Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Annotation of Specific Genes in Olimarabidopsis pumila

Bailey Johnson

jbaileyi@g.ucla.edu

University of California, Los Angeles

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11702485/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Annotation of Specific Genes in *Olimarabidopsis pumila*

Bailey Johnson

Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA, United States.

Abstract

Olimarabidopsis pumila (*O. pumila*) is an annual spring ephemeral mustard plant that possesses adaptations to its semi-arid and salinized desert environments, affording it high salinity tolerance and the ability to maintain efficient photosynthesis under high-intensity solar radiation [1]. Four genes were identified between base pairs 7,500,000 and 8,000,000 on contig tig00002140 using the Genomic Sequence Annotation Server (GenSAS) and AUGUSTUS [2, 3]. Sequence conservation was assessed using the National Center for Biotechnology Information (NCBI) Protein Basic Local Alignment Search Tool (BLASTP) and Multiple Sequence Alignment (MSA) Viewer [4, 5]. Localization was determined using WoLF PSORT, SignalP 6.0, and Transmembrane Hidden Markov Models (TMHMM) 2.0 [6, 7, 8]. Lastly, structural analysis was configured using InterProScan, AlphaFold, STRING, and the combination of UniProt and ColabFold [9, 10, 11, 12, 13]. The four gene models were predicted to encode a DnaJ-family chaperone that assists in protein folding, an indole-3-glycerol phosphate synthase used in tryptophan biosynthesis, a ribosomal RNA-processing protein involved in rRNA maturation, and a protein with a GEF domain that is annotated as a ubiquitin ligase [9]. These four gene models are candidates for roles in protein translation, cellular homeostasis, signaling, and hormone biosynthesis that may contribute to plant growth and stress resilience.

Introduction

Olimarabidopsis pumila (*O. pumila*) is an annual, spring ephemeral brassicaceous plant from central and southwest Asia [14]. *O. pumila*, known for its rapid life cycle and high-intensity light utilization, is closely related to *Arabidopsis thaliana*, the first plant to have its genome fully sequenced, with approximately 27,000 protein-coding genes, and a model organism used to study plant biology and genetics [15, 16]. Notably, the Brassicaceae family is well-suited for studying the relationship between pericentromere evolution and genome reorganization due to the recent occurrence of both hybridization and polyploidization in its species [17].

Gene annotation and correction are essential for the identification of functional genes encoding proteins or regulatory RNAs. High-precision gene models reduce analytical error, enabling more accurate gene expression measurements, more reliable protein function predictions, and more thorough reconstruction of evolutionary history across lineages. In particular, gene annotation of *O. pumila*, a species often found in arid or semi-arid environments, can provide insight into potential genes that promote stress tolerance, such as tolerance to nutrient-poor soil or high salinity [1, 18]. Promotion of stress tolerance characteristics isolated from these genes can inform decisions going forward regarding crop improvement under conditions of abiotic stress.

A synergistic implementation of bioinformatics tools was used to predict gene structure and function between base pairs 7,500,000 and 8,000,000 on contig tig00002140. Gene selection and annotation were performed on the Genomic Sequence Annotation Server (GenSAS) [2]. Using the National Center for Biotechnology Information (NCBI) Protein Basic Local Alignment Search Tool (BLASTP) and Multiple Sequence Alignment (MSA) Viewer, sequence comparisons between selected genes and

known sequences from other organisms informed revisions to gene boundaries [4, 5]. Protein localization was analyzed using WoLF PSORT, SignalP 6.0, and Transmembrane Hidden Markov Models (TMHMM) 2.0. Protein domain prediction using InterProScan assisted in the identification of any additional gene boundary adjustments [9]. Finally, AlphaFold and ColabFold were used to assess whether corrected genes yield plausible three-dimensional structures that align with previously established protein folds [10, 13].

Claude Sonnet 5.5 (Anthropic) was used solely for refining the prose [19]. All analytical procedures and gene-model refinements were performed exclusively with established computational tools.

Methods

Gene Annotation

O. pumila genetic scaffolds were imported into the web-based platform, GenSAS, for initial automated annotation [2]. The Apollo browser was used to manually curate gene models using the ab initio gene prediction tool, AUGUSTUS [3]. Predicted genes were selected between base pairs 7,500,000 and 8,000,000 on contig tig00002140. Gene models were developed based on transcript alignment and protein homology tracks, which were integrated through Evidence Modeler consensus models.

Alignment

Sequence similarity searches were performed using BLASTP on the AUGUSTUS-derived FASTA sequences [3, 4]. Gene model correction involved extending exons based on alignment evidence in MSA Viewer [5]. Insertions were defined by a concurrence of missing amino acids

identified in the protein alignment. Deletions were defined by regions absent from homologous reference sequences or segments misaligned due to premature stop codons. These corrections were manually made on GenSAS [2]. The annotated gene was then run through BLASTP again to confirm the validity of the annotation based on changes in query coverage and percent identity relative to the original FASTA protein sequence [4].

Localization

General localization of sorting signals and global amino acid composition was conducted using WoLF PSORT, implementing a k-nearest neighbor algorithm [6]. N-terminal signal peptides and cleavage sites were predicted using SignalP 6.0 to differentiate between Sec-translocated signals and other secretory pathways [7]. Additionally, to represent the membrane topology of the protein derived from the gene model, TMHMM 2.0 was employed to quantify transmembrane helices and distinguish protein loop orientation with respect to the lipid bilayer [8].

Phylogeny

Taxonomic distributions were visualized using the NCBI BLAST Tree View [4]. The BLAST Tree View tool clustered homologous protein hits based on similarity to the query sequence, producing a distance-related dendrogram. This enabled the identification of conserved lineages and their evolutionary proximity to related taxa.

Structure

InterProScan was used to conduct structural analysis, identifying the conserved protein domains [9]. STRING was used to score and visualize protein-protein interaction networks through

Bayesian integration combining various evidence channels, such as gene neighborhood and curated databases [11]. The predicted three-dimensional structures were then built using AlphaFold [10]. The AlphaFold predictions were then compared to the InterProScan domain annotations [9, 10]. Structural features, such as active sites and binding pockets, were then evaluated according to protein function. Multimeric status was further assessed using ColabFold, with query sequences taken directly from the gene model annotations and partner proteins selected based on STRING-predicted interactions, with all reference sequences obtained from UniProt [11, 12, 13, 27].

Data Availability

Sequence data and annotations are stored within GenSAS and associated repositories [2]. The corrected gene models and evidence tracks are publicly accessible in relevant databases.

Results

Prediction and Alignment of Four Genes

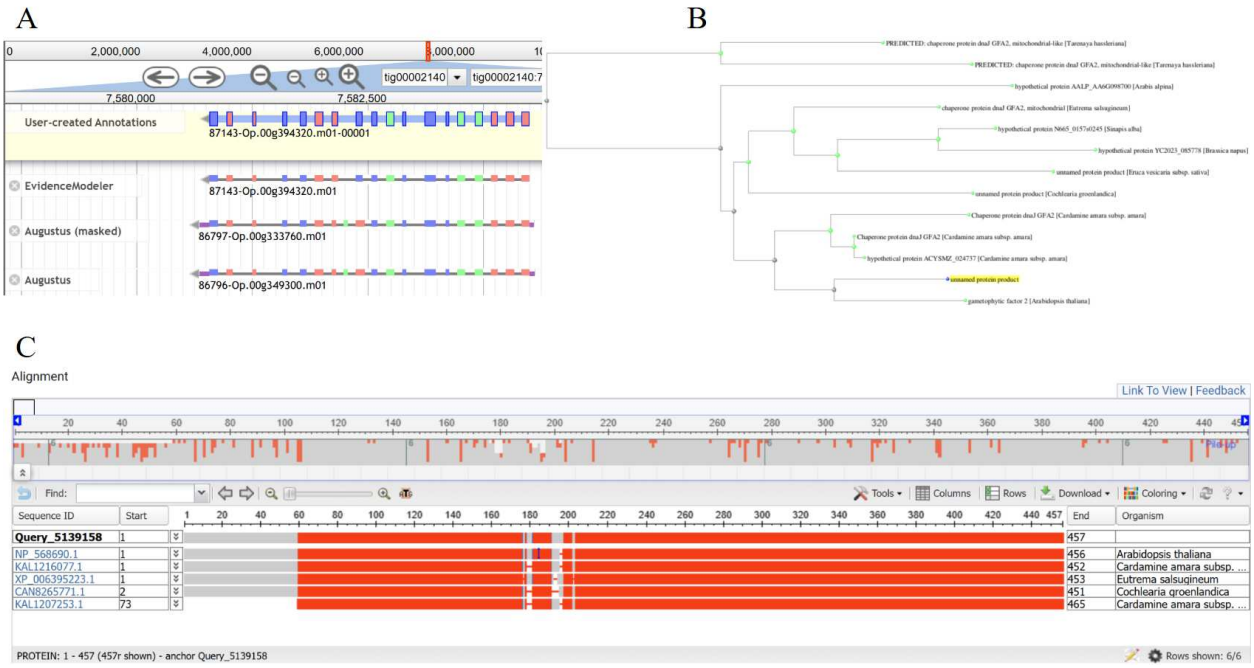


Figure 1. Comparative annotation, phylogenetic placement, and sequence conservation of the first candidate gene.

(A) Genome browser view showing predicted gene models from EvidenceModeler and AUGUSTUS across the target locus [2, 3]. (B) Phylogenetic reconstruction clusters the predicted protein with DnaJ/GFA2-family homologs from *Arabidopsis thaliana*, *Cardamine amara*, *Eutrema salsugineum*, *Brassica* species, and other *Brassicales* [4]. (C) Multiple sequence alignment of the candidate protein with close homologs reveals extensive conservation across the ~455 amino acid length as well as a lack of alignment support at residues 198 and 199 [5].

Gene models predicted by EvidenceModeler and AUGUSTUS were compared across the target locus (**Figure 1A**), and the predicted protein clustered with DnaJ/GFA2-family homologs from Brassicales species (**Figure 1B**). In the first gene model, the MSA viewer identified a consensus substitution at residues 198 and 199 (**Figure 1C**) [5]. The removal of these two amino acid residues from the exon raised the percent identity after re-blasting from 92.14% to 92.54%, with full query coverage [4].

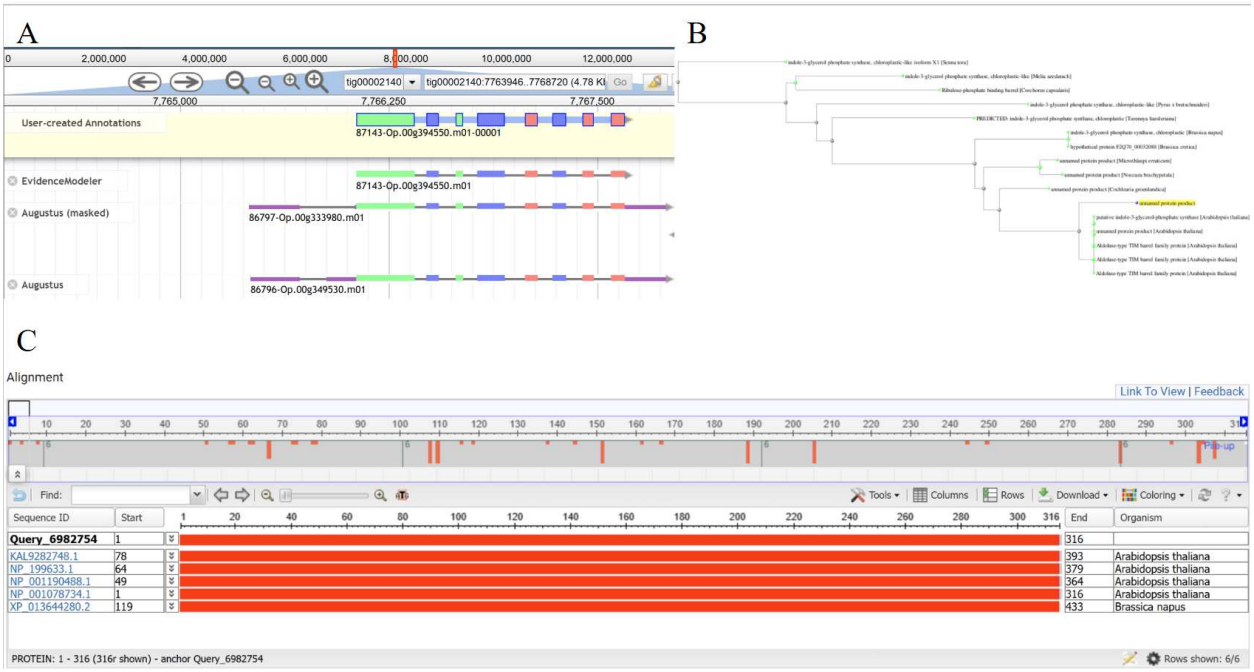


Figure 2. Comparative annotation, phylogenetic placement, and sequence conservation of the second candidate gene.

(A) Genome browser view showing de novo gene model predictions from EvidenceModeler and AUGUSTUS across the target locus [2, 3]. (B) BLAST Tree View of the predicted protein places it within the TIM-barrel family, grouping with homologs from *Arabidopsis thaliana*, *Brassica napus*, and related species [4]. (C) Multiple sequence alignment of the candidate protein and close homologs reveals strong conservation across the core catalytic region, with highly conserved residues spanning positions ~1–316 [5].

De novo gene model predictions from EvidenceModeler and AUGUSTUS were compared at the second locus (**Figure 2A**). In the MSA viewer, the second gene model was highly conserved, with no genomic annotations required (**Figure 2C**) [5]. The phylogenetic tree identified the closest related genes as the Aldolase-type TIM barrel family protein and the indole-3-glycerol-phosphate synthase (**Figure 2B**). The percent identity using BLASTP was 97.15%, with full query coverage [4].

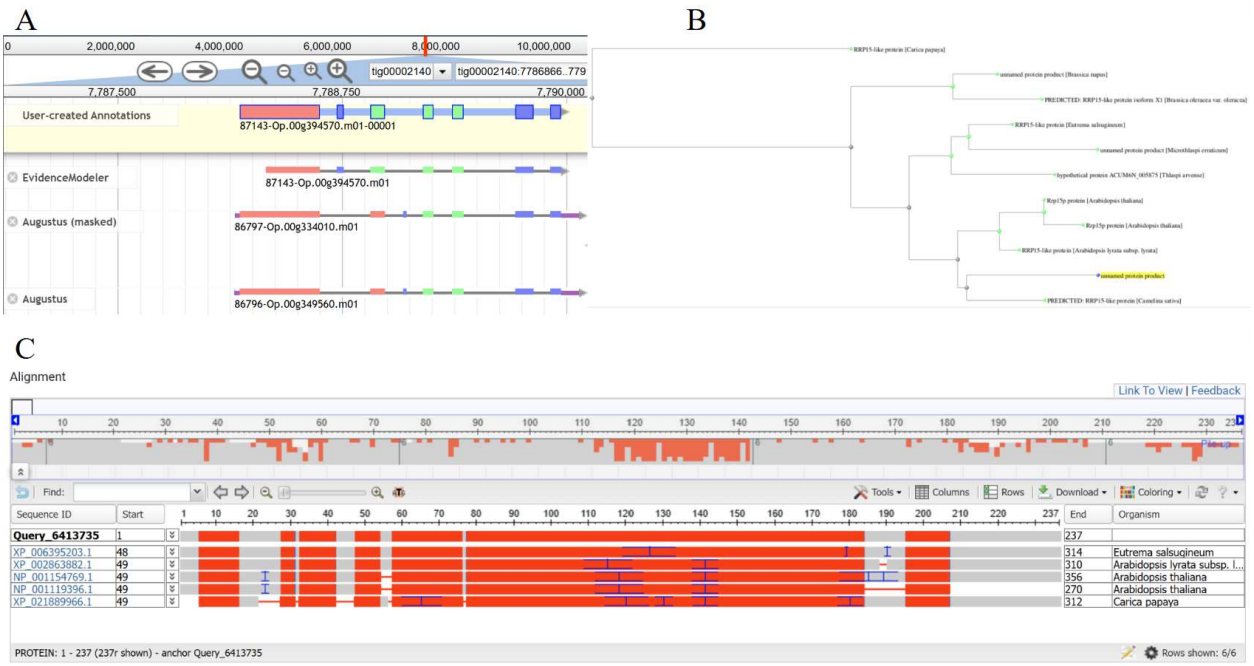


Figure 3. Comparative annotation, phylogenetic placement, and sequence conservation of the third candidate gene.

(A) Genome browser view showing predicted gene models from EvidenceModeler and AUGUSTUS across the target locus, with AUGUSTUS providing a greater percent identity match than EvidenceModeler [2, 3]. (B) Phylogenetic analysis places the predicted protein within a conserved RRP15-like clade, grouping with homologs from *Arabidopsis thaliana*, *Brassica* species, *Carica papaya*, and other angiosperms [4]. (C) Multiple sequence alignment of the candidate protein indicates consensus in a 48-amino-acid extension to the start site of the gene model [5].

The third gene model required the start site to be extended by 48 amino acid residues based on MSA viewer consensus (**Figure 3C**) [5]. AUGUSTUS provided a greater percent identity match than EvidenceModeler at this locus (Figure 3A), and the predicted protein grouped within a conserved RRP15-like clade (**Figure 3B**). The extension increased the percent identity from 72.45% to 75.72% upon re-blasting [4]. Additionally, unknown amino acids were filled in using the consensus of the top five relatives indicated in the phylogeny (**Figure 3B**), addressing the unknown amino acids from site 65 extending to site 74, as well as from site 194 extending to site 200, with the corrected amino acid sequences DEEEEDESTK and GVVKLFN, respectively. This raised the percent identity once re-blasted to 80.45%. Because these residues were inferred

from homologs, part of this increase reflects the imputation itself and is not independent support for the model.

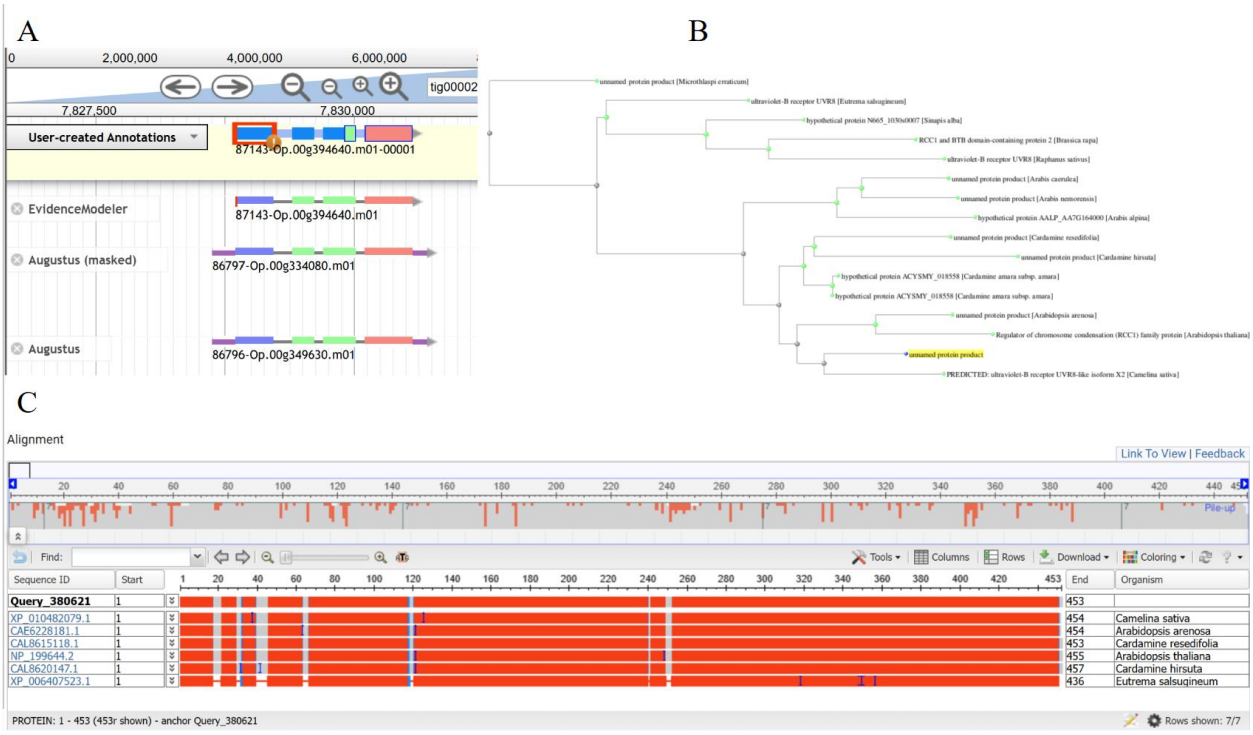


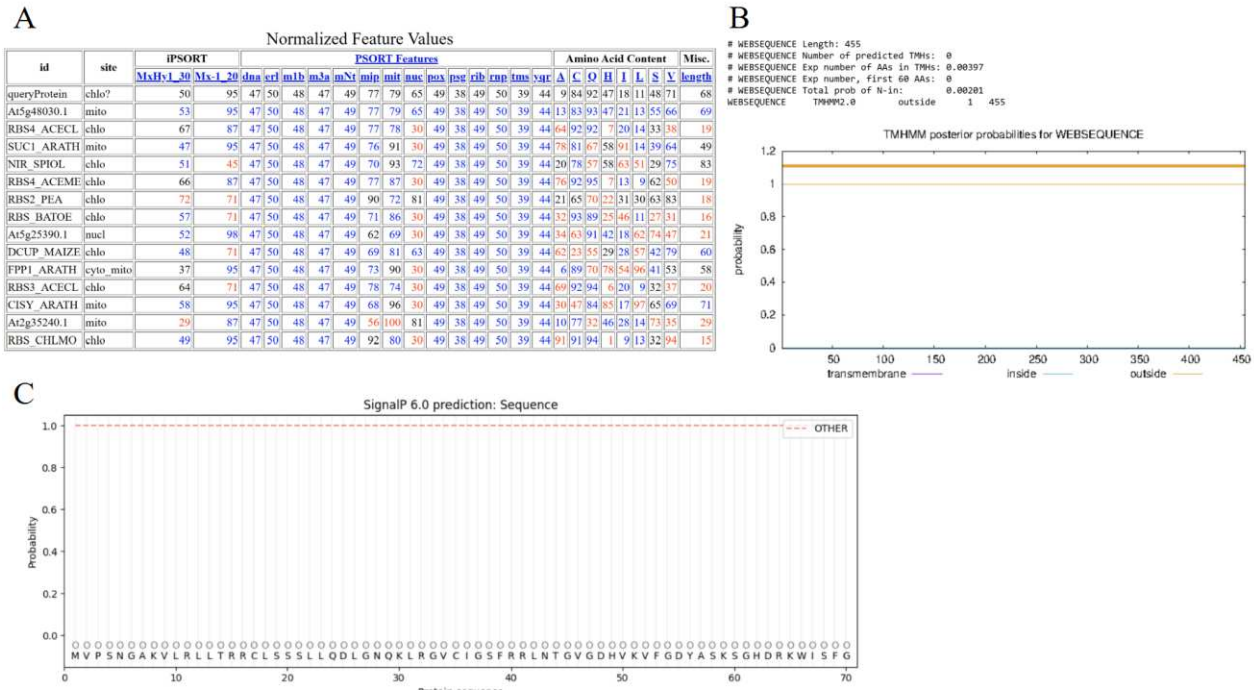
Figure 4. Comparative annotation, phylogenetic placement, and sequence conservation of the fourth candidate gene.

(A) Genome browser view showing predicted gene models from EvidenceModeler and AUGUSTUS across the target locus, with annotation of an additional serine at the end of the exon [2, 3]. (B) Phylogenetic analysis clusters the predicted protein within the UVR8/RCC1-family clade, grouping with homologs from *Arabidopsis thaliana*, *Camelina sativa*, *Eutrema salsugineum*, *Cardamine* species, and other Brassicales [4]. (C) Multiple sequence alignment of the candidate protein with close homologs reveals extensive conservation across the ~454 amino acid length and a deletion near residue 123 [5].

The fourth gene model was edited by adding a serine near residue 123 (**Figure 4A**), following the identification of a deletion in the MSA viewer (**Figure 4C**) [5]. The predicted protein clustered within the UVR8/RCC1-family clade (**Figure 4B**). This edit, which extended the exon by one amino acid residue, raised the BLASTP max score from 826 to 832 [4]. The edit lies at an

exon end and does not coincide with canonical GT and AG splice sites, so it should be regarded as provisional.

Localization



id	site	iPSORT		PSORT Features													Amino Acid Content							Misc.
		MxHy1_30	Mx-1_20	dna	crl	m1b	m3a	mNt	mlp	mit	nuc	pox	psg	rib	rnp	tms	vqr	A	C	Q	H	I	L	S	V	length
queryProtein	chlo?	50	95	47	50	48	47	49	77	79	65	49	38	49	50	39	44	9	84	92	47	18	11	48	71	68
At5g48030.1	mito	53	95	47	50	48	47	49	77	79	65	49	38	49	50	39	44	13	83	93	47	21	13	55	66	69
RBS4_ACECL	chlo	67	87	47	50	48	47	49	77	78	30	49	38	49	50	39	44	64	92	92	7	20	14	33	38	19
SUC1_ARATH	mito	47	95	47	50	48	47	49	76	91	30	49	38	49	50	39	44	78	81	67	58	91	14	39	64	49
NIR_SPIOL	chlo	51	45	47	50	48	47	49	70	93	72	49	38	49	50	39	44	20	78	57	58	63	51	29	75	83
RBS4_ACEME	chlo	66	87	47	50	48	47	49	77	87	30	49	38	49	50	39	44	76	92	95	7	13	9	62	50	19
RBS2_PEA	chlo	72	71	47	50	48	47	49	90	72	81	49	38	49	50	39	44	21	65	70	22	31	30	63	83	18
RBS_BATOE	chlo	57	71	47	50	48	47	49	71	86	30	49	38	49	50	39	44	32	93	89	25	46	11	27	31	16
At5g25390.1	nucl	52	98	47	50	48	47	49	62	69	30	49	38	49	50	39	44	34	63	91	42	18	62	74	47	21
DCUP_MAIZE	chlo	48	71	47	50	48	47	49	69	81	63	49	38	49	50	39	44	62	23	55	29	28	57	42	79	60
FPP1_ARATH	cyto_mito	37	95	47	50	48	47	49	73	90	30	49	38	49	50	39	44	6	89	70	78	54	96	41	53	58
RBS3_ACECL	chlo	64	71	47	50	48	47	49	78	74	30	49	38	49	50	39	44	69	92	94	6	20	9	32	37	20
CISY_ARATH	mito	58	95	47	50	48	47	49	68	96	30	49	38	49	50	39	44	30	47	84	85	17	97	65	69	71
At2g35240.1	mito	29	87	47	50	48	47	49	56	100	81	49	38	49	50	39	44	10	77	32	46	28	14	73	35	29
RBS_CHLMO	chlo	49	95	47	50	48	47	49	92	80	30	49	38	49	50	39	44	91	91	94	1	9	13	32	94	15

Figure 5. Computational feature profiling of the first predicted protein.

(A) WoLF PSORT assigned equal top scores to the chloroplast and mitochondrial classes (6.5 each; combined chlo_mito score of 7) [6]. (B) TMHMM predicts zero transmembrane helices, with the full 455-residue sequence assigned a high probability of being non-membrane-embedded [8]. (C) SignalP 6.0 detects no signal peptide [7].

WoLF PSORT returned a tied prediction for protein 1 between the chloroplast and mitochondrial classes (chlo, 6.5; mito, 6.5), with a combined chlo_mito score of 7 and a nuclear score of 1. The result therefore does not favor either organelle and is consistent with chloroplast targeting, mitochondrial targeting, or dual targeting. This profile resembles that of the annotated mitochondrial protein At5g48030.1 (**Figure 5A**) [6]. TMHMM predicts no transmembrane helices, with an expected transmembrane residue count of 0.00397 and residues 1–455 assigned

a probability of approximately 1.0 of residing on the non-membrane side (**Figure 5B**) [8]. SignalP 6.0 likewise detects no signal peptide, assigning the entire N-terminus a probability of approximately 1.0 (**Figure 5C**) [7].

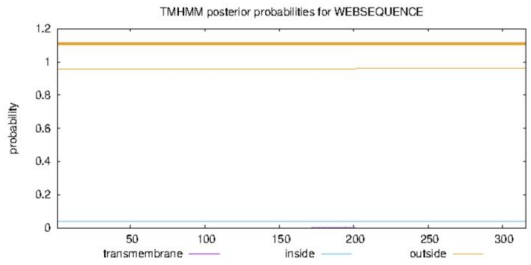
A

Normalized Feature Values

id	site	iPSORT		PSORT Features														Amino Acid Content							Misc.
		MxHy1_30	Mx-1_20	dna	erl	m1b	m3a	mNt	mip	mit	nuc	pox	psg	rib	rnp	tms	vgr	A	C	Q	H	I	L	S	V	length
queryProtein	cyto?	11	95	47	50	48	47	49	21	28	63	49	38	49	50	39	44	44	32	60	28	94	87	5	67	44
At1g67990.1	cyto	24	71	47	50	48	47	49	21	31	30	49	38	49	50	39	44	39	12	58	29	97	97	1	18	29
At1g20575.1	E.R.	24	87	47	50	48	47	49	21	28	30	49	38	49	50	39	44	8	11	41	83	95	90	18	77	31
ENO_ALNGL	cyto	16	71	47	50	48	47	49	21	37	30	49	38	49	50	39	44	85	30	58	41	79	35	11	75	66
At4g26220.1	cyto	19	45	47	50	48	47	49	21	22	30	49	38	49	50	39	44	55	12	21	46	84	97	4	35	29
ERFA_ARATH	cyto	18	87	47	50	48	47	49	21	5	30	49	38	49	50	39	44	29	20	59	25	65	79	34	21	66
At1g67980.1	cyto	6	45	47	50	48	47	49	21	38	30	49	38	49	50	39	44	52	31	8	82	86	88	4	59	25
IPYR_MAIZE	cyto	13	71	47	50	48	47	49	21	31	30	49	38	49	50	39	44	58	31	67	70	56	48	32	75	25
KADC_MAIZE	chlo	61	71	47	50	48	47	49	21	12	30	49	38	49	50	39	44	45	13	74	15	74	94	20	51	27
GPD6_ARATH	cyto	14	71	47	50	48	47	49	21	27	30	49	38	49	50	39	44	6	43	89	55	90	70	19	34	75
ERFB_ARATH	cyto	19	87	47	50	48	47	49	21	27	30	49	38	49	50	39	44	9	20	71	33	50	72	39	21	66
ERFC_ARATH	cyto	18	87	47	50	48	47	49	21	14	30	49	38	49	50	39	44	11	20	80	33	60	79	22	21	66
FPPS_HELAN	cyto	21	87	47	50	48	47	49	21	35	30	49	38	49	50	39	44	11	43	75	85	42	96	30	67	48
At4g13940.1	mito	17	71	47	50	48	47	49	21	43	30	49	38	49	50	39	44	52	81	61	68	71	61	6	61	72
MDHG_ORYSA	pero	14	87	47	50	48	47	49	21	25	30	49	38	49	50	39	44	90	80	23	23	57	79	20	87	51

B

```
# WEBSEQUENCE Length: 316
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.01478
# WEBSEQUENCE Exp number, first 60 AAs:  0
# WEBSEQUENCE Total prob of N-in:        0.04089
WEBSEQUENCE      TMHMM2.0        outside      1   316
```



C

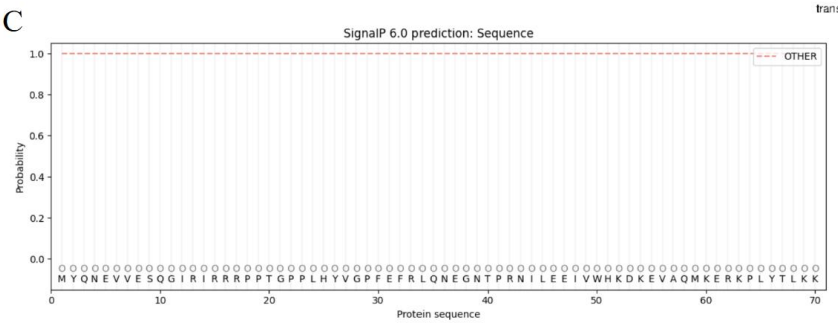
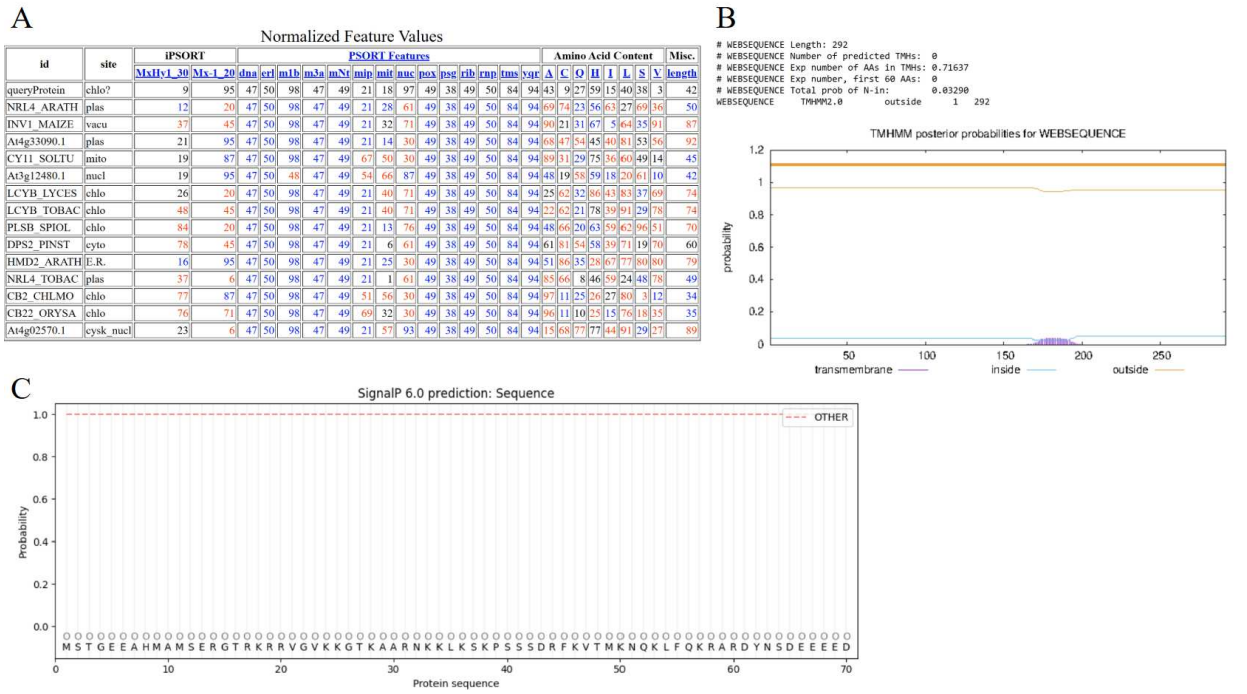


Figure 6. Computational feature profiling of the second predicted protein.

(A) WoLF PSORT assigned its highest score (10) to the cytosol [6]. (B) TMHMM predicts zero transmembrane helices, with the full 316-residue sequence assigned a high probability of being non-membrane-embedded [8]. (C) SignalP 6.0 detects no signal peptide [7].

WoLF PSORT assigned the highest score to cytosolic localization (10), with only minor scores for chloroplast (1), mitochondrial (1), ER (1), and peroxisomal (1) compartments (**Figure 6A**) [6]. TMHMM predicts no transmembrane helices, with an expected transmembrane residue count of 0.01478 and residues 1–316 assigned a probability of approximately 1.0 of residing on the non-membrane side (**Figure 6B**) [8]. SignalP 6.0 likewise detects no signal peptide, assigning the entire N-terminus a probability of approximately 1.0 (**Figure 6C**) [7].



id	site	iPSORT		PSORT Features														Amino Acid Content								Misc.
		MxIIy1_30	Mx-1_20	dna	erl	m1b	m3a	mNt	mip	mit	nuc	pox	psg	rib	rnp	tms	yqt	Δ	C	Q	H	I	L	S	V	length
queryProtein	chlo?	9	95	47	50	98	47	49	21	18	97	49	38	49	50	84	94	43	9	27	59	15	40	38	3	42
NRL4_ARATH	plas	12	20	47	50	98	47	49	21	28	61	49	38	49	50	84	94	69	74	23	56	63	27	69	36	50
INV1_MAIZE	vacu	37	45	47	50	98	47	49	21	32	71	49	38	49	50	84	94	90	21	31	67	5	64	35	91	87
At4g33090.1	plas	21	95	47	50	98	47	49	21	14	30	49	38	49	50	84	94	68	47	54	45	40	81	53	56	92
CY11_SOLTU	mito	19	87	47	50	98	47	49	67	50	30	49	38	49	50	84	94	89	31	29	75	36	60	49	14	45
At3g12480.1	nucl	19	95	47	50	48	47	49	54	66	87	49	38	49	50	84	94	48	19	58	59	18	20	61	10	42
LCYB_LYCES	chlo	26	20	47	50	98	47	49	21	40	71	49	38	49	50	84	94	25	62	32	86	43	83	37	69	74
LCYB_TOBAC	chlo	48	45	47	50	98	47	49	21	40	71	49	38	49	50	84	94	22	62	21	78	39	91	29	78	74
PLSB_SPIOL	chlo	84	20	47	50	98	47	49	21	13	76	49	38	49	50	84	94	48	66	20	63	59	62	96	51	70
DPS2_PINST	cyto	78	45	47	50	98	47	49	21	6	61	49	38	49	50	84	94	61	81	54	58	39	71	19	70	60
HMD2_ARATH	E.R.	16	95	47	50	98	47	49	21	25	30	49	38	49	50	84	94	51	86	35	28	67	77	80	80	79
NRL4_TOBAC	plas	37	6	47	50	98	47	49	21	1	61	49	38	49	50	84	94	85	66	8	46	59	24	48	78	49
CB2_CHLMO	chlo	77	87	47	50	98	47	49	51	56	30	49	38	49	50	84	94	97	11	25	26	27	80	3	12	34
CB22_ORYSA	chlo	76	71	47	50	98	47	49	69	32	30	49	38	49	50	84	94	96	11	10	25	15	76	18	35	35
At4g02570.1	cysk_nucl	23	6	47	50	98	47	49	21	57	93	49	38	49	50	84	94	15	68	77	77	44	91	29	27	89

Figure 7. Computational feature profiling of the third predicted protein.

(A) WoLF PSORT assigned its highest score (5) to the chloroplast [6]. (B) TMHMM predicts zero transmembrane helices, with the 292-residue sequence assigned a high probability of being outside any membrane [8]. (C) SignalP 6.0 detects no signal peptide [7].

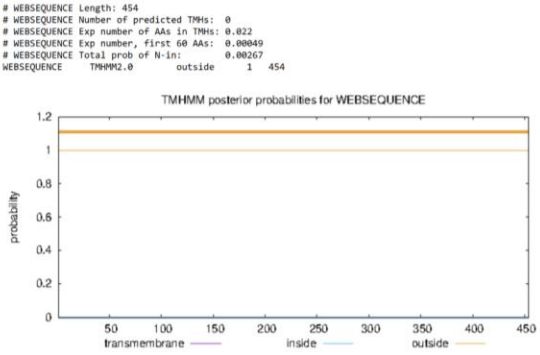
WoLF PSORT predicted chloroplast localization as the top class (score 5), with lower scores for plasma membrane (3), nuclear (1.5), cytoskeleton–nuclear (1.5), cytosolic (1), mitochondrial (1), vacuolar (1), and endoplasmic-reticulum (1) compartments (**Figure 7A**) [6]. TMHMM predicts no transmembrane helices, with an expected transmembrane amino-acid count of 0.71637 and a total probability of N-in of 0.033, while assigning residues 1–292 a probability of approximately 1.0 of residing on the non-membrane side (**Figure 7B**) [8]. SignalP 6.0 again identifies no signal peptide, with a probability of approximately 1.0 (**Figure 7C**) [7].

A

Normalized Feature Values

id	site	iPSORT		PSORT Features													Amino Acid Content							Misc.
		MxHy1_30	Mx-1_20	dna	erl	m1b	m3a	mNr	mip	mit	nuc	pox	psg	rib	rnp	tms	vqr	A	C	Q	H	I	L	S	V	length
queryProtein	nucl?	50	95	118	50	48	47	49	52	74	30	49	38	49	50	39	44	50	50	10	72	50	69	88	75	68
TT2_ARATH	nucl	1	87	100	50	48	47	49	51	69	88	49	38	49	50	39	44	0	87	36	91	44	93	60	0	34
VRN2_ARATH	nucl	53	87	97	50	48	47	49	49	64	98	49	38	49	50	39	44	5	94	44	85	17	61	71	23	66
At5g53810.1	cyto	10	71	97	50	48	47	49	21	31	30	49	38	49	50	39	44	50	83	10	63	61	90	24	76	56
H2A_CICAR	nucl	9	98	97	50	48	47	49	83	78	30	49	38	49	50	39	44	92	4	42	35	28	78	67	67	7
ATH5_ARATH	nucl	34	71	97	50	48	47	49	68	66	82	49	38	49	50	39	44	57	32	83	41	6	52	82	12	44
H2A_PICAB	nucl	6	98	97	50	48	47	49	82	72	30	49	38	49	50	39	44	96	4	64	36	28	80	57	68	7
At4g39100.1	nucl	14	98	97	50	48	47	49	50	71	30	49	38	49	50	39	44	8	95	72	91	10	1	79	20	28
HSF1_ARATH	nucl	78	87	97	50	48	47	49	54	56	68	49	38	49	50	39	44	3	7	96	58	15	47	81	16	73
HSF2_LYCPE	nucl	43	71	97	50	48	47	49	45	71	94	49	38	49	50	39	44	15	33	65	4	7	31	82	31	43
H2A_CHLRE	nucl	29	98	97	50	48	47	49	78	65	73	49	38	49	50	39	44	96	4	10	40	53	88	1	46	6
H2A1_WHEAT	nucl	40	98	97	50	48	47	49	75	79	99	49	38	49	50	39	44	97	4	19	33	24	83	8	73	7
H2A3_VOLCA	nucl	29	98	97	50	48	47	49	78	62	73	49	38	49	50	39	44	98	4	10	40	53	88	1	46	6
H2A4_VOLCA	nucl	20	98	97	50	48	47	49	78	65	73	49	38	49	50	39	44	97	4	10	40	53	88	3	46	6
SUP_ARATH	nucl	6	87	97	50	48	47	49	21	42	30	49	38	49	50	39	44	12	84	44	85	15	82	98	0	24

B



C

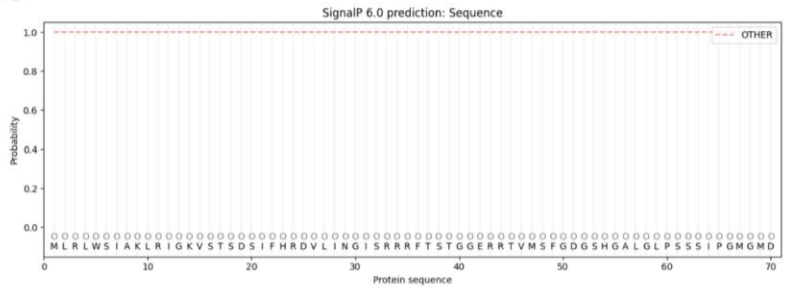


Figure 8. Computational feature profiling of the fourth predicted protein.

(A) WoLF PSORT assigned its highest score (13) to the nucleus [6]. (B) TMHMM predicts zero transmembrane helices, with the 454-residue sequence assigned a high probability of being outside any membrane [8]. (C) SignalP 6.0 detects no signal peptide [7].

WoLF PSORT gave a nuclear score of 13, with a score of 1 for the cytosol (**Figure 8A**) [6]. TMHMM predicts no transmembrane helices, with an expected transmembrane residue count of 0.022 and a total probability of N-in of 0.00267, and assigns residues 1–454 a probability of approximately 1.0 of residing on the non-membrane side (**Figure 8B**) [8]. SignalP 6.0 likewise detects no signal peptide (**Figure 8C**) [7].

Structural and Functional Analysis

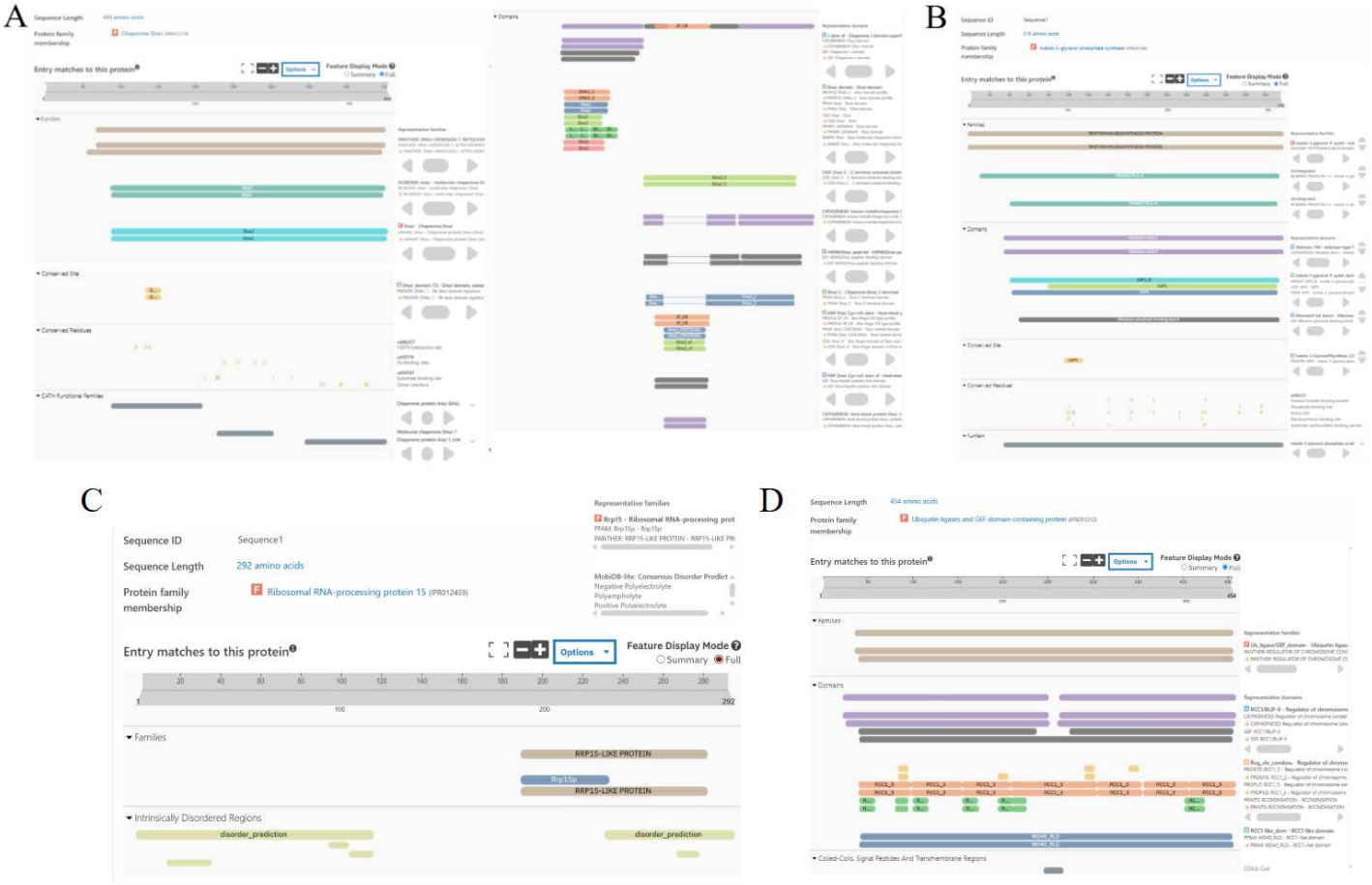


Figure 9. Predicted protein functions for candidate genes.

(A) InterProScan prediction of chaperone DnaJ for model protein one [9]. (B) InterProScan prediction of indole-3-glycerol phosphate synthase for model protein two [9]. (C) InterProScan prediction of ribosomal RNA processing for model protein three [9]. (D) InterProScan prediction of ubiquitin ligase and GEF domain-containing protein for model protein four [9].

InterProScan predicted that the gene 87143-Op.00g394320.m01-0001 encodes a DnaJ-family chaperone (**Figure 9A**), gene 87143-Op.00g394550.m01-0001 encodes an indole-3-glycerol phosphate synthase (**Figure 9B**), gene 87143-Op.00g394570.m01-0001 encodes a ribosomal RNA-processing protein (**Figure 9C**), and gene 87143-Op.00g394640.m01-0001 encodes a protein annotated as a ubiquitin ligase containing a GEF domain (**Figure 9D**) [9].

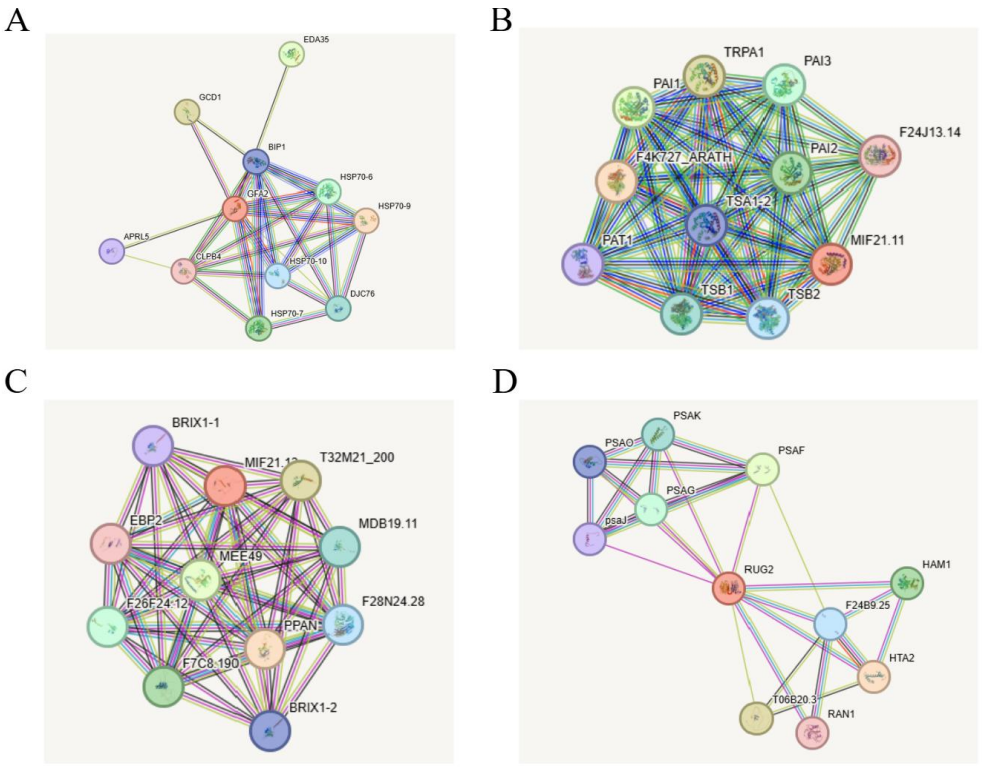


Figure 10. STRING-predicted protein–protein interaction networks reveal conserved functional partners across candidate genes.

(A) GFA2 network, clustering with BIP1, HSP70-6/7/9/10, CLPB4, DJC76, GCD1, EDA35, and APRL5 [11]. (B) MIF21.11 network, clustering with TSA1-2, TRPA1, PAI1/PAI2/PAI3, TSB1/TSB2, PAT1, F4K727_ARATH, and F24J13.14 [11]. (C) MIF21.13 network, clustering with BRIX1-1/BRIX1-2, EBP2, MEE49, MDB19.11, PPAN, F26F24-12, F28N24.28, F7C8-190, and T32M21_200 [11]. (D) RUG2 network, clustering with HTA2, RAN1, HAM1, F24B9.25, T06B20.3, and photosystem I subunits PSAK, PSAO, PSAG, PSAF, and psaJ [11].

For the first protein prediction, STRING analysis identified interactions with multiple heat-shock proteins and co-chaperones, including BIP1, HSP70-6, HSP70-7, HSP70-9, HSP70-10, CLPB4, DJC76, GCD1, EDA35, and APRL5, forming a chaperone-rich interaction network with confidence scores ranging from 0.745 to 0.875 (**Figure 10A**) [11]. The second model prediction indicated associations with enzymes in the tryptophan biosynthetic pathway, including TSA1-2, TRPA1, PAI1, PAI2, PAI3, TSB1, TSB2, PAT1, F4K727_ARATH, and F24J13.14, with confidence scores of 0.991–0.998 (**Figure 10B**) [11]. The third model prediction identified associations (scores 0.982–0.998) with nucleolar ribosome-biogenesis factors, including BRIX1-1, BRIX1-2, EBP2, MEE49, MDB19.11, PPAN, F26F24-12, F28N24.28, F7C8-190, and

T32M21_200 (**Figure 10C**) [11]. In the fourth model, STRING analysis identified interactions with proteins involved in chromatin organization and photosystem I components, including HTA2, RAN1, HAM1, F24B9.25, T06B20.3, and the photosystem I subunits PSAK, PSAO, PSAG, PSAF, and psaJ, with confidence scores ranging from 0.613 to 0.792 (**Figure 10D**) [11].

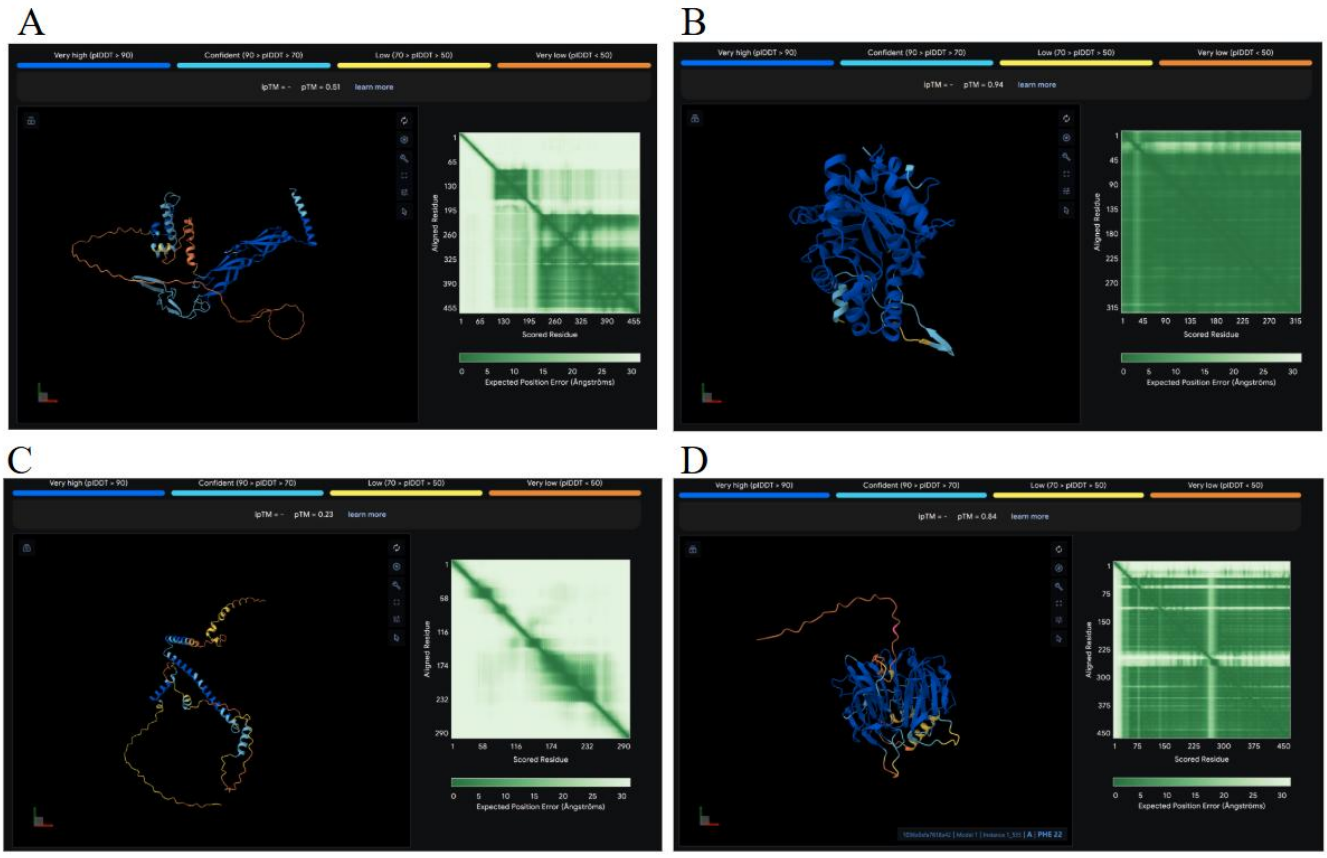


Figure 11. AlphaFold-Multimer homomer predictions and monomer baseline comparisons for all four proteins [10].

(A) Protein 1: the two-copy run produced an ipTM of 0.41 and a pTM of 0.44; the three-copy run produced an ipTM of 0.18 and a pTM of 0.28. The monomer model showed a pTM of 0.51. (B) Protein 2: two copies yielded an ipTM of 0.28 and a pTM of 0.50, and three copies yielded an ipTM of 0.25 and a pTM of 0.44. The monomer pTM was 0.94. (C) Protein 3: two copies gave an ipTM of 0.18 and a pTM of 0.22, and three copies gave an ipTM of 0.13 and a pTM of 0.17. The monomer pTM was 0.23. (D) Protein 4: two copies gave an ipTM of 0.21 and a pTM of 0.52, and three copies gave an ipTM of 0.17 and a pTM of 0.40. The monomer pTM was 0.84.

AlphaFold-Multimer was used to evaluate potential homomer formation for the four proteins [10]. For Protein 1, the two-copy model produced an ipTM of 0.41 and a pTM of 0.44, while the three-copy run had an ipTM of 0.18 and a pTM of 0.28; the monomer pTM was 0.51 (**Figure**

11A). Protein 2 yielded ipTM/pTM values of 0.28/0.50 for two copies and 0.25/0.44 for three copies, with a monomer pTM of 0.94 (**Figure 11B**). Protein 3 produced ipTM/pTM values of 0.18/0.22 for two copies and 0.13/0.17 for three copies, with a monomer pTM of 0.23 (**Figure 11C**). Protein 4 generated ipTM/pTM values of 0.21/0.52 for two copies and 0.17/0.40 for three copies, with a monomer pTM of 0.84 (**Figure 11D**).

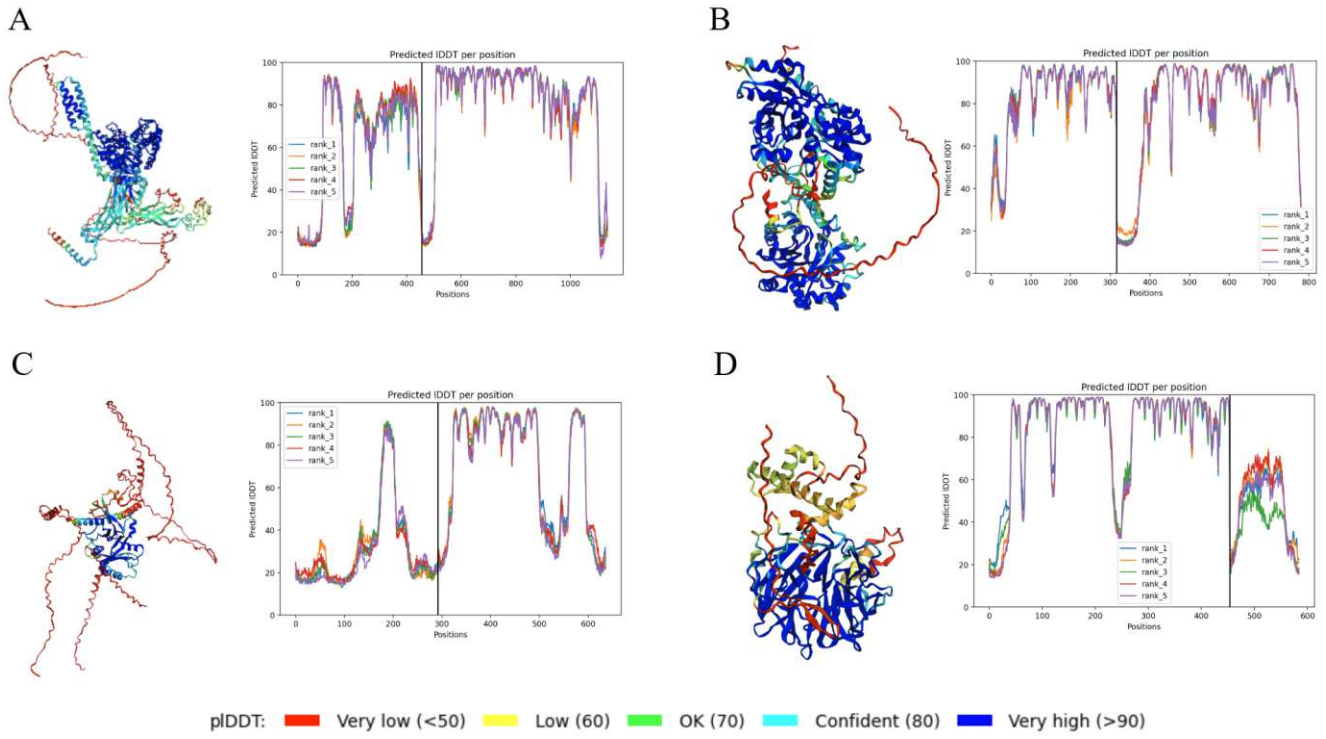


Figure 12. Multimer analysis with associated proteins predicted from STRING, using ColabFold.

(A) Predicted structure and pLDDT plot for Protein 1 paired with HSP70-9 [10, 13]. (B) Predicted structure and pLDDT plot for Protein 2 paired with tryptophan synthase (F4K727_ARATH) [10, 13]. (C) Predicted structure and pLDDT plot for Protein 3 paired with a Peter Pan-like protein [10, 13]. (D) Predicted structure and pLDDT plot for Protein 4 paired with histone HTA2 [10, 13].

AlphaFold-Multimer, run via ColabFold, produced five ranked heterodimeric models for each STRING-predicted interaction pair [10, 11, 13]. HSP70-9 reached pLDDT 85–95 across nearly its full length, and Protein 1 reached pLDDT 75–90 across its central region, dropping to 15–20

at both termini (**Figure 12A**). Protein 2 and tryptophan synthase, F4K727_ARATH, both reached pLDDT 85–100 across most of their length, with brief dips to 60–70 at the Protein 2 N-terminus and the tryptophan synthase chain start (**Figure 12B**). Protein 3 remained at pLDDT 20–50 across most of its length, with one region reaching 90, while its partner, the Peter Pan-like protein, showed two regions of pLDDT 85–95 separated by a region near 30 (**Figure 12C**). Protein 4 reached pLDDT 90–95 in several discrete regions separated by regions of 35–50, and histone HTA2, At4g27230, remained at pLDDT 40–70 across its full length (**Figure 12D**).

Discussion and Conclusions

Protein One: DnaJ Chaperone for Protein Folding

Protein One matches its closest homologs at 92.54% identity and clusters with *Arabidopsis* GFA2 in **Figure 1B** [4]. InterProScan identified a J domain, which is what J-proteins use to bind HSP70 and drive ATP hydrolysis, along with conserved sites for HSP70 interaction, zinc binding, and substrate binding, as shown in **Figure 9A** [9, 20]. Based on a 0.7 high-confidence cutoff, as depicted in **Figure 10A** [11], the closest *Arabidopsis* counterpart using STRING connects to HSP70-family chaperones and DJC76 with a confidence range of 0.745 to 0.875. The ColabFold model of Protein One with HSP70-9 points to a functional association. **Figure 12A** indicates a pLDDT score of 85 to 95 for the HSP70-9 model, and more specifically a score of 75 to 90 for the central region with a score of only 15 to 20 for both termini [10, 13]. Homomer runs returned an ipTM of 0.41 for two copies and 0.18 for three, with a monomer pTM of 0.51, **Figure 11A** [27]. This evinces no predictable stable homo-oligomer.

WoLF PSORT tied the chloroplast and mitochondrial classes at 6.5 each, assigning the combined chlo_mito class a score of 7 and the nucleus score of 1 [6]. The neighbor list in **Figure 5A** resolves the tie. Most chloroplast neighbors are RuBisCO small subunits, which have nothing to do with J-proteins, whereas the mitochondrial neighbor At5g48030.1 is GFA2, a protein targeted to mitochondria that partly rescues a yeast *mdj1* mutant [28]. HSP70-9 and HSP70-10, both in the STRING network, are annotated as mitochondrial [12]. Plastids also hold J-protein and HSP70 pairs, so family membership alone cannot rule out a chloroplast location [22]. The first 45 residues of Protein One include six arginines, two lysines, one aspartate, and five serines. That basic, acid-poor profile is typical of mitochondrial presequences, despite the fact that mitochondrial and chloroplast targeting peptides are both highly enriched in positively charged amino acids, stemming from shared endosymbiotic ancestry [21]. Mitochondria are therefore the better-supported location because chloroplast targeting peptides are generally neutral or marginally basic. TargetP 2.0 [29] or a fluorescent fusion of the first 60 residues would investigate this. If this prediction stands, Protein One may also act in synergid cell death, the process in which GFA2 was first identified [28].

Protein Two: Indole-3-Glycerol Phosphate Synthase for Tryptophan Biosynthesis

Protein Two is the best supported of the four. It needed no correction and matches *Arabidopsis* IGPS at 97.15% identity, and InterProScan independently assigns it IGPS function [4, 9]. Its monomer pTM of

0.94 was the highest in the set. The homomer ipTM values are 0.28 and 0.25, which is expected for a single-chain enzyme (**Figure 11B**) [27]. **Figure 10B** highlights how STRING partners are tryptophan-pathway enzymes, scoring 0.991 to 0.998 [11]. The modeled partner F4K727, also known as At5g28237, is annotated as a beta-type tryptophan synthase, not the alpha subunit, with a short N-terminus and a predicted cytosolic location [12]. Tryptophan synthase alpha and beta subunits form a complex in maize, so association among enzymes of this pathway is plausible [23]. Both chains in the model folded well, with pLDDT 85 to 100 in **Figure 12B**.

WoLF PSORT assigned Protein Two to the cytosol with a score of 10 [6]. However, tryptophan biosynthesis in *Arabidopsis* is chloroplastic [24]. The query has 316 residues, and four of the five homologs in **Figure 2C** run 364 to 433 residues and align only from residues 49 to 119 onward. Furthermore, this means they carry an N-terminal sequence that the query lacks [5]. The fifth, NP_001078734.1, is also 316 residues and aligns end-to-end. Protein Two may be a naturally short cytosolic form, which would suit its cytosolic partner F4K727. Alternatively, the gene model may be truncated at the 5’ end, as with the case of Protein Three in **Figure 3C**.

Protein Three: Ribosomal RNA-Processing Protein

After a 48-residue start-site extension and two gap fills taken from the homolog consensus, Protein Three reached 80.45% identity with its RRP15-like relatives [4, 5]. InterProScan assigns to the predicted

protein a classification of the RRP15 family (**Figure 9C**) [9]. The top STRING partners are centered around ribosomal subunit assembly and nucleolar RNA processing with scores between 0.982 and 0.998, namely BRIX1-1, BRIX1-2, EBP2, MEE49, and PPAN (**Figures 10C**) [11]. This partner set is coherent and points to a role in pre-ribosome assembly.

WoLF PSORT predicted the chloroplast with a score of 5, against 3 for the plasma membrane and 1.5 for the nucleus [6], which does not fit that role. The program has no nucleolar class, and the chloroplast neighbors in **Figure 7A**, including lycopene cyclase and chlorophyll-binding proteins, have no link to ribosome biogenesis. Nucleolar localization is a better fit for this data. Independently, the weak structural confidence metrics, specifically the monomer pTM of 0.23 and most residues at pLDDT 20 to 50, as presented in **Figures 11C and 12C** [10, 13], likely derive from intrinsic disorder. InterProScan places the family match in the C-terminal region and flags the N-terminal region and the far C-terminus as disordered (**Figure 9C**). This disorder is common among ribosomal proteins [25]. The single region at pLDDT 90, near residue 190, sits next to the start of the family match. The acidic stretch recovered from the consensus, DEEEEDESTK at residues 65 to 74, lies within the disordered N-terminal region. The quality of having a negatively charged N-terminal region with a well-folded core domain suggests Protein Three has a role in nucleolar liquid-liquid phase separation.

Protein Four: RCC1-Family Protein with a GEF-Related Domain and a Ubiquitin Ligase Label

Protein Four groups with UVR8/RCC1-family proteins in **Figure 4B** [4]. InterProScan reports RCC1 repeats and an RCC1-like domain across most of its 454 residues, as illustrated in **Figure 9D** [9], which

forms the RCC1 β-propeller architecture. The ubiquitin-ligase annotation in **Figure 9D** comes from the family entry IPR051210, titled Ubiquitin ligases and GEF domain-containing protein [9]. This annotation does not contain a specific catalytic E3 ligase domain. The label therefore records general sequence family membership and does not predict E3 activity here. The guanine nucleotide exchange factor (GEF) label is inherited the same way. The current sequence does not demonstrate true nucleotide exchange activity for Protein Four, even though canonical RCC1 proteins act as the standard Ran GEF across eukaryotes.

WoLF PSORT provided a localization score of 13 for the nucleus 13 against 1 for the cytosol, the highest score in this study [6]. STRING links the protein to RAN1 and the histone gene *HTA2* (**Figure 10D**) [11]. RCC1 itself binds the nucleosome [26]. Against this, the named neighbors in **Figure 8A** are histones and transcription factors, so the score reflects sequence composition and not family membership. An organellar role finds support in the STRING network, which was built on *Arabidopsis* RUG2, locus At5g48330, and contains five photosystem I subunits [12]. The photosystem I subunits include PSAK, PSAO, PSAG, PSAF, and psaJ [12]. Sequence-based prediction annotated RUG2 to the chloroplast [12]. The N-terminus of Protein Four in **Figure 8C** is also rich in arginine, serine, and threonine, with ten arginine or lysine residues and three acidic residues in the first 45, which parallels an organellar targeting peptide. The network scores, 0.613 to 0.792, are the lowest of the four proteins. A role as a UVR8-like UV-B photoreceptor is the least likely determination, since the network matches RUG2 and not UVR8, locus At5g63860. Protein Four is, thereby, an RCC1-family protein whose

location and activity remain unresolved. TargetP 2.0 [29] on the N-terminus and a fluorescent fusion would separate nuclear from organellar, and a RAN1 exchange assay would test the GEF label directly.

Limitations

Every assignment here is a computational prediction. High pLDDT and pTM values solely inform confidence in a model, not an in vivo interaction [10, 27]. STRING scores reflect shared function, and the networks were built on the *Arabidopsis* counterparts; therefore, physical interactions in *O. pumila* remain untested [11]. WoLF PSORT scores are nearest-neighbor counts, not probabilities [6]. All of these results are hypotheses to be tested with reporter-fusion localization and co-immunoprecipitation.

Acknowledgments

The author would like to thank Dr. Pellegrini and Charlotte Hu for supporting this research through thoughtful feedback on genome annotation and bioinformatics analysis.

Conflicts of Interest

The author declares no competing financial or commercial interests.

References

1. Al-Shehbaz, I. A., O'Kane, S. L., Jr., & Price, R. A. (1999). Generic placement of species excluded from Arabidopsis (Brassicaceae). Novon, 9(3), 296–307. https://doi.org/10.2307/3391724

2. Humann, J. L., Lee, T., Ficklin, S., & Main, D. (2019). Structural and functional annotation of eukaryotic genomes with GenSAS. In M. Kollmar (Ed.), Gene Prediction: Methods and Protocols (Methods in Molecular Biology, Vol. 1962, pp. 29–51). Humana. https://doi.org/10.1007/978-1-4939-9173-0_3

3. Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research, 33, W465–W467.

4. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389–3402.

5. Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., et al. (2016). MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics, 32, 3501–3503.

6. Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Research, 35, W585–W587.

7. Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., et al. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. Nature Biotechnology, 40, 1023–1025.

8. Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology, 305(3), 567–580.

9. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236–1240.

10. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.

11. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 51, D638–D646.

12. The UniProt Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531.

13. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nature Methods, 19, 679–682.

14. Huang, X., Yang, L., Jin, Y., Lin, J., & Liu, F. (2017). Generation, annotation, and analysis of a large-scale expressed sequence tag library from Arabidopsis pumila to explore salt-responsive genes. Frontiers in Plant Science, 8, 955. https://doi.org/10.3389/fpls.2017.00955

15. Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408, 796–815.

16. Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal, 89(4), 789–804.

17. Hall, A. E., Kettler, G. C., & Preuss, D. (2006). Dynamic evolution at pericentromeres. Genome Research, 16(3), 355–364. https://doi.org/10.1101/gr.4399206

18. Tang, A. J., Tian, M. H., & Long, C. L. (2009). Environmental control of dormancy and germination in the seeds of short-lived Olimarabidopsis pumila (Brassicaceae). Journal of Arid Environments, 73(3), 385–388.

19. Anthropic. (2026). Claude Sonnet 5.5 [Large language model]. https://claude.ai

20. Kampinga, H. H., & Craig, E. A. (2010). The Hsp70 chaperone machinery: J-proteins as drivers of functional specificity. Nature Reviews Molecular Cell Biology, 11(8), 579–592.

21. Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology, 300(4), 1005–1016.

22. Liu, C., Willmund, F., Whitelegge, J. P., Hawat, S., Knapp, B., Lodha, M., & Schroda, M. (2005). J-domain protein CDJ2 and HSP70B are a plastidic chaperone pair that interacts with vesicle-inducing protein in plastids 1. Molecular Biology of the Cell, 16(3), 1165–1177. https://doi.org/10.1091/mbc.E04-08-0736

23. Kriechbaumer, V., Weigang, L., Fießelmann, A., Letzel, T., Frey, M., Gierl, A., & Glawischnig, E. (2008). Characterisation of the tryptophan synthase alpha subunit in maize. BMC Plant Biology, 8, 44. https://doi.org/10.1186/1471-2229-8-44

24. Radwanski, E. R., & Last, R. L. (1995). Tryptophan biosynthesis and metabolism: biochemical and molecular genetics. The Plant Cell, 7(7), 921–934.

25. Peng, Z., Oldfield, C. J., Xue, B., Mizianty, M. J., Dunker, A. K., Kurgan, L., & Uversky, V. N. (2014). A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cellular and Molecular Life Sciences, 71(8), 1477–1504. https://doi.org/10.1007/s00018-013-1446-6

26. Makde, R. D., England, J. R., Yennawar, H. P., & Tan, S. (2010). Crystal structure of the chromatin factor RCC1 bound to the nucleosome core particle. Nature, 467(7315), 562–566.

27. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2021). Protein complex prediction with AlphaFold-Multimer. bioRxiv. https://doi.org/10.1101/2021.10.04.463034

28. Christensen, C. A., Gorsich, S. W., Brown, R. H., Jones, L. G., Brown, J., Shaw, J. M., & Drews, G. N. (2002). Mitochondrial GFA2 is required for synergid cell death in Arabidopsis. The Plant Cell, 14(9), 2215–2232. https://doi.org/10.1105/tpc.002170

29. Almagro Armenteros, J. J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. Life Science Alliance, 2(5), e201900429. https://doi.org/10.26508/lsa.201900429.