Supplementary Figures - ViennaRNA-predicted secondary-structure comparison of selected CURIA human-mouse pairs

This exploratory analysis is supplementary to, and is not used to support, the manuscript's primary performance claims. The examples are selected case studies, not a systematic benchmark.

Software: ViennaRNA (RNA) 2.7.2, pyrion 0.4.0, python 3.14.2.

Human and mouse sequences (the exact Figure-5 short-ncRNA intervals and the accepted MALAT1 island matches) are folded with ViennaRNA and their predicted base-pair-probability matrices are compared in sequence-alignment coordinates (affine-gap Smith-Waterman), so small insertions/deletions do not misregister the two matrices.

Short ncRNAs are similar-length and folded whole. MALAT1 cores are folded on length-matched windows around the aligned span (core, +-20 nt, +-40 nt) with both the global partition function and local RNAplfold; the raw, highly unequal island intervals are not folded.

Sequence metrics shown in the figures (nucleotide identity, normalized nucleotide-SW) are the CURIA baseline-table values, so the supplement stays consistent with the manuscript; they are labelled separately from the embedding metric. The pipeline matching distance is the short-branch MMD for the short ncRNAs and the island embedding-SW distance computed from RiNALMo embeddings for MALAT1. A separate local-alignment identity computed on the analysis window is retained only in summary.tsv, not shown in the figures.

Supplementary methods - structural-overlap metrics, limitation, and contents

Structural-overlap metrics (all exploratory; not validated structure-similarity scores). Both are computed on the alignment-mapped base-pair-probability matrices P_human and P_mouse. Exact structural weighted Jaccard = sum over aligned column pairs of min(P_human(i,j), P_mouse(i,j)) / sum of max(P_human(i,j), P_mouse(i,j)).

+/-2-column tolerant weighted overlap. Candidate matches are the high-probability predicted pairs (P>=0.1) in each matrix. A human pair (i,j) and a mouse pair (k,l) are compatible when BOTH aligned endpoints differ by at most two alignment columns (|i-k|<=2 and |j-l|<=2). Compatible pairs are matched ONE-TO-ONE by maximum-weight bipartite matching (scipy.optimize.linear_sum_assignment over the compatibility-restricted weight matrix; incompatible entries have weight 0 and are excluded from the reported matches). The weight of a matched pair is the smaller of the two base-pair probabilities, min(P_human, P_mouse). The score is the sum of matched weights normalized by the smaller of the two total high-probability masses, min(sum P_human, sum P_mouse), yielding a value in [0,1]. The metric is symmetric in the two species.

Limitation. Structural comparison is conditional on the nucleotide alignment used to define shared coordinates. Weak structural overlap may therefore reflect either divergent predicted pairing or misregistration of structurally corresponding positions, especially at low sequence similarity.

These structural-overlap values do not establish structural conservation, RNA function, or evolutionary orthology.

Contents: S1-S3, short-ncRNA case studies (RNU6-7, SNORD57, vault RNA); S4-S7, MALAT1 recurrent cores R5/R4/R3/R2 on the core matched window; Table S1, MALAT1 boundary-stability summary. Per-window / per-method values and the window-stability diagnostic galleries are provided as source data under analysis/vienna_structure_examples/.

RNU6-7.1 | aligned span 103 nt | nucleotide identity 1.00 | normalized nucleotide-SW 0.96 | short-branch MMD 0.000
Structural overlap (exploratory): exact wJaccard 0.94 | +/-2 tolerant overlap 1.00

