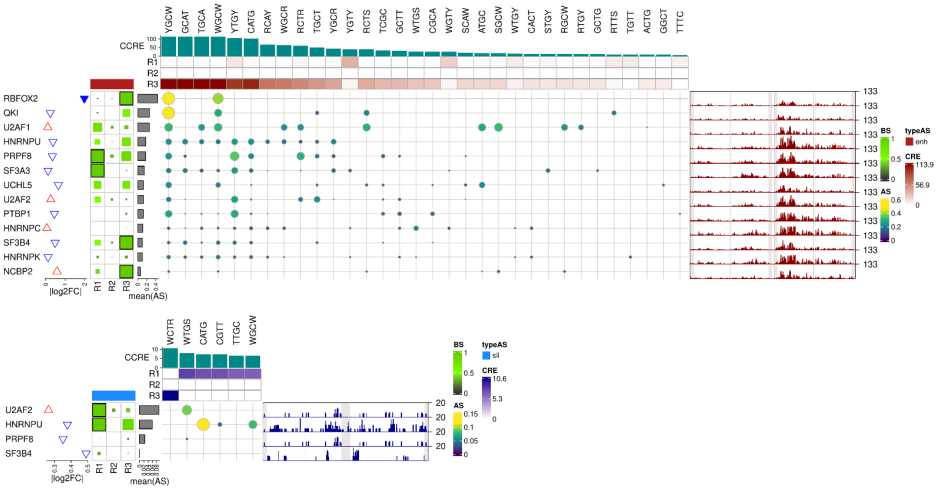
Complete RNAmotifs 2.0 outputs for PTPB1-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S16. RNAmotifs 2.0 results for QKI-regulated exons in HepG2 cells.



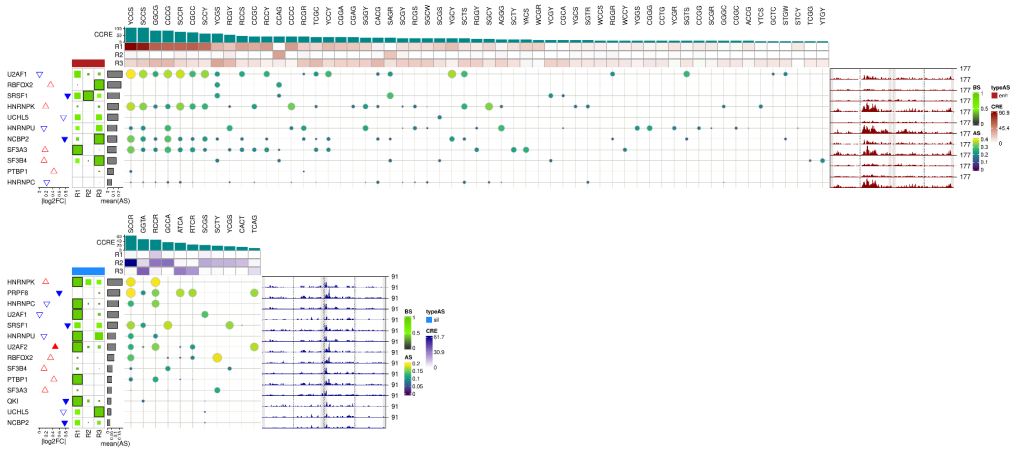
Complete RNAmotifs 2.0 outputs for QKI-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S17. RNAmotifs 2.0 results for RBFOX2-regulated exons in HepG2 cells.



