frontiers in Bioinformatics

TYPE Technology and Code
PUBLISHED 01 June 2026
DOI 10.3389/fbinf.2026.1821531

OPEN ACCESS

EDITED BY
Jiming Liu,
Nanyang Technological University, Singapore

REVIEWED BY
John E. Carlson,
The Pennsylvania State University (PSU), United States
Felix Langschied,
Goethe University Frankfurt, Germany

*CORRESPONDENCE
Unai López de Heredia,
unai.lopezdeheredia@upm.es

RECEIVED 02 March 2026
REVISED 16 April 2026
ACCEPTED 27 April 2026
PUBLISHED 01 June 2026

CITATION
Mora-Márquez F, Hurtado M and López de Heredia U (2026) QUERCUSTOA: integrating functional annotations and comparative genomics across oak lineages.
Front. Bioinform. 6:1821531.
doi: 10.3389/fbinf.2026.1821531

QUERCUSTOA: integrating functional annotations and comparative genomics across oak lineages

Fernando Mora-Márquez[1], Mikel Hurtado[2,3] and Unai López de Heredia[1]*

[1]GI en Desarrollo de Especies y Comunidades Leñosas (WooSP), Dpto. Sistemas y Recursos Naturales, ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Madrid, Spain, [2]Dep. of Immunology and Cell Biology, Université de Sherbrooke, Sherbrooke, QC, Canada, [3]Dep. Biología Vegetal y Ecología. Facultad de Ciencia y Tecnología. Euskal Herriko Unibertsitatea (EHU), Leioa, Spain

Oaks (*Quercus* L.) are key components of Northern Hemisphere Forest ecosystems, yet the integration of their rapidly growing genomic resources remains challenging. Here, we present QUERCUSTOA, a genomic and functional resource that integrates nine *Quercus* genome assemblies. By combining automated functional annotation (INTERPROSCAN, EGGNOG-MAPPER) with comparative genomics via genomic lift-over, we have developed a relational database designed to link protein-centric annotations with positional genomic data. Our results demonstrate that this integration supports cross-species ortholog identification and synteny analysis. To facilitate data access and exploration, we provide the QUERCUSTOA-APP, a user-friendly interface that streamlines database management and specific bioinformatic tasks. These include functional annotation, sequence-based homology searches, and multiple sequence alignments. Furthermore, the application enables the construction of phylogenetic trees for individual orthologous genes and proteins, allowing for the study of specific sequence evolution across the included assemblies. QUERCUSTOA provides a standardized and scalable framework for evolutionary and functional studies in *Quercus*, offering a consistent approach to maintain genomic coordinate synchronization across the genus.

KEYWORDS
comparative genomics, forest genetics, functional annotation, genomic lift-over, orthology, *Quercus*, SQLite database

1 Introduction

Oaks (*Quercus* L., Fagaceae) are a woody angiosperm clade including about 400–500 species. They are among the most widespread and diverse tree genera in Northern Hemisphere Forest ecosystems, including temperate deciduous forests, subtropical and tropical savannas, cloud forests, and tropical montane or Mediterranean forests (Nixon, 2006). The economic significance of oaks stems from fruit exploitation for food and extensive livestock farming, as well as from the extraction of wood and other sustainable products like cork, for various uses. The infrageneric classification of the genus is a subject of controversy (Denk et al., 2017), partly due to the extraordinary hybridization ability of these allogamous species (Whittemore and Schaal, 1991). For

all these reasons, oaks have garnered significant attention from botanists, evolutionary biologists, forests breeders, and biotechnologists with a special focus on molecular data in recent decades. Indeed, genome and transcriptome sequencing is shedding light on oak evolution and diversification (Larson et al., 2025), historical and contemporary hybridization (López de Heredia et al., 2017), and their response to stress (Escandón et al., 2021), among other aspects. Furthermore, from a more applied perspective, it has driven the development of molecular markers associated with traits of interest (Nagamitsu et al., 2024).

Genomic resources for oak species are continuously expanding due to the ongoing production of various research groups. Since the seminal *Quercus lobata* genome assembly draft became publicly available (Sork et al., 2016), numerous oak genome assemblies of diverse quality have been released (Zhang et al., 2025). While some of these genome assemblies are drafts, presenting only data at the scaffold level -e.g., *Q. suber* (Ramos et al., 2018)-, others have reached chromosome level and are classified as high-quality references. This progress is possible by the continuous development of sequencing technologies, such as HI-C conformation capture (Belton et al., 2012) and the combination of short and long read sequencing (Wang et al., 2023). Additionally, the optimization of assemblies, gene discovery, and annotation has been enhanced by advances in novel bioinformatic algorithms (Li and Durbin, 2024).

This impressive explosion of whole-genome sequencing studies in oaks has yielded a vast number of genomic resources. These are essential for advancing our understanding of oak genetics and epigenetics, not only to unravel the evolution of the genus but also to enable more applied research on specific genes or gene families. However, public access to sequence data and associated information is often limited. This is due to the scattered storage of genomic data across separate repositories. While some data is available through major public archives like the National Center for Biotechnology Information (NCBI) (Sayers et al., 2026), the European Nucleotide Archive (ENA) (Burgin et al., 2023), and the National Genomics Data Center of the China National Center for Bioinformation (CNCB-NGDC) (CNCB–NGDC Members and Partners, 2026), finding comprehensive resources can be challenging. Furthermore, specialized, single-species databases, such as CorkOakDB for *Q. suber* (Arias-Baldrich et al., 2020), and broader tree genomics resources like TreeGenes (Falk et al., 2018), which often contain genetic maps and comparative genomics data, also host valuable oak genomic information, along with various institutional or project-specific repositories.

Unfortunately, this fragmentation of genomic data across public repositories often leads to critical issues for the research community. Firstly, there is variable assembly quality, meaning the integrity and accuracy of available genome assemblies can differ significantly across platforms (Amarasinghe et al., 2020). Secondly, information redundancy is a frequent problem, where identical or highly similar data might be present in multiple, disparate locations, complicating data retrieval and increasing the risk of inconsistencies (Chen et al., 2017). Furthermore, outdated or incorrect annotations pose a significant challenge, as gene annotations might be incomplete, erroneous, or not regularly updated, hindering downstream analyses (Vuruputoor et al., 2023). For example, Larson et al. (2025) suggested that previously reported expansions and contractions of disease-resistance gene families in oaks might be artifacts of technical differences in repeat annotation and gene identification rather than true biological variation. Lastly, access is hampered by limited centralized resources; for instance, the Quercus Portal (Quercus Portal, 2024), while intended as a central hub, may suffer from outdated links or incomplete coverage of the most recent genomic resources. The absence of a standardized bioinformatic framework that synchronizes physical genomic coordinates with functional metadata remains a major technological barrier, as it prevents researchers from performing seamless synteny studies or large-scale evolutionary analyses of gene families. Moreover, only a subset of available assemblies provides publicly accessible gene annotation and protein files. This significantly restricts the utility of the associated genomic and functional information, thereby impeding collaborative research and comprehensive analyses within the broader scientific community.