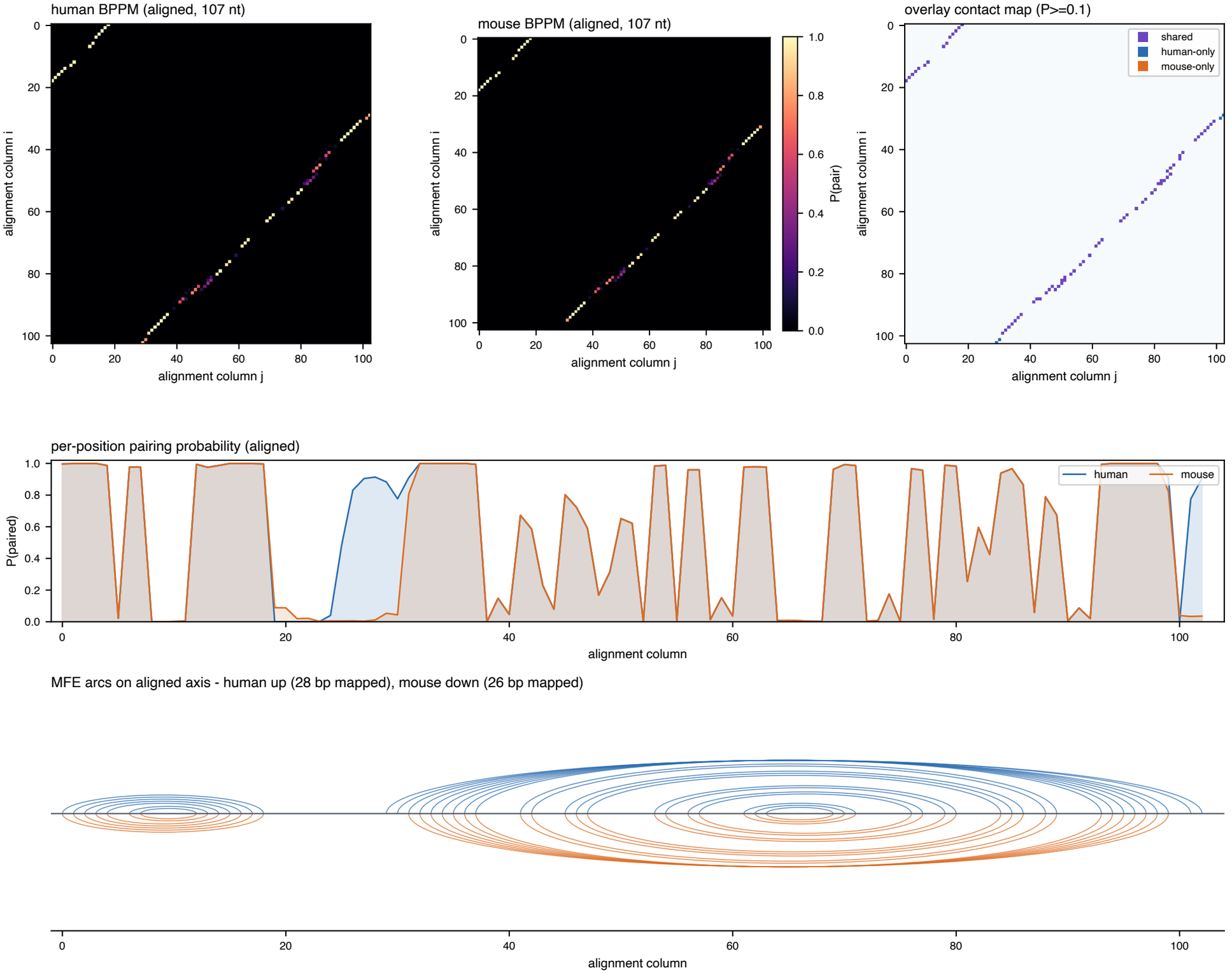

Figure S1. RNU6-7.1 (human vs mouse), full predicted interval; alignment-aware comparison of ViennaRNA-predicted pairing (aligned base-pair-probability matrices, overlay contact map at P>=0.1, per-position pairing probability, mirrored MFE arc diagrams). Sequence metrics (CURIA baseline): nucleotide identity 1.00, normalized nucleotide-SW 0.96; short-branch MMD 0.000 (embedding-SW distance not applicable to the short-ncRNA branch). Exploratory structural overlap: exact weighted Jaccard 0.94, +/-2-column tolerant weighted overlap 1.00. Near-identical ViennaRNA-predicted pairing after alignment-aware mapping; included as a positive control. Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

SNORD57.7 | aligned span 69 nt | nucleotide identity 0.80 | normalized nucleotide-SW 0.67 | short-branch MMD 0.000
Structural overlap (exploratory): exact wJaccard 0.10 | +/-2 tolerant overlap 0.44