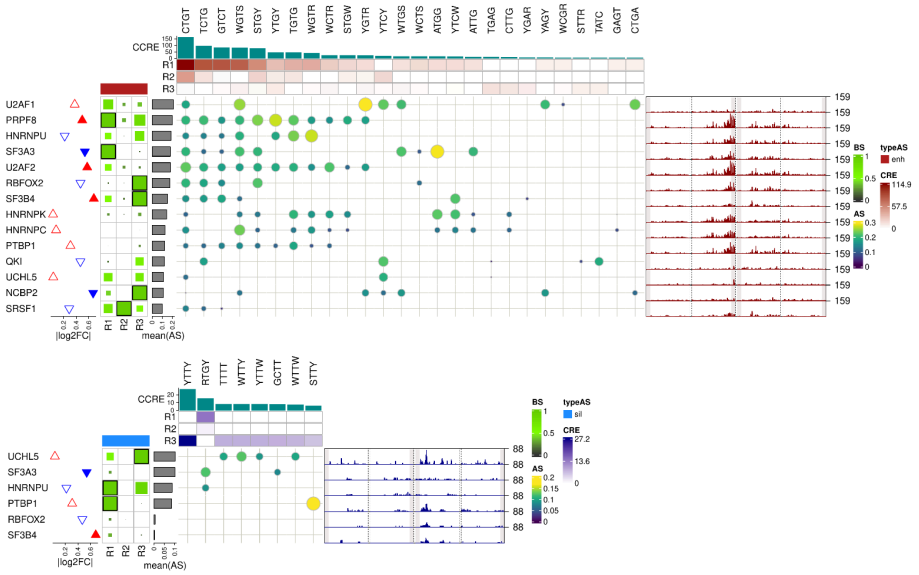
Complete RNAmotifs 2.0 outputs for RBFOX2-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S18. RNAmotifs 2.0 results for RBM22-regulated exons in HepG2 cells.



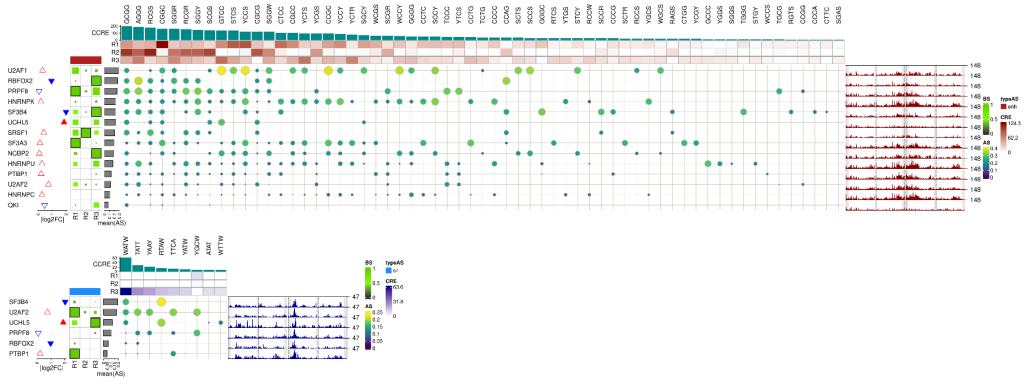
Complete RNAmotifs 2.0 outputs for RBM22-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S19. RNAmotifs 2.0 results for SF3A3-regulated exons in HepG2 cells.



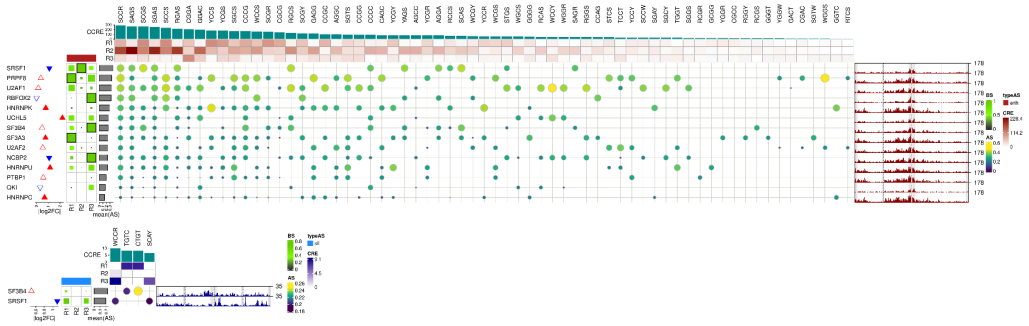
Complete RNAmotifs 2.0 outputs for SF3A3-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S20. RNAmotifs 2.0 results for SF3B4-regulated exons in HepG2 cells.



