

SUPPLEMENTARY FILE S1

Public sequence archives reveal overlooked orchid barcode resources for biodiversity monitoring

Operational definitions, complete tables, extraction audit and reproducibility evidence.
Revision R1, 9 September 2026.

Table S1. Units and operational definitions.

Term	Definition
Archive frame	ENA sequence catalogue for tax_tree(4747), retrieved 3 September 2026; candidate inclusion follows the frozen configuration.
Taxid–marker unit	A distinct archive taxonomy identifier associated with one operational marker category; not necessarily an accepted species.
Standalone	Whole uniquely labelled short parent records and eligible features from parents of 150–5,000 bp.
Embedded	Eligible 150–5,000 bp annotated features extracted from parent records longer than 5,000 bp.
Hidden-only	A taxid–marker unit in E but not S; overlap is excluded from the numerator and counted once in the union.
Hidden fraction / increase	$H = 100 \times \text{hidden} / \text{union}$; $R = 100 \times \text{hidden} / \text{standalone}$. Core results sum separate <i>matK</i> and <i>rbcL</i> units.
COI-5P	Operational category matching COI, cox1 and related cytochrome-c-oxidase-I labels. A 5' barcode segment was not independently verified.
Minimum readiness	Syntactic pass plus a non-empty taxonomic label, material identifier, contextual field and accession.version; not a taxonomic validation.

COMPLETE MARKER RESULTS

Table S2. Quality-passing taxid–marker combinations by category.

Marker	Standalone	Embedded	Overlap	Hidden-only	Union	Hidden (%)
COI-5P	109	1	0	1	110	0.909
ITS	8,794	0	0	0	8,794	0.000
ITS1	4,492	54	12	42	4,534	0.926
ITS2	4,401	51	10	41	4,442	0.923
<i>atpF-atpH</i>	230	0	0	0	230	0.000
<i>matK</i>	7,243	925	695	230	7,473	3.078
<i>matK-like</i>	22	0	0	0	22	0.000
<i>psbK-psbI</i>	153	0	0	0	153	0.000
<i>rbcL</i>	3,226	911	528	383	3,609	10.612
<i>rbcL-like</i>	5	0	0	0	5	0.000
<i>rbcLa</i>	226	0	0	0	226	0.000
<i>rpoB</i>	500	905	104	801	1,301	61.568
<i>rpoC1</i>	704	898	183	715	1,419	50.388
<i>trnH-psbA</i>	2,670	0	0	0	2,670	0.000
<i>trnL-F</i>	4,088	0	0	0	4,088	0.000
<i>ycf1</i>	2,435	463	99	364	2,799	13.005

All markers combined: standalone 39,298; embedded 4,208; overlap 1,631; union 41,875; hidden-only 2,577 (6.154% of the union; 6.558% relative increase over standalone). *matK* + *rbcL*: union 11,082; hidden-only 613 (5.531%; relative increase 5.855%).

The combined core count does not indicate paired loci from the same individual. ITS/ITS1/ITS2 and explicit “-like” categories follow the frozen parser and are not merged after extraction.

QUALITY-PASSING REGIONS

Table S2b. Region counts and relative increases before collapse across marker categories.

Marker	Standalone regions	Embedded regions	Increase over standalone (%)
COI-5P	126	3	0.917
ITS	22,041	0	0.000
ITS1	8,493	58	0.935
ITS2	8,696	55	0.932
<i>atpF-atpH</i>	935	0	0.000
<i>matK</i>	18,340	1,594	3.175
<i>matK-like</i>	62	0	0.000
<i>psbK-psbI</i>	338	0	0.000
<i>rbcL</i>	7,767	1,555	11.872
<i>rbcL-like</i>	15	0	0.000
<i>rbcLa</i>	485	0	0.000
<i>rpoB</i>	1,153	1,541	160.200
<i>rpoC1</i>	1,669	1,674	101.562
<i>trnH-psbA</i>	5,997	0	0.000
<i>trnL-F</i>	9,186	0	0.000
<i>ycf1</i>	4,434	746	14.949

Region counts are not counts of specimens or species. The relative-increase denominator is the marker-specific standalone taxid set, not the number of regions.

SENSITIVITY ANALYSIS

Table S3. All 18 configured scenarios; the minimum of 180 ACGT bases remains fixed.