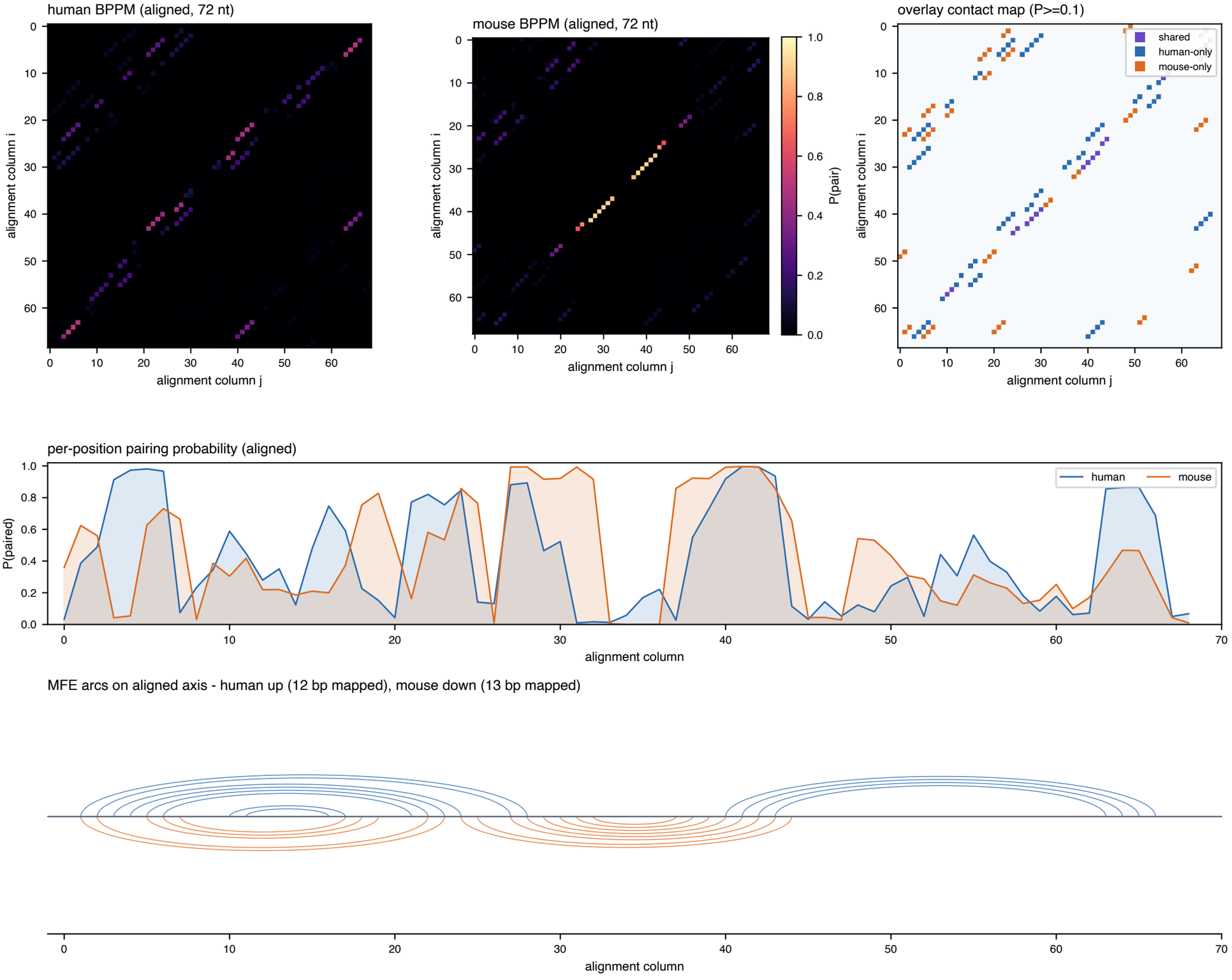

Figure S2. SNORD57.7 (human vs mouse), full predicted interval; alignment-aware comparison of ViennaRNA-predicted pairing (aligned base-pair-probability matrices, overlay contact map at P>=0.1, per-position pairing probability, mirrored MFE arc diagrams). Sequence metrics (CURIA baseline): nucleotide identity 0.80, normalized nucleotide-SW 0.67; short-branch MMD 0.000 (embedding-SW distance not applicable to the short-ncRNA branch). Exploratory structural overlap: exact weighted Jaccard 0.10, +/-2-column tolerant weighted overlap 0.44. Partial predicted structural agreement. Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

VTRNA(vaultRNA).57 | aligned span 83 nt | nucleotide identity 0.63 | normalized nucleotide-SW 0.35 | short-branch MMD 0.137
Structural overlap (exploratory): exact wJaccard 0.00 | +/-2 tolerant overlap 0.00

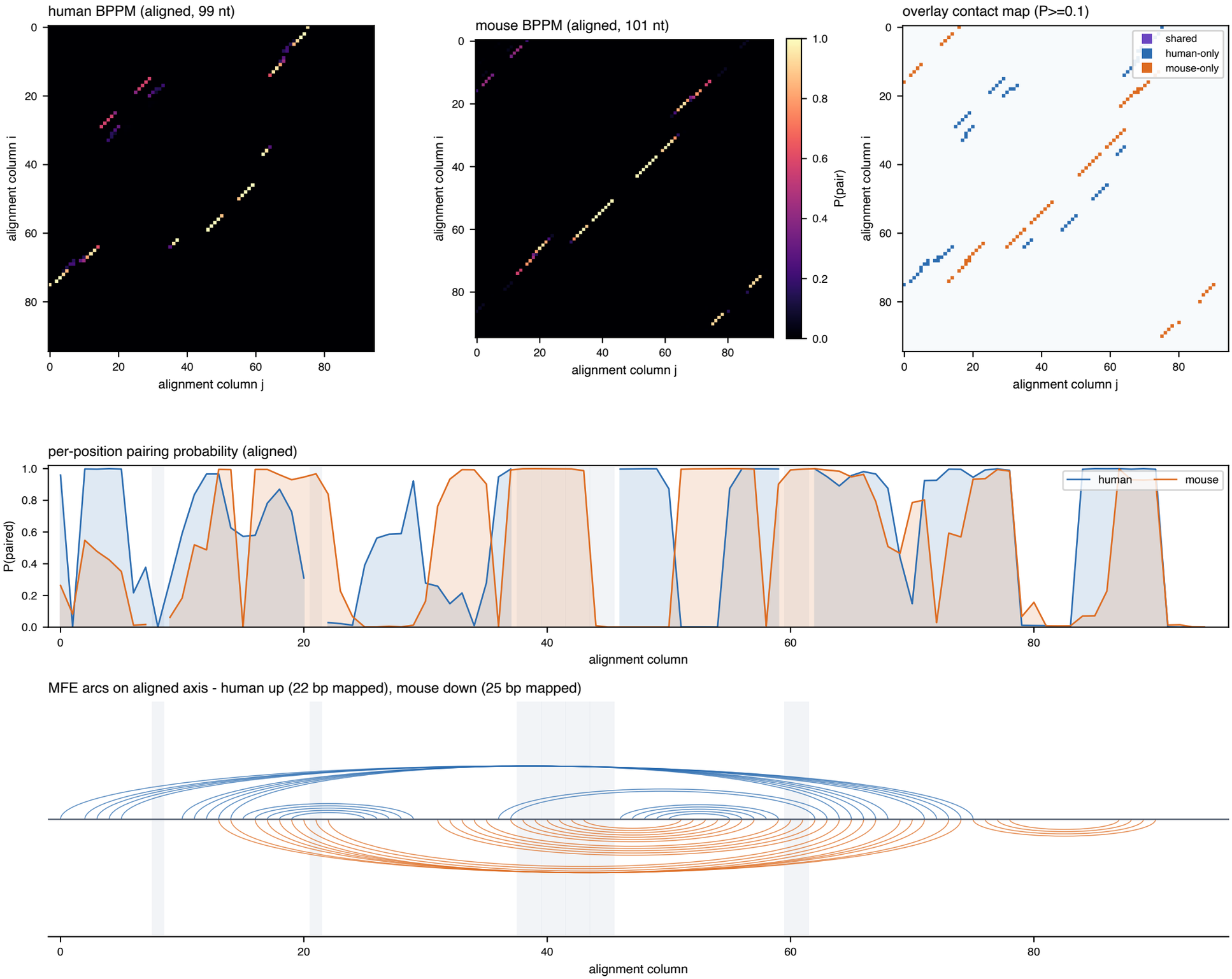

Figure S3. VTRNA(vaultRNA).57 (human vs mouse), full predicted interval; alignment-aware comparison of ViennaRNA-predicted pairing (aligned base-pair-probability matrices, overlay contact map at P>=0.1, per-position pairing probability, mirrored MFE arc diagrams). Sequence metrics (CURIA baseline): nucleotide identity 0.63, normalized nucleotide-SW 0.35; short-branch MMD 0.137 (embedding-SW distance not applicable to the short-ncRNA branch). Exploratory structural overlap: exact weighted Jaccard 0.00, +/-2-column tolerant weighted overlap 0.00. Both ViennaRNA-predicted folds contain prominent long-range pairing, but alignment-aware contact correspondence remains weak even after allowing small register shifts. Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