Complete RNAmotifs 2.0 outputs for SF3B4-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

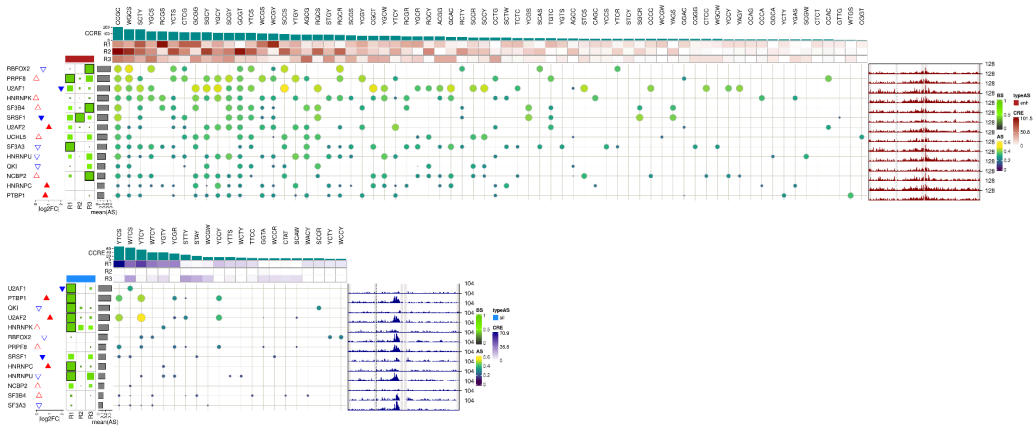
Figure S21. RNAmotifs 2.0 results for SRSF1-regulated exons in HepG2 cells.



Complete RNAmotifs 2.0 outputs for SRSF1-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein

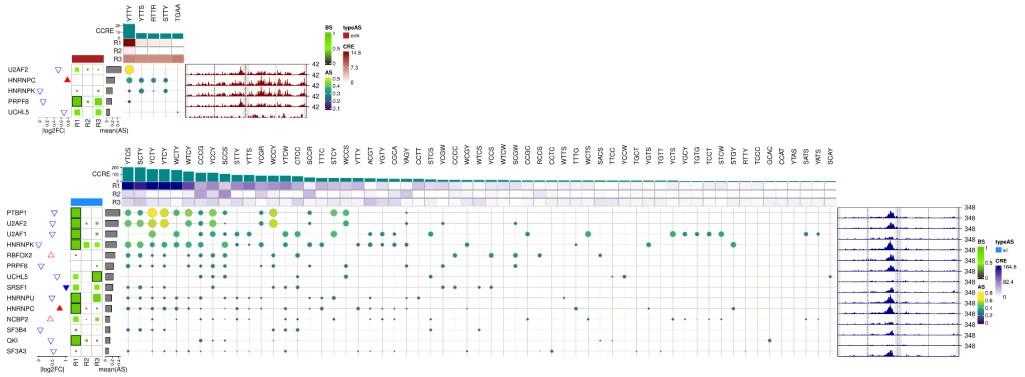
(RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S22. RNAmotifs 2.0 results for U2AF1-regulated exons in HepG2 cells.



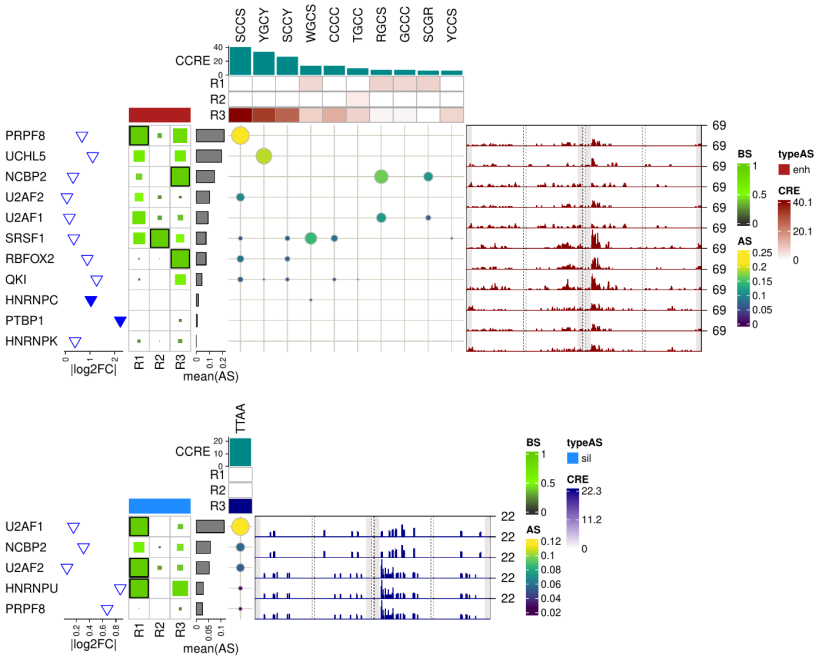
Complete RNAmotifs 2.0 outputs for U2AF1-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S23. RNAmotifs 2.0 results for U2AF2-regulated exons in HepG2 cells.