Length (bp)	Ambig . (%)	Label rule	Regions	Standalone	Embedded	Union	Hidden-only	Hidden (%)
150	0	Retain	90,144	37,006	4,203	39,601	2,595	6.553
150	0	Exclude	82,890	33,006	4,080	35,489	2,483	6.997
150	1	Retain	96,963	39,298	4,208	41,875	2,577	6.154
150	1	Exclude	89,184	35,004	4,084	37,468	2,464	6.576
150	5	Retain	97,870	39,583	4,209	42,149	2,566	6.088
150	5	Exclude	89,986	35,243	4,085	37,696	2,453	6.507
300	0	Retain	75,916	30,080	4,081	32,580	2,500	7.673
300	0	Exclude	69,803	26,786	3,958	29,174	2,388	8.185
300	1	Retain	81,831	32,085	4,084	34,565	2,480	7.175
300	1	Exclude	75,252	28,531	3,960	30,898	2,367	7.661
300	5	Retain	82,616	32,334	4,084	34,803	2,469	7.094
300	5	Exclude	75,956	28,737	3,960	31,093	2,356	7.577
500	0	Retain	66,150	26,638	4,041	29,275	2,637	9.008
500	0	Exclude	60,900	23,764	3,919	26,288	2,524	9.601
500	1	Retain	71,697	28,558	4,044	31,179	2,621	8.406
500	1	Exclude	66,031	25,438	3,921	27,945	2,507	8.971
500	5	Retain	72,427	28,789	4,044	31,401	2,612	8.318
500	5	Exclude	66,694	25,631	3,921	28,129	2,498	8.881

Ambig. = maximum ambiguous-base fraction. “Exclude” applies the literal uncertain-label token screen described in the main text. Zero gaps and valid IUPAC symbols are required throughout. The parent architecture threshold remains 5,000 bp. The nominal 150 bp floor is subordinate to the 180-ACGT requirement; it is not a 150-ACGT threshold.

AUDIT AND IMPLEMENTATION NOTES

Table S4. Reproduced quality-control and retrieval totals.

Check	Observed result
Candidates requested / satisfied	555,683 / 555,683
Missing candidates / version mismatches	0 / 0 (original retrieval provenance)
Extracted regions / accession versions	98,534 / 71,358
Syntactic quality pass	96,963 / 98,534 (98.406%)
Minimum readiness pass	22,099 / 98,534 (22.428%)
Emitted regions without sequence / without accession version	0 / 0; applies to emitted regions, not all parent records
Analytical replay	QC matrix identical after decompression; primary, sensitivity and readiness CSV tables identical.
Historical targeted extraction check	Eight fixed-stratum regions matched source hashes; superseded for technical extraction verification by the full R1 audit (Table S6).

Extraction boundaries

The parser emitted 68,891 whole-short-record, 22,400 short-record-feature and 7,243 long-record-feature regions before QC. It recorded 484,325 parents with no unambiguous eligible marker region and no parse errors. Feature-level exclusion events included 78,372 inexact-coordinate features, 3 non-extractable features and 2 ambiguous multimarker features; there were also 2 ambiguous record-label events. These event categories are not mutually exclusive and must not be summed as excluded parent records.

Ambiguous record descriptions do not suppress an independently resolvable eligible feature. Whole-record assignment uses description text; feature assignment uses annotation text. Neither is an independent homology test. Exact Biopython locations support joined and reverse-strand extraction. Stored start/end coordinates are outer bounds, not a serialization of joined locations.

Replay scope

Supplementary File S2 is retained unchanged and supports exact analytical replay from the frozen extracted inputs. Supplementary File S3 contains the original candidate EMBL records and the evidence needed to verify extraction independently at the code level. S3 is prepared but is not attached to this resubmission because it exceeds the portal upload limit. Its delivery is pending confirmation of a transfer method by the Editorial Office. The record index identifies source batches and uncompressed byte offsets; the region index records parent hashes, original locations, feature identity, sequence hashes and comparison outcomes. The full internal re-extraction matched all 98,534 regions. This establishes correspondence to the preserved source records, not biological correctness of their annotations.

The original extraction environment used Python 3.13.5, Biopython 1.85, pandas 2.3.2 and requests 2.32.5. The previously completed derived-input replay and the new standard-library extraction audit used Python 3.12.14. The protocol was frozen locally before extraction but was not externally preregistered. The analyses in Tables S5 and S7 were added after review and must not be described as predeclared.

POST-REVIEW OPERATIONAL SENSITIVITY

Table S5. Primary and post-review estimates within the frozen candidate frame.

Scenario	Regions passing QC	Standalone	Embedded	Union	Embedded-only	Hidden (%)
Primary, 5,000 bp	96,963	39,298	4,208	41,875	2,577	6.154
Joint 2,500 bp	95,171	39,213	3,350	41,024	1,811	4.414
Joint 10,000 bp	98,792	39,415	4,575	42,237	2,822	6.681
Structured features only	24,483	10,280	4,193	14,100	3,820	27.092
Without generic COI/ <i>cox1</i>	96,834	39,189	4,207	41,765	2,576	6.168
<i>matK</i> and <i>rbcL</i> only	29,256	10,469	1,836	11,082	613	5.531

The two threshold scenarios jointly change the parent-record architecture boundary and maximum accepted region length. The candidate frame remains the original 555,683 accession versions. Applying the candidate-selection rules anew would select 547,596 records at 2,500 bp and 561,611 at 10,000 bp; the latter includes 5,928 records not present in the frozen snapshot. These results are therefore not complete new inventories at alternative thresholds.

