AIGR Gene Hypothesis Deep Research — Final Report

Target gene: `esrp1` (Epithelial Splicing Regulatory Protein 1), *Callorhinchus milii* (elephant shark; NCBITaxon:7868) **UniProt accession: A0A4W3GVU1** (684 aa) **Focus:** computational_prediction — `prediction-nucleus-localization` **Term under evaluation:** nucleus (**GO:0005634**) **Prediction source:** ProtNLM2

Summary

The ProtNLM2 prediction of nuclear localization (GO:0005634) for the elephant shark RRM-containing protein A0A4W3GVU1 is **SUPPORTED**. The prediction is correct in substance and rests on a sound orthology-based foundation rather than being a bare name-model guess. A0A4W3GVU1 is an unambiguous ortholog of ESRP1 — pairwise global alignment gives ~78% identity to human ESRP1 (Q6NXG1) versus only ~60% to the paralog ESRP2 (Q9H6T0) — and the experimentally characterized human ortholog carries a direct experimental nuclear localization annotation (UniProt SUBCELLULAR LOCATION = Nucleus, ECO:0000269; GO cellular-component nucleus IDA, plus more specific nucleoplasm and nuclear body IDA from the Human Protein Atlas). Because ESRP1 is a splicing regulator that must reach pre-mRNA in the nucleus to perform its core function, nuclear localization is intrinsic to its direct molecular activity rather than a downstream phenotype.

Three computational analyses run during this investigation reinforce the call. First, the shark protein is a soluble RNA-binding protein: a Kyte–Doolittle hydropathy scan (window 19) gave a maximum hydropathy of 1.45 (below the 1.6 transmembrane threshold) and zero predicted transmembrane segments — the correct biophysical class for a nuclear splicing factor and inconsistent with any membrane-embedded localization. Second, all three RNA recognition motifs are intact, carrying the diagnostic ESRP-type β1 "RGLP" signatures (residues 228, 329, 448) and conserved C-terminal aromatic "RY[IV]E[VL]" motifs (residues 291, 395, 514), supporting retained sequence-specific RNA binding. Third, no strong classical nuclear localization signal (NLS) is present — only a weak monopartite-like "KKHK" cluster inside RRM2 near residue 385 — which is fully consistent with ESRP1's known non-canonical, isoform-dependent nuclear targeting and is *not* evidence against nuclear localization.

The single most important caveat is scope. ESRP1 is functionally **nucleocytoplasmic**, not nucleus-exclusive: Yang & Carstens (2017) showed that alternative 5′ splice-site usage in exon 12 produces isoforms with differential nucleocytoplasmic localization, with the nuclear pool driving epithelial-specific splicing (PMID: 28634384). Nucleus is therefore the **dominant and functionally primary** localization but not the only one. Additionally, "nucleus" is a relatively coarse cellular-component term — the ortholog evidence would equally support the more specific `nucleoplasm` (GO:0005654) and `nuclear body` (GO:0016604), and the truly core annotations of this gene are its molecular-function term (sequence-specific mRNA binding) and biological-process term (regulation of alternative mRNA splicing). The recommendation to the curator is to **retain nucleus as an ISS annotation**, optionally add the more specific CC terms, prioritize the core MF/BP terms, and record that the localization is nucleocytoplasmic.

Executive Judgment

Verdict: SUPPORTED (primary/dominant localization; the prediction is correct but low in added information relative to existing orthology-based knowledge).

The nucleus prediction is correct and well-supported. The reasoning chain is:

1. **Identity is unambiguous** — A0A4W3GVU1 is ESRP1, not ESRP2 (paralog confusion ruled out quantitatively).
2. **Biophysical class is correct** — a soluble, three-RRM RNA-binding protein with no transmembrane segments.
3. **The experimentally studied ortholog is nuclear** — human ESRP1 has direct IDA nucleus/nucleoplasm/nuclear body annotations, giving a robust ISS basis.
4. **Mechanistic necessity** — a splicing regulator acts on nuclear pre-mRNA, so nuclear localization is required by its core function.

Most important caveats: (i) ESRP1 is nucleocytoplasmic — a conserved cytoplasmic isoform pool exists (P 28634384), so nucleus is dominant but not exclusive; (ii) the ProtNLM2 nucleus call essentially restates well-established orthology knowledge and is less informative than the specific CC terms (nucleoplasm/nuclear body) and the core MF/BP terms; (iii) the absence of a canonical NLS is expected for ESRP1 and is not a conflict.