Complete RNAmotifs 2.0 outputs for U2AF2-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S24. RNAmotifs 2.0 results for UCHL5-regulated exons in HepG2 cells.

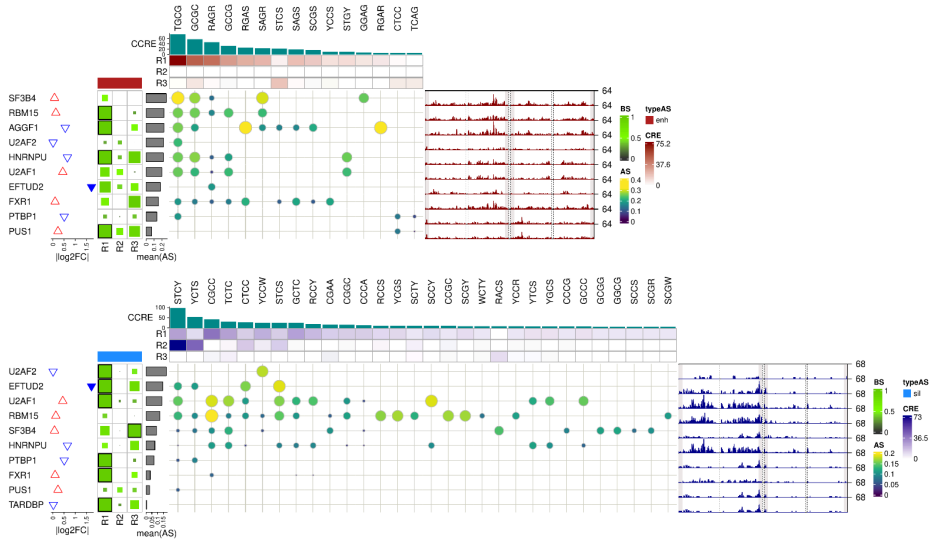


Complete RNAmotifs 2.0 outputs for UCHL5-regulated enhanced and silenced exons in HepG2 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S25. RNAmotifs 2.0 results for AGGF1-regulated exons in K562 cells.

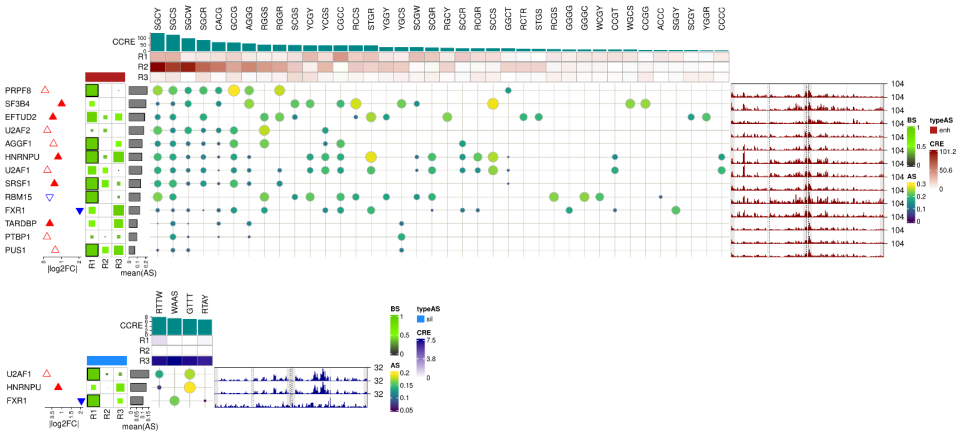
Complete RNAmotifs 2.0 output for AGGF1-regulated enhanced in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S26. RNAmotifs 2.0 results for EFTUD2-regulated exons in K562 cells.