Table S7 separates the effect of parent-length classification by keeping the original region pool and all QC and marker assignments fixed. It does not replace the joint re-extraction results in Table S5.

The structured-feature-only scenario uses gene, product and standard_name qualifiers at 5,000 bp and excludes description-based whole-record assignments. Its smaller standalone set changes the denominator; the 27.092% value must not be interpreted as a measured annotation error rate or stronger biological validation. The COI-exclusion scenario removes the entire historically named COI-5P category because generic COI/*cox1* labels do not establish a specific 5' interval.

The primary estimate meets the original 5% descriptive effect criterion. The conditional 2,500-bp estimate does not meet that criterion, although it remains above the separate 2.5% robustness reference. These are descriptive criteria, not statistical significance thresholds. No sampling confidence intervals are assigned to this exact frozen-frame census.

COMPLETE SOURCE-RECORD VERIFICATION

Table S6. Scope and outcomes of the internal independent-code audit.

Check	Verified outcome
Original gzip batches and historical SHA-256	3,705; all matched
Candidate accession.version records	555,683; none missing
Candidate selection from the frozen catalogue	Exact match to original 555,683-record frame
Parent records without inline sequence	159,369; assembly-component records preserved
Regions compared against original records	98,534; no discrepancies in tested fields
Reverse-strand regions	4,848
Compound-location regions	71
QC-pass regions and primary set counts	96,963; union 41,875; embedded-only 2,577
Extracted FASTA versus frozen QC hashes	98,534 matched; no extra or missing FASTA identifiers
Non-extractable exact-location events	3, all at KZ504779.1; excluded in both implementations
Biological annotation and taxonomy validation	Not performed; outside this technical audit

The earlier eight-region targeted check has been superseded for technical extraction verification by exhaustive comparison of all emitted regions. The new parser does not import the original parser or Biopython, but deliberately retains the frozen scientific vocabulary, precedence and quality-control definitions. Agreement between implementations is therefore not independent evidence that the original biological annotation is correct.

The 159,369 assembly-component parent records contain original CO descriptions instead of inline SQ sequence. Their records are present, but the external component sequences are not reconstructed. KZ504779.1 declares 357 bp and refers to JSDN02142735.1:1..357; its three feature events at 64..219 cannot be extracted from an absent inline sequence. The audit preserved this exclusion rather than silently fetching or inferring sequence. All 98,534 emitted regions were verified against inline source sequence.

S3 includes a portable standard-library verifier, original logical batches and retrieval sidecars, the catalogue and candidate table, batch checksums, record and region indices, full audit results, and a member-level SHA-256 manifest. Historical catalogue queries and accessions are documented, but a live re-retrieval, updated taxonomic reconciliation, homology-based marker validation, contamination assessment and specimen authentication were not performed.

FIXED-REGION PARENT-LENGTH SENSITIVITY

Table S7. Post-review parent-length reclassification of the same QC-passing regions.

Parent cutoff (bp)	Short-parent combinations	Long-parent combinations	Union	Long-parent-only	Long-parent-only (%)
2,500	39,213	4,310	41,875	2,662	6.357
5,000	39,298	4,208	41,875	2,577	6.154
10,000	39,382	4,102	41,875	2,493	5.953

The frozen pool contains 98,534 extracted regions, including 96,963 that pass primary QC. Each row partitions those same passing regions by parent length; no region is added, removed, shortened or re-extracted. Short-parent means parent length at or below the cutoff; long-parent means parent length above it. Counts are unique NCBI taxon ID–marker combinations, not species or specimens. The long-parent-only percentage is $100 \times \text{long-parent-only} / \text{union}$.

At 2,500 bp, 164 passing whole-record regions are retained on the long-parent side. This is a classification consequence, not evidence of embedded feature origin. At 5,000 bp the original source-architecture sets are recovered exactly. The long-parent-only fraction is therefore distinguished from the hidden fraction after conditional re-extraction in Table S5.

The candidate frame, extracted sequences, marker labels, taxon IDs and QC rules remain unchanged. The original 150–5,000-bp extraction window and 180-ACGT, zero-gap, valid-symbol, at-most-1%-ambiguity QC rules are retained. These are exact frozen-frame counts; no sampling tests or confidence intervals are applied.

Supplementary File S3 includes `architecture_only_sensitivity.py` and `verification/architecture_only_sensitivity.csv`, `architecture_only_by_marker.csv` and `architecture_only_checks.json`. The script reconstructs the original 5,000-bp source labels, checks that the union is invariant, and verifies the set counts with a separate SQL aggregation. It requires only the supplied frozen QC matrix and the Python standard library. These checks establish numerical consistency, not biological marker or taxonomic validity.