

Public sequence archives reveal overlooked orchid barcode resources for biodiversity monitoring

José Kalión Guerra Lu^{1*} , Alan Guillermo Gallo Álvarez² ,
Leiwer Flores³ , Manuel Emilio Reategui Inga⁴ 

¹ Facultad de Recursos Naturales Renovables, Universidad Nacional Agraria de la Selva, Carretera Central km 1.21, Tingo María 10131, Huánuco, Peru

² Facultad de Ingeniería y Ciencias, Universidad Nacional Autónoma de Alto Amazonas, Prolongación Libertad 1220–1228, Yurimaguas, Alto Amazonas, Loreto, Peru

³ Facultad de Ciencias Agrarias, Universidad Nacional de Cajamarca – Filial Jaén, Jr. Bolívar 1342, Jaén, Cajamarca, Peru

⁴ Escuela Profesional de Ingeniería Ambiental, Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa, Calle 7 de Junio s/n, I.E. Santiago Antúnez de Mayolo, Pichanaqui, Chanchamayo, Junín, Peru

* Corresponding author's e-mail: jose.guerra@unas.edu.pe

ABSTRACT

Public nucleotide archives contain annotated marker loci within long sequence records, but inventories based only on short records can overlook them. We quantified this difference for Orchidaceae using a frozen European Nucleotide Archive catalogue and a predeclared, annotation-based extraction protocol. The catalogue contained 578,690 records; 555,683 candidate accession versions were retrieved. Parsing yielded 98,534 marker regions from 71,358 parent records, of which 96,963 (98.406%) passed syntactic quality control. The primary counting unit was the NCBI taxon ID–marker combination, not an accepted species or specimen. Short-record extraction represented 39,298 combinations and long-record features represented 4,208, yielding a union of 41,875. Of these, 2,577 were present only as embedded features: 6.154% of the union and a 6.558% increase over the short-record inventory. For the core plant markers *matK* and *rbcL*, 613 of 11,082 combinations (5.531%) were embedded-only. The hidden fraction ranged from 6.088% to 9.601% across 18 predeclared quality-control scenarios. Post-review parent-length reclassification with a fixed region pool yielded 5.953–6.357%, whereas jointly changing the boundary and maximum region length yielded 4.414% at 2,500 bp and 6.681% at 10,000 bp. An internal audit using independently implemented code verified all 98,534 extracted regions against the preserved EMBL records without discrepancies. The complete candidate-record snapshot and region-level verification evidence have been preserved. These findings identify a methodological source of undercounting in archive inventories. They do not establish species-level barcode coverage, biological correctness of deposited annotations, or improved identification performance.

Keywords: biodiversity monitoring, molecular reference library, Orchidaceae, plant identification, public sequence archive, sequence annotation.

INTRODUCTION

Molecular surveys increasingly complement conventional biodiversity monitoring by detecting organisms from specimens, mixed samples, or environmental DNA. Interpreting these sequences requires reference libraries that link target regions

to deposited taxonomic labels. When a relevant reference is absent from the searchable library, an otherwise valid environmental sequence may remain unassigned or may only be resolved at a broad taxonomic level. The completeness of these libraries thus affects what biodiversity surveys can identify (Deiner et al., 2017; Weigand et al., 2019).

Public sequence resources are maintained in complementary systems. The International Nucleotide Sequence Database Collaboration preserves nucleotide records and their submitted annotations, while specialized platforms such as the Barcode of Life Data System organize barcode records with specimen and taxonomic information (Arita et al., 2021; Ratnasingham et al., 2024). What users can find depends on both the deposited sequences and the structure of their records. A marker may be deposited as a short independent record, or it may occur as an annotated feature inside a complete plastome or another long record.

This distinction matters for plants. The plastid loci *matK* and *rbcL* were proposed as a core two-locus barcode, while other plastid and nuclear regions can provide complementary taxonomic information (CBOL Plant Working Group, 2009). Genome skimming and organellar sequencing can recover barcode regions without generating separate short accessions for every locus (Nevill et al., 2020). Complete plastomes have also been used to extract and compare candidate markers such as *ycf1* (Dong et al., 2015). A census restricted to short, marker-specific accessions can therefore miss loci that are already present in longer records.

Several approaches are available for building and auditing reference libraries. Automated tools can create marker-specific databases from public archives and apply taxonomic, primer, and length filters (Keller et al., 2020). Audit systems can grade sequence libraries according to taxonomic congruence, metadata, and sequence properties (Fontes et al., 2021). Curated resources have been produced for plant internal transcribed spacer sequences and other molecular targets (Banchi et al., 2020; Kolter and Hebert, 2025). Recent studies have also quantified global plant-barcode availability and developed compact reference-library representations for downstream analysis (Kartzinel et al., 2025; Prosser et al., 2025).

One question remains: how many deposited taxon–marker combinations are missed when an inventory counts only short independent records and ignores explicitly annotated loci in longer sequences? The answer depends on whether the inventory counts parent records or annotated marker regions. This question concerns the representation of deposited information, rather than accepted-species coverage, taxonomic validation, primer performance or discrimination accuracy.

Orchidaceae is a useful group for examining this question. Barcode libraries have been

assembled for applied orchid collections, and *matK* and *rbcL* have been evaluated using curated regional material (Raskoti and Ale, 2021; Worthly et al., 2022). Documented reliability problems in deposited orchid sequences also show that availability alone does not guarantee taxonomic or metadata quality (Wu et al., 2021). Regional plant libraries further illustrate how traceable barcode resources can support floristic surveys and monitoring (Jin et al., 2023; Song et al., 2023).