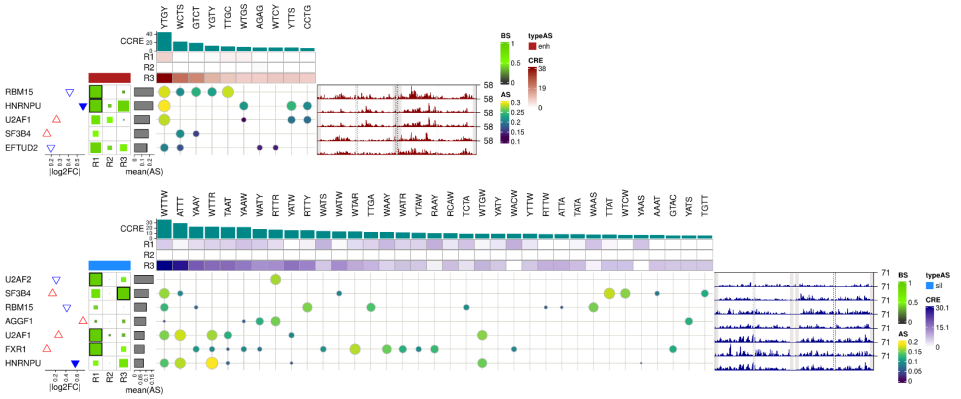
Complete RNAmotifs 2.0 outputs for EFTUD2-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S27. RNAmotifs 2.0 results for FXR1-regulated exons in K562 cells.



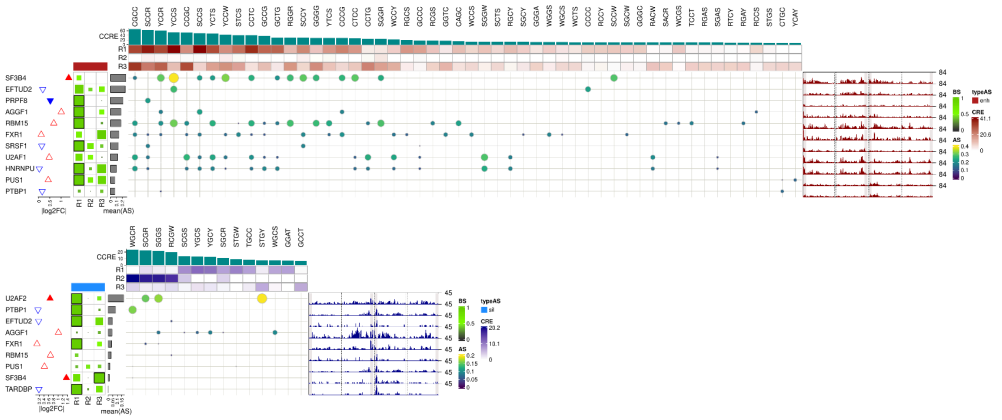
Complete RNAmotifs 2.0 outputs for FXR1-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S28. RNAmotifs 2.0 results for HNRNPU-regulated exons in K562 cells.



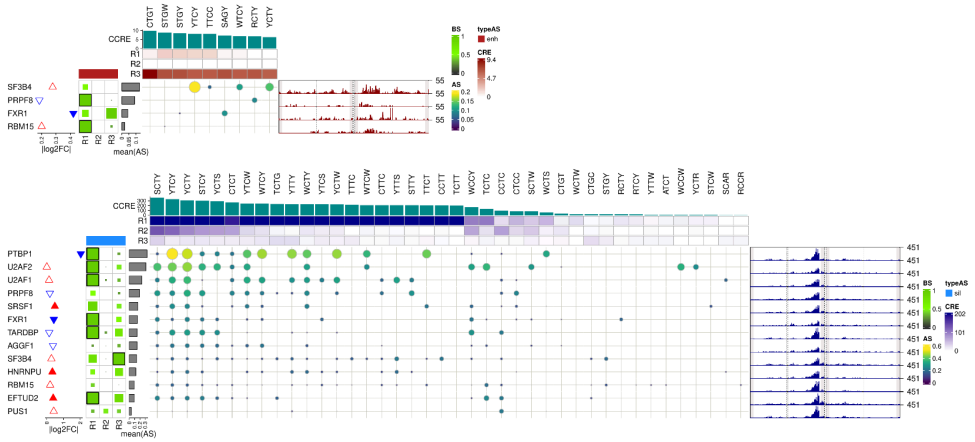
Complete RNAmotifs 2.0 outputs for HNRNPU-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S29. RNAmotifs 2.0 results for PRPF8-regulated exons in K562 cells.