MALAT1 R5 | aligned span 285 nt | nucleotide identity 0.73 | normalized nucleotide-SW 0.61 | island embedding-SW distance 0.013
Structural overlap (exploratory): exact wJaccard 0.19 | +/-2 tolerant overlap 0.37

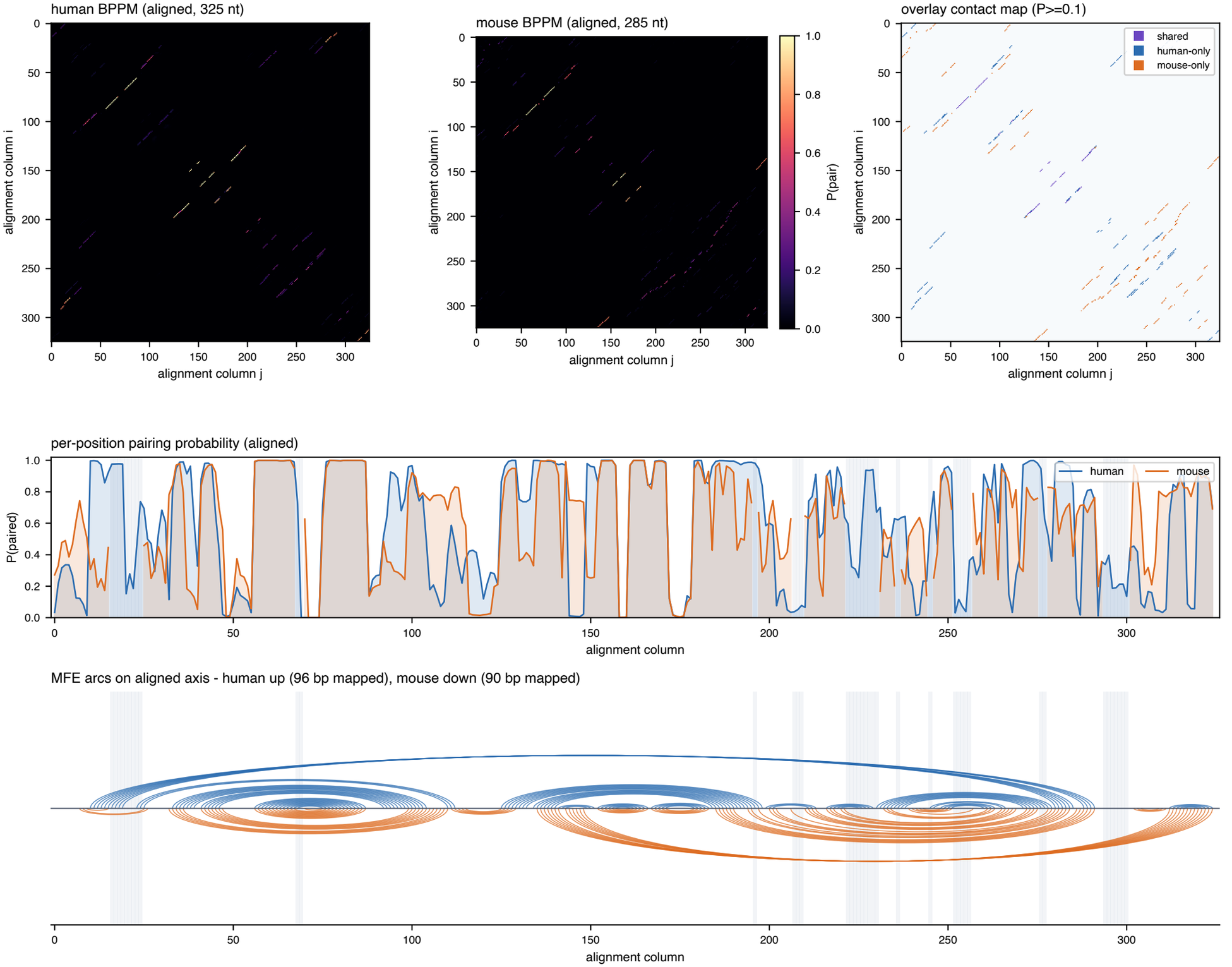

Figure S4. MALAT1 core R5 matched window (human 325 nt vs mouse 285 nt; alignment coverage 0.88/1.00) - the length-matched homologous window, not the raw unequal island intervals. Sequence metrics (CURIA baseline): nucleotide identity 0.73, normalized nucleotide-SW 0.61; island embedding-SW distance 0.013 (MMD is the short-branch term and is not used here). Exploratory structural overlap: exact weighted Jaccard 0.19 (range across core/+-20/+-40 nt 0.19-0.23); +/-2-column tolerant weighted overlap 0.37. This core showed the strongest and most stable moderate predicted structural agreement across the tested boundary windows, and was consistent across the tested folding procedures in the source data (malat1_windows.tsv). Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

MALAT1 R4 | aligned span 156 nt | nucleotide identity 0.86 | normalized nucleotide-SW 0.74 | island embedding-SW distance 0.017
Structural overlap (exploratory): exact wJaccard 0.02 | +/-2 tolerant overlap 0.00

