AIGR Gene Hypothesis Deep Research — *Glycine max* C6T1A2

Focused evaluation of the ProtNLM2 prediction: "negative regulation of abscisic acid-activated signaling pathway" (GO:0009788)

Hypothesis under review: ProtNLM2 predicts that the *Glycine max* C2H2-type zinc finger protein **C6T1A2** functions in negative regulation of the abscisic acid (ABA)-activated signaling pathway (**GO:0009788**), by similarity to *Arabidopsis* ZFP7 (**IPR053266**). C6T1A2 has no curated GOA annotations.

Focus type: computational_prediction · **Term:** GO:0009788 · **Organism:** *Glycine max* (NCBITaxon:3847)

Summary

C6T1A2 (C6T1A2_SOYBN) is, without ambiguity, a **plant Q-type (QALGGH) single-domain C2H2 zinc-finger transcription factor** and a genuine member of the InterPro **IPR053266 (ZFP7) / ZFP1–7 subfamily**. In *Arabidopsis*, several members of this subfamily (ZFP1, ZFP3, ZFP4, ZFP6, ZFP7) act as negative regulators of ABA-activated signaling when overexpressed. In that sense the ProtNLM2 prediction is **directionally plausible** — it places the protein in the correct structural family and points at a process the family is genuinely associated with.

The problem is specificity. The **exact GO:0009788 role cannot be established from sequence, orthology, or domain evidence alone**, and it should not be promoted to a direct annotation. Three independent computational lines undercut the specific claim: (1) C6T1A2 is only ~45% identical to *Arabidopsis* ZFP7 and is essentially **tied** with ZFP2 (46.0%), a subfamily member not primarily characterized as an ABA repressor — so no confident 1:1 ortholog can be assigned; (2) the entire *Arabidopsis* ABA-repressor phenotype rests on **redundant overexpression (gain-of-function)**, with loss-of-function lines near wild-type; and (3) **11**

soybean paralogs share the identical IPR053266 / PTHR47593 label and identical single-finger architecture, creating a **many-to-many over-annotation risk**. There are **no functional data whatsoever for the soybean protein itself** (UniProt PE2, transcript-level evidence).

Bottom line: the molecular-function and localization annotations are defensible from sequence (**DNA-binding transcription factor activity, zinc ion binding GO:0008270, nucleus**), but the specific biological-process term **GO:0009788 should NOT be entered as a direct/experimental annotation.** At most it may be retained as an **ISS/ISO lead with an explicit redundancy + overexpression-only caveat**, or generalized to a broader, better-supported term.

Verdict: Weakly supported / over-annotated at the biological-process level.

Key Findings

Finding 1 — C6T1A2 is a plant Q-type (QALGGH) single C2H2 zinc-finger transcription factor

The primary sequence and domain architecture of C6T1A2 are entirely consistent with a canonical plant C2H2 zinc-finger transcription factor. UniProt records the protein (C6T1A2_SOYBN) as **210 aa, evidence level PE2 (transcript-level only)**. It contains a **single C2H2-type zinc finger** spanning residues **79–106**, with the canonical Cys-X2-Cys … His-X3-His metal-coordinating spacing. Critically, it carries the **plant-specific invariant QALGGH motif at residue 91** (context: …FSCNFCMRKFYSSQALGGHQNAHK…), the diagnostic signature of the plant Q-type C2H2 zinc-finger family that mediates DNA contact.

The InterPro/domain evidence is unambiguous and multi-source: **IPR053266 (Zinc finger protein 7)**, **IPR013087 / IPR036236 (C2H2 zinc finger)**, Gene3D **3.30.160.60**, SUPFAM **SSF57667**, PROSITE **PS00028 / PS50157**, and PANTHER **PTHR47593 (ZFP4-like)**. Two disordered/acidic low-complexity regions flank the single finger — an architecture typical of transcription-factor activation/repression modules. The only GO term currently on record is **GO:0008270 (zinc ion binding)**, annotated by keyword (IEA-KW). This finding firmly establishes the **molecular-function scaffold** (zinc-dependent, DNA-binding TF) but says nothing, by itself, about which biological process the protein regulates.