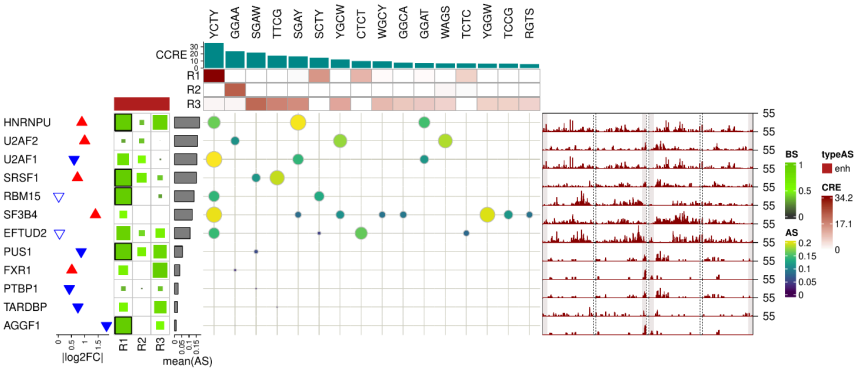
Complete RNAmotifs 2.0 outputs for PRPF8-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S30. RNAmotifs 2.0 results for PTBP1-regulated exons in K562 cells.



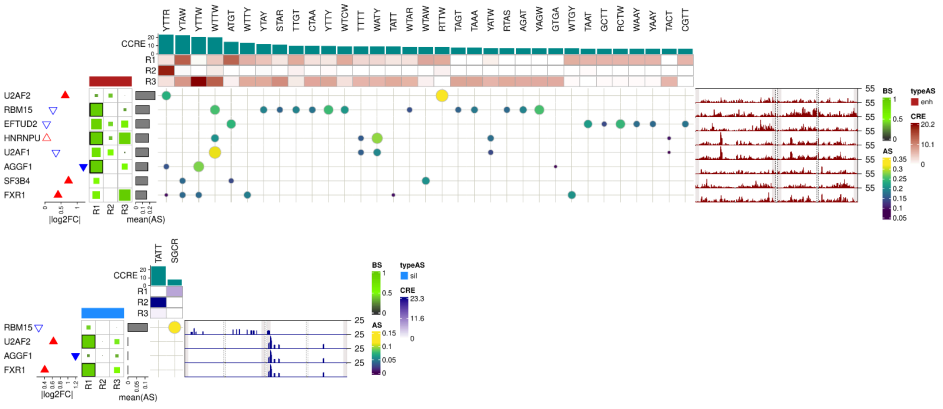
Complete RNAmotifs 2.0 outputs for PTBP1-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S31. RNAmotifs 2.0 results for PUS1-regulated exons in K562 cells.



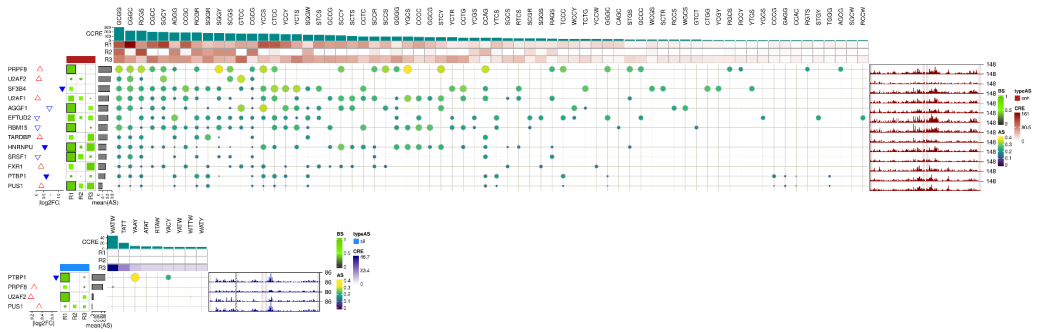
Complete RNAmotifs 2.0 output for PUS1-regulated enhanced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S32. RNAmotifs 2.0 results for RBM15-regulated exons in K562 cells.