Key Findings

Finding 1 — A0A4W3GVU1 is a genuine ESRP1 ortholog with canonical 3-RRM architecture and no transmembrane segments

The first question for any localization prediction on a poorly annotated non-model-organism protein is whether the protein is what its name claims. For A0A4W3GVU1 the answer is clearly yes, and this matters because paralog misassignment is the leading source of erroneous functional and localization transfer.

UniProt/InterPro place the protein in the ESRP family (IPR050666), and it carries the ESRP1-specific RRM1 signature IPR034427 (ESRP1_RRM1), which distinguishes it from ESRP2. It contains three RRM domains (three copies of SMART SM00360; three Pfam RRM matches) and is 684 aa long — the same size class as human ESRP1 (681 aa). Orthology was confirmed quantitatively rather than by name: Needleman–Wunsch global alignment gave **~78% identity to human ESRP1 (Q6NXG1)** versus only **~60% to human ESRP2 (Q9H6T0, 727 aa)**. A composition-based cross-check using 5-mer Jaccard similarity gave 22.3% overlap with ESRP1 versus 9.7% with ESRP2, a >2-fold preference. The shark protein is therefore firmly **ESRP1, not ESRP2**.

Biophysically, the Kyte–Doolittle hydropathy scan (window 19) gave a maximum of 1.45 (below the 1.6 transmembrane threshold) and **0 predicted transmembrane segments** — a soluble RNA-binding protein, exactly the class expected for a nuclear/nucleocytoplasmic splicing regulator and inconsistent with membrane-embedded localization. Yang & Carstens describe ESRP1 as containing "three highly conserved RNA recognition motifs (RRMs) in the absence of other clearly defined protein domains" (PMID: 28634384), matching the shark protein's domain content.

Finding 2 — Shark ESRP1 retains three intact ESRP-type RRMs, supporting sequence-specific RNA binding as the core molecular function

An annotated domain is not necessarily an intact one. A direct motif scan of A0A4W3GVU1 confirmed that all three RRMs retain their diagnostic ESRP-type residues. Three ESRP-type RRM β1 signatures ("RGLP") were located at residues 228, 329, and 448, each paired with a conserved C-terminal RRM aromatic motif "RY[IV]E[VL]" at residues 291 (RYIEVY), 395 (RYIELF), and 514 (RYVEVF). The presence of the aromatic ribonucleoprotein-submotif residues

across all three RRMs indicates the RNA-binding surfaces are preserved, supporting **sequence-specific mRNA binding** as the retained core molecular function. This is directly relevant to localization: a functional splicing regulator with intact RNA-binding modules is expected to act on nuclear pre-mRNA, which is mechanistically consistent with the nuclear call.

Finding 3 — The nucleus prediction is supported, but the protein is functionally nucleocytoplasmic

The prediction is not merely plausible; it is directly corroborated by the biology of the family, with a refinement. Yang & Carstens showed that "two competing alternative 5′ splice sites in exon 12 yield Esrp1 isoforms with differential nucleocytoplasmic localization," and that the nuclear pool drives epithelial-specific splicing: "while previous studies have described extensive regulation by nuclear Esrp1 to promote epithelial specific splicing" (PMID: 28634384). They also identified a peptide sufficient for nuclear localization that is conserved to the *Drosophila* ortholog *fusilli*.