We measured how much extracting explicit marker features from long European Nucleotide Archive records added to Orchidaceae taxon–marker representation. We compared those features with short standalone records from the same frozen archival frame. We estimated the overall embedded-only fraction, evaluated the core plant barcodes *matK* and *rbcL* separately, described marker-level heterogeneity, and tested whether the result persisted under alternative sequence and taxonomic-label filters. We hypothesized that extracting annotated features would reveal a measurable set of taxon–marker combinations missed by standalone-record counting.

MATERIALS AND METHODS

Study design and frozen protocol

This retrospective archive analysis used a locally frozen protocol and machine-readable configuration (version 1.0.0; Supplementary File S2). These files define the archive frame, marker vocabulary, record architecture, quality rules and descriptive effect-size criteria. They document the operational rules but are not an independently time-stamped public preregistration. We make no claim of prospective registration.

The primary molecular unit was an accession.version–marker region. Representation was measured using unique National Center for Biotechnology Information (NCBI) taxon identifier–marker combinations. Sequences, accessions and deposited taxon labels were not treated as accepted species. The same taxon identifier could contribute once to each marker but not more than once within a marker and record-architecture set. This rule avoided counting the same marker repeatedly when it occurred in multiple accessions with the same deposited identifier.

The local configuration defines descriptive decision criteria of at least 100 taxid–marker

combinations in the union, at least 25 embedded-only combinations and a primary hidden fraction of at least 5%; the robustness criterion requires at least 2.5% in every configured sensitivity scenario. These are protocol-specific decision rules, not statistical significance tests or general thresholds of biological importance.

Archive query and sequence retrieval

On 3 September 2026, the European Nucleotide Archive sequence domain was queried with `tax_tree(4747)`, where 4747 is the NCBI taxonomic identifier used for Orchidaceae. The resulting catalog contained 578,690 records. Candidate selection retained records through one of three predeclared routes: a parent-record length between 150 and 5,000 base pairs, a marker-bearing description, or a long-record description indicating chloroplast, plastid, plastome, mitochondrial-genome, or mitogenome context. The selection yielded 555,683 unique candidate accessions.

Records were retrieved in EMBL format through official ENA and EMBL-EBI sequence-retrieval services, with request endpoints retained in the batch provenance. Retrieval used bounded logical batches, gzip transport, a 180-second timeout, and up to six attempts. Every identifier was resolved and frozen as `accession.version`. We checked retrieval completeness against the candidate list and recorded SHA-256 checksums for the catalog, source responses, configurations and derived outputs. The European Nucleotide Archive and the wider international collaboration provide the provenance and version information needed to reconstruct these records (Harrison et al., 2021; Arita et al., 2021).

Marker detection and record-architecture classification

The controlled marker vocabulary comprised *matK*, *matK-like*, *rbcL*, *rbcLa*, *rbcL-like*, ITS, ITS1, ITS2, *trnH-psbA*, *trnL-F*, *atpF-atpH*, *psbK-psbI*, *rpoB*, *rpoC1*, *ycf1*, and COI-5P. Similar labels were retained as separate categories to avoid inflating coverage through semantic merging. For example, *rbcLa* was not merged into *rbcL*, and *matK-like* was not treated as *matK*.

ITS was treated as a broad locus only when ITS1 or ITS2 did not provide a more specific assignment. Labels such as *matK-like* and *rbcL-like* were retained as separate operational categories,

and *rbcLa* was not merged with *rbcL*. The frozen category named COI-5P accepts generic COI/*cox1* annotations; this name does not establish a verified 5' barcode interval. We retain the category name only to preserve correspondence with the frozen files and report a separate analysis excluding it entirely. These rules classify archive annotations, not independently confirmed homologous loci.

A uniquely resolved marker label in the description of a 150–5,000 bp parent record allowed inclusion of the whole record as a stand-alone candidate. Eligible annotated features were also extracted separately. Features from parents of 150–5,000 bp were classified as standalone, whereas eligible 150–5,000 bp features from parents longer than 5,000 bp were classified as embedded. Feature extraction required exact annotation locations and used Biopython sequence extraction, including strand and compound-location handling. Ambiguous record-level labels did not prevent extraction of independently unambiguous eligible features. Multiple representations of a marker within an accession were ultimately collapsed at the taxid-marker level. Figure 1 summarizes the workflow.

The region metadata record `accession.version`, marker, source architecture, outer 1-based coordinate bounds and strand. Sequence extraction uses compound locations where present. The two bounding-coordinate columns do not retain their full joined topology, but the FASTA file preserves the exact reconstructed sequences.

Sequence quality and metadata readiness

Primary syntactic quality control required at least 180 ACGT bases, no gaps, an ambiguous-base fraction no greater than 1% and no symbols outside the configured IUPAC alphabet. The `accession.version` field and sequence availability were checked independently across the extracted library; neither check failed. These rules assess sequence syntax, not contamination, reading-frame integrity, taxonomic accuracy or identification performance.