In this manuscript, we introduce quercusTOA, an integrative framework designed to centralize and enhance access to the dispersed genomic resources of the genus *Quercus*. quercusTOA bridges existing data repositories by synchronizing protein-centric functional metadata with positional genomic data across nine genome assemblies from NCBI and CNCB-NGDC. Building upon the architectural principles of TOA-Taxonomy Oriented Annotation (Mora-Márquez et al., 2021a) and gymnoTOA (Mora-Márquez et al., 2025), this resource incorporates a taxonomy-aware approach that enables the classification of gene clusters based on their evolutionary conservation. Specifically, the database allows researchers to distinguish between 'core orthologs' shared across the genus and lineage-specific genes unique to particular infrageneric sections. Accompanied by the quercusTOA-app, a user-friendly interface for local data exploration, this system ensures scalability to accommodate the rapidly expanding volume of oak genomic resources.

2 Methods

2.1 Data acquisition

quercusTOA was constructed by integrating data from the increasing number of *Quercus* genome assemblies available through the NCBI Genomes database and the CNCB-NGDC Plant Genome Warehouse repository. In the first stage of the pipeline, *Quercus* genome and protein FASTA sequences were fetched for species with complete, publicly available genome assembly records (Table 1). The most recent genome assemblies for *Q. suber* (Usié et al., 2023), *Q. robur* (Plomion et al., 2016), *Q. lobata* (Sork et al., 2016), and *Q. rubra* (Kapoor et al., 2023) were downloaded from the NCBI Genomes database. The retrieved data included the genome assemblies in FASTA format, their corresponding annotation GFF files, and the protein FASTA sequences. Similarly, equivalent data were downloaded from the CNCB-NGDC Plant Genome Warehouse for the assemblies of *Q. variabilis* (Liang et al., 2025), *Q. dentata* (Wang et al., 2023), *Q. gilva* (Zhou et al., 2022), *Q. longispica* (Ren et al., 2025) and *Q. acutissima* (Liu et al., 2022). The genome and protein sequences were then loaded into the

TABLE 1 Species and genome assemblies integrated in QuercusTOA.

Species	Assembly version	Source	Data citation
Quercus acutissima	GWHBGBO00000000	CNCB-NGDC	CNCB–NGDC (2025a)
Quercus dentata	GWHBRAD00000000	CNCB-NGDC	CNCB–NGDC (2025b)
Quercus gilva	GWHDOCW00000000	CNCB-NGDC	CNCB–NGDC (2025c)
Quercus longispica	GWHESEV00000000	CNCB-NGDC	CNCB–NGDC (2025d)
Quercus variabilis	GWHEQCV00000000	CNCB-NGDC	CNCB–NGDC (2025e)
Quercus lobata	GCF_001633185.2_ValleyOak3.2	NCBI	NCBI Genome (2019)
Quercus robur	GCF_932294415.1_dhQueRobu3.1	NCBI	NCBI Genome (2022)
Quercus rubra	GCA_035136125.1_Qrubra_687_v2.0	NCBI	NCBI Genome (2024a)
Quercus suber	GCF_002906115.3_Cork_oak_2.0	NCBI	NCBI Genome (2024b)

List of the nine *Quercus* species included, specifying assembly version, accession numbers (NCBI/CNCB-NGDC), and specific dataset citation. Most genome assemblies are at the chromosome level, except for the *Quercus suber* assembly, which is at the scaffold level.

`sequences.db` SQLite database file (Figure 1) using the `load-species-seqs.py` utility from the NGSHelper suite (Mora-Márquez et al., 2021b).

2.2 Genomic mapping and redundancy filtering

To take advantage of the information retrieved from genome repositories, the FASTA genome files and their corresponding GFFs included in the sequence dataset were then analyzed using Liftoff (Shumate and Salzberg, 2021) to set up Coding Sequences (CDS) and gene correspondence among the species. We relied on original methods of gene inference employed by the authors of the genome assemblies rather than using a single protocol to perform unified gene inference. However, the GFF file was previously processed by `remove-gff-redundancies.py` (NGSHelper) to obtain a GFF without redundancies (mRNAs with the same concatenated CDS sequence and their corresponding features within a gene). This redundancy filtering was particularly critical for assemblies with high levels of alternative transcript predictions, such as *Q. robur* and *Q. lobata*, ensuring that the subsequent mapping and ortholog clustering were based on a single, representative protein per locus when identical sequences were present. Subsequently, we employed Liftoff to map Gene and Coding Sequence (CDS) coordinates across genomes. These correspondences were loaded into the `comparative-genomics.db` SQLite database using the `get-liftoff-gff-data.py` and `load-liftoff-gff-data.py` scripts (NGSHelper). Separately, the `extract-gff-feature-seqs.py` script (NGSHelper) was used to extract gene, CDS, and concatenated CDS, which were subsequently clustered using MMseqs2. The resulting cluster sequences were loaded into the database by the `split-mmseqs2-concds-clusters.py` and `load-mmseqs2-concds-clusters.py` scripts (NGSHelper). Finally, genomic relationships and homologous proteins were retrieved using the `get-liftoff-genome-relationships.py` and `get-liftoff-homologous-proteins.py` scripts (NGSHelper).

2.3 Functional annotation and multi-source integration

The protein sequences were concatenated and clustered using MMseqs2 (Steinegger and Söding, 2017), Mafft (Katoh and Standley, 2013) and Emboss (Rice et al., 2000). This process yielded a 'clustered consensus proteins' file that was then used to obtain functional annotation features and to build Blast+ (Camacho et al., 2009) and Diamond (Buchfink et al., 2015) databases using `makeblastdb` and `diamond makedb` commands. Blast+ databases were also constructed to store homologs to the clustered sequences in *Arabidopsis* (TAIR10) (Berardini et al., 2015). These homologs were loaded into the `functional-annotation.db` SQLite database using the `load-tair10-orthologs.py` script (NGSHelper). In addition, known plant long non-coding RNA (lncRNA) sequences for *Q. lobata*, *Q. suber*, and *A. thaliana* were sourced from CANTATA (Szcześniak et al., 2016) and processed to construct the 'lncRNA' Blast+ database. Finally, the database was populated with a comprehensive set of functional annotations and orthologous groups. These annotations, derived from InterProScan (Jones et al., 2014) and eggNOG-mapper (Cantalapiedra et al., 2021) searches, included: InterPro (Hunter et al., 2009), PANTHER (Thomas et al., 2003), and eggNOG (Hernández-Plaza et al., 2023) Gene Ontology (GO) terms; Enzyme Commission (E.C.) numbers; KEGG KOs, pathways, modules, classes, and reactions (Kanehisa et al., 2016); CAZy (Drula et al., 2022) and Pfam (Mistry et al., 2021) records; and MetaCyc metabolic pathway information (Caspi et al., 2014). Additionally, orthologous groups of proteins and genes corresponding to each consensus protein sequence were loaded. The overall quality of the consensus sequences was performed using BUSCO with the *embryophyta_odb10* gene set (Simão et al., 2015). Functional annotation features were loaded into the `functional-annotation.db` SQLite database using the `load-interproscan-annotations.py` and `load-emapper-annotations.py` scripts (NGSHelper).