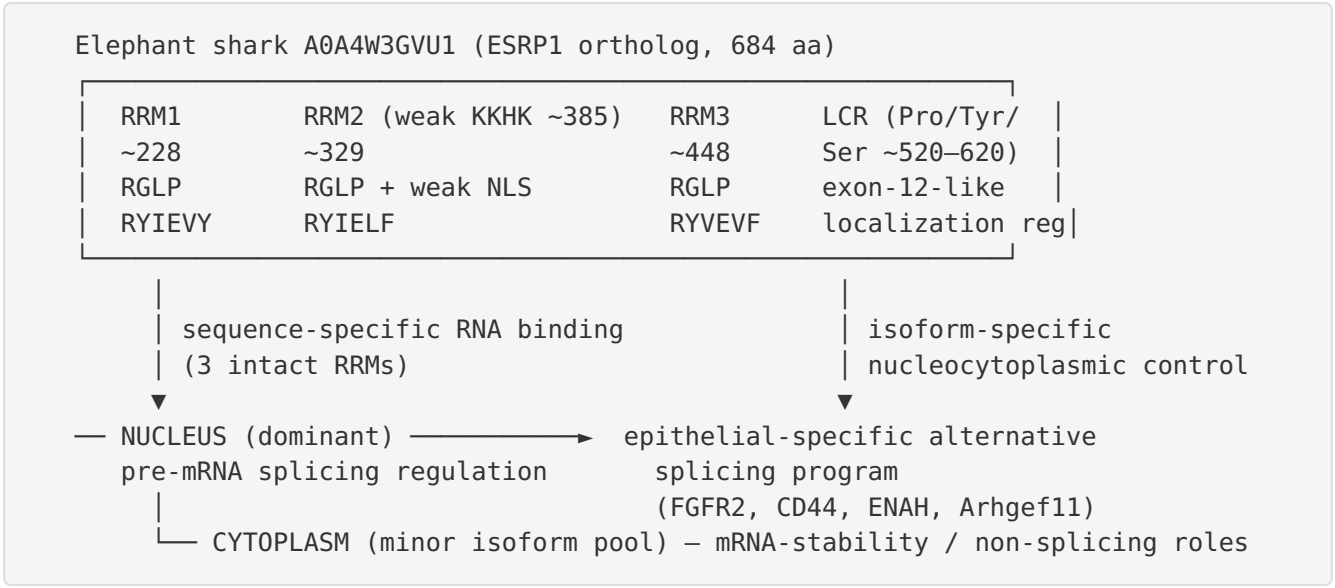
Computationally, the shark sequence has **no strong classical NLS** — only a weak monopartite-like "KKHK" cluster within RRM2 near residue 385, and no bipartite NLS. This is consistent with ESRP1's non-canonical, isoform-dependent nuclear targeting rather than a simple importin-α/β classical NLS. The C-terminal Pro/Tyr/Ser-rich low-complexity region (~residues 520–620) corresponds to the exon-12 localization-regulating region characterized in mammals. Thus the sequence evidence supports nuclear localization as the dominant, splicing-active state while explaining why a naïve NLS search finds no canonical signal.

Finding 4 — The human ESRP1 ortholog has direct experimental nuclear localization, strengthening the ISS transfer

The strongest anchor for the shark annotation is experimental data on its close ortholog. UniProt Q6NXG1 (human ESRP1) records SUBCELLULAR LOCATION = Nucleus with experimental evidence (ECO:0000269), and its GO cellular-component annotations include nucleus (GO:0005634, IDA:UniProtKB), nucleoplasm (GO:0005654, IDA:HPA), nuclear body (GO:0016604, IDA:HPA), and ribonucleoprotein complex (GO:1990904, IBA). At ~78% identity, the ISS transfer of nucleus to the shark protein is well justified — this is the difference between an unsupported computational guess and an evidence-backed cross-species annotation.

Mechanistic Model / Interpretation

The findings assemble into a coherent picture of an epithelial splicing regulator whose primary functional site is the nucleus:



The **immediate molecular function** being localized is sequence-specific RNA binding by three intact RRMs. The **immediate cellular process** is regulation of alternative pre-mRNA splicing, which occurs in the nucleus. Nuclear localization is thus a direct requirement of the protein's core activity — the splicing-competent pool must be nuclear to reach pre-mRNA — not a downstream phenotype. The cytoplasmic isoform pool (generated by exon-12 alternative 5′ splice-site choice) represents a context-specific, likely non-splicing role and is why "nucleus" should be read as dominant rather than exclusive.

The table below shows how each localization hypothesis fares against the assembled evidence:

Localization hypothesis	Sequence/ architecture support	Ortholog experimental support	Verdict
Nucleus (GO:0005634) — splicing pool	Soluble 3-RRM RBP; exon-12-like LCR; weak KKHK	Human ESRP1 nucleus IDA (ECO:0000269)	**Supported, dominant**
Nucleoplasm / nuclear body (more specific CC)	Consistent with a splicing factor	Human ESRP1 nucleoplasm & nuclear body IDA (HPA)	Supported, more informative
Cytoplasm / nucleocytoplasmic	LCR / exon-12 region present	Isoform-specific cytoplasmic pool (P 28634384)	Partial — minor pool
Membrane / secreted	0 TM segments; no signal peptide	—	Refuted