Metadata readiness was evaluated as a secondary descriptive profile. A region was considered ready when it had a nonempty deposited taxonomic label, passed syntactic sequence quality, contained at least one traceability field among specimen voucher, bio-material, or culture collection, and contained at least one contextual field among country or collection date. We assessed readiness separately from the primary

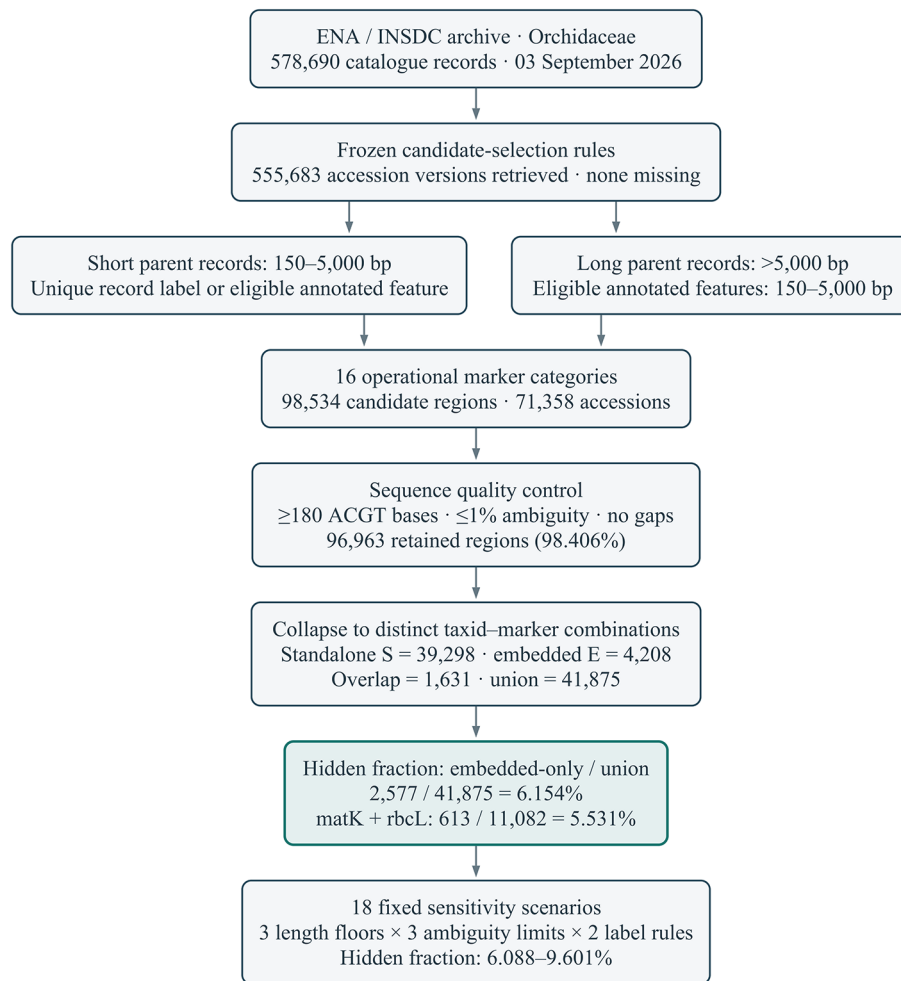


Figure 1. Archive selection, extraction and set-based analysis. The standalone and embedded sets comprise distinct archive taxid–marker combinations, not species. The sensitivity grid keeps the 180-ACGT quality floor and the 5,000 bp parent-record threshold fixed. Full rules and operational marker definitions are provided in Supplementary File S1

representation estimate to distinguish the presence of a sequence in the archive from the completeness of its specimen documentation.

Representation estimands

For marker m , let S_m be the set of archive taxonomy identifiers represented by at least one quality-passing standalone region and E_m the corresponding set represented by an embedded region. An identifier present in both sets is counted once in their union. The same identifier associated with two different markers constitutes two distinct taxid–marker combinations. The hidden fraction H and relative increase R were defined as:

$$H = [\sum_m |E_m \setminus S_m| / \sum_m |S_m \cup E_m|] \times 100 \quad (1)$$

$$R = [\sum_m |E_m \setminus S_m| / \sum_m |S_m|] \times 100 \quad (2)$$

where: $E_m \setminus S_m$ denotes identifiers represented only by embedded regions for marker m . H uses the union as its denominator; R instead uses the standalone set.

Summation is across the 16 marker categories, or across *matK* and *rbcL* for the core-marker analysis. These estimates describe the observed archive frame under the stated rules. They do not measure the percentage of orchid species lacking barcodes. No sampling-based confidence intervals or hypothesis-test P values were calculated for this archive census.

Sensitivity analyses

The configured grid crosses minimum extracted-region lengths of 150, 300 and 500 bp; maximum ambiguous-base fractions of 0%, 1%

and 5%; and retention or exclusion of labels matching the specified uncertain-name tokens (sp., cf., aff., hybrid, × or a standalone x). All 18 scenarios retain the fixed minimum of 180 ACGT bases, zero gaps and valid symbols. Consequently, the nominal 150 bp floor does not admit sequences with fewer than 180 ACGT bases. Label filtering is a literal token screen, not an accepted-name or species-level taxonomic reconciliation. The 18 predeclared scenarios held the 5,000-bp architecture boundary and maximum accepted region length fixed. Following review, we added conditional analyses at 2,500 and 10,000 bp, changing both parameters together while retaining the original 555,683-record candidate frame and primary quality-control rules. These analyses are not complete archive inventories under new candidate-selection thresholds. We also applied a stricter assignment rule using only gene, product and standard_name feature qualifiers, with no description-based whole-record assignments, at 5,000 bp. Finally, we removed the entire generic COI/*cox1* category and separately retained only *matK* and *rbcL*. These post-review analyses test operational dependence; they do not estimate annotation error rates or changes in accepted taxonomy.