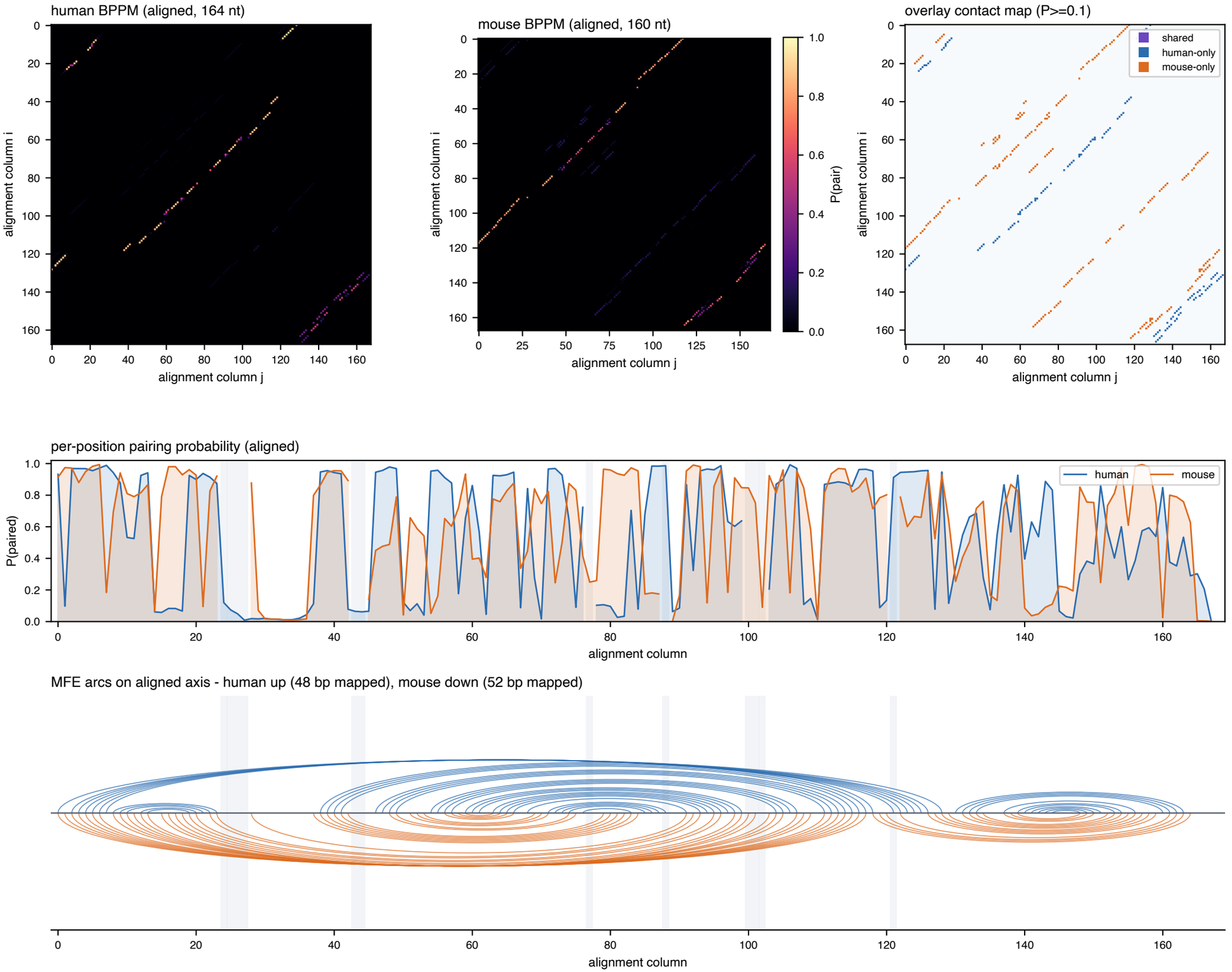

Figure S5. MALAT1 core R4 matched window (human 164 nt vs mouse 160 nt; alignment coverage 0.95/0.97) - the length-matched homologous window, not the raw unequal island intervals. Sequence metrics (CURIA baseline): nucleotide identity 0.86, normalized nucleotide-SW 0.74; island embedding-SW distance 0.017 (MMD is the short-branch term and is not used here). Exploratory structural overlap: exact weighted Jaccard 0.02 (range across core/+-20/+-40 nt 0.02-0.14); +/-2-column tolerant weighted overlap 0.00. Weak and boundary-dependent predicted agreement; global and local folding results differ. Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

MALAT1 R3 | aligned span 157 nt | nucleotide identity 0.63 | normalized nucleotide-SW 0.38 | island embedding-SW distance 0.046
Structural overlap (exploratory): exact wJaccard 0.05 | +/-2 tolerant overlap 0.15