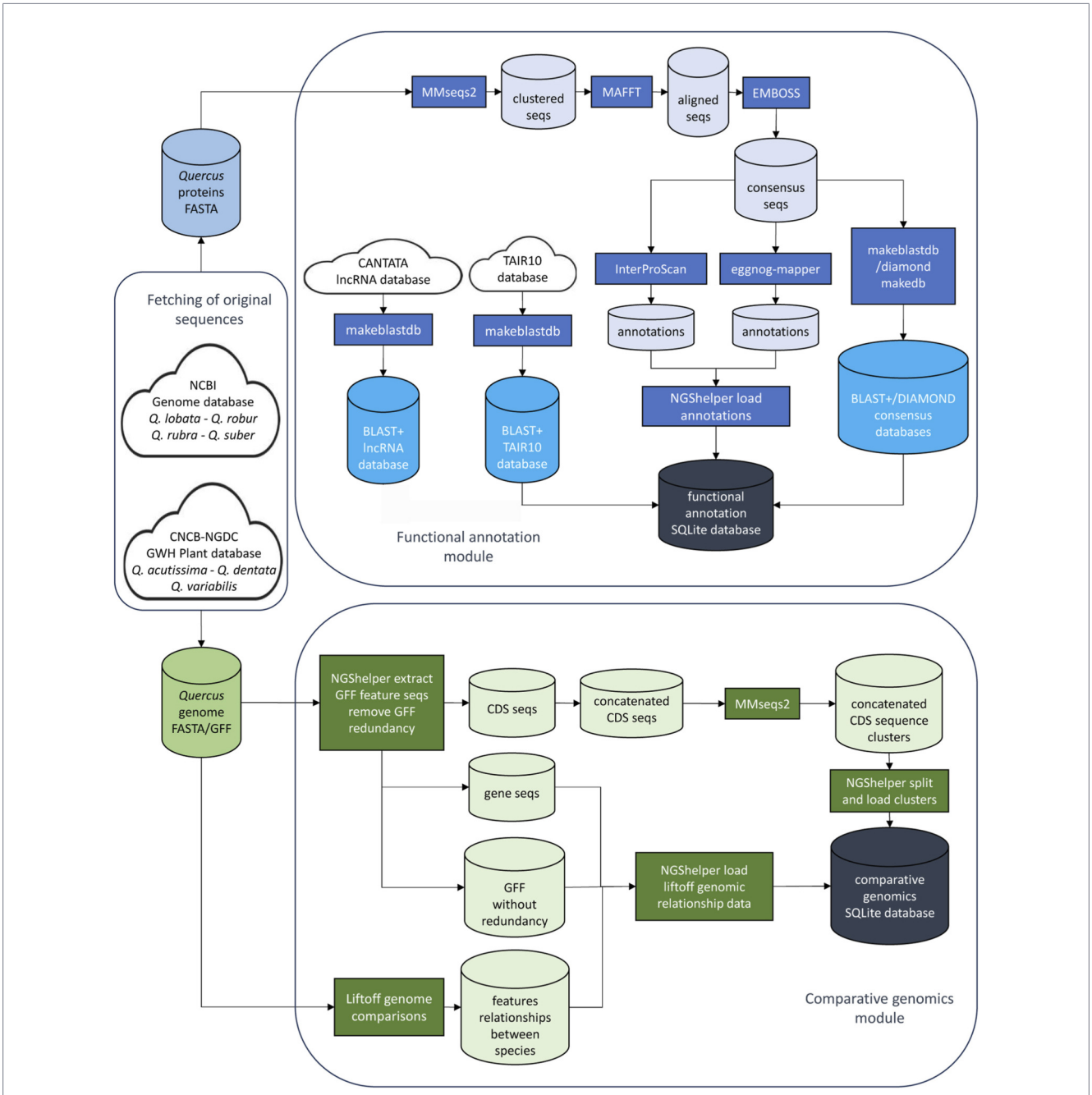

FIGURE 1
QUERCUSTOA construction workflow. Schematic representation of the database pipeline. The functional annotation module includes processing of *Quercus* protein sequences through clustering (MMSEQS2), consensus generation (EMBOSS), and functional assignment using INTERPROSCAN and EGGNOG-MAPPER against specialized databases. The comparative genomics module includes workflow for genome-wide analysis, including GFF normalization (NGSHELPER), gene mapping (LIFTOFF), and ortholog clustering (MMSEQS2). Processed data are integrated into modular SQLite databases for downstream analysis.

2.4 Computational infrastructure

To ensure reproducibility, the entire construction process was automated through a BASH script executed on an *AWS r5.8xlarge* instance. This script, along with the QUERCUSTOA-APP -a graphical interface developed to facilitate the exploration and analysis of the resulting resources (see Section 3.2)- are publicly available at the GITHUB (https://github.com/GGFHF/quercusTOA-app) and ZENODO repositories (Mora-Márquez et al., 2026a). All SQLite databases were built and loaded using Python scripts from NGSHELPER.

2.5 Technical validation

2.5.1 Sequence integrity and comparative genomics module

The QUERCUSTOA approach enables a comprehensive comparison of the genomic space across the selected reference *Quercus* assemblies. To validate the architectural consistency of the processed data, we calculated basic statistics for each species, including the total number of genes, protein-coding genes, and the percentage of redundant mRNA identified during GFF analysis. Furthermore, we assessed the efficiency of the gene mapping process by calculating the mean percentage of genes successfully mapped with LIFTOFF after GFF correction. These metrics will provide a baseline for the comparative analysis of gene relationships and transcript redundancies across the genus. Technical validation of the comparative genomics module was performed by assessing the consistency of gene mapping and the identification of redundant mRNA sequences across nine *Quercus* species.

In addition, to evaluate the genomic consistency and conservation of QUERCUSTOA across the nine *Quercus* species, we performed a comparative distribution analysis of shared gene clusters and orthogroups (OGs) to detect core orthologs and species-specific OGs and gene clusters using five distinct methodologies: First, "oak gene clusters" were defined based on the homology relationships between the nine genomes. Functional orthology was assessed by mapping *Quercus* protein clusters to the EGGNOG and TAIR databases within the QUERCUSTOA framework. We identified orthologs based on homology with the *A. thaliana* proteome (TAIR). Additionally, protein clusters were assigned to EGGNOG OGs at two hierarchical taxonomic levels: Streptophyta (representing broad plant conservation) and order (representing lineage-specific conservation). Finally, we used the alignment-based *de novo* approach of ORTHOFINDER (Emms and Kelly, 2019) on the original protein files of each genome. To minimize redundancy and ensure computational efficiency, the input protein sets were filtered to include only the longest isoform per gene for each species. For each method, we calculated the frequency of gene clusters/OGs shared by a specific number of species, ranging from 1 (species-specific) to 9 (core genome). The resulting distribution was expressed as the percentage of the total gene clusters/OGs identified by each method to allow for direct comparison of genomic stability and database curation quality.

2.5.2 Functional annotation and benchmarking

We evaluated two database groups: DB1, generated using a protein-centric protocol analogous to GYMNOTOA-DB (Mora-Márquez et al., 2025), and DB2, constructed using annotated genome assemblies from NCBI and CNCB-NGDC repositories. While DB1 relies on a broad retrieval of *Quercus* protein sequences (taxid:3511) from multiple sources -including UNIPROT/SWISSPROT (The Uniprot Consortium, 2017), PIR (George et al., 1997), PRF (Protein Research Foundation, 2026), PDB (Burley et al., 2025), GENBANK (Sayers et al., 2025), and REFSEQ (O'Leary et al., 2016)- DB2 integrates protein/GFF data from genome assemblies to enable advanced comparative genomics features.