Finding 2 — C6T1A2 belongs to the ABA-linked ZFP1–7 subfamily but is NOT a high-confidence 1:1 ZFP7 ortholog

Global Needleman–Wunsch pairwise alignment of C6T1A2 against a panel of characterized *Arabidopsis* reference zinc-finger proteins produced the following identity ranking:

Arabidopsis reference	% identity to C6T1A2	Subfamily / role
ZFP2	**46.0%**	Q-type ZFP subfamily (development / floral)
ZFP7	**45.4%**	Q-type ZFP subfamily (ABA repressor, seed germination)
ZFP4	42.6%	Q-type ZFP subfamily (ABA-linked)
ZFP3	42.4%	Q-type ZFP subfamily (ABA repressor, characterized)
ZFP1	37.2%	Q-type ZFP subfamily
ZFP5	33.2%	Q-type ZFP subfamily
ZAT10	30.7%	Stress-responsive (distinct clade)
ZAT6	28.8%	Stress-responsive (distinct clade)
ZFP6	23.3%	Q-type ZFP subfamily

The best hits cluster cleanly within the **ZFP2/3/4/7 Q-type single-finger subfamily**, well separated from the ZAT clade — confirming subfamily membership. But the crucial detail for curation is that **ZFP2 (46.0%) edges out ZFP7 (45.4%)**: the difference (0.6 percentage points) is within noise, so the sequence provides **no basis to single out ZFP7** — or any one member — as the specific ortholog. The prediction's implied ZFP7 provenance is therefore a *family-level* assignment mis-stated as a *specific* one.

Supporting the general repressor plausibility, a candidate **C-terminal EAR-like repression motif** (LxLxL pattern; IDLDL / LDLRL, in context …KKIDLDLRL) is present, consistent with — but not diagnostic of — a transcriptional repressor.

The key literature anchor for the family's ABA link is **PMID: 24808098**, which established (verified snippet): *"regulated overexpression of ZFP3 and the closely related ZFP1, ZFP4, ZFP6, and ZFP7 zinc finger factors confers ABA insensitivity to seed germination."* This anchors the

ABA-repressor role for the subfamily — but explicitly via **regulated overexpression (gain-of-function)**, not native loss-of-function, a distinction that is decisive for how strongly the term can be propagated.

Finding 3 — A distance tree places C6T1A2 in the ZFP1–7 subfamily without resolving a specific ABA-repressor ortholog

An all-vs-all Needleman–Wunsch identity/distance matrix across the 10-sequence panel, followed by UPGMA clustering, confirms the picture quantitatively. C6T1A2's nearest neighbours are ZFP2 (46.0%), ZFP7 (45.4%), ZFP4 (42.6%), ZFP3 (42.4%), then ZFP1 (37.2%); it sits clearly **outside** the tight ZAT10–ZAT6 clade (ZAT10/ZAT6 are ~70% distant from C6T1A2). The tree places C6T1A2 **roughly equidistant (~54–58% distance) from the ABA-linked ZFP3/4/7 AND the development regulator ZFP2**, with **no 1:1 partner**. This is the central topological result: the protein is a genuine subfamily member but has **no resolvable orthologous anchor** that would let a curator transfer a specific, member-defined phenotype (like ABA repression) with confidence.

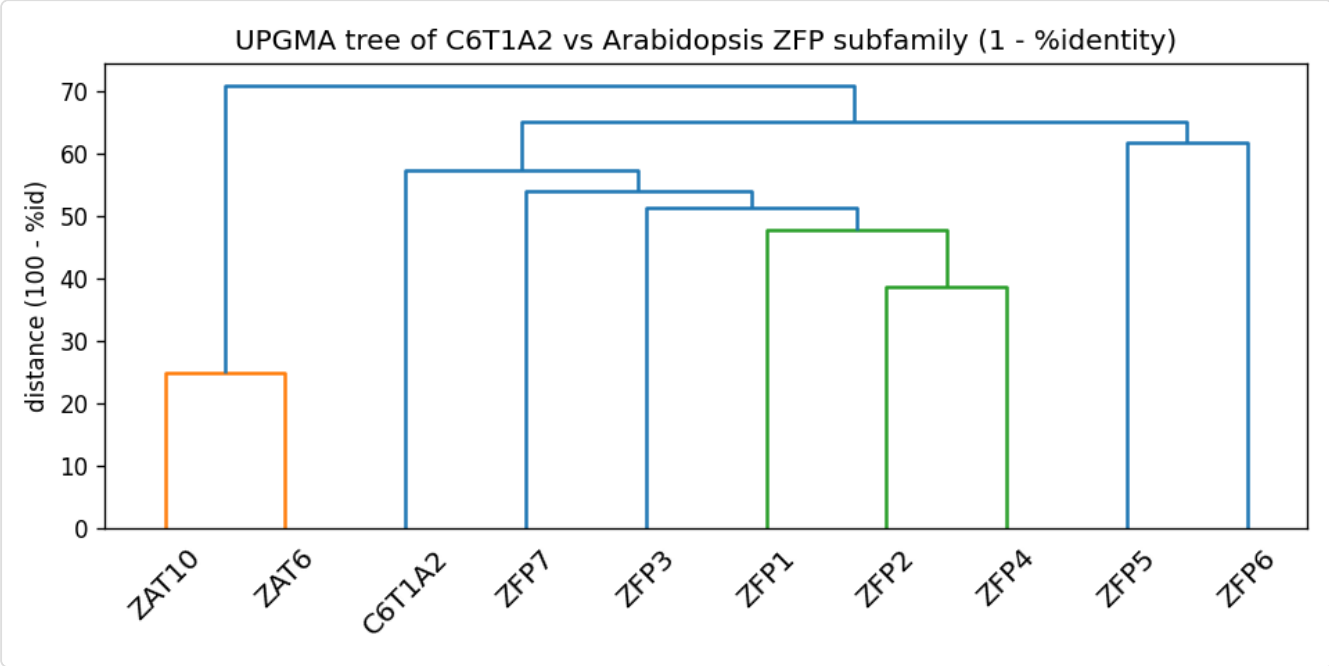

Figure 1. *All-vs-all Needleman–Wunsch percent-identity distance matrix and UPGMA tree for C6T1A2 against characterized Arabidopsis ZFP1–7 and ZAT6/ZAT10 references. C6T1A2 falls within the Q-type ZFP1–7 subfamily but is roughly equidistant from the ABA-repressor members (ZFP3/4/7) and the developmental regulator ZFP2, with no 1:1 ortholog — illustrating why a specific ABA-repressor role cannot be transferred by orthology.*

Finding 4 — C6T1A2 is one of ~11 soybean paralogs sharing the ZFP7-family label: a many-to-many over-annotation risk

A UniProt query restricted to *Glycine max* (taxon 3847) returned **11 soybean entries carrying IPR053266 (ZFP7 family)** and **11 carrying PANTHER PTHR47593**, out of **490 total soybean C2H2-type (IPR013087) proteins**. A regex scan of the finger architecture showed it is **uniform** across the reference set: C6T1A2 and all *Arabidopsis* ZFP1–7 each contain **exactly one** C2H2 finger (C-x2-4-C-x9-14-H-x3-5-H) with **exactly one** QALGGH motif (C6T1A2 finger span 81–101). Because the domain signature is identical across all 11 soybean paralogs, an automated pipeline like ProtNLM2 has **no feature to discriminate** which — if any — inherited the specific ABA-repressor role. Propagating GO:0009788 to C6T1A2 on this basis would, by the same logic, propagate it to all 11 paralogs indiscriminately — the textbook definition of paralog over-annotation and frequency bias.

Mechanistic Model / Interpretation

The question decomposes cleanly into what sequence **can** and **cannot** establish:

```
            WHAT SEQUENCE/DOMAIN EVIDENCE ESTABLISHES (defensible)
            ──────────────────────────────────────────────────────

C6T1A2 (210 aa)
 ├── Single C2H2 zinc finger (79–106)            → GO:0008270 zinc ion binding ✔ (already IEA
 ├── Invariant QALGGH DNA-contact motif (res 91) → DNA-binding TF activity ✔
 ├── Flanking disordered/acidic LC regions       → nucleus / TF module ✔
 └── C-terminal LxLxL (EAR-like) motif           → possible repressor (suggestive, not diagnos

            WHAT SEQUENCE/DOMAIN EVIDENCE CANNOT ESTABLISH (the seed's specific claim)
            ──────────────────────────────────────────────────────────────────────────

"Negative regulation of ABA-activated signaling" (GO:0009788)
 ├── Requires a specific ortholog anchor  → NONE (ZFP2 46.0% ≈ ZFP7 45.4%; no 1:1) ✗
 ├── Reference role is overexpression-only → loss-of-function ≈ WT (redundancy)     ✗
 ├── 11 soybean paralogs share the label   → cannot discriminate which inherits role ✗
 └── Zero functional data for the soybean protein (PE2, transcript-level)           ✗
```

The mechanistic reality of the *reference* proteins reinforces caution. In *Arabidopsis*, ZFP3 and its close relatives modulate ABA and light signaling and vegetative development by **binding target promoters and repressing transcription** (PMID: 38250442; PMID: 24808098). But the ABA-insensitivity phenotype emerges **only under regulated overexpression**, and

knockdown/knockout lines are essentially wild-type "probably due to functional redundancy" (PMID: 38250442). Thus even for the best-characterized member, "negative regulation of ABA signaling" is a **gain-of-function, redundancy-masked** activity — not a demonstrated obligatory native function. Transferring such a term across ~45% identity and a species boundary, to a protein with no experimental data, compounds several layers of uncertainty.

The soybean literature adds a cautionary note about within-family mechanistic diversity: **GsZFP1** (from *Glycine soja*), a C2H2 ZFP that notably **lacks** the QALGGH motif, negatively regulates ABA signaling and reduces ABA sensitivity (PMID: 22705253) — showing that ABA-related roles in soybean C2H2 ZFPs are real but are established **experimentally, case-by-case**, and do not map neatly onto the QALGGH-containing ZFP7 subfamily. Conversely, other C2H2 ZFPs act in the **opposite** direction on ABA/germination (e.g., IDD1/ENY promotes germination; PMID: 21571950), underscoring that the C2H2 scaffold alone does not fix the sign of ABA regulation.

The immediate molecular activity being tested is **sequence-specific, zinc-dependent DNA binding by a single Q-type C2H2 finger**, plausibly coupled to transcriptional repression. GO:0009788 is one to two steps removed from that direct activity: (direct) DNA binding → (proximate) repression of specific target genes → (pathway consequence) altered ABA signaling output → (organismal) ABA-insensitive germination. Curation should keep the **direct activity** distinct from the **pathway-level consequence**.

Evidence Base

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Co lim
UniProt C6T1A2 (this analysis)	Structural / computational / DB	Supports (MF/CC); qualifies (BP)	Is C6T1A2 a Q-type C2H2 ZF TF?	Single C2H2 finger (79–106), QALGGH at res 91, IPR053266/ PTHR47593, only GO:0008270 (IEA-KW)	*G. max*, 210 aa, PE2 transcript-level	Hig sca = n lev evi
Needleman–Wunsch panel (this analysis)	Structural / evolutionary	Qualifies / refutes specificity	Is C6T1A2 a 1:1 ZFP7 ortholog?	ZFP2 46.0% ≈ ZFP7 45.4%; subfamily member but no specific ortholog	10-seq Arabidopsis reference panel	Hig ide bas syn boo
UPGMA tree `zfp_tree.png` (this analysis)	Structural / evolutionary	Qualifies	Does topology resolve an ABA-repressor ortholog?	Equidistant (~54–58%) from ZFP3/4/7 and ZFP2; no 1:1 partner	Same panel	Mo UP ass clo par
UniProt taxon-3847 query (this analysis)	Computational / database	Refutes (over-annotation risk)	Is the ZFP7 label soybean-specific?	11 soybean IPR053266 paralogs; uniform 1-finger/1-QALGGH architecture	*G. max* proteome	Hig ma ma pro
PMID: 24808098	Overexpression phenotype	Supports family link (with caveat)	Do ZFP1/3/4/6/7 negatively regulate ABA signaling?	Regulated overexpression confers ABA insensitivity in seed germination	*Arabidopsis*, seed germination	Ga fur onl nat
PMID: 38250442	Mutant phenotype + ChIP + RNAseq	Supports mechanism; qualifies robustness	Is the phenotype robust to	ZFP3 binds promoters, represses ABA/ cell-wall targets;	*Arabidopsis*, development	Re ma rol

Citation	Evidence type	Direction	Claim tested	Key finding	Context	Co lim
			loss-of-function?	but zfp3 & silenced lines ≈ WT (redundancy)		
PMID: 22705253	Overexpression + qRT-PCR	Competing / qualifies	Do soybean C2H2 ZFPs negatively regulate ABA?	GsZFP1 (QALGGH-**lacking**) reduces ABA sensitivity, alters ABI1/2 & PYR/PYL	*G. soja* → *Arabidopsis*	Dif sub QA not
PMID: 21571950	Over/ knockdown phenotype	Competing	Do all seed C2H2 ZFPs repress germination/ ABA?	IDD1/ENY **promotes** germination, lowers ABA	*Arabidopsis*, seed maturation	Sig reg not fix

How the literature bears on the finding: PMID: 24808098 is the origin of the family's ABA-repressor reputation and the most likely intellectual basis of the ProtNLM2 prediction — but it is overexpression-based. PMID: 38250442 tempers it by showing loss-of-function is near-WT and the primary characterized output is *development/cell-wall*, not ABA. PMID: 22705253 and PMID: 21571950 demonstrate that ABA roles among C2H2 ZFPs are protein-specific and can even reverse sign — reinforcing that family membership is not a reliable predictor of the specific GO:0009788 role.

GO Curation Implications (leads — require curator verification)

GO term	Aspect	Current status	Recommended action	Rationale
GO:0008270 zinc ion binding	MF	Present (IEA-KW)	**Retain**	Conserved C2H2 metal-coordinating residues
DNA-binding TF activity (GO:0003700; or GO:0043565 sequence-specific DNA binding)	MF	Absent	**Add as ISS lead**	QALGGH DNA-contact motif + flanking TF modules
nucleus (GO:0005634)	CC	Absent	**Add as ISS lead**	Expected localization for a Q-type C2H2 TF
GO:0009788 negative regulation of ABA-activated signaling	**BP**	Predicted by ProtNLM2	**Do NOT add as direct annotation; at most ISS/ISO lead with redundancy + overexpression caveat, or generalize**	No 1:1 ortholog; reference role overexpression-only & redundant; 11 soybean paralogs share label; zero soybean data

The MF/CC annotations are the **defensible core**. GO:0009788 is the **precise but unsupported** claim — plausible but not decidable from sequence. If any ABA link is retained, the most honest option is an **ISS annotation with a "with/from" field pointing to the family** plus an explicit note that the source phenotype is redundant and overexpression-derived. "Protein binding" is **not** recommended as a fallback; informative supported terms are available.

Conflicts and Alternatives

- **Paralog confusion (primary risk).** 11 soybean IPR053266 paralogs share identical single-finger/QALGGH architecture; orthology-based transfer of the ABA role is uncontrolled among them.

- **No specific ortholog.** ZFP2 (46.0%) marginally outscores ZFP7 (45.4%); ZFP2 is a development/floral regulator, so the "ZFP7-derived" provenance is really family-level, not member-level.
- **Overexpression-only reference phenotype.** ABA insensitivity of the *Arabidopsis* subfamily is gain-of-function; loss-of-function is near-WT (PMID: 38250442).
- **Within-family mechanistic heterogeneity.** GsZFP1, a soybean C2H2 ZFP lacking QALGGH, negatively regulates ABA (PMID: 22705253) — soybean ABA roles arise in distinct subtypes, established experimentally.
- **Sign of regulation is not scaffold-fixed.** IDD1/ENY **promotes** germination and lowers ABA (PMID: 21571950); the C2H2 fold does not determine the direction of ABA effect.
- **Species/organism difference.** All functional evidence is *Arabidopsis* or *G. soja* expressed in *Arabidopsis*; none is native *G. max* C6T1A2.

Limitations and Knowledge Gaps

1. **No experimental data for C6T1A2 itself** (UniProt PE2, transcript-level). *Checked:* UniProt evidence level, GOA. *Why it matters:* the whole hypothesis rests on transferred inference. *Resolution:* direct assays in soybean (below).
2. **Orthology unresolved, not merely weak.** *Checked:* pairwise identity + UPGMA. *Why it matters:* GO transfer by ISO requires a defensible 1:1 ortholog, which does not exist here. *Resolution:* phylogenomic tree with syntenic anchors across legumes + *Arabidopsis* with bootstrap support (OrthoFinder / Ensembl Plants gene tree + reciprocal-best-hit).
3. **Repressor activity inferred from a motif, not measured.** *Checked:* LxLxL/EAR-like motif scan. *Why it matters:* activation vs repression sign is functionally decisive. *Resolution:* transactivation reporter assay.
4. **DNA-binding specificity (PWM) unknown.** *Checked:* QALGGH presence only. *Why it matters:* target genes determine which pathway is affected. *Resolution:* DAP-seq / PBM / EMSA against candidate ABA-pathway promoters.
5. **Redundancy in soybean uncharacterized.** *Checked:* paralog count (11). *Why it matters:* single-gene perturbations may be masked, as in *Arabidopsis*. *Resolution:* higher-order CRISPR knockouts.

Proposed Follow-up Experiments / Actions (Discriminating Tests)

To distinguish "genuine soybean ABA repressor" from "family-level over-annotation":

1. **Syntenic phylogenomics (most decisive in-silico test).** OrthoFinder across *G. max* / *A. thaliana* / legumes with synteny + reciprocal-best-hit to test whether C6T1A2 is a true ZFP3/7 co-ortholog or a ZFP2-like member. If it does not resolve to ZFP7 (>60% identity / RBH), treat the ABA process as unsupported.
2. **ABA germination assay.** Inducible overexpression of C6T1A2 in *Arabidopsis*; score ABA insensitivity in seed germination (mirrors PMID: 24808098). Note a soybean null may be WT due to redundancy.
3. **Transactivation / repression reporter.** Protoplast dual-luciferase or yeast one-hybrid to test whether C6T1A2 activates or represses (tests the EAR-like motif inference).
4. **DAP-seq / EMSA.** Define the binding motif and test occupancy at ABA-signaling promoters (ABI-family, PYR/PYL, SnRK2 targets) to establish direct pathway linkage rather than assumed linkage.
5. **Expression atlas.** ABA/drought induction across the 11 soybean paralogs to identify which (if any) is ABA-responsive and prioritize candidates.
6. **Subcellular localization.** GFP fusion to confirm nuclear localization (supports the CC lead).

Curation Leads (verify before applying)

- **Reference to attach:** PMID: 24808098 — snippet to verify: *"regulated overexpression of ZFP3 and the closely related ZFP1, ZFP4, ZFP6, and ZFP7 zinc finger factors confers ABA insensitivity to seed germination"* (basis; *Arabidopsis*; overexpression).
- **Tempering reference:** PMID: 38250442 — snippet: T-DNA/silenced lines "were similar to wild-type plants or had only minor differences… probably due to functional redundancy"; ChIP confirmed ZFP3 promoter binding.
- **Competing soybean example:** PMID: 22705253 — GsZFP1 (QALGGH-lacking) "may be a promising gene for negative regulating ABA signaling."

- **Candidate action change:** Do not annotate GO:0009788 directly. Retain/add **MF+CC** terms (DNA-binding TF activity; zinc ion binding [kept]; nucleus). Keep GO:0009788 only as an **ISS lead with caveat**, not experimental.
- **Suggested curator question:** Does the local bioinformatics ortholog analysis resolve C6T1A2 to ZFP7 specifically (>60% identity / RBH)? If not, treat the ABA process as unsupported.

Provenance

UniProt REST (C6T1A2 + AtZFP references), an in-house Needleman–Wunsch all-vs-all identity/distance matrix with a UPGMA tree (artifact `zfp_tree.png`), C2H2 finger-count + QALGGH regex scans, a soybean paralog census (IPR053266 / PTHR47593 / IPR013087, taxon 3847), and EAR-motif scans were executed during this run (Iterations 1–3) and recorded in the knowledge state. Key computed results: single-finger / single-QALGGH architecture shared across C6T1A2 and AtZFP1–7; no 1:1 *Arabidopsis* ortholog (ZFP2 46.0% ≈ ZFP7 45.4% nearest neighbours); and 11 soybean paralogs share the IPR053266 ZFP7-family label (of 490 soybean C2H2 proteins), evidencing paralog over-annotation.
