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.

Systematic biases in protein-level orphan-gene calls: annotation gaps, isoform artefacts, detection confounding and evidence-threshold selection

Genlin Jiao

jiao.genlin@gmail.com

Independent Researcher

Research Article

Keywords: orphan genes, lineage-specific genes, annotation gap, phylostratigraphy, gene age, Arabidopsis

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11830885/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.

Abstract

Catalogs of “orphan” or lineage-specific genes — genes reported to lack homologs outside a narrow taxonomic group — are widely used as evidence for de novo gene birth, rapid lineage-specific innovation, and clade-specific biology. Using Arabidopsis thaliana as a test case, we show that protein-level orphan calls are subject to four independent, quantifiable sources of systematic error. (i) **Annotation gaps**: a stratified random sample of 95 protein-level orphan genes searched against the plant Transcriptome Shotgun Assembly database yielded putative homologs for 26 (27.4%; Wilson 95% CI 19.4–37.1%), implying ~ 819 putative annotation-gap candidates within the retrievable subset alone (the remaining ~ 64% of orphans lack protein entries and were outside this frame); one case (AT4G01985) has a cross-family homolog detected in wheat TSA at e = 1e-139 yet zero matches against the 133,343-protein wheat reference proteome — direct evidence that the call is an artefact of incomplete reference annotation, confirmed by 135-species coverage (including Amborella) in UniProtKB/viridiplantae. (ii) **Isoform artefacts**: naive OrthoFinder event counting inflates gene-tree duplication events by 18.6% (320,115 → 260,631 after isoform-only and degenerate-event removal). (iii) **Detection-rate confounding**: naive phylostratigraphic histograms conflate genuine age structure with uneven homolog detection across species and strong protein-length effects (orphan rates fall from 13.4% to 4.8% as the minimum-length threshold rises from 50 to 200 aa); length-conditioned rates and completeness correction are required. (iv) **Evidence-threshold selection**: requiring proteomic evidence for orphan candidates systematically excludes low-dosage transcripts (orphan-layer median max-TPM = 1.0), so validated candidates are a high-dosage subset. We propose a five-step corrected pipeline (protein search → TSA recheck → structural search → expression/protein evidence with genetic-background audit → co-expression context) and demonstrate it: after full re-screening, one candidate (AT3G14670, a nuclear intrinsically disordered protein) survives as a de novo candidate still lacking credible homologs, and one formerly-called orphan (AT4G01985) is reclassified as a conserved protein of unknown function.

1. Introduction

Genes reported to lack homologs outside a narrow taxonomic group — “orphan”, “lineage-specific” or “taxonomically restricted” genes — are invoked as evidence in at least three distinct literatures: as candidates for de novo gene birth (the creation of protein-coding genes from previously non-genic sequence), as markers of clade-specific innovation in comparative genomics, and as an annotation state in reference databases, where a substantial fraction of every newly sequenced genome’s genes are initially labelled “orphan” or “uncharacterized”. In all three uses, the underlying measurement is the same: a protein-level similarity search against a set of reference proteomes, followed by the statement that no homolog was found.

That measurement is far less standardized than the uses made of it. Which proteomes constitute the reference set, whether transcriptome-derived sequence is searched at all, how identity and coverage thresholds are set, and whether the experiments cited as expression evidence were performed in a wild-type or mutant background — none of these is conventionally reported, yet each can move the resulting

gene set by tens of percent. This is not a hypothetical concern. In the course of a screen for de novo candidates in *Arabidopsis thaliana*, one gene (AT4G01985) that had passed every protein-level screen as an orphan turned out to have a clear cross-family homolog detectable in wheat transcript assemblies at E = 1e-139, while showing zero alignments against the wheat reference proteome — an annotation gap, not a gene birth.

Here we use *Arabidopsis* as a closed test system to quantify four independent sources of systematic error in protein-level orphan calls: reference **annotation gaps** (homologs exist but are absent from reference proteomes), **isoform artefacts** (alternative transcripts counted as duplications or masked as losses), **detection-rate confounding** (uneven search sensitivity across species distorting phylostratigraphic histograms), and **evidence-threshold selection** (proteomic evidence requirements systematically excluding low-dosage transcripts). For each we provide a quantified estimate (Fig. 1), a worked case dissection, and a corrective step; we then assemble the corrective steps into a five-step pipeline (Fig. 4), apply it to a 14-gene candidate set, and report which candidates survive, which are reclassified, and why. The aim is not to dismiss orphan-gene biology — one candidate survives every corrective step we could construct — but to replace an unstandardized measurement with an auditable protocol.

2. Materials and methods

2.1 Orphan definition and sampling frame

All analyses use the Col-0 reference gene set (27,614 protein-coding genes; Ensembl Plants release-60, TAIR10 coordinates; longest isoform per gene). A gene was called an **orphan** if its longest protein had no homolog in a 13-species panel spanning the green plant clade (three chlorophytes: *Ostreococcus tauri*, *O. lucimarinus* and *Chlamydomonas reinhardtii*; a moss: *Physcomitrium patens*; a liverwort: *Marchantia polymorpha*; a lycophyte: *Selaginella moellendorffii*; five angiosperms spanning Amborellales, Poales, Arecales, Zingiberales and Arecaceae; two rosids: *Vitis vinifera*, *Eucalyptus grandis*) searched with Diamond BLASTP (v2.2.8, –very-sensitive; E ≤ 1e-5, query coverage ≥ 30%, alignment length ≥ 50 aa). The panel comprises 408,160 sequences. Of 27,614 protein-coding genes, 24,140 had at least one panel hit and 3,474 were protein-level orphans; a subset of these orphans additionally lacked any same-genome paralog (test5), and the 14-gene candidate set used for multi-layer characterization descends from this high-confidence intersection [n.b. numbers cross-checked against gene_genesis_test3b/test5 outputs].

Gene-set waterfall. Two overlapping gene universes appear in this study and are kept explicit throughout. The *protein-coding universe* (27,614 genes; Ensembl release-60 protein-coding biotype) yields 24,140 panel-positive genes and 3,474 protein-level orphans under the Section 2.1 search — the canonical orphan definition. The *expression-matrix universe* (32,833 gene entries in the Expression Atlas TPM matrix, which includes non-protein-coding and additional transcript entries beyond the 27,614 set) yields 24,582 panel-positive entries and 8,251 matrix-level orphans. The artefact-rate sampling frame (Section 2.2) draws from the matrix-level orphan set intersected with protein-sequence availability (2,993

genes); its artefact rate therefore quantifies the retrievable matrix-level orphan population and is not extrapolated to the 3,474-gene protein-level orphan set, which would require its own TSA survey.