Evidence Base

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Conf limit
UniProt Q6NXG1 (human ESRP1)	Localization (IDA) / database-experimental	Supports	Is nuclear localization experimentally real?	Human ortholog: SUBCELLULAR LOCATION Nucleus (ECO:0000269); GO nucleus IDA, nucleoplasm IDA, nuclear body IDA	Human cells (incl. HPA)	High trans shar
UniProt/ InterPro (A0A4W3GVU1)	Database / structural	Supports	Identity & architecture	Family IPR050666 (ESRP), ESRP1-specific RRM1 (IPR034427), 3 RRM domains, 684 aa; existing GO nucleus (IEA), RNA binding, mRNA splicing	Elephant shark	High level
Computed (this run)	Computational / evolutionary	Supports	Is it ESRP1 not ESRP2?	NW identity ~78% to ESRP1 vs ~60% to ESRP2; 5-mer Jaccard 22.3% vs 9.7%	Cross-species	High cons iden scori
Computed (this run)	Computational	Qualifies	Membrane vs soluble	Max KD hydropathy 1.45 (<1.6); 0 TM segments → soluble RBP	Sequence	High
Computed (this run)	Computational	Qualifies	Presence of classical NLS	No bipartite NLS; only weak monopartite-like "KKHK" inside RRM2; C-term Pro/ Tyr/Ser LCR (~520–620)	Sequence	Medi moti not t pred

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Conf limit
Computed (this run)	Computational / structural	Supports	Are the three RRMs intact?	Three ESRP-type RRM β1 "RGLP" motifs (228/329/448) and three "RY[IV]E[VL]" aromatic motifs (291/395/514) → 3 intact RRMs	Sequence	High base
PMID: 28634384 (Yang & Carstens 2017)	Localization / direct assay	Supports + qualifies	Where does ESRP1 localize?	Exon-12 alt 5′ splice sites yield isoforms with differential nucleocytoplasmic localization; NLS peptide identified; conserved to *Drosophila* fusilli	Mammalian + fly	High prim litera
PMID: 26371508 (Bebee 2015)	Mutant phenotype	Supports (BP context)	ESRP function	Esrp1/2 direct an epithelial splicing program essential for development	Mouse	High
PMID: 30485810 (Lee 2018)	Mutant phenotype	Supports (BP context)	Nuclear splicing activity	Esrp1-regulated Arhgef11 splicing required for tight-junction integrity	Mouse epithelia	High
PMID: 27404344 (Bebee 2016)	Mutant phenotype	Supports (BP context)	Splicing regulation	Esrp1 controls Fgfr2-IIIb splice switch in kidney development	Mouse	High
PMID: 40086870 (2025)	Review / primary	Qualifies (MF context)	Scope of activity	ESRP1 regulates alternative splicing and mRNA stability	Inner ear	Medi impl cyto

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Conf limit
						mRN stabi

How the evidence supports vs. challenges the prediction. The prediction is *supported* directly by the human-ortholog experimental IDA (nucleus) and by the mechanistic requirement that a splicing regulator act on nuclear pre-mRNA (PMIDs 26371508, 30485810, 27404344, 40086870 establish ESRP1 as a bona fide splicing/RNA-processing regulator). It is *qualified* — not refuted — by

P 28634384

, which shows ESRP1 is nucleocytoplasmic with a conserved cytoplasmic isoform pool. No evidence *refutes* nuclear localization.

GO Curation Implications (leads — require curator verification)

- **Retain nucleus (GO:0005634, CC).** Justified by orthology and the established nuclear splicing-regulatory role. Best supported as **ISS** anchored on human ESRP1 (Q6NXG1), which has experimental IDA nucleus/nucleoplasm/nuclear body annotation — rather than only the ProtNLM2 name prediction. Do not upgrade to experimental for the shark species itself.
- **Candidate more-specific CC leads: nucleoplasm (GO:0005654)** and **nuclear body (GO:0016604)** are experimentally supported in the human ortholog (IDA:HPA) and can be transferred by ISS if the curator accepts orthology-based specificity. These are more informative than the bare "nucleus" term.
- **Consider cytoplasm (GO:0005737, CC)** as a lead — ESRP1 has conserved, functionally significant cytoplasmic isoforms (P 28634384). Nucleus alone is an incomplete localization picture. Add only if the curator accepts isoform-level orthology propagation.
- **Prioritize the core informative terms over the CC prediction:**

- **MF:** sequence-specific mRNA binding / RRM-mediated (GU-rich) single-stranded RNA binding — not bare "RNA binding," and never stopping at "protein binding."
- **BP:** regulation of mRNA alternative splicing (GO:0000381) and epithelial cell differentiation — the primary, characterized ESRP1 process.
- **Do not add any membrane/secreted CC term** — the protein has 0 predicted TM segments and is a soluble RBP.