To determine the optimal configuration, we benchmarked five combinations of MMSEQS2 coverage (`-c`) and identity (`--min-seq-id`) thresholds for both groups. Sensitivity and completeness were assessed by comparing total consensus sequences, functional assignments (INTERPROSCAN and EGGNOG-MAPPER), *A. thaliana* (TAIR10) homology, and BUSCO scores.

2.5.3 Functional recovery in non-model oak species

To evaluate the practical utility of QUERCUSTOA for species lacking a reference genome, we simulated a "degenerated" transcript set using 111 EST sequences from *Q. pubescens* available in TREEGENES (Treegenes Database, 2026). We employed the `debase-transcript-sequences.py` script from the NGSHELPER suite to introduce realistic biological and sequencing noise, including fragmentation (*prob* = 0.3), single-nucleotide mutations (*prob* = 0.5), and indels (*prob* = 0.25). The resulting sequences were annotated using the QUERCUSTOA-APP on the QUERCUSTOA-DB2-a database and cross-referenced with TRAPID2.0 (Van Bel et al., 2013) (using *PLAZA Dicots 4.5*) (Van Bel et al., 2018) to benchmark performance against a standard plant genomics platform.

A critical challenge in benchmarking these pipelines is the lack of standardized nomenclature across protein domain databases. While TRAPID provides assignments primarily based on INTERPRO (IPR) accessions, QUERCUSTOA integrates signatures from PFAM, EGGNOG, and PANTHER. To ensure a fair comparison, Gene Ontology (GO) terms were adopted as the universal functional language for validation.

3 Results

3.1 Database architecture and content

The QUERCUSTOA database is hosted at the UPMDRIVE cloud repository (Universidad Politécnica de Madrid, 2014–2026) and is permanently archived in the ZENODO repository (Mora-Márquez et al., 2026b). The data can be accessed and downloaded via the project website (https://blogs.upm.es/quercustoa/) or using the QUERCUSTOA-APP software available at the GITHUB (https://github.com/GGFHF/quercusTOA-app) and ZENODO repositories (Mora-Márquez et al., 2026a). The dataset consists of three relational SQLite databases (`sequences.db`, `functional-annotation.db`, and `comparative-genomics.db`), two BLAST+ databases ('lncRNA', and 'clustered consensus oak proteins'), and one DIAMOND database ('clustered consensus oak proteins') (Table 2). To facilitate data integration, the three SQLite databases share a common relational structure where the protein identification acts as the unique connector across all datasets (Supplementary Material 1). While this modular architecture allows for advanced SQL querying, the QUERCUSTOA-APP provides a user-friendly interface to exploit these relationships. Beyond simple data retrieval, the application allows users to input protein or transcript sequences to identify homologous genes and proteins across the nine *Quercus* species, automatically performing multiple sequence alignments of

TABLE 2 Summary of data records and database architecture within QuercusTOA.

Data record	File name	Format	Description
Raw sequences	`sequences.db`	SQLite	Contains raw protein and gene sequences for available *Quercus* species
Functional annotation	`functional-annotation.db`	SQLite	Includes protein clusters, GO terms, metabolic pathways, and *A. thaliana* orthologs
Comparative genomics	`comparative-genomics.db`	SQLite	Stores gene/CDS correspondences, homologous proteins, and unmapped genes
lncRNA BLAST+ DB	`lncRNA`	BLAST+	Database for long non-coding RNA sequences
Clustered proteins	`clustered_consensus_oak_proteins`	BLAST+/DIAMOND	Consensus protein sequences for clustered oak orthologs

The database integrates structured genomic and proteomic resources for *Quercus* species across three primary SQLite modules: sequences.db, containing raw protein and gene sequences; functional-annotation.db, which includes protein clusters, GO terms, metabolic pathways, and *Arabidopsis thaliana* orthologs; and comparative-genomics.db, storing gene/CDS correspondences and homologous protein relationships. Additionally, the repository provides specialized BLAST+ and DIAMOND databases for long non-coding RNA (lncRNA) sequences and consensus protein sequences derived from clustered oak orthologs.

the recovered orthologs, and retrieving functional data. This functionality effectively bridges the gap between raw genomic data and comparative evolutionary analysis.

The SQLite databases are organized into functional modules as follows:

Sequence Database (`sequences.db`)

This database serves as the repository for all raw sequence data:

- `species_protein_seqs`: Stores the full amino acid sequences for each protein ID and species.
- `species_gene_seqs`: Stores the nucleotide sequences for the identified genes.

Comparative Genomics Database (`comparative-genomics.db`)

This record stores the results of the genomic lift-over and comparative analysis among species:

- `liftoff_gff_cds_data` and `liftoff_gff_gene_data`: Detail the genomic coordinates (start, end, strand) and correspondences for CDS and genes between reference and target species.
- `liftoff_homologous_proteins`: Maps protein identifications between reference and target species.
- `liftoff_unmapped_genes`: Records genes from the reference species that could not be mapped to the target genome.
- `mmseqs2_concatenated_cds_clusters`: Links concatenated CDS sequences to their respective clusters, genes, and protein IDs.

Functional Annotation Database (`functional-annotation.db`)

This database contains the core functional information for the protein clusters. It is structured into the following tables:

- `mmseq2_protein_clusters`: Maps protein sequences to their respective cluster IDs and species.
- `interproscan_annotations`: Stores GO terms and metabolic pathway IDs derived from InterPro, PANTHER, MetaCyc, and Reactome.
- `emapper_annotations`: A comprehensive table including eggNOG orthologous groups, COG categories, KEGG information (KOs, pathways, modules, reactions), CAZymes, and Pfam families.
- `tair10_info` and `tair10_orthologs`: Contain *A. thaliana* peptide descriptions and their corresponding homology relationships with the oak protein clusters.
- `go_ontology`: A reference table for Gene Ontology term names and namespaces (biological process, molecular function, or cellular component).

3.2 The QuercusTOA-app: a graphical interface for data exploration

To facilitate the exploration and analysis of the integrated resources, we developed the QuercusTOA-app. Built in Python 3, this application features a cross-platform graphical user interface (GUI) compatible with Linux, Windows, and macOS, while also providing BASH scripts for high-throughput analysis on Linux servers. The application inherits robust functional annotation and enrichment utilities from the GymnoTOA-app framework (Mora-Márquez et al., 2025), including integrated BLAST/DIAMOND search engines and functional enrichment modules.

To ensure accessibility for users with limited bioinformatics expertise, QuercusTOA-app automates the installation of required third-party dependencies and manages the direct download of the latest database versions from remote repositories. A key enhancement in this iteration is the Comparative Genomics Module, which enables homology searches via protein ID or sequence to retrieve consensus proteins alongside original sequences from all nine oak species. The tool automatically generates gene/protein FASTA files, multiple sequence alignments, and phylogenetic trees via Mafft. The application also allows users to supply their own genome assemblies to map annotations -using any database species as a reference through Liftoff producing outputs that are fully compatible with LiftoffTools (Shumate and Salzberg, 2024) for

TABLE 3 Summary of genomic features and LIFTOFF mapping statistics for the nine *Quercus* species.