To isolate the effect of the parent-length boundary, we also held the original extracted regions, primary quality-control outcomes, marker assignments and taxon IDs fixed, and partitioned the same 96,963 passing regions at 2,500, 5,000 and 10,000 bp. We counted taxid–marker combinations represented only on the long-parent side of each partition. Unlike conditional re-extraction, this analysis changes no region-eligibility rule. Its labels refer only to parent length: a retained whole-record region can cross to the long-parent side at 2,500 bp without becoming an annotated embedded feature. This post-review diagnostic is reported separately in Supplementary File S1, Table S7.

Computation and reproducibility

The original analysis environment recorded Python 3.13.5, Biopython 1.85, pandas 2.3.2 and requests 2.32.5 on Windows 11. Numbered scripts separate catalogue querying, candidate selection, sequence retrieval, parsing, quality control and set-based analysis. Supplementary File S2 preserves the frozen derived inputs, original code and offline replay. The previously

completed replay under Python 3.12.14 reproduced the decompressed quality-control matrix and the primary, 18-scenario sensitivity and readiness tables exactly.

For this revision, we implemented a separate EMBL parser and extraction verifier using the Python 3.12.14 standard library, without importing the original parser or Biopython. The verifier uses the same frozen scientific definitions, including marker vocabulary, precedence, deduplication and quality-control rules. It checks every gzip batch against the historical checksum and accession membership, independently evaluates exact feature locations including complements and compound intervals, and compares every emitted region with the frozen sequence hash, length, outer coordinates, strand, taxon ID, source architecture and quality-control outcome. It also reconstructs the primary candidate selection from the frozen catalogue. Source-record and region indices make these checks traceable in Supplementary File S3.

This is an internal independent-code verification against preserved archival annotations, distinct from both the earlier analytical replay and external biological validation. It does not independently authenticate the original annotations, taxonomic identifications or specimens, and it does not re-run a historical query against the live service. Assembly-component records without an inline sequence are preserved as deposited; remote components are not fetched or reconstructed. Artificial-intelligence tools were not used to assign taxa or generate nucleotide sequences.

RESULTS

Retrieval and quality control

Figure 1 and Table 1 summarize sequence retrieval and processing. The archive query returned 578,690 records, and 555,683 candidate accession versions met at least one predeclared selection route. All 555,683 parent records were retrieved, with no missing identifiers or version discrepancies. Parsing yielded 98,534 marker regions from 71,358 accession versions. Of these regions, 96,963 (98.406%) passed primary syntactic quality control. The stricter sequence-and-metadata readiness profile was met by 22,099 regions (22.428%). Retrieval completeness refers

Table 1. Frozen-frame processing summary

Processing stage	Result
ENA catalog records	578,690
Candidate accession versions	555,683
Candidate versions retrieved	555,683
Missing versions	0
Version discrepancies	0
Marker regions emitted	98,534
Source accession versions represented	71,358
Regions passing syntactic quality control	96,963 (98.406%)
Regions meeting metadata-readiness profile	22,099 (22.428%)

Note: ENA, European Nucleotide Archive.

to archived records, not to the availability of in-line nucleotide sequence in every parent record.

Overall hidden representation

Standalone records represented 39,298 taxid–marker combinations, whereas embedded loci represented 4,208. Their union comprised 41,875 combinations, including 2,577 found only among embedded regions (Table 2). The overall hidden fraction was 6.154%, and annotation-aware inclusion increased observable representation by 6.558% relative to the standalone count. The primary result exceeded all configured descriptive decision criteria.

Core plant barcodes

The analysis restricted to the core plant barcodes *matK* and *rbcL* also identified combinations found only in embedded loci. For *matK*, 230 of 7,473 observed taxon–marker combinations were embedded-only, corresponding to a hidden fraction of 3.078%. For *rbcL*, 383 of 3,609 combinations were embedded-only, corresponding

to 10.612%. Jointly, the two markers contained 613 embedded-only combinations within a union of 11,082, producing a hidden fraction of 5.531% and a 5.855% increase relative to the 10,469 standalone combinations.

Marker-level heterogeneity

The size of the effect varied considerably among markers (Figure 2). *rpoB* contributed 801 embedded-only combinations and *rpoC1* contributed 715; together they accounted for 1.516 of the overall 2.577 embedded-only combinations (58.828%). Adding *rbcL*, *ycf1*, and *matK* increased the combined contribution of these five plastid loci to 2.493 combinations (96.740%). Marker-specific hidden fractions were 61.568% for *rpoB*, 50.388% for *rpoC1*, 13.005% for *ycf1*, 10.612% for *rbcL*, and 3.078% for *matK*. In contrast, several markers had no embedded-only representation under the specified annotation rules. These counts reflect how sequences were deposited. They do not rank the markers by discrimination ability or laboratory performance.