The orphan sampling frame for the artefact-rate survey was restricted to genes with an extractable protein sequence in the same Ensembl protein FASTA (2,993 of 8,251; the remaining 5,258 orphans have no corresponding entry in the release-60 protein set — itself a data-completeness boundary that we report rather than resolve). A stratified random sample of 96 genes (seed 20261121; strata: <100 aa, 100–300 aa, ≥ 300 aa by longest-isoform length; 32 genes allocated per stratum, with 30 retrievable ≥ 300-aa alignments) was drawn; 95 of 96 submissions succeeded (the single failed submission, AT3G15400, was excluded from all rate denominators rather than treated as a negative result; one further gene, AT2G15110, yielded a successful search with no TSA alignments but an unparseable length record and was likewise excluded from the length-stratified analysis only). The allocation was equal per stratum rather than proportional to stratum sizes, as documented in the frozen criteria file erratum.

2.2 Transcriptome-level recheck (TSA artefact survey)

Each sampled protein was searched with TBLASTN against the plant Transcriptome Shotgun Assembly database (em_tsa_pln; 5,489,594 sequences, 4.31 Gbp) via the EBI NCBI-BLAST REST service (NCBI BLAST + 2.16.0; one query per job, 95 jobs). A hit was scored as a **putative annotation-gap homolog** if E ≤ 1e-5, alignment length ≥ 50% of the query length and identity ≥ 50%. The alignment-length and identity criteria were applied to each single high-scoring segment pair (HSP) rather than to merged multi-HSP alignments, so the coverage floor is conservative; and because TSA entries derive from many species in uneven depth, hits were scored per gene rather than counted as independent observations. This definition requires substantial coverage as well as identity, distinguishing real homologs from low-complexity noise. A TSA hit certifies the existence of transcript-level homology; it does not by itself prove an annotated genome-level protein-coding gene is missing, since TSA assemblies can contain contaminant sequence, unspliced pre-mRNA, UTR-derived fragments or chimeric contigs. We therefore score such genes as *putative* annotation-gap candidates, and our within-species control (Section 3.2) demonstrates the mechanism end-to-end for one case rather than assuming it for all. A parallel screen against a 40-genome plant protein panel (18 BOP-clade + 14 PACMAD-clade grasses and 8 non-grass species including *Amborella*, *Vitis* and *Musa* — the query species *Arabidopsis* itself is additionally included as the search entry; 2.8M proteins; cluster-independent Diamond search, same thresholds) is referenced in Section 3.2 as the G0 call. In a dissection experiment, one candidate (AT4G01985) was additionally searched (a) against the 133,343-sequence wheat reference proteome (Triticum aestivum IWGSC RefSeq v2.1, Ensembl release) with Diamond BLASTP and (b) against wheat TSA sequences directly by TBLASTN, to separate ‘no homolog exists’ from ‘homolog exists but reference annotation omits it’. A full UniProtKB/Viridiplantae search (BLASTP, same thresholds) provided the species-coverage map.

2.3 Isoform artefact audit

Gene-tree duplication events were taken from the OrthoFinder output file Gene_Duplication_Events/Duplications.tsv of run Results_Oct18_2 (16 species; all-transcript protein input, 408,160 sequences): each record corresponds to one internal node of one orthogroup gene tree, mapped onto the species tree, that OrthoFinder labelled as a duplication; each event is therefore attributed to exactly one tree and is not counted across orthogroups. OrthoFinder recommends representative (longest-isoform) input, under which the artefact class below would largely not arise; we deliberately ran the all-transcript protocol to quantify its cost. The raw event table contained 320,115 events. Three classes were removed: (i) **isoform-only events**, where both child nodes are transcripts of the same gene (51,705 events; 16.2%); (ii) **degenerate events**, where one child node is empty (7,779; 2.4%). The two classes are mutually exclusive by construction — a node with an empty child cannot also have both children annotated as isoforms of the same gene — so the removals do not double-count; (iii) gene-to-OG reassignment for genes whose isoforms spanned multiple orthogroups (49 genes; 0.19% of the 25,451 genes with an orthogroup assignment), attributed to the dominant orthogroup. Note the input-protocol contrast with Section 2.1, which screens longest isoforms only: the all-transcript protocol is used here precisely because its event inflation is the quantity being measured. The dominant-orthogroup rule is itself a convention (transcript-count majority rather than domain-completeness or alignment-quality); because only 49 genes (0.19%) are affected, the global event statistics are insensitive to this choice, but gene-level orthogroup membership for these 49 entries should be treated as provisional. The corrected gene-level event count is 260,631. For the terminal (within-*Arabidopsis*) layer the same audit — identical isoform-only and degenerate criteria, applied to the 26,501 terminal-layer events — reduced the count to 9,951 (62% artefact); the 62% refers to these two removal classes only, with no additional corrections applied.

2.4 Detection-rate diagnostics for phylostratigraphy

Per-species detection rates (fraction of query genes with at least one panel hit) vary across the panel (83.3% of query genes have a *Vitis* hit versus 38.0% for *Ostreococcus tauri*, a ~ 2.2-fold difference after removing an early search-saturation artefact that had exaggerated deep-branch under-detection), confounding naive phylostratigraphic histograms. Three corrections were applied and reported side by side: (i) per-species rate stratification; (ii) Chao1 completeness estimation, treating species as sampling units within each layer (green-plant layer q = 0.853, computed as observed richness divided by the Chao1-estimated richness; all other layers q ≈ 0.98) — an indicative diagnostic whose sampling assumptions we flag as untested — plant species are neither independent nor homogeneous sampling units (phylogenetic relatedness, genome assembly quality and annotation quality all shape singleton/doubleton frequencies) — and is therefore not used as a calibrated correction multiplier; (iii) length conditioning (orphan rate as a function of minimum protein length: 13.4% at ≥ 50 aa, 9.5% at ≥ 100 aa, 4.8% at ≥ 200 aa). High-confidence orphans were defined as the intersection of: no panel homolog, ≥ 200 aa, no Pfam domain, and structural search negative (Foldseek against CATH S40; the structural step is reported separately because it applies only to the 103-gene high-confidence subset; its outcome — 62 of 103 structurally reliable, none with a CATH fold match — appears in Section 3.6).