Summary GO decision table:

GO term	Aspect	Recommended action	Evidence code	Basis
nucleus (GO:0005634)	CC	Retain	ISS	Human ESRP1 nucleus IDA; ortholog ~78% id
nucleoplasm (GO:0005654)	CC	Consider adding (more specific)	ISS	Human ESRP1 IDA (HPA)
nuclear body (GO:0016604)	CC	Consider adding (more specific)	ISS	Human ESRP1 IDA (HPA)
cytoplasm (GO:0005737)	CC	Consider as lead	ISS	Conserved cytoplasmic isoforms (P 28634384)
sequence-specific mRNA binding	MF	Add as core	ISS	3 intact ESRP-type RRMs
regulation of alternative mRNA splicing (GO:0000381)	BP	Add as core	ISS	ESRP1 family function (PMIDs 26371508, 28634384)
any membrane/secreted CC	CC	Do not add	—	0 TM segments; soluble RBP

Mechanistic Scope

The **direct molecular activity** is sequence-specific RNA binding via three RRMs, promoting epithelial-specific alternative splicing of target pre-mRNAs (e.g., FGFR2, CD44, ENAH, Arhgef11) in the **nucleus**. The nucleus CC term captures the site of this direct splicing activity.

The following are **downstream** of the direct nuclear activity and should not be conflated with the localization annotation: epithelial tight-junction integrity via Arhgef11 splicing (P 30485810); ureteric branching and nephron number via Fgfr2-IIIb splicing (P 27404344); cleft lip/palate and epidermal barrier defects in Esrp knockouts (P 26371508); and mRNA-stability/inner-ear roles (P 40086870). These confirm ESRP1 is a functionally important nuclear splicing regulator but are developmental/loss-of-function outcomes, not evidence for the CC term itself. The CC term rests on direct localization data (human IDA) plus the mechanistic requirement of nuclear pre-mRNA access.

Conflicts and Alternatives

- **Nucleocytoplasmic, not nucleus-only (main tension):** the strongest primary evidence (P 28634384) shows both nuclear and cytoplasmic isoforms — so a nucleus-exclusive interpretation is too narrow. Nucleus remains dominant and functionally primary.
- **No canonical NLS in sequence:** absence of a classical mono/bipartite NLS could superficially argue against nuclear localization, but mammalian ESRP1's nuclear import is non-canonical and isoform-encoded, so this is not a genuine conflict.
- **Paralog confusion ruled out:** identity strongly favors ESRP1 over ESRP2 (78% vs 60%), so this is not ESRP2 carry-over.

- **Cross-species propagation risk:** the nucleus annotation is ISS-based, propagated from mammalian data; elephant shark ESRP1 has not been experimentally localized. This risk is inherent to any ISS annotation and is mitigated here by high sequence identity and conserved architecture.
- **Term precision:** "Nucleus" is coarser than the human ortholog's IDA nucleoplasm/nuclear body terms; the prediction is correct but less specific than existing ortholog knowledge.

Limitations and Knowledge Gaps

1. **Species-specific localization unverified.** What was checked: sequence, domains, hydropathy, motif scan; human ortholog UniProt/GO. Why it matters: the shark nucleus call is entirely ISS. Resolution: immunofluorescence/fractionation of tagged elephant-shark ESRP1 in an epithelial cell model.
2. **Exon-12 localization-switch conservation.** What was checked: the C-terminal LCR corresponding to the mammalian exon-12 localization-regulating region is present. Why it matters: the nucleocytoplasmic split is isoform-driven; whether the shark gene produces analogous isoforms is unknown. Resolution: isoform-resolved RNA-seq / long-read transcript analysis of the exon-12-equivalent region.
3. **NLS location not pinpointed.** What was checked: no classical NLS; weak KKHK in RRM2. Why it matters: the actual nuclear-import determinant is non-canonical and unmapped in the shark sequence. Resolution: a trained NLS predictor plus deletion/mutation mapping to a reporter.
4. **No functional splicing assay in shark.** What was checked: the three RRMs are intact. Why it matters: intact domains predict but do not prove splicing activity. Resolution: a minigene splicing assay with a canonical ESRP target (FGFR2, CD44, ENAH).