Predeclared and post-review sensitivity analyses

The overall result persisted across all 18 configured sensitivity scenarios (Figure 3). The hidden fraction ranged from 6.088% to 9.601%, and the embedded-only count ranged from 2.356 to 2.637 combinations. Every scenario remained above the 2.5% robustness threshold. Increasing the minimum accepted region length generally reduced both standalone and embedded counts but increased the hidden fraction because the standalone denominator contracted more strongly. Excluding labels matching the specified uncertain-name tokens did not reverse the direction of the effect. The exhaustive source-record audit verified

Table 2. Primary and core-barcode representation estimands

Scope	Standalone	Embedded	Hidden-only	Union	Hidden (%)	Increase (%)
All 16 marker categories	39,298	4,208	2,577	41,875	6.154	6.558
<i>matK</i>	7,243	925	230	7,473	3.078	3.175
<i>rbcL</i>	3,226	911	383	3,609	10.612	11.872
<i>matK</i> and <i>rbcL</i> combined	10,469	1,836	613	11,082	5.531	5.855

Note: hidden fraction – embedded-only combinations divided by the observed union. Increase – embedded-only combinations divided by standalone combinations. The combined *matK*–*rbcL* row sums marker-specific combinations and does not collapse the two loci into one unit.

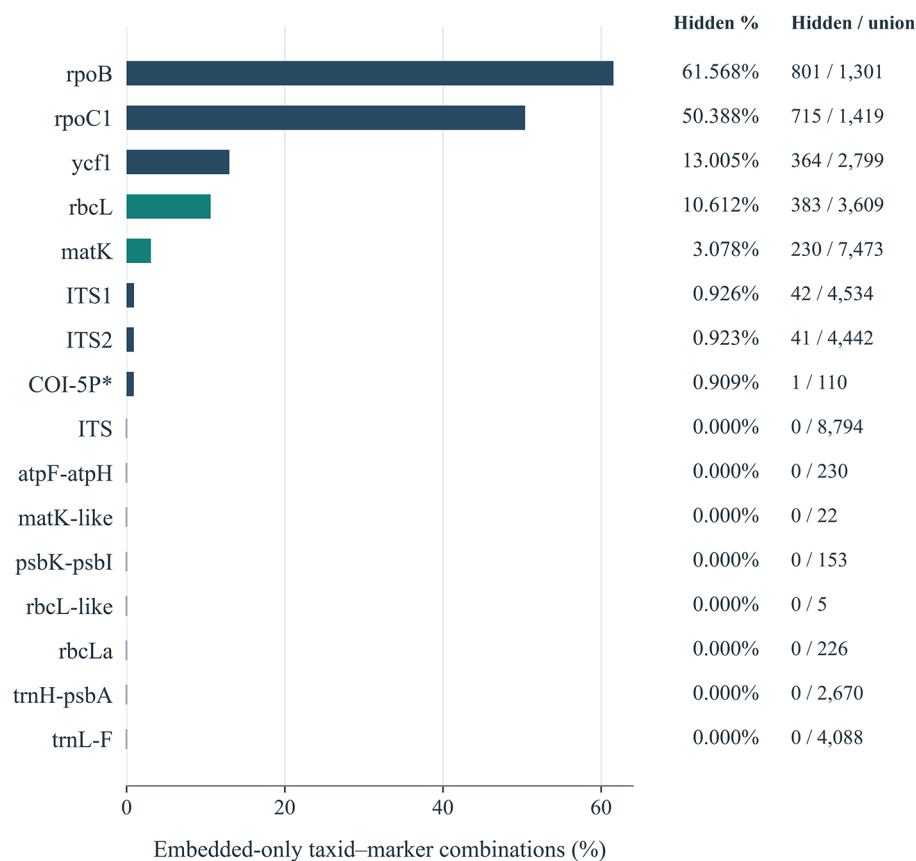


Figure 2. Core markers (teal): *matK* + *rbcL* = 613 / 11,082 (5.531%)* Operational COI-5P category; the 5' position was not verified. Embedded-only representation by marker, ranked by the hidden fraction of each marker-specific union. Values and denominators are shown directly; teal highlights *matK* and *rbcL*. The combined core estimate sums marker-specific taxid units and does not require both markers from the same specimen. COI-5P is the historical computational category for generic COI/*cox1* annotations, not a verified 5' interval. Zero values indicate no embedded-only contribution under the frozen extraction rules, not biological absence. No confidence intervals are shown

all 555,683 candidate records in 3.705 historical batches and all 98,534 extracted regions, with no discrepancies in the tested fields. This included 4,848 reverse-strand and 71 compound-location regions. The primary estimate was reproduced exactly. Of the parent records, 159,369 contained assembly-component descriptions without inline nucleotide sequences. All emitted regions were supported by inline sequence in their parent records. Three non-extractable feature events at KZ504779.1 were excluded in both implementations (Supplementary File S1, Table S6).

Within the unchanged candidate frame, the joint 2.500-bp threshold yielded 1.811 embedded-only combinations among 41,024 (4.414%); at 10.000 bp the corresponding counts were 2.822 of 42,237 (6.681%). The lower-threshold result falls below the original 5% descriptive effect criterion. Applying a new 10.000-bp candidate rule to the

frozen catalogue would select 5,928 additional records absent from the preserved candidate snapshot, so these estimates describe conditional re-extraction rather than complete alternative-threshold inventories. With the region pool and QC fixed, the long-parent-only counts were 2.662, 2.577 and 2.493 at 2.500, 5.000 and 10.000 bp, respectively, each within the same union of 41,875 combinations (6.357, 6.154 and 5.953%; Supplementary File S1, Table S7). At 2.500 bp, 164 passing whole-record regions remained in the fixed pool but moved to the long-parent side. They were not re-extracted or reinterpreted as embedded features. The lower 4.414% result from joint re-extraction therefore cannot be attributed to parent-length classification alone. The structured-feature-only analysis yielded 3.820 embedded-only combinations among 14,100 (27.092%). Its standalone denominator fell sharply because description-based records were excluded;