Species	#Genes	#Proteins in coding genes	Redundant mRNA (n)	Redundant mRNA (%)	Mean% targeted maps liftoff (s.d.)	Mean% reference maps liftoff (s.d.)
Quercus acutissima	31,490	31,040	0	0.00	86.91 (3.62)	92.81 (1.10)
Quercus dentata	28,626	35,836	2,906	8.11	90.32 (2.55)	88.14 (1.18)
Quercus gilva	40,166	40,166	0	0.00	88.93 (2.94)	92.18 (0.87)
Quercus lobata	41,698	53,226	8,126	15.27	89.58 (2.72)	87.32 (1.87)
Quercus longispica	35,901	35,901	0	0.00	90.12 (3.18)	84.79 (1.10)
Quercus robur	41,859	53,760	10,303	19.16	89.59 (2.10)	84.54 (1.84)
Quercus rubra	33,247	47,694	6,205	13.01	87.84 (3.07)	89.20 (1.24)
Quercus suber	39,985	44,993	6,179	13.73	88.02 (3.91)	88.91 (0.77)
Quercus variabilis	34,681	34,681	0	0.00	89.38 (4.77)	92.80 (0.95)

The table summarizes the total gene counts, protein-coding genes, and total proteins within the GFF for each assembly. "Redundant mRNA" (n and %) refers specifically to mRNA, features within a single gene that encode for identical protein sequences, which were identified and normalized during our GFF analysis to ensure a non-redundant dataset. Mapping efficiency is presented via two metrics: "Mean targeted maps" represents the average percentage of genes successfully mapped to the species when it acts as the target for the other eight species. "Mean reference maps" represents the average success rate of the species' own genes when mapped onto the other eight target species. Standard deviations (SD) are provided in parentheses.

further genomic analysis. Furthermore, QUERCUSTOA-app can also run on Linux servers using the BASH scripts included in the software.

3.3 Technical validation

3.3.1 Sequence integrity and comparative genomics module

The high success rate of gene mapping via LIFTOFF, with mean targeted map percentages consistently exceeding 86% (SD < 4.8) and mean reference maps above 84% (SD < 1.9), demonstrates the robustness of the cross-species annotation transfer (Table 3). Notably, the reference map percentages reached their highest values (>92%) in *Q. acutissima*, *Q. gilva*, and *Q. variabilis*, indicating that these chromosome-level assemblies provide highly reliable source annotations for the genus. Additionally, the quantification of redundant mRNAs -specifically defined as multiple mRNA features within a gene encoding identical protein sequences-highlights the effectiveness of our GFF analysis in normalizing disparate assembly formats. By identifying and removing 10,303 such redundant sequences in *Q. robur* and 8,126 in *Q. lobata*, the pipeline ensures that comparative metrics are not biased by identical duplicate entries in the original datasets.

The comparative analysis of the nine *Quercus* genomes included in QUERCUSTOA revealed a characteristic U-shaped distribution in the sharing of orthogroups and gene clusters across all five methodologies. This pattern suggests a genomic architecture composed of a substantial conserved core alongside a fraction of lineage-specific or low-frequency clusters (Figure 2). The proportion of core orthogroups (n = 9) varied depending on the taxonomic scale and clustering stringency of each method. The highest conservation was identified at the EGGNOG Streptophyta level (60.66%), followed by TAIR-based mapping (48.39%) and the oak gene clusters (43.70%). In contrast, the more lineage-specific EGGNOG order level and the *de novo* ORTHOFINDER approach identified 29.77% and 26.71% of the clusters as core components, respectively. Regarding low-frequency clusters, the combined percentage of gene clusters/OGs present in only one or two species remained within a similar range across several independent methods: 21.12% for ORTHOFINDER, 24.80% for oak gene clustering, and 29.26% for EGGNOG order. While the internal distribution between these two categories varied, with ORTHOFINDER showing a higher frequency at n = 2 (14.86%) compared to n = 1 (6.26%), and oak genes peaking at n = 1 (21.10%), the aggregate values for the accessory component were consistent across methodologies.

3.3.2 Functional annotation and benchmarking

Our results demonstrate that DB2 achieves an annotation quality equivalent to DB1 while significantly expanding the sequence space (Table 4). The DB2-a parameter combination (`MMseqs2 -c 1.0, --min-seq-id 1.000`) was identified as the most comprehensive resource. While both configurations achieved high functional coverage, the DB2-a setup proved superior by enabling the direct synchronization of functional annotations with genomic coordinates through the LIFTOFF-derived mapping. This integration allows for a more comprehensive analysis than the protein-centric DB1 approach, as it facilitates the study of gene synteny and the identification of non-redundant genomic loci across the *Quercus* genus.

Unlike clustered versions that prioritize size reduction, DB2-a preserves maximum functional diversity, identifying 7,584 clusters with distinct EGGNOG-based assignments and 4,711 with unique INTERPROSCAN-based features. Remarkably, this functional breadth is highly comparable to the gold standard provided by the TAIR10 proteome, which identifies 7,734 and 5,709 clusters in the same categories, respectively.

A critical observation from the benchmarking process is the trade-off between database size reduction and functional

10.3389/fbinf.2026.1821531

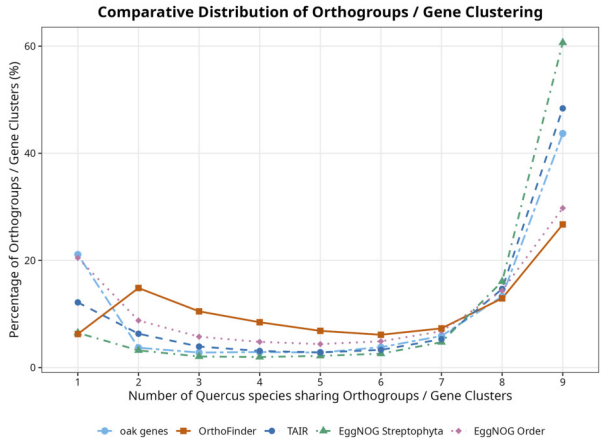

FIGURE 2
Comparative distribution of shared orthogroups and gene clusters across nine *Quercus* species. The plot illustrates the percentage of orthogroups/gene clusters (Y-axis) shared by an increasing number of species (X-axis), ranging from species-specific (1) to core orthogroups (9). Five different methodologies are compared: oak gene clusters, EggNOG OGs (OrthoGroups) at two taxonomic levels (Streptophyta and Order), TAIR (Arabidopsis-based orthology), and OrthoFinder orthogroups (*de novo* clustering).

information retention. As clustering parameters become more permissive -specifically in the DB2-c configuration where the coverage threshold is removed (`-c 0.0`)- the total number of clusters drops significantly to 84,705. However, this reduction in size is accompanied by a sharp increase in the percentage of unannotated clusters, which rises from 6.21% in DB2-a to 17.00% in DB2-c. This trend suggests that aggressive clustering leads to the loss of unique functional signatures.

Finally, DB2-a ensures full proteome representation, recovering 1,612 complete BUSCO genes (99.8% of the 1,614 searched). These metrics, combined with the identification of 288,743 clusters showing TAIR10 homology, reinforce the selection of DB2-a as the optimal balance for providing a high-resolution functional and comparative resource for *Quercus* genomics.

3.3.3 Functional recovery in non-model oak species

The comparative analysis demonstrated that QUERCUSTOA significantly increases the functional recovery rate, as evidenced by assigning metadata to 74.8% (83/111) of the transcripts compared to the 62.2% (69/111) achieved by TRAPID. This improvement is further reflected in the reduction of "dark" sequences, where our tool left only 25.2% (28/111) without functional assignment, whereas TRAPID left 37.8% (42/111) unannotated. Beyond mere recovery, QUERCUSTOA provided a much higher descriptive granularity; as shown in the distribution analysis, the pipeline achieved a significantly higher mean GO density of approximately 81 terms per sequence, nearly doubling the ~37 terms per sequence produced by TRAPID (Table 5; Figure 3A).