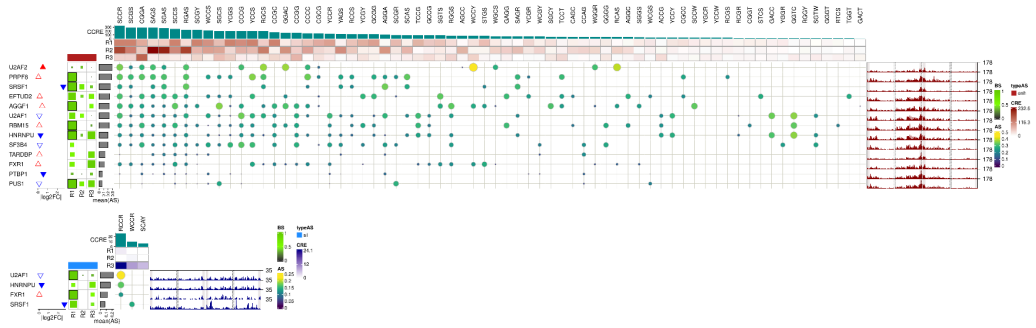
Complete RNAmotifs 2.0 outputs for RBM15-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S33. RNAmotifs 2.0 results for SF3B4-regulated exons in K562 cells.



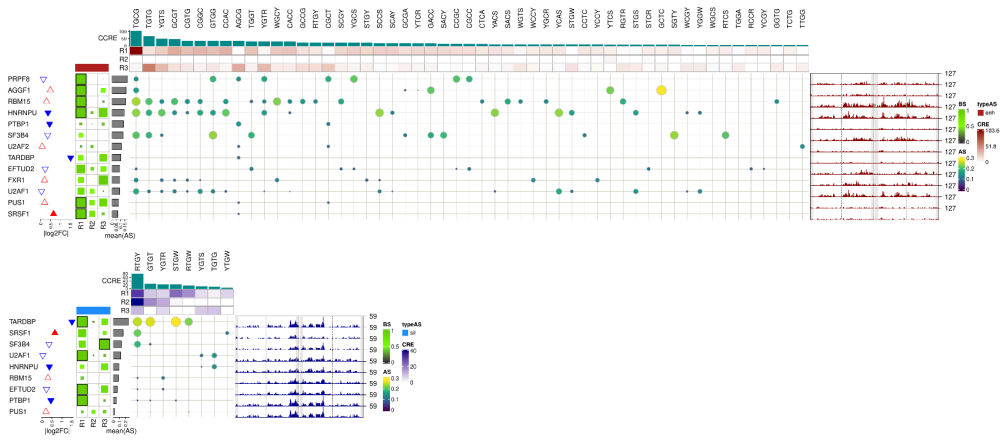
Complete RNAmotifs 2.0 outputs for SF3B4-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S34. RNAmotifs 2.0 results for SRSF1-regulated exons in K562 cells.



Complete RNAmotifs 2.0 outputs for SRSF1-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

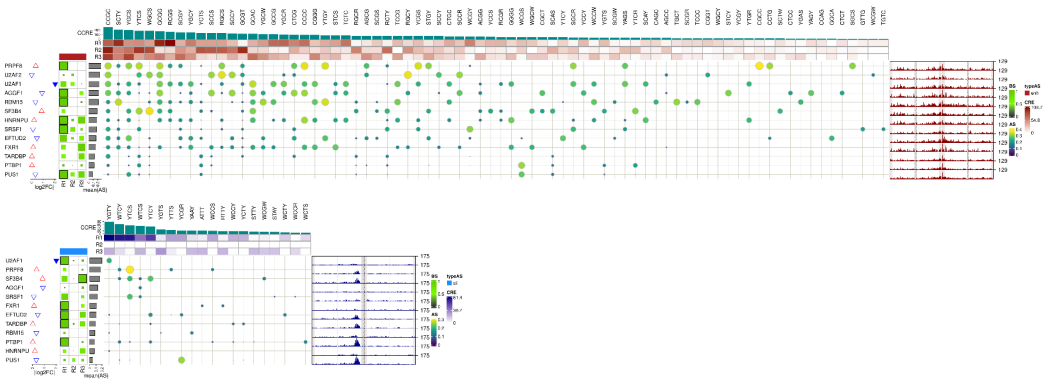
Figure S35. RNAmotifs 2.0 results for TARDBP-regulated exons in K562 cells.