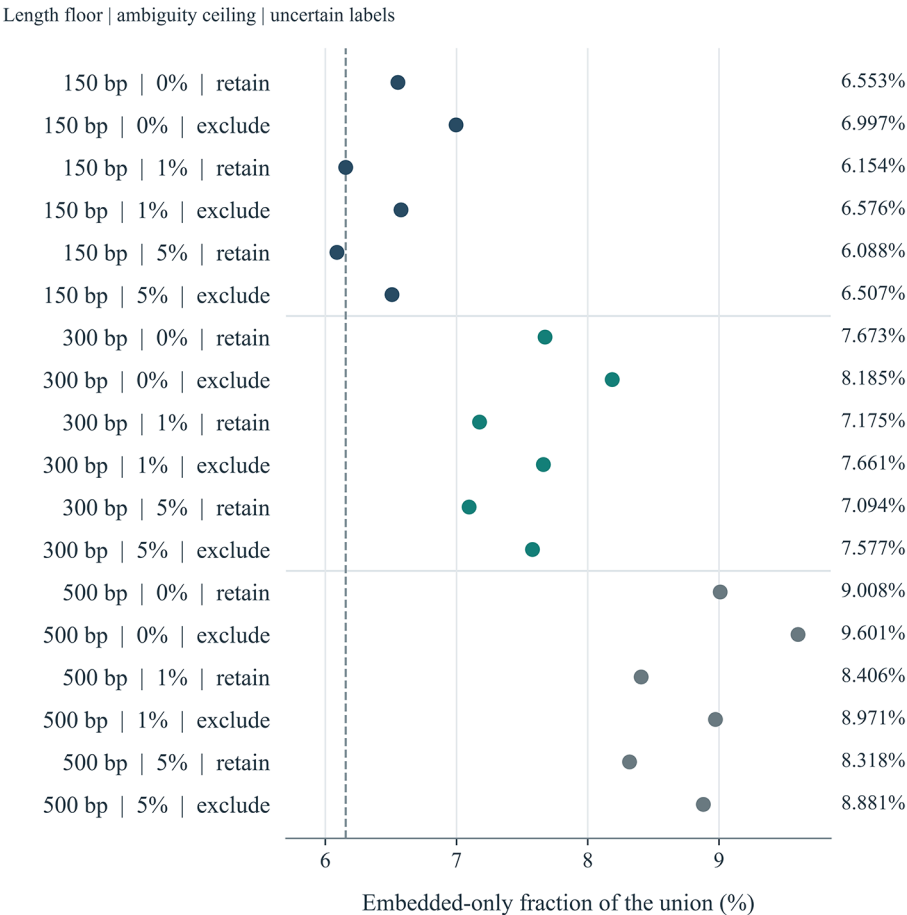


Figure 3. Dashed line: primary estimate, 6.154%. Points are scenario estimates, not confidence intervals. All scenarios retain 2:180 ACGT bases, no gaps and valid symbols; parent threshold remains 5,000 bp. Hidden fractions under all 18 configured sensitivity scenarios. Labels specify minimum region length, maximum ambiguity and whether uncertain taxon-label tokens are retained or excluded. All scenarios require at least 180 ACGT bases, valid symbols and no gaps. The dashed line marks the primary estimate (6.154%). Each point is an exact scenario estimate for the frozen archive frame, not an uncertainty interval

the percentage is therefore not a direct annotation-accuracy comparison. Excluding the generic COI/*cox1* category yielded 2.576 of 41,765 combinations (6.168%). The *matK/rbcL*-only analysis reproduced 613 of 11,082 (5.531%). Full post-review counts are reported in Supplementary File S1, Table S5.

DISCUSSION

Implications for biodiversity-monitoring infrastructure

The contribution of this study is methodological: it quantifies an inventory difference caused by the searchable unit used in a public sequence archive. It is not a biological discovery or a new set of nucleotide sequences. Counting

short records alone omitted 2.577 taxid–marker combinations that were recoverable from annotated features in longer records. Those loci had already been deposited; the added representation arose from extraction and accounting, not new sampling.

This affects how reference libraries are assembled for molecular biodiversity monitoring. Environmental DNA and metabarcoding workflows can match observed sequences only against target regions included in those libraries. A locus embedded in a plastome cannot support assignment if a database-building procedure never extracts it. Incomplete extraction can therefore leave a reference library incomplete even when a sequence has been deposited. This is a separate problem from a taxon being biologically absent, a specimen not being sequenced or suitable meta-data being unavailable.

The primary hidden fraction was 6.154%, and the direction of the inventory difference persisted across the 18 predeclared quality-control scenarios and the post-review threshold analyses. Its magnitude was not invariant to the extraction definition. At 2.500 bp, the conditional estimate fell to 4.414%, below the predeclared 5% descriptive effect criterion, although it remained above the separate 2.5% robustness reference. The primary 5.000-bp result still meets the original descriptive criteria. These comparisons support a reproducible inventory effect under specified rules, not a universal percentage for Orchidaceae.

The fixed-region analysis separates parent-length classification from region eligibility. Its long-parent-only fraction ranged from 5.953% to 6.357% while the denominator remained unchanged. The wider variation under joint re-extraction also reflects which regions are retained. Neither comparison validates marker identity or taxonomy, and the lowered-cutoff partition must not be interpreted as evidence that whole-record regions are biologically embedded.