These results highlight QUERCUSTOA's ability to leverage multiple databases evidence to overcome annotation bottlenecks. By doubling the number of functional descriptors per sequence

TABLE 4 Clustering and functional annotation statistics for *Quercus* genome assemblies.

Version	MMseqs2 (-c, --msi)	Original-seqs (n)	Clusters (n)	InterProScan	EggNOG-mapper	TAIR10	% w/o annot	BUSCO
DB1-a	1.0, 1.000	297,142	196,486	157,597 (3,967)	188,413 (7,410)	179,957 (18,999)	3.45%	1,614 (1,612)
DB1-b	1.0, 0.900	297,142	143,408	111,457 (3,949)	135,863 (7,373)	128,057 (18,837)	4.46%	1,614 (1,612)
DB1-c	0.0, 0.900	297,142	60,281	43,075 (3,406)	54,833 (7,011)	50,104 (16,671)	7.89%	1,614 (1,585)
DB1-d	0.9, 0.750	297,142	85,700	62,635 (3,855)	79,096 (7,214)	74,020 (18,166)	6.65%	1,614 (1,612)
DB1-e	0.9, 0.900	297,142	99,870	73,662 (3,895)	92,857 (7,316)	85,917 (18,421)	6.02%	1,614 (1,612)
DB2-a	1.0, 1.000	377,297	324,105	250,406 (4,711)	301,337 (7,584)	288,743 (20,038)	6.21%	1,614 (1,612)
DB2-b	1.0, 0.900	377,297	218,143	159,315 (4,697)	196,492 (7,539)	185,632 (19,834)	8.82%	1,614 (1,612)
DB2-c	0.0, 0.900	377,297	84,705	53,004 (3,645)	68,758 (7,036)	62,641 (17,049)	17.00%	1,614 (1,552)
DB2-d	0.9, 0.750	377,297	135,483	90,130 (4,595)	115,762 (7,401)	108,248 (19,254)	13.08%	1,614 (1,612)
DB2-e	0.9, 0.900	377,297	155,888	105,790 (4,646)	135,162 (7,480)	125,544 (19,477)	11.90%	1,614 (1,612)

The table summarizes the performance of the functional annotation module across different database versions and MMseqs2 clustering parameters (sequence identity and coverage). Statistics include the total number of input sequences and resulting non-redundant clusters, alongside functional assignment counts and Gene Ontology (GO) terms retrieved via InterProScan and eggNOG-mapper (unique records in parentheses). Homology results against the TAIR10 (*Arabidopsis thaliana*) reference (unique records in parentheses) and the percentage of unannotated clusters are provided. Assembly and annotation completeness are represented by BUSCO (*embryophyta_odb10*) scores, showing the total groups searched and the count of recovered complete genes (in parentheses).

TABLE 5 Comparative performance of QUERCUSTOA and TRAPID2.0 in the functional annotation of *Quercus pubescens* transcripts.

Metric	QUERCUSTOA	TRAPID
Total sequences (N)	111	111
Global annotation (any metadata)	83 (74.8%)	69 (62.2%)
GO term assignment (functional)	65 (58.6%)	63 (56.7%)
Unannotated (no metadata)	28	42
Transcripts w/o GO terms	46	48
Mean GO density	High (~81)	Mean (~37)

The table summarizes the annotation metrics for a reference set of 111 simulated "degenerated" transcripts from *Quercus pubescens*. Global Annotation indicates the number and percentage of sequences receiving at least one piece of metadata (e.g., protein signatures, families, or GO terms). GO Term Assignment specifically denotes the subset of sequences with at least one associated Gene Ontology term. Unannotated sequences represent the "dark" transcriptomic fraction with no functional hits in the respective databases. Mean GO density reflects the average number of GO terms assigned per annotated sequence.

while maintaining a high rate of global assignment (Figure 3B), the platform proves to be a robust resource for the functional characterization of non-model oak species.

4 Discussion

The exponential growth of genomic resources for the genus *Quercus* has led to a situation where valuable information remains scattered across fragmented repositories (Zhang et al., 2025). QUERCUSTOA addresses this challenge by providing a centralized and scalable framework that synchronizes functional metadata with genomic positioning, offering a more cohesive biological context for oak research. A key advantage of our integrated architecture is the implementation of genomic lift-over via LIFTOFF, which represents a significant enhancement over purely protein-centric databases. By maintaining the correspondence between physical genomic coordinates and functional annotations, QUERCUSTOA facilitates comparative analyses across different infrageneric sections. This positional information allows the identification of orthologous relationships and syntenic blocks that are often obscured when analyzing single sequences. Furthermore, our taxonomy-aware approach allows for the characterization of both core and lineage-specific gene clusters at multiple taxonomic levels, providing the necessary resolution to explore the molecular basis of the high adaptive diversity that defines the genus *Quercus*.

The observed U-shaped distribution of orthogroup/cluster sharing across the nine oak genomes reflects a genomic architecture defined by a conserved core alongside a fraction of species-specific or accessory genes. This pattern is consistent with recent evidence for other species complexes, such as *Populus* (Shi et al., 2024), *Oryza* (Guo et al., 2025), and *Solanum* (Benoit et al., 2025). The identification of a significant core genome, representing 43.7% of our oak gene clusters, aligns with the high chromosomal synteny and stable genome size reported for the genus (Plomion et al., 2018; Larson et al., 2025). This structural stability suggests that the core component of QUERCUSTOA captures the shared evolutionary heritage of these species. Furthermore, the significant fraction of lineage-specific or low-frequency clusters (24.8% for oak gene clusters at n = 1 and n = 2) reflects the evolutionary lability observed in specific gene families. As noted by Larson et al. (2025), while the overall oak genome structure is conserved, certain regions, such as those involved in environmental adaptation or disease resistance, exhibit rapid turnover and diversification. Our results, derived from both *de novo* and homology-based strategies, support the overall coherence of QUERCUSTOA and are consistent with established genomic conservation patterns described for the *Quercus* genus (Plomion et al., 2018; Kremer and Hipp, 2020; Larson et al., 2025).

Beyond the database itself, the public availability of our database building scripts represents a significant contribution to the community. By providing the complete automated pipeline in an open format, we empower other researchers to build their own customized databases, allowing them to experiment with different parameters -such as sequence identity thresholds or coverage- and to even incorporate additional genomes as they are released. Moreover, the database generation scripts can be easily adapted to other species complexes at the infrageneric level, provided that enough high-quality genomes from different closely related species are available - i.e., genus *Populus* (Shi et al., 2024). This open-science approach makes it possible that researchers can independently update and adapt the resource to accommodate the continuous influx of new genomic data. This scalability is complemented by the QUERCUSTOA-APP, which ensures that these complex resources are easily operable for a broad range of users. By providing a graphical interface that automates local BLAST searches, alignments, and phylogenomic reconstructions, we lower the technical barrier for researchers who may not have extensive command-line expertise, facilitating the biological interpretation of large-scale genomic datasets through an intuitive analytical framework.