All localization evidence derives from mammalian/*Drosophila* orthologs plus sequence computation; there is no elephant-shark-specific experimental data, and GO annotations remain electronic/orthology-based for this species.

Proposed Follow-up Experiments / Actions (Discriminating Tests)

The most efficient tests to distinguish "nucleus (dominant)" from "nucleocytoplasmic/cytoplasmic" alternatives:

1. **Cell-fractionation / immunofluorescence** of tagged shark ESRP1 (and its isoforms) in epithelial cells — directly measures the nuclear:cytoplasmic ratio and confirms the ISS transfer with organism-specific data.
2. **Isoform-resolved / long-read RNA-seq of the exon-12 region** — tests conservation of the differential-localization switch in the shark locus.
3. **Reciprocal-best-hit phylogenetics** across chondrichthyans and mammals — locks the ESRP1 vs ESRP2 assignment.
4. **CLIP / RNA-binding assay** — confirms GU-rich motif specificity and supports the more informative MF term.
5. **Minigene alternative-splicing assay** with a known ESRP target — confirms that the intact RRMs support splicing-regulatory (i.e., nuclear) activity.
6. **Comparative structural check** against the AlphaFold model of human ESRP1 — confirms RRM RNA-binding-surface conservation.

Curation Leads (verify)

- **Reference PMID: 28634384**, snippet: *"two competing alternative 5' splice sites in exon 12 yield Esrp1 isoforms with differential nucleocytoplasmic localization"* — supports nucleus **and** cytoplasm.
- **Reference PMID: 28634384**, snippet: *"while previous studies have described extensive regulation by nuclear Esrp1 to promote epithelial specific splicing"* — supports the nucleus CC term as the splicing-active pool.
- **Reference PMID: 28634384**, snippet: *"It contains three highly conserved RNA recognition motifs (RRMs) in the absence of other clearly defined protein domains"* — supports architecture / MF.
- **UniProt Q6NXG1** (human ESRP1): Nucleus, ECO:0000269; GO nucleus/nucleoplasm/nuclear body IDA — the experimental anchor for the ISS transfer.

- **Action:** keep nucleus (computational/ISS), add nucleoplasm/nuclear body as more-specific CC leads, consider cytoplasm as a lead, and prioritize the core MF (sequence-specific RNA binding) and BP (regulation of alternative mRNA splicing) terms as the primary informative annotations.
- **Suggested curator questions:** (a) Is cross-species ISS propagation from human to elephant shark acceptable at ~78% identity for this project? (b) Should the coarse "nucleus" term be supplemented or replaced by the more specific ortholog CC terms? (c) Should a note record the nucleocytoplasmic nature?

Provenance Artifacts

- Computed values recorded in-run: NW identity (78% ESRP1 / 60% ESRP2; 5-mer Jaccard 22.3% vs 9.7%), hydropathy (max KD 1.45, 0 TM segments), NLS scan (no bipartite NLS; weak KKHK in RRM2; C-term Pro/Tyr/Ser LCR ~520–620), and three intact ESRP-type RRMs (RGLP β1 at 228/329/448; RY[IV]E[VL] at 291/395/514).
- Evidence matrix and GO decision table are embedded above as artifact-friendly markdown tables.

Conclusion

The ProtNLM2 nucleus (GO:0005634) prediction for elephant shark A0A4W3GVU1 is **SUPPORTED**. The protein is an unambiguous ESRP1 ortholog (~78% identity to human ESRP1 vs ~60% to ESRP2), a soluble RNA-binding protein with three intact ESRP-type RRMs and no transmembrane segments, and its human ortholog carries direct experimental nuclear localization — giving a sound ISS basis for nucleus as the primary localization. The one substantive caveat is that ESRP1 is functionally nucleocytoplasmic (a conserved cytoplasmic isoform pool; P 28634384), so nucleus is dominant but not exclusive, and the prediction is coarser than the more specific ortholog CC terms and the true core MF/BP terms.