Complete RNAmotifs 2.0 outputs for TARDBP-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across

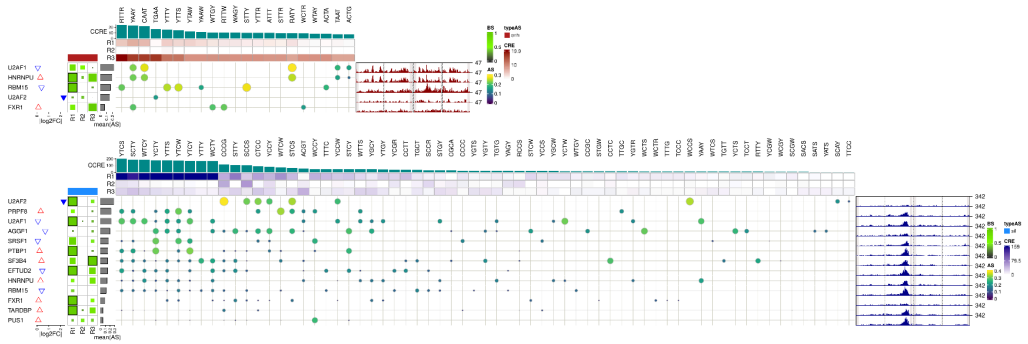
splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S36. RNAmotifs 2.0 results for U2AF1-regulated exons in K562 cells.



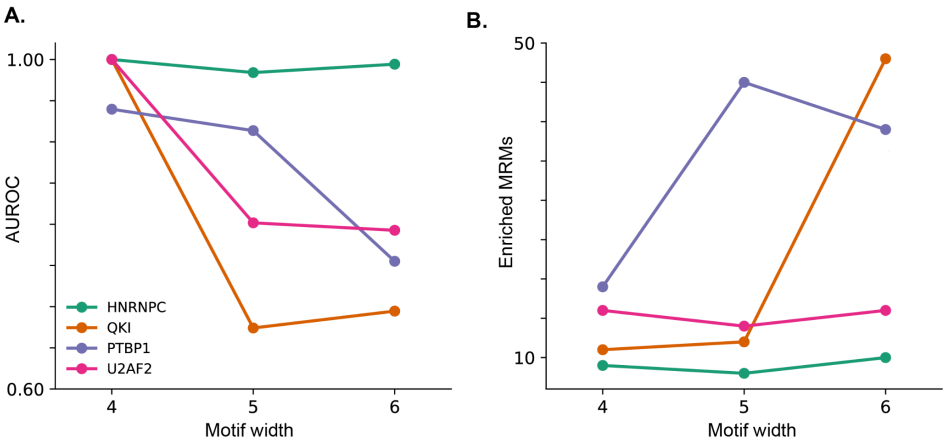
Complete RNAmotifs 2.0 outputs for U2AF1-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S37. RNAmotifs 2.0 results for U2AF2-regulated exons in K562 cells.



Complete RNAmotifs 2.0 outputs for U2AF2-regulated enhanced and silenced exons in K562 cells. The visualization integrates normalized binding scores (BS) across splice-proximal regions (R1-R3) (left), multivalent RNA motif (MRM)-RNA-binding protein (RBP) association scores (AS), cumulative combined regional enrichment (CCRE) across enriched MRMs (top), and the corresponding RNA splicing maps (right). Differential gene expression (DEG) values of corresponding RBP upon knockdown are shown as triangles. The upper and lower vertices indicated up- and down-regulation, respectively, and filled triangles denoted differential expression changes (*i.e.*, |log2FC| ≥0.1 and adjusted p-values ≤0.1). Barplots indicate the mean AS per the candidate regulator.

Figure S38. Effect of motif width on RBP identification and enriched-MRM counts



A. RNA-binding protein (RBP)-identification performance, quantified as the AUROC of the full association score against the HepG2 eCLIP panel, as a function of motif width. **B.** Number of significantly enriched multivalent RNA motifs (MRMs) as a function of motif width.