Despite these strengths, the utility of the database is inherently influenced by the heterogeneous quality of the original genome assemblies and the diverse pipelines employed for gene inference. Variations in the number of repetitive elements (Rey et al., 2023), high levels of heterozygosity (Kremer et al., 2012) or assembly contiguity (Sork et al., 2022) and annotation depth among the source genomes (Plomion et al., 2018) can impact the success rate of genomic lift-over and the density of functional assignments in oaks. Although most genomes included in QUERCUSTOA share standard protocols for gene modeling, including repeat identification, RNA-seq integration, and homology searches, the specific methods have evolved significantly over time. We chose to rely on the original gene models provided by the authors of each assembly rather than imposing a unified re-annotation protocol. Researchers should, therefore, consider these underlying technical differences when interpreting comparative results across species with divergent assembly levels, such as those at the scaffold versus chromosome level. Future updates of QUERCUSTOA will focus on the continuous integration of new high-quality assemblies and the refinement of functional mapping to maintain the tool as a robust reference. Ultimately, by reconciling fragmented information into a standardized, taxonomy-aware framework, QUERCUSTOA constitutes a valuable resource for future studies in oak evolution and molecular breeding.

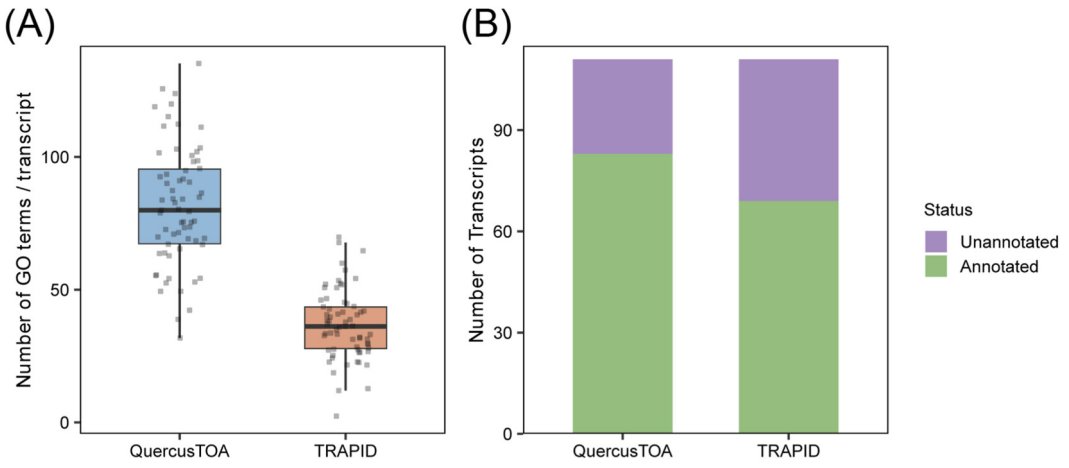

FIGURE 3
Benchmarking of annotation efficiency and functional richness. **(A)** Boxplot showing the distribution of GO term percentages by annotated transcripts. Despite the challenges in ontology mapping, QUERCUSTOA provides a denser functional profile per sequence **(B)** Comparative annotation success rate. The chart illustrates the percentage of transcripts successfully assigned to any functional metadata ("Annotated") versus those remaining uncharacterized ("Unannotated"). QUERCUSTOA reduced the number of uncharacterized transcripts to 28, compared to 42 in the TRAPID pipeline.

Data availability statement

The QUERCUSTOA database is hosted at the UPMDRIVE cloud repository (Universidad Politécnica de Madrid, 2014–2026) and is permanently archived in the ZENODO repository (Mora-Márquez et al., 2026b). The data can be accessed and downloaded via the project website (https://blogs.upm.es/quercustoa/) or using the QUERCUSTOA-APP software available at the GITHUB (https://github.com/GGFHF/quercusTOA-app) and ZENODO repositories (Mora-Márquez et al., 2026a).

Author contributions

FM-M: Conceptualization, Data curation, Formal Analysis, Methodology, Resources, Software, Validation, Writing – review and editing. MH: Conceptualization, Data curation, Methodology, Validation, Writing – review and editing, Formal Analysis. UH: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was funded by the Spanish Ministerio de Ciencia e Innovación-Agencia Estatal de Investigación (grant number PID 2023–152249NB-I00 MICIU/AEI/10.13039/501100011033) and by an AWS research program supported by the UPM Cloud Unit for Research (AWS grant number ID3-ETSIM-SuberOmics).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. The authors declare that Gen AI was used in the creation of this manuscript. Gemini was used to improve the quality of the English language.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2026.1821531/full#supplementary-material

References

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21, 30. doi:10.1186/s13059-020-1935-5

Arias-Baldrich, C., Silva, M. C., Bergeretti, F., Chaves, I., Miguel, C., Saibo, N. J. M., et al. (2020). CorkOakDB-The Cork oak genome database portal. *Database* 2020, baaa114. doi:10.1093/database/baaa114

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276. doi:10.1016/j.ymeth.2012.05.001

Benoit, M., Jenike, K. M., Satterlee, J. W., Ramakrishnan, S., Gentile, S., Hendelman, A., et al. (2025). *Solanum* pan-genetics reveals paralogues as contingencies in crop engineering. *Nature* 640, 135–145. doi:10.1038/s41586-025-08619-6

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* 53, 474–485. doi:10.1002/dvg.22877

Buchfink, B., Xie, C., and Huson, D. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176

Burgin, J., Ahamed, A., Cummins, C., Devraj, R., Gueye, K., Gupta, D., et al. (2023). The European nucleotide archive in 2022. *Nucleic Acids Res.* 51, D121–D125. doi:10.1093/nar/gkac1051

Burley, S. K., Bhatt, R., Bhikadiya, C., Bi, C., Biester, A., Biswas, P., et al. (2025). Updated resources for exploring experimentally-determined PDB structures and computed structure models at the RCSB protein data bank. *Nucleic Acids Res.* 53, D564–D574. doi:10.1093/nar/gkae1091

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf* 10, 421. doi:10.1186/1471-2105-10-421

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829. doi:10.1093/molbev/msab293

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 42, D459–D471. doi:10.1093/nar/gkt1103

Chen, Q., Zobel, J., and Verspoor, K. (2017). Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database* 2017, baw163. doi:10.1093/database/baw163

CNCB–NGDC (2025a). *Quercus acutissima genome assembly*. GWH. Available online at: https://download.cncb.ac.cn/gwh/Plants/Quercus_acutissima_Quercus_a_v1.0_GWHBGBO00000000/ (Accessed February 27, 2026).

CNCB–NGDC (2025b). *Quercus dentata genome assembly*. GWH. Available online at: https://download.cncb.ac.cn/gwh/Plants/Quercus_dentata_NA_GWHBRAD00000000/ (Accessed February 27, 2026).

CNCB–NGDC (2025c). *Quercus gilva genome assembly*. GWH. Available online at: https://download.cncb.ac.cn/gwh/Plants/Quercus_gilva_Quercus_gilva_V1_GWHDOCW00000000/ (Accessed February 27, 2026).

CNCB–NGDC (2025d). *Quercus longispica genome assembly*. GWH. Available online at: https://download.cncb.ac.cn/gwh/Plants/Quercus_longispica_LKM1_GWHESEV00000000/ (Accessed February 27, 2026).

CNCB–NGDC (2025e). *Quercus variabilis genome assembly*. GWH. Available online at: https://download.cncb.ac.cn/gwh/Plants/Quercus_variabilis_Qv_SD_1.0_GWHEQCV00000000/ (Accessed February 27, 2026).