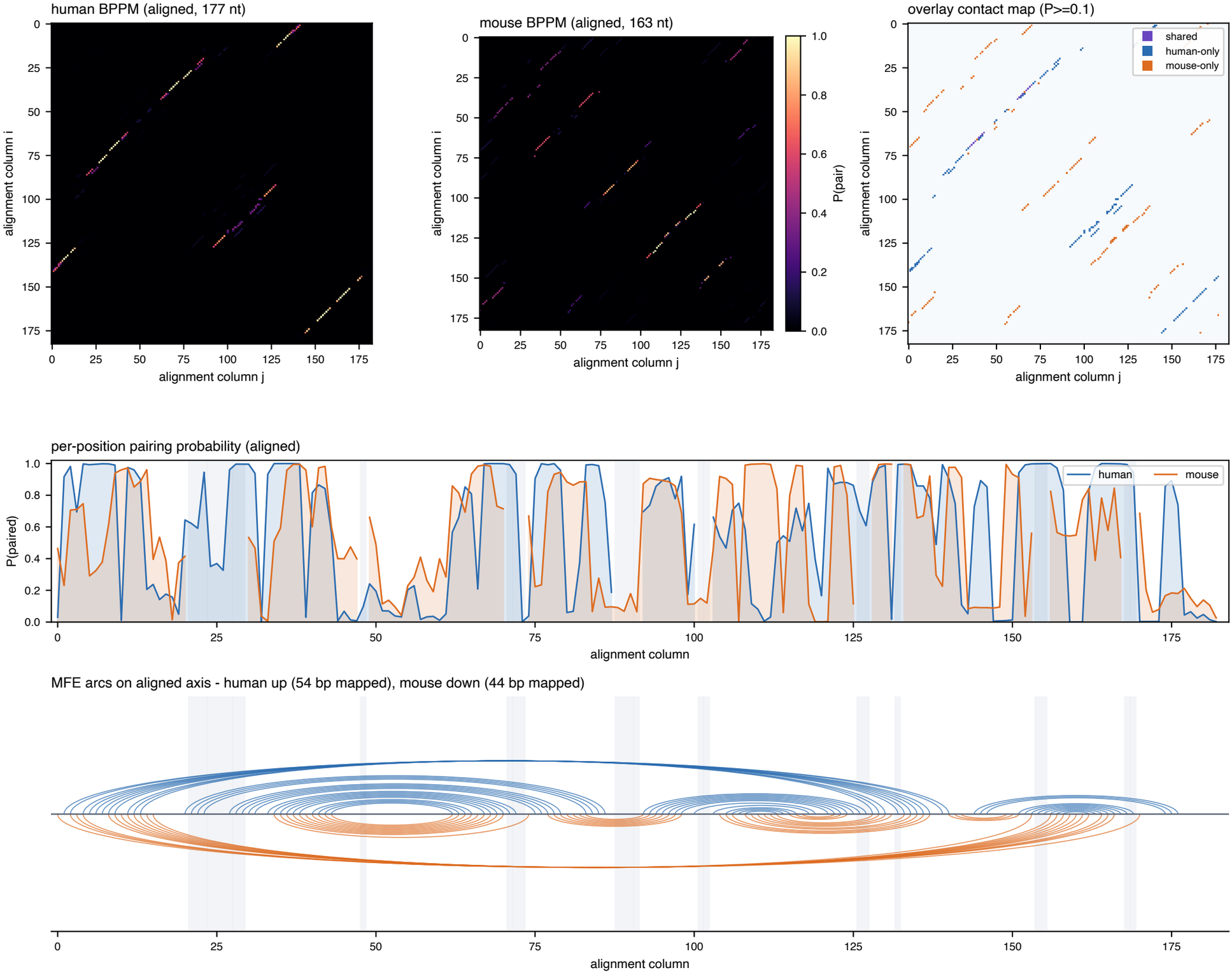

Figure S6. MALAT1 core R3 matched window (human 177 nt vs mouse 163 nt; alignment coverage 0.89/0.96) - the length-matched homologous window, not the raw unequal island intervals. Sequence metrics (CURIA baseline): nucleotide identity 0.63, normalized nucleotide-SW 0.38; island embedding-SW distance 0.046 (MMD is the short-branch term and is not used here). Exploratory structural overlap: exact weighted Jaccard 0.05 (range across core/+-20/+-40 nt 0.04-0.05); +/-2-column tolerant weighted overlap 0.15. Weak predicted agreement, inconsistent between global and local folding. Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

MALAT1 R2 | aligned span 227 nt | nucleotide identity 0.60 | normalized nucleotide-SW 0.34 | island embedding-SW distance 0.033
Structural overlap (exploratory): exact wJaccard 0.01 | +/-2 tolerant overlap 0.04