Record architecture and marker heterogeneity

The concentration of embedded-only representation in *rpoB*, *rpoC1*, *rbcL*, *ycf1*, and *matK* is consistent with their occurrence as annotated components of plastid records. *rpoB* and *rpoC1* together contributed nearly three fifths of the overall embedded-only count because they had substantial representation in long records but comparatively limited standalone representation. Their frequency in the archive reflects past sequencing strategies, annotation practices and record submission. It does not establish their suitability for identification, which requires independent taxonomic validation and sequence-comparison experiments.

The hidden fraction for *rbcL* was more than three times that of *matK*. An archive inventory based only on short records would consequently understate the deposited *rbcL* resource more strongly in this frame. However, the study did not evaluate amplification success, read recovery, barcode gaps, taxonomic resolution, or error rates. The difference is therefore limited to discoverability and representation; analytical performance remains untested.

Relation to previous library-building approaches

Existing tools assemble and filter barcode libraries from public databases, and auditing systems assess taxonomic congruence, metadata and sequence quality (Keller et al., 2020; Fontes et al., 2021). Plant-focused resources also emphasize marker-aware curation and standardized reference data (Banchi et al., 2020; Kolter and Hebert, 2025). Our analysis complements these approaches by distinguishing the structure of parent records before assessing the quality of the recovered regions.

The choice of counting unit also matters when assessing coverage. National plastome-readiness analyses show that provenance and representation gaps can remain even when whole plastomes are available (Szczecińska et al., 2026). Surveys of plant-barcode availability quantify important taxonomic and geographic gaps, but their conclusions depend on how the searchable molecular unit is defined (Kartzinel et al., 2025). Annotation-aware reconstruction cannot solve missing-specimen or missing-sequence problems, yet it can prevent deposited loci from being misclassified as absent. Feature extraction identifies suitable regions before compact-library methods address redundancy and computational scale (Prosser et al., 2025).

Finding a sequence in an archive does not remove the need to trace it to a specimen. Libraries developed from vouchered collections provide ecological and geographic context that bare sequence records may lack (Jin et al., 2023; Song et al., 2023). In the present archive frame, 98.406% of extracted regions passed syntactic sequence checks, but only 22.428% met the stricter readiness profile combining sequence, traceability, and geographic or temporal fields. Recovering additional annotated loci increases observable sequence representation without resolving gaps in specimen documentation.

Robustness and limitations

The original 18-scenario analysis tests quality-control and uncertain-label filters, not all sources of biological uncertainty. The joint threshold analyses show that changing the architecture boundary and maximum region length together changes the hidden fraction. Because the candidate frame remained frozen, the 10.000-bp estimate does not include the 5.928 additional records selected by a newly applied 10.000-bp candidate

rule. The structured-feature-only analysis changes the ascertainment of standalone records and its denominator; its higher hidden fraction cannot be read as stronger biological evidence. A different parsing implementation also cannot detect annotation mistakes shared by the source records and the frozen vocabulary.

The estimand remains archive representation at the NCBI taxon ID–marker level. Taxon IDs are not a verified list of accepted Orchidaceae species, so 6.154% is not the proportion of orchid species previously lacking barcode data. Synonymy, misidentification, contamination, duplicate specimens and taxonomic correctness were not assessed. Comparing stored taxon IDs and feature coordinates verifies faithful transfer from the source annotations, not their biological validity. COI/*cox1* assignments do not verify a specific 5' interval; excluding that category leaves the aggregate result nearly unchanged. Some parent records contain only remote assembly-component descriptions, and those components were not reconstructed. Live retrieval, archive updates, independent homology assessment and environmental assay performance were not tested.

Future archive audits can distinguish parent records from marker regions using this reproducible approach. Reporting standalone, embedded, overlapping and embedded-only components separately shows how much information was recovered and makes clear that it came from existing deposits.

CONCLUSIONS

Annotation-aware extraction from long public sequence records added 2.577 Orchidaceae NCBI taxon identifier–marker combinations that were not observed in short standalone records. These combinations constituted 6.154% of the observed union and increased representation by 6.558% relative to standalone counting.

For the core plant barcodes *matK* and *rbcL*, 613 of 11,082 combinations were embedded-only, corresponding to a hidden fraction of 5.531%. Across the full marker set, the effect varied among loci and was concentrated in plastid markers.

Across 18 predeclared quality-control scenarios, the hidden fraction ranged from 6.088% to 9.601%. Post-review analyses within the frozen candidate frame yielded 4.414% and 6.681% when the architecture boundary and maximum

region length were jointly set to 2.500 and 10,000 bp. With the region pool fixed, changing parent-length classification alone yielded 5.953–6.357%. The archive inventory difference is therefore measurable and depends on the extraction definition.

Acknowledgements

The authors acknowledge the researchers and institutions that generated, annotated, and submitted the nucleotide records preserved by the International Nucleotide Sequence Database Collaboration. No new biological specimens were collected for this study.