CNCB–NGDC Members and Partners (2026). Database resources of the national genomics data center, China national center for bioinformation in 2026. *Nucleic Acids Res.* 54, D28–D47. doi:10.1093/nar/gkaf1172

Denk, T., Grimm, G. W., Manos, P. S., Deng, M., and Hipp, A. L. (2017). An updated infrageneric classification of the oaks: review of previous taxonomic schemes and synthesis of evolutionary patterns. *Tree Physiol.* 7, 13–38. doi:10.1007/978-3-319-69099-5_2

Drula, E., Garron, M. L., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* 50, D571–D577. doi:10.1093/nar/gkab1045

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20 (1), 238. doi:10.1186/s13059-019-1832-y

Escandón, M., Castillejo, M. Á., Jorrín-Novo, J. V., and Rey, M. D. (2021). Molecular research on stress responses in *Quercus* spp.: from classical biochemistry to systems biology through omics analysis. *Forests* 12, 364. doi:10.3390/f12030364

Falk, T., Herndon, N., Grau, E., Buehler, S., Richter, P., Zaman, S., et al. (2018). Growing and cultivating the forest genomics database, TreeGenes. *Database* 2018, 84–111. doi:10.1093/database/bay084

George, D. G., Dodson, R. J., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., et al. (1997). The protein information resource (PIR) and the PIR-international protein sequence database. *Nucleic Acids Res.* 25, 24–27. doi:10.1093/nar/25.1.24

Guo, D., Li, Y., Lu, H., Zhao, Y., Kurata, N., Wei, X., et al. (2025). A pangenome reference of wild and cultivated rice. *Nature* 642, 662–671. doi:10.1038/s41586-025-08883-6

Hernández-Plaza, A., Szklarczyk, D., Botas, J., Cantalapiedra, C. P., Giner-Lamia, J., Mende, D. R., et al. (2023). eggNOG 6.0: enabling comparative genomics across 12,535 organisms. *Nucleic Acids Res.* 51, D389–D394. doi:10.1093/nar/gkac1022

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi:10.1093/nar/gkn785

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070

Kapoor, B., Jenkins, J., Schmutz, J., Zhebentyayeva, T., Kuelheim, C., Coggeshall, M., et al. (2023). A haplotype-resolved chromosome-scale genome for *Quercus rubra* L. provides insights into the genetics of adaptive traits for red oak species. *G3* 13, jkad209. doi:10.1093/g3journal/jkad209

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010

Kremer, A., and Hipp, A. L. (2020). Oaks: an evolutionary success story. *New Phytol.* 226, 987–1011. doi:10.1111/nph.16274

Kremer, A., Abbott, A. G., Carlson, J. E., Manos, P. S., Plomion, C., Sisco, P., et al. (2012). Genomics of fagaceae. *Tree Genet. Genomes* 8, 583–610. doi:10.1007/s11295-012-0498-3

Larson, D. A., Staton, M. E., Kapoor, B., Islam-Faridi, N., Zhebentyayeva, T., Fan, S., et al. (2025). A haplotype-resolved reference genome of *Quercus alba* sheds light on the evolutionary history of oaks. *New Phytol.* 246, 331–348. doi:10.1111/nph.20463

Li, H., and Durbin, R. (2024). Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* 25, 658–670. doi:10.1038/s41576-024-00718-w

Liang, Y. Y., Liu, H., Lin, Q. Q., Shi, Y., Zhou, B. F., Wang, J. S., et al. (2025). Pan-genome analysis reveals local adaptation to climate driven by introgression in oak species. *Mol. Biol. Evol.* 42, msaf088. doi:10.1093/molbev/msaf088

Liu, D., Xie, X., Tong, B., Zhou, C., Qu, K., Guo, H., et al. (2022). A high-quality genome assembly and annotation of *Quercus acutissima* Carruth. *Front. Plant Sci.* 13, 1068802. doi:10.3389/fpls.2022.1068802

López de Heredia, U., Vázquez, F. M., and Soto, Á. (2017). The role of hybridization on the adaptive potential of mediterranean sclerophyllous oaks: the case of the *Quercus ilex* x *Q. suber* complex. *Tree Physiol.* 7, 551–576. doi:10.1007/978-3-319-69099-5_17

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi:10.1093/nar/gkaa913

Mora-Márquez, F., Chano, V., Vázquez-Poletti, J. L., and López de Heredia, U. (2021a). TOA: a software package for automated functional annotation in non-model plant species. *Mol. Ecol. Resour.* 21, 621–636. doi:10.1111/1755-0998.13285

Mora-Márquez, F., Vázquez-Poletti, J. L., and López de Heredia, U. (2021b). NGScloud2: optimized bioinformatic analysis using Amazon Web Services. *PeerJ* 9, e11237. doi:10.7717/peerj.11237

Mora-Márquez, F., Hurtado, M., and López de Heredia, U. (2025). gymnotoa-db: a database and application to optimize functional annotation in gymnosperms. *Database* 2025, baaf019. doi:10.1093/database/baaf019

Mora-Márquez, F., Hurtado, M., and López de Heredia, U. (2026a). quercusTOA-APP: a desktop application for oak functional annotation and comparative genomics. *Zenodo*. doi:10.5281/zenodo.18740268

Mora-Márquez, F., Hurtado, M., and López de Heredia, U. (2026b). quercusTOA database: integrating functional annotations and comparative genomics across oak lineages. *Zenodo*. doi:10.5281/zenodo.18712996

Nagamitsu, T., Uchiyama, K., Izuno, A., and Shimizu, H. (2024). Common-garden study of introgression at loci associated with traits adaptive to coastal environment from *Quercus dentata* into *Quercus mongolica* var. *crispula*. *Plant Species Biol.* 39, 231–248. doi:10.1111/1442-1984.12472

NCBI Genome (2019). *Quercus lobata (valley oak) genome assembly*. National Center for Biotechnology Information. Available online at: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/633/185/GCF_001633185.2_ValleyOak3.2/ (Accessed February 27, 2026).

NCBI Genome (2022). *Quercus robur genome assembly*. NCBI. Available online at: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/932/294/415/GCF_932294415.1_dhQueRobu3.1/ (Accessed February 27, 2026).

NCBI Genome (2024a). *Quercus rubra genome assembly*. NCBI. Available online at: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/035/136/125/GCA_035136125.1_Qrubra_687_v2.0/ (Accessed February 27, 2026).

NCBI Genome (2024b). *Quercus suber (cork oak) genome assembly*. NCBI. Available online at: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/906/115/GCF_002906115.3_Cork_oak_2.0/ (Accessed February 27, 2026).

Nixon, K. C. (2006). Global and neotropical distribution and diversity of oak (genus *quercus*) and oak forests. *Ecological Studies* 185, 3–13. doi:10.1007/3-540-28909-7_1

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi:10.1093/nar/gkv1189

Plomion, C., Aury, J.-M., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C., et al. (2016). Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Mol. Ecol. Resour.* 16, 254–265. doi:10.1111/1755-0998.12425

Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., et al. (2018). VOak genome reveals facets of long lifespan. *Nature Plants*4, 440–452. doi:10.1038/s41477-018-0172-3

Protein Research Foundation (2026). Protein research foundation (PRF) database. Available online at: http://www.prf.or.jp/