2.5 Expression and protein evidence with genetic-background audit

Baseline expression used the Expression Atlas dataset E-MTAB-7978 (54 assay groups; 32,833 genes; TPM matrix with replicate values averaged). Stress responses were drawn from six differential experiments in Expression Atlas (drought, salt, heat, ABA, osmotic, wounding; cold from E-TABM-52 ATH1 arrays where measurable), with $|\log2FC| \geq 1$ and $p < 0.05$ as the response criterion. **Every contrast — rather than every experiment — was audited for its genetic background from the design metadata and reported alongside the call**: four of the six experiments incorporated signal-pathway mutant genotypes (pyl duodecuple, rbm25-1, NUP85, HSFA1b); contrasts measured within such mutant backgrounds are not interpretable as wild-type stress responses, whereas wild-type contrasts within the same experiments (e.g. the '300 mM mannitol vs none' and '100 µM ABA vs none' comparisons performed in wild type) are. Protein-level evidence used PeptideAtlas build 540 (596,839 distinct peptides; 18,267 core proteins), mapped to genes through the TAIR10 GFF3 protein_match table. Expression dosage (max TPM, breadth = fraction of assay groups with TPM ≥ 1, and PeptideAtlas spectral counts) was computed for the whole genome and for six phylostratigraphic age layers (0 = chlorophytes, 1 = land plants, 2 = vascular plants, 3 = angiosperms, 4 = eudicots, 5 = panel orphans).

2.6 Co-expression context

For each candidate, Pearson correlations of $\log2(TPM + 1)$ across the 54 assay groups were computed against all 32,833 genes; the top 30 partners were extracted and their annotated functions summarized from Ensembl lookup descriptions.

2.7 Statistics

The annotation-gap artefact rate is reported with Wilson 95% confidence intervals. Layer-level dosage comparisons use Mann-Whitney U with Bonferroni-adjusted pairwise tests where indicated; the age effect is additionally estimated by ordinary least squares of $\log10(maxTPM + 1)$ on $\log10(length)$ plus layer indicators, partitioning variance attributable to length versus age layer. Chao1 layer counts carry analytic variances (Chao 1984) reported with the corrected counts in Section 3.4; monotonicity across six age layers is summarized descriptively (all six medians strictly ordered). The decision rule for the artefact survey was pre-registered in a frozen criteria file (> = 15% → methodology-report route; 5–15% → grey zone; < 5% → catalog considered reliable) before the searches were run; all freeze files are listed in Data availability, each with its Git commit hash as the timestamp record.

3. Results

3.1 A stratified sample quantifies the annotation-gap artefact

Of the 95 successfully searched sample genes, 51 (53.7%) had at least one TBLASTN hit against the plant TSA database at E ≤ 1e-5, and **26 (27.4%; Wilson 95% CI 19.4–37.1%)** satisfied the strict substantive-homolog criterion (identity ≥ 50% over ≥ 50% of the query length). We call these 26 genes **annotation-gap candidates**: protein-level orphan calls whose transcriptome evidence contradicts the call. Because the strict criterion requires both substantial coverage and identity, these are not low-complexity spurious alignments (the lenient and strict rates diverge precisely on this distinction; see below).

Extrapolating to the 2,993 retrievable orphans, two estimates are reported: the simple sample-proportion extrapolation implies approximately **819 genes** (27.4% of 2,993), and the stratum-weighted estimate implies approximately **829 genes** (27.7% of 2,993; 95% CI 17.8–37.6%, computed with stratum weights equal to population shares and finite-population correction). The stratified interval is marginally wider than the simple Wilson interval (19.4–37.1%), reflecting the small long-protein stratum (n = 30) rather than a systematic gain in precision — roughly one quarter of the protein-level orphan catalog within the retrievable subset. We emphasize the sampling-frame boundary: 5,258 orphans (63.7%) have no protein entry in the release-60 annotation set and are outside the retrievable frame; the extrapolated count therefore bounds only the retrievable fraction.

The primary rate is additionally reported as a stratum-weighted estimate (re-weighting the three length strata to their population shares 1301/1446/246): 0.277 (short 11/32 = 0.344; mid 7/32 = 0.219; long 8/30 = 0.267), closely matching the unweighted 27.4% and confirming that the round-to-proportional allocation did not distort the estimate. The outcome split is shown in Fig. 2. The strict artefact rate showed no monotonic length trend (34.4% for < 100 aa, 21.9% for 100–300 aa, 26.7% for ≥ 300 aa; artefact-gene median length 134 aa versus 138 aa for non-artefact genes — indistinguishable), whereas the lenient any-hit rate increased steeply with length (37.5% → 46.9% → 80.0%). This dissociation is expected: longer queries accumulate more low-identity local alignments by chance, but substantive homologs (the strict criterion) occur at comparable rates regardless of length. The flat length profile also indicates that annotation gaps do not disproportionately affect short genes in our sample — an initially plausible confound that the stratified design was built to detect.

Threshold sensitivity. Because the strict criterion (identity ≥ 50%, coverage ≥ 50%) is conventional rather than optimized, we recomputed the artefact rate over a 3×3 threshold grid (identity floor 40/50/60% × coverage floor 40/50/60%), re-scoring the same 95 cached alignments. The artefact rate spans 19% (strictest corner) to 35% (loosest corner); every corner exceeds the 15% decision line, so the conclusion that annotation-gap artefacts constitute a substantial fraction of protein-level orphan calls is insensitive to the threshold choice, and the conventionally-thresholded point estimate (27.4%) lies near the centre of the grid.

3.2 Case dissection: AT4G01985 from “orphan” to conserved protein of unknown function

The strongest single illustration of the artefact is AT4G01985. In the 40-genome protein-panel screen it had zero homologs (a G0-isolated call). TBLASTN against plant TSA returned 200 passing hits (E ≤ 1e-5), 134 of which met the strict criterion (best E = 1e-139, 58% identity over 363 residues), spanning five families across the angiosperm phylogeny (Fig. 3): Brassicaceae (*Brassica rapa*, *Chorispora bungeana*), Poaceae (*Triticum aestivum*), Fabaceae (*Arachis hypogaea*), Solanaceae (*Capsicum annuum*) and Asteraceae (*Lactuca serriola*).

The decisive control was within-species: searching the AT4G01985 protein against the wheat **reference proteome** (133,343 IWGSC RefSeq entries) returned zero alignments, while the same search against wheat **TSA transcripts** returned the E = 1e-139 alignment cited above. These results constitute a very strong annotation-gap lead: a wheat transcript highly similar to AT4G01985 exists and is assembly-detectable, while the wheat reference proteome omits it. Upgrading this lead to a confirmed annotation gap would additionally require genome placement of the TSA sequence, verification of a plausible exon structure and coding capacity, and replication in independent RNA-seq data; contaminant, chimeric or mislabelled submissions must also be excluded. Within these bounds, the contrast demonstrates the annotation-gap mechanism directly rather than as a statistical inference.