REFERENCES

1. Arita, M., Karsch-Mizrachi, I., Cochrane, G. (2021). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 49(D1), D121–D124. <https://doi.org/10.1093/nar/gkaa967>
2. Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., Pallavicini, A. (2020). PLANiTS: A curated sequence reference dataset for plant ITS DNA metabarcoding. *Database*, 2020, baz155. <https://doi.org/10.1093/database/baz155>
3. CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31), 12794–12797. <https://doi.org/10.1073/pnas.0905845106>
4. Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872–5895. <https://doi.org/10.1111/mec.14350>
5. Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., Cheng, T., Guo, J., Zhou, S. (2015). *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific Reports*, 5(1), 8348. <https://doi.org/10.1038/srep08348>
6. Fontes, J. T., Vieira, P. E., Ekrem, T., Soares, P., & Costa, F. O. (2021). BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. *Molecular Ecology Resources*, 21(2), 573–583. <https://doi.org/10.1111/1755-0998.13262>
7. Harrison, P. W., Ahamed, A., Aslam, R., Alako, B. T. F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Balavenkataraman Kadhivelu, V., Kumar, M., Lopez, R., Kay, S., Leinonen, R., ... Cochrane, G. (2021). The European nucleotide

- archive in 2020. *Nucleic Acids Research*, 49(D1), D82–D85. <https://doi.org/10.1093/nar/gkaa1028>
8. Jin, L., Shi, H. Y., Li, T., Zhao, N., Xu, Y., Xiao, T. W., Song, F., Ma, C. X., Li, Q. M., Lin, L. X., Shao, X. N., Li, B. H., Mi, X. C., Ren, H. B., Qiao, X. J., Lian, J. Y., Du, H., Ge, X. J. (2023). A DNA barcode library for woody plants in tropical and subtropical China. *Scientific Data*, 10(1), 819. <https://doi.org/10.1038/s41597-023-02742-7>
9. Kartzinel, T. R., Hoff, H. K., Divoll, T. J., Littelford-Colquhoun, B. L., Anderson, H., Burak, M. K., Kuzmina, M. L., Musili, P. M., Rogers, H., Troncoso, A. J., Kartzinel, R. Y. (2025). Global availability of plant DNA barcodes as genomic resources to support basic and policy-relevant biodiversity research. *Molecular Ecology*, 34(7), e17712. <https://doi.org/10.1111/mec.17712>
10. Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., Ankenbrand, M. J. (2020). BCdatabaser: On-the-fly reference database creation for (meta-)barcoding. *Bioinformatics*, 36(8), 2630–2631. <https://doi.org/10.1093/bioinformatics/btz960>
11. Kolter, A., Hebert, P. (2025). Automating the curation of DNA barcode databases for vascular plants. *Environmental DNA*, 7(3), e70125. <https://doi.org/10.1002/edn3.70125>
12. Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., van Leeuwen, S., Boykin, L. M., Small, I. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods*, 16(1), 1. <https://doi.org/10.1186/s13007-019-0534-5>
13. Prosser, S. W. J., Floyd, R. M., Thompson, K. A., Monckton, S. K., Hebert, P. D. N. (2025). BOLD-istilled: Automated construction of comprehensive but compact DNA barcode reference libraries. *Molecular Ecology Resources*, 25(8), e70043. <https://doi.org/10.1111/1755-0998.70043>
14. Raskoti, B. B., Ale, R. (2021). DNA barcoding of medicinal orchids in Asia. *Scientific Reports*, 11(1), 23651. <https://doi.org/10.1038/s41598-021-03025-0>
15. Ratnasingham, S., Wei, C., Chan, D., Agda, J., Agda, J., Ballesteros-Mejia, L., Boutou, H. A., El Bastami, Z. M., Ma, E., Manjunath, R., Rea, D., Ho, C., Telfer, A., McKeowan, J., Rahulan, M., Steinke, C., Dorsheimer, J., Milton, M., Hebert, P. D. N. (2024). BOLD v4: A centralized bioinformatics platform for DNA-based biodiversity data. In R. DeSalle (Ed.), *DNA barcoding: Methods and protocols* (Methods in Molecular Biology, Vol. 2744) (pp. 403–441). Humana. https://doi.org/10.1007/978-1-0716-3581-0_26
16. Song, F., Deng, Y. F., Yan, H. F., Lin, Z. L., Delgado, A., Trinidad, H., Gonzales-Arce, P., Riva, S., Cano-Echevarría, A., Ramos, E., Aroni, Y. P., Rivera, S., Arakaki, M., Ge, X. J. (2023). Flora diversity survey and establishment of a plant DNA barcode database of Lomas ecosystems in Peru. *Scientific Data*, 10(1), 294. <https://doi.org/10.1038/s41597-023-02206-y>
17. Szczecińska, M., Garbowska, A., Pupek, N., Czołchór, W., Koczwara, K., Szablińska, J., Sawicki, J. (2026). Plastome super-barcoding readiness of the flora of Poland: Coverage, provenance, and the resolution gap. *Applied Biosciences*, 5(3), 71. <https://doi.org/10.3390/applbiosci5030071>
18. Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, S. A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., ... Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of The Total Environment*, 678, 499–524. <https://doi.org/10.1016/j.scitotenv.2019.04.247>
19. Worthy, S. J., Bucalo, K., Perry, E., Reynolds, A., Cruse-Sanders, J., Pérez, Á. J., Burgess, K. S. (2022). Ability of *rbcL* and *matK* DNA barcodes to discriminate between montane forest orchids. *Plant Systematics and Evolution*, 308(3), 19. <https://doi.org/10.1007/s00606-022-01809-z>
20. Wu, H. Y., Chan, K. T., But, G. W. C., Shaw, P. C. (2021). Assessing the reliability of medicinal *Dendrobium* sequences in GenBank for botanical species identification. *Scientific Reports*, 11(1), 3439. <https://doi.org/10.1038/s41598-021-82385-z>