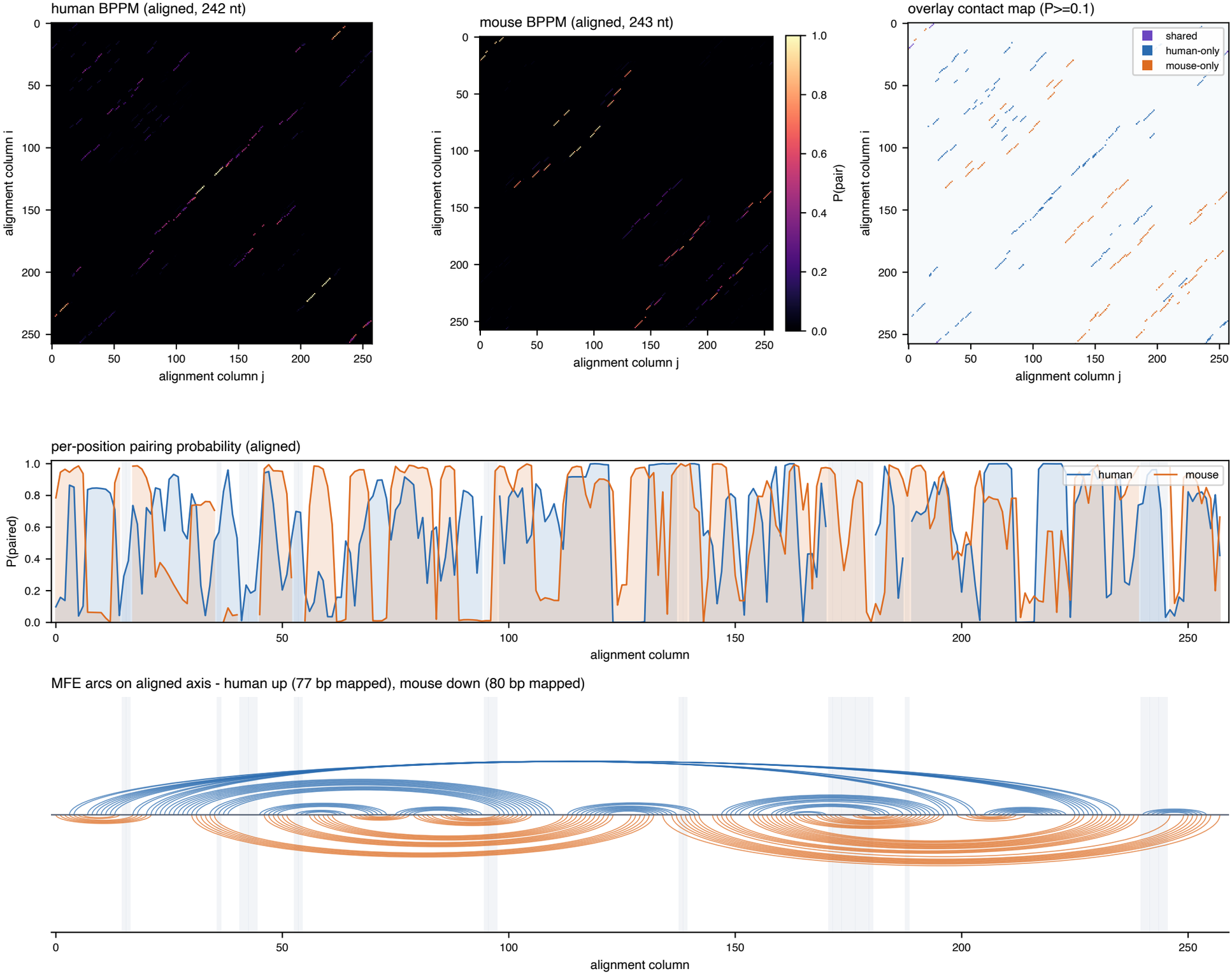

Figure S7. MALAT1 core R2 matched window (human 242 nt vs mouse 243 nt; alignment coverage 0.94/0.93) - the length-matched homologous window, not the raw unequal island intervals. Sequence metrics (CURIA baseline): nucleotide identity 0.60, normalized nucleotide-SW 0.34; island embedding-SW distance 0.033 (MMD is the short-branch term and is not used here). Exploratory structural overlap: exact weighted Jaccard 0.01 (range across core/+-20/+-40 nt 0.01-0.01); +/-2-column tolerant weighted overlap 0.04. Weak predicted agreement, consistently near zero across the tested windows. Exploratory structural-overlap metric; see the introductory note for definitions and caveats.

Table S1. MALAT1 recurrent cores - boundary stability & sequence metrics

Core	Nucleotide identity	Normalized nt-SW	Embedding-SW distance	Exact structural wJaccard (core; range)	+/-2 tolerant structural overlap (core; range)	Assessment
R5	0.73	0.61	0.013	0.19 (0.19-0.23)	0.37 (0.37-0.46)	moderate; boundary-stable
R4	0.86	0.74	0.017	0.02 (0.02-0.14)	0.00 (0.00-0.33)	weak; boundary-dependent
R3	0.63	0.38	0.046	0.05 (0.04-0.05)	0.15 (0.11-0.15)	weak; inconsistent
R2	0.60	0.34	0.033	0.01 (0.01-0.01)	0.04 (0.04-0.05)	weak; consistently near zero

Table S1. MALAT1 recurrent cores on length-matched windows (aligned span +/- 0/20/40 nt). Sequence metrics are the CURIA baseline nucleotide identity and normalized nucleotide-SW; the embedding-SW distance is the RiNALMo island metric (nucleotide-SW and embedding-SW are distinct and reported separately). The exploratory structural-overlap metrics (exact weighted Jaccard; +/-2-column tolerant weighted overlap) are given as the core-window value and its range across the three windows; they are descriptive measures, not validated structure-similarity scores. RNAplfold local-folding results were used as an additional boundary-sensitivity check; complete per-window and per-method values are provided in malat1_windows.tsv.

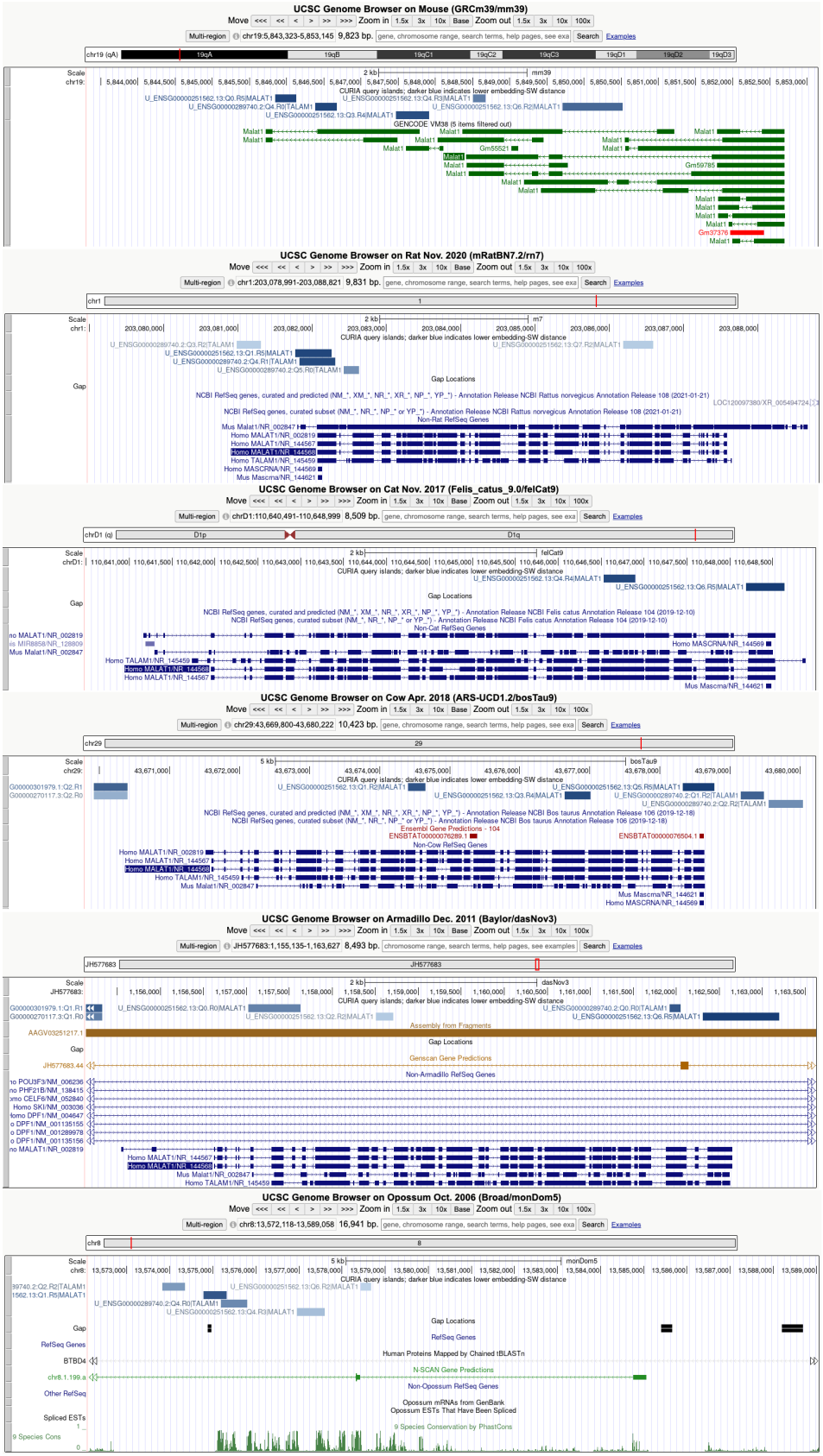

Figure S8. Representative locus-level recovery of MALAT1 across mammalian query genomes. Stacked UCSC Genome Browser views show accepted CURIA query islands for mouse, rat, cat, cow, armadillo, and opossum. Darker blue indicates lower embedding-Smith-Waterman distance. The same localized terminal reference cores are recurrently recovered across the species panel, including genomes in which a native MALAT1 annotation is absent or incomplete. Gene annotations and cross-species transcript tracks are shown where available. These views are intended as qualitative locus-level examples and do not by themselves establish conserved molecular function.

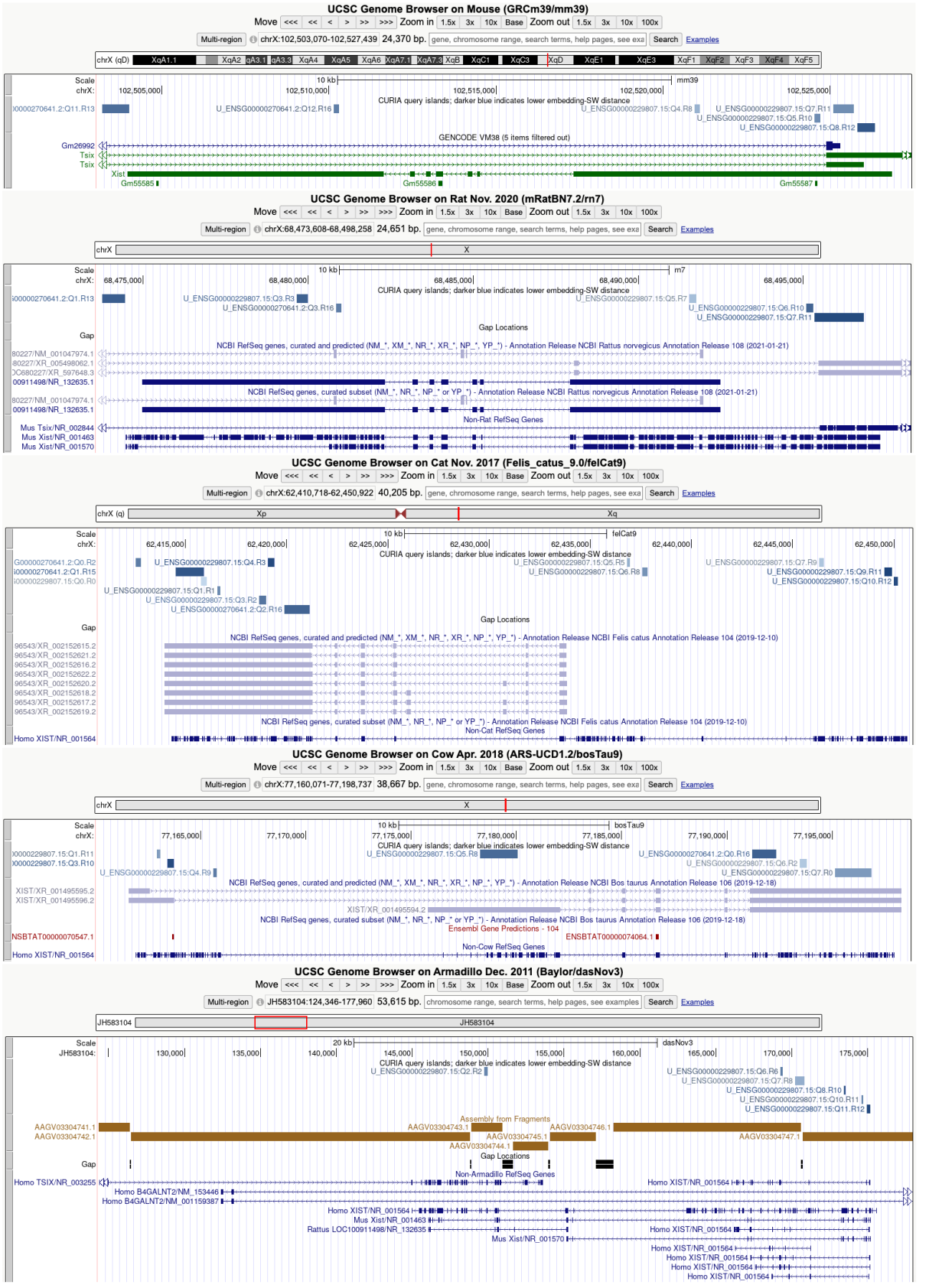

Figure S9. Representative locus-level recovery of XIST across placental mammalian query genomes. Stacked UCSC Genome Browser views show accepted CURIA query islands for mouse, rat, cat, cow, and armadillo. Darker blue indicates lower embedding-Smith-Waterman distance. Several recurrent reference cores cluster within a localized portion of the XIST locus across species, whereas other internal regions are recovered less consistently. Gene annotations and cross-species transcript tracks are shown where available. No correspondence to a specific named XIST repeat or functional domain is asserted here.