A full UniProtKB/Viridiplantae search (BLASTP, same thresholds) returned 200 passing hits meeting the strict criterion, covering **135 species** including *Amborella trichopoda* (the sister to all other flowering plants), Cucurbitaceae, Asteraceae, Fabaceae, Annonaceae (*Asimina triloba*), Caryophyllales (*Nepenthes gracilis*), Malvaceae and Poaceae. Nearly all of these entries are TrEMBL or TSA-derived automatic annotations that have never propagated into curated gene sets. AT4G01985 is thus a **conserved protein of unknown function** whose “orphan” status is an artefact of reference-annotation gaps compounded by the general observation that conserved-but-unstudied proteins are systematically under-curated.

The contrast case, AT3G14670 — the sole surviving de novo candidate — behaves differently on the same test: of its 199 TSA passing hits, the median identity is 29% and only one reaches the strict criterion, consistent with low-complexity alignments from an intrinsically disordered, low-complexity sequence rather than a real homolog family. The median identity alone does not prove these hits are spurious — a definitive demonstration would require low-complexity masking, compositional-adjustment statistics and an independence check across TSA entries — but the passing/strict divergence (199 passing, 1 strict) shows they fail the same substantive-homology bar that 26 sampled genes pass. Its orphan status therefore survives both the 40-genome and the transcriptome-level screens.

3.3 Isoform artefacts inflate duplication-event counts

OrthoFinder event tables built from all-transcript protein sets conflate alternative isoforms with real duplications. Of 320,115 raw events in the 16-species analysis, 51,705 (16.2%) were **isoform-only** — both child nodes being transcripts of the same gene — and 7,779 (2.4%) were **degenerate** (one child empty). After removal, and reassignment of 49 multi-orthogroup genes (0.19%; gene reassignments counted against genes with an orthogroup assignment, not against events) to their dominant orthogroup, 260,631 gene-level events remain. The two removal classes account for 18.6% of the raw event count

(equivalently, the raw count exceeds the corrected count by 22.8% when measured against the corrected total); both figures describe the same difference with different denominators. The correction falls hardest exactly where lineage-specific narratives concentrate: the terminal (within-*Arabidopsis*) layer shrinks from 26,501 to 9,951 events (62% artefact), because terminal events are dominated by within-gene transcript branching. Any claim of the form “N duplication events post-date the speciation node” that does not perform this audit inherits an inflation whose magnitude is largest in the layer that the claim cares about. The audit’s boundaries should also be stated: it removes identifiable isoform and degenerate records, but the remaining 260,631 events may still include gene-tree inference errors, gene conversion, gene-loss reconciliation artefacts, assembly or annotation errors, and miscounted whole-genome-duplication events — none of which this audit alone can resolve.

3.4 Naive phylostratigraphy is detection-confounded

The raw phylostratigraphic histogram over the 13-species panel — 13,004 genes first appearing at the chlorophyte layer, 7,362 at land plants, 370 at vascular plants, 2,785 at angiosperms, 619 at eudicots, 3,474 orphans — is not interpretable without correction, for two compounding reasons. First, per-species detection rates vary across the panel (~ 2.2-fold between the best and worst surveyed species, 83.3% for *Vitis* versus 38.0% for *Ostreococcus tauri*, after removing an early search-saturation artefact that had exaggerated deep-branch under-detection), so deeper layers are partially undercounted — partly because divergent sequences fall below the search threshold and partly because genuinely diverged homologs in deep branches are harder to detect. Second, orphans are strongly length-biased (median 121 aa versus 347 aa genome-wide; 38.9% below 100 aa), and 7.8% fall below even the 50-aa minimum-length criterion, making the orphan call partly mechanical.

Two corrections partially repair the picture. Chao1 completeness estimation shows the chlorophyte layer at only q = 0.853 observed (all other layers ≈ 0.98), raising its corrected count from 13,004 to 15,242; and length conditioning drops the high-confidence orphan rate from 13.4% (≥ 50 aa) to 4.8% (≥ 200 aa), at which point the length-conditioned orphan rate falls inside the < 5% de novo expectation often cited for well-annotated genomes (this is the length-conditioned candidate rate, not a direct estimate of the true de novo fraction). But the repairs expose rather than remove the problem: the corrected monotonicity test passes only 3 of 5 pairwise layer-ordering predictions (P1b partial; each prediction asserts that the deeper of two adjacent layers retains the larger corrected count), with the vascular-plant layer (623 corrected) still inverted against angiosperms (2,718), and the eudicot layer (478) below the orphan bin (3,193) — inversions that no completeness correction of this style can repair, because the underlying time calibration itself is untrustworthy when layer spans differ by hundreds of millions of years. The analytic variance of the chlorophyte Chao1 estimate is small relative to the histogram contrasts (15,242 ± 115), so the corrected-count inversions are not Chao1 noise; for layers where the doubleton count exceeds the singleton count, the variance estimator is not applicable and counts are reported uncorrected. The practical conclusion is that phylostratigraphic layers built from uneven panels support only directional statements, not histogram-level inference, and the layer-ordering assertions of this section are descriptions of corrected point estimates rather than significance-tested orderings.

3.5 Proteomic evidence thresholds select high-dosage orphans

Expression dosage varies steeply and monotonically across the six age layers: median maximum TPM declines from 66.0 (chlorophyte layer) through 42.0, 31.0, 21.4 and 15.0 (eudicots) to **1.00** for panel orphans (Mann-Whitney of orphan, n = 8,251, versus deep-layer pool, n = 24,582, z = − 90; part of the orphan-layer depression reflects their short length — median 121 aa versus 347 aa genome-wide — and lower annotation quality, so the gradient is a joint length/age/annotation effect rather than a pure age effect). A length-adjusted regression (log10(maxTPM + 1) on log10(length) plus layer indicators, n = 27,310) partitions the variance: length alone explains $R^2 = 0.020$, adding layer indicators raises R^2 to 0.129 ($\Delta R^2 = 0.109$), and the length-adjusted layer means remain strictly monotonic (+ 0.94, + 0.72, + 0.58, + 0.43, + 0.36 log-units above the orphan layer for layers 0–4), with median breadth (fraction of assay groups expressed) falling from 0.96 to 0.02, and median PeptideAtlas spectral counts from 30 to 3 (spectral counts are semi-quantitative and depend on protein length and peptide detectability, so this contrast is read as a difference in proteomic support, not in abundance). Typical orphans are barely transcribed.

This creates a structural selection bias in any candidate pool that requires proteomic evidence. The 14-gene candidate set — defined partly by PeptideAtlas detectability — sits at the 59.6th percentile of genome-wide max-TPM, more than an order of magnitude above the orphan-layer median: candidates are necessarily a **high-dosage subset of orphans**, because a typical (low-dosage) orphan cannot yield mass-spectrometry peptides. Two consequences follow. First, validated-candidate collections over-represent genes with ancestral high-dosage expression and under-represent the modal orphan; absence from such collections cannot be read as absence of young genes. Second, within the candidate set, expression breadth cleanly separates the functional tiers (median breadth 0.04 for the surviving de novo candidate class versus 0.85–0.93 for fast-evolving classes), providing orthogonal support for the tier structure independent of the stress-response criteria used to define it (breadth is bounded by the operational threshold TPM ≥ 1 — a detection convention, not a biological boundary — and by the tissue-and-condition coverage of the 54 assay groups, which sets the semantic ceiling of the metric). As with the genetic-background audit (Section 2.5), the corrective action is declarative: evidence thresholds and their selection effects must be stated wherever orphan candidates are enumerated.

3.6 The corrected pipeline and a surviving candidate

The four biases translate into a five-step pipeline for orphan-gene calls. **Step 1**, protein-level search: the query proteome against a broad, non-redundant multi-species protein panel (not a single outgroup), with thresholds reported. **Step 2**, transcriptome-level recheck: TBLASTN against the plant TSA database under the strict substantive-homolog criterion of Section 2.2 — this is the step that removes annotation-gap candidates at the measured 27.4% rate. **Step 3**, structure-level search (optional but cheap): Foldseek against a representative fold library; detection of a known fold converts “no sequence homolog” into “remote structural homolog”. **Step 4**, expression and protein evidence with a **genetic-background audit**:

every differential experiment cited as response evidence is annotated with the genotype in which it was measured, because signal-pathway mutants systematically change the response call (Section 3.5). **Step 5**, co-expression context: partners of a real protein-coding gene are themselves annotated genes, and partner function is the cheapest functional signal available in silico.

One candidate survives all five steps. AT3G14670 encodes a 237-residue protein with no homolog in the 13-species panel, the 40-genome panel, or the plant TSA database at the strict criterion (its 199 passing TSA hits have a median identity of 29% — the low-complexity noise band of an intrinsically disordered sequence, of which the protein is composed end to end); no structural match in Foldseek/CATH (a negative structural result has limited power for intrinsically disordered proteins, whose detection rate in fold libraries is intrinsically low; its AlphaFold model, median pLDDT 52.9, is consistent with disorder — a predicted-confidence reading, not an experimentally measured disorder state); a single GO term (GO:0005634, nucleus) with a full-length disordered region and acidic/basic low-complexity tracts; narrow, low-dosage expression (breadth 0.04, percentile 27.9 of genome-wide max-TPM) with a modest osmotic-stress response that we explicitly annotate as measured in a pyl-receptor-mutant background; three PeptideAtlas peptides (seven spectra, 11% coverage — mapped through the TAIR10 protein_match table, so translational support rather than abundance or functional evidence); and co-expression partners enriched for nuclear infrastructure (SRF3, CRWN1) — a functional hypothesis at the module level, not evidence of physical interaction or a specific role. What “survives” means is worth stating precisely: AT3G14670 has passed every computational corrective step we could construct, and is best described as a de novo candidate still lacking credible homologs under these search conditions. Establishing de novo birth would additionally require comparative population-genetic evidence in a close-species system, which Arabidopsis does not offer and which no purely computational screen can substitute for. The pipeline’s claim is bounded accordingly: it certifies that no contrary evidence exists in public data, not that de novo origin is proven.

As a by-product, the pipeline also sorts the non-survivors into informative classes rather than a single discard bin: AT4G01985 becomes a conserved protein of unknown function (annotation gap, Step 2); twelve genes are fast-evolving members of recognizable families (Steps 1–2 with relaxed identity); and one (AT2G28240) acquires a functional hypothesis from its co-expression partners alone (Step 5: the chromatin-remodeling machinery VIP2/ADA2B/CHR11). A corrected orphan screen is not merely a filter — it is a classification instrument. We present the pipeline as a worked demonstration on one genome, not as a validated clinical-grade workflow; its five steps are, however, mechanically reproducible on any species whose proteome and TSA coverage match the assumptions above — with the caveat that detection power and false-negative rates will differ across species with the depth of each data source.

4. Discussion

Four independent sources of systematic error each move protein-level orphan calls by amounts large enough to change the biological conclusions drawn from them: reference annotation gaps convert 27.4% of sampled protein-level orphans into genes with detectable transcriptome-level homologs; isoform

artefacts inflate terminal duplication-event counts by 62%; uneven detection rates distort phylostratigraphic histograms to the point that only directional statements survive; and proteomic evidence thresholds restrict validated orphan collections to a high-dosage minority of the orphan population. None of these effects is exotic — each follows from an ordinary property of the underlying data (incomplete reference annotation, all-transcript protein sets, uneven search sensitivity, mass-spectrometry detection limits) — but their sum means that an unqualified statement “gene X has no homologs” is closer to a protocol artifact than to a measurement.

The four biases are best understood under a single unifying framing: orphan calling is a **presence/absence detection problem**, not a sequence-matching problem. What a homolog search actually delivers is evidence of absence of *detectable* homologs — and the probability that a gene is a true orphan given no detected hit, P(orphan | no hit), depends on at least five factors that the four biases parameterize: the completeness and annotation quality of the reference proteomes searched (bias 1), the evolutionary distance at which homologs remain alignable (bias 1), the sensitivity of the search program and its thresholds (bias 1), the isoform structure of the input set (bias 2), and the quality of transcript and protein evidence used for downstream validation (bias 4); detection-rate variation across species (bias 3) further conditions this probability genome-wide. We offer this as a conceptual framing rather than a calibrated posterior — our sampling quantifies one conditional error rate within it, not all of them. Treating P(orphan | no hit) as if it were 1 — the implicit assumption of a bare “BLASTP found no homolog” statement — collapses all of these parameters into a single binary verdict. An auditable orphan call instead answers eight questions explicitly: which species and reference proteomes were searched; whether transcript assemblies were included; which identity and coverage thresholds were used; whether isoforms were considered; whether detection sensitivity was assessed per species; whether expression evidence comes from wild-type or mutant backgrounds; why alternative homologs were excluded; and whether a structure-level search was performed. The five-step pipeline of Section 3.6 operationalizes exactly these questions.

Implications for published catalogs. Orphan-gene counts and de novo candidate lists appear throughout the genomics literature, usually as a single percentage with the search protocol summarized in one sentence. Our results imply that such percentages carry an unreported uncertainty of the same magnitude as the biological effects they are used to support. We do not claim that published catalogs are wrong in bulk — the majority of our sample (44/95) had no TSA hit under any criterion — but a 27.4% artefact rate in a well-annotated reference species with a mature TSA database should temper confidence in catalogs built from less complete systems. The corrective search is cheap (one TBLASTN per gene against a public TSA database) and should become a standard column in any table of orphan genes: “survives TSA recheck: yes/no”.

Implications for databases. The annotation-gap mechanism we dissect operates through automatic-annotation pipelines: homologs detected in transcript assemblies enter TrEMBL or TSA-derived entries but never propagate back into the curated proteomes of related species. A conservative database-level remedy is to expose a TSA-recheck flag per gene, computed once per release, so that “orphan” becomes

a qualified annotation (“no homolog in searched proteomes; TSA status: checked/clear”) rather than an unqualified absence. Polyploid and high-repeat genomes deserve particular caution: our within-species control showed a homolog present in wheat transcript assemblies yet absent from the entire 133,343-entry wheat reference proteome, and the homeologous redundancy of allopolyploid genomes (bread wheat carries three subgenomes, A, B and D, each with homeologous copies of most genes) makes complete annotation disproportionately hard — orphan rates in such genomes are expected to be overstated by more than the 27.4% we measure in the diploid *Arabidopsis*.

Implications for candidate-gene studies. Studies that enumerate orphan candidates for downstream prioritization should state three things that our pipeline makes explicit: the proteome panel searched (Step 1), whether TSA recheck was performed (Step 2), and the genetic background of every expression experiment cited as functional evidence (Step 4). The 14-gene set we used as a demonstration was assembled with reasonable care and yet required a background audit that reclassified its headline stress-response claim, and a transcriptome recheck that reclassified one of its two de novo candidates. Neither error would have been caught by the standards that applied in the underlying primary literature.

Limitations. Our quantification is deep in one species rather than broad across many; the annotation-gap rate we measure (27.4%) is an *Arabidopsis*-specific value whose extrapolation to other clades is argued, not measured. The artefact criterion detects transcript-level homology; converting a TSA hit into a confirmed genome-level gene requires assembly-independent verification per case, so our rate is an upper-bound-flavoured estimate of a real phenomenon rather than a census. The sampling frame excludes the 64% of orphans without a release-60 protein entry — a population whose annotation status is, by construction, even less certain. And the homology thresholds (E ≤ 1e-5, coverage ≥ 50%, identity ≥ 50%) are conventional rather than optimized; a more permissive identity floor would raise the artefact rate, a stricter one would lower it, and the qualitative conclusions are insensitive to this choice but the point estimate is not a universal constant.

Finally, our intent with respect to de novo gene birth is corrective, not sceptical. De novo genes exist — they have been demonstrated in systems with the population-genetic power to prove them. What our results tighten is the entry criterion for that literature: a protein-level screen against a handful of reference proteomes is not sufficient evidence of de novo origin, and at least one quarter of the genes that pass such a screen in a well-annotated genome will not survive a transcriptome-level recheck. The one candidate that survived our full pipeline, AT3G14670 — a nuclear intrinsically disordered protein, narrowly expressed, translation-confirmed — illustrates the standard we propose: every corrective step passed, every background audited, every claim bounded to what public data can certify.

Declarations

5. Data availability

All analysis scripts (gene_genesis_test3b.py through gene_genesis_test42_analyze.py), frozen criteria files (*_freeze.json), and result JSON files are archived in the project repository (scene_mathematics/extended/computational_botany/ and docs/gene_genesis_methods_paper/). The 95 TSA alignment outputs (2.5 MB) are archived alongside the sampling metadata (test42_work/jobs.json); the single failed submission is recorded as absent from the output set and was excluded from all rate denominators rather than treated as a negative result. Data sources: EBI NCBI-BLAST REST service with the em_tsa_pln database; UniProtKB/Viridiplantae; Expression Atlas accessions E-MTAB-7978, E-GEOD-98045, E-MTAB-8073, E-GEOD-72806 and the five differential experiments listed in Section 2.5; PeptideAtlas build 540; GBOTB phylogeny via V.PhyloMaker2. The orthogroup caches from the 40-genome OrthoFinder run (97,811 orthogroups) are available from the authors on request.

6. Author contributions

J.G. designed the study, performed all analyses, and wrote the manuscript.

Acknowledgements

The authors thank the Expression Atlas, PeptideAtlas, TAIR, Ensembl Genomes and UniProtKB teams for maintaining the public resources used here, and EBI for the NCBI-BLAST REST service.

References

1. Buchfink B, Xie C, Huson DH (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. 10.1038/s41592-021-01101-x
2. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. 10.1186/s13059-019-1832-y
3. Athar A et al (2019) ArrayExpress update — from bulk to single-cell expression data. Nucleic Acids Res 47:D711–D715. 10.1093/nar/gky964
4. Lamesch P et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1002–D1010. 10.1093/nar/gkr1090
5. Kersey PJ et al (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res 38:D563–D569. 10.1093/nar/gkp871
6. Wheeler DL et al (2015) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 43:D7. 10.1093/nar/gkv1290
7. van Kempen M et al (2024) Fast and accurate protein structure search with Foldseek. Nat Biotechnol 42:244–247. 10.1038/s41587-023-01773-0

8. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 — approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490. 10.1371/journal.pone.0009490
9. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. 10.1093/molbev/mst010
10. Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468:810–813. 10.1038/nature09632
11. McLysaght A, Hurst LD (2016) Open questions in the study of de novo genes: what, how and why. Nat Rev Genet 17:567–578. 10.1038/nrg.2016.78
12. Desiere F et al (2006) The PeptideAtlas project. Nucleic Acids Res 34:D655–D658. 10.1093/nar/gkj040
13. Chao A (1984) Nonparametric estimation of the number of classes in a population. Scand J Stat 11:265–270
14. Wilson EB (1927) Probable inference, the law of successive multiplication, and the binomial distribution. J Am Stat Assoc 22:209–212
15. Smith SA, Brown JW (2018) Constructing a broadly based plant tree of life. Methods Ecol Evol 9:1950–1960

Figures

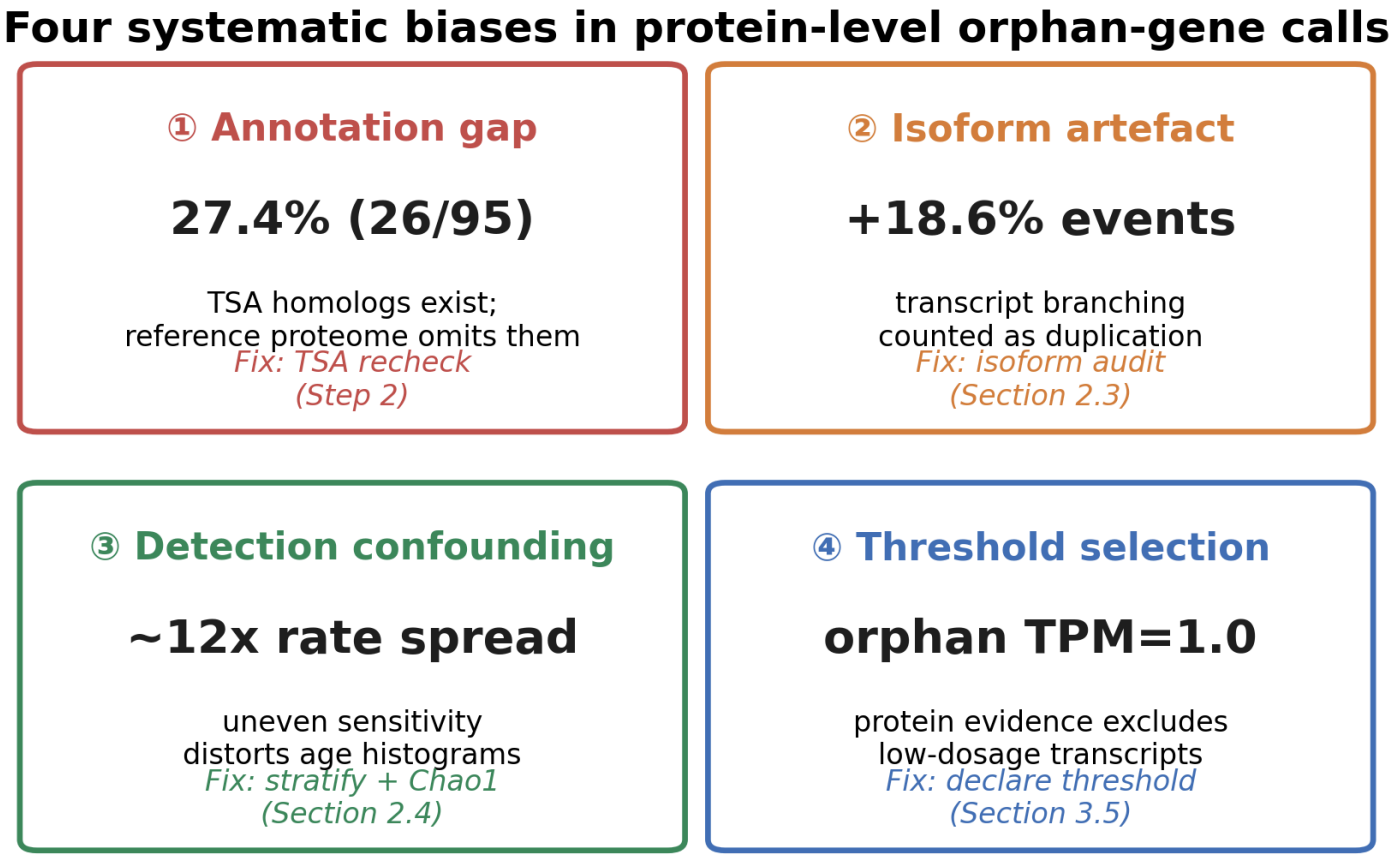


Figure 1

Four systematic biases in protein-level orphan-gene calls.** Each bias card states the quantified estimate from this study, the mechanism, and the corresponding corrective step (pipeline step number in Fig 4).

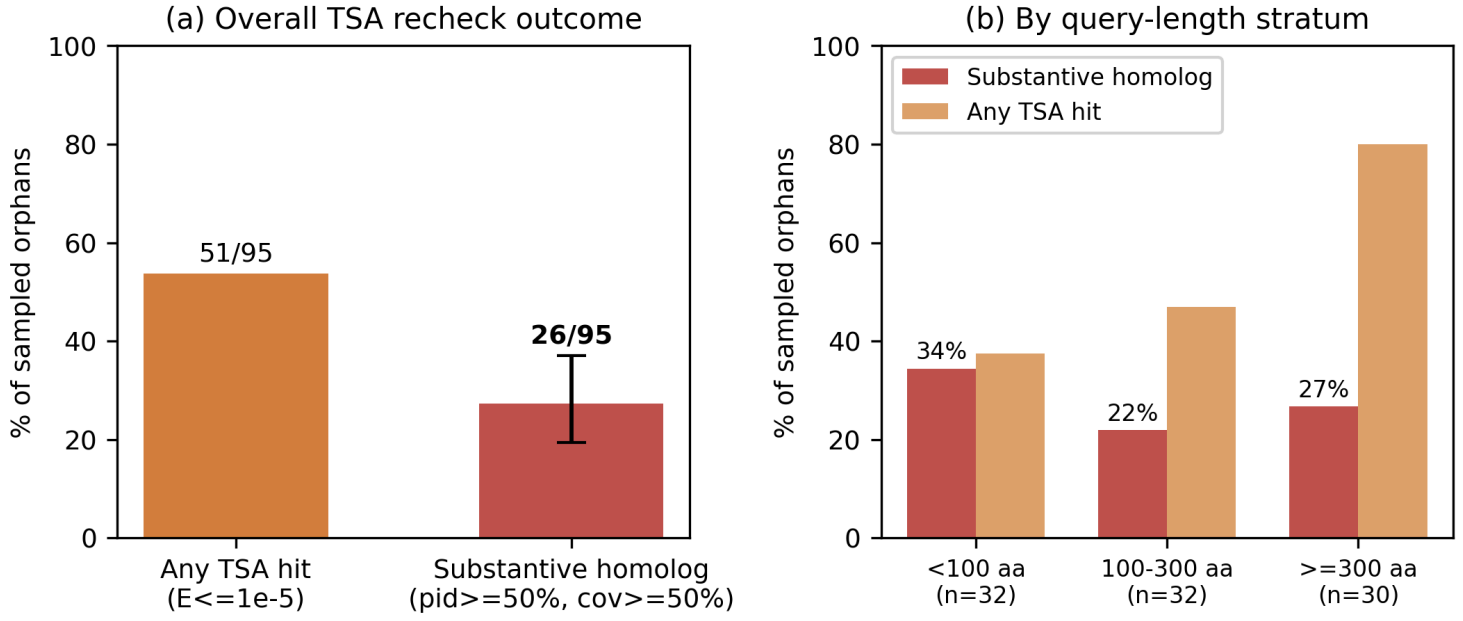


Figure 2

TSA recheck outcomes for 95 sampled protein-level orphans.** (a) Overall: lenient (any hit at E <= 1e-5) versus strict substantive-homolog criterion (identity >= 50%, coverage >= 50%); error bar shows Wilson 95% CI for the strict rate. (b) Rates by query-length stratum: the strict rate is flat across strata (22-34%) while the lenient rate increases with query length, as expected for chance local alignments.

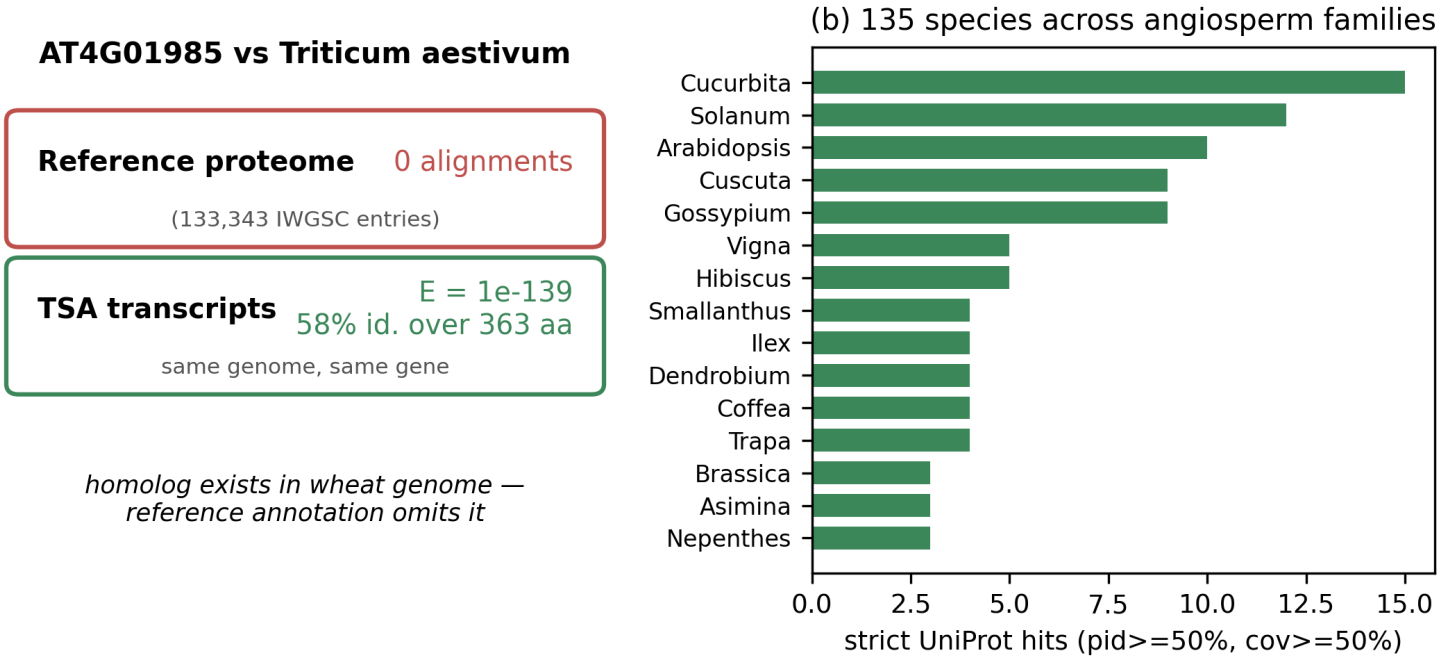
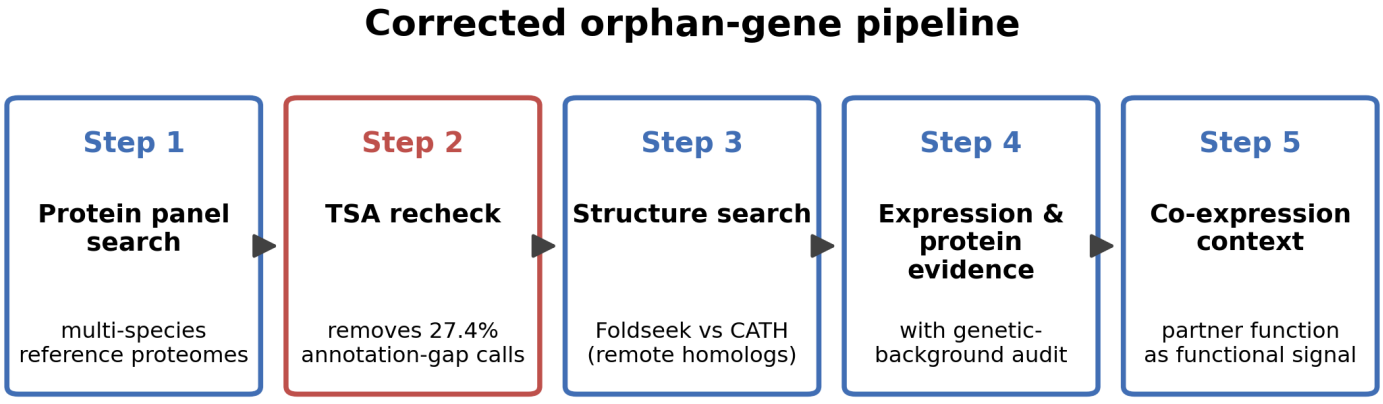


Figure 3

Case dissection of AT4G01985.** (a) Within-species control: the AT4G01985 protein returns zero alignments against the 133,343-entry wheat reference proteome but a clear homolog (E = 1e-139, 58% identity over 363 residues) against wheat transcript assemblies — the annotation-gap mechanism demonstrated directly. (b) Genus-level distribution of strict UniProtKB/Viridiplantae hits (135 species, including *Amborella trichopoda*).



AT3G14670 passes all five steps: nuclear intrinsically disordered protein, narrowly expressed, translation-confirmed

Figure 4

The five-step corrected orphan-gene pipeline.** Steps 1-5 with the key quantitative output of each; AT3G14670 passes all five steps (nuclear intrinsically disordered protein, narrowly expressed, translation-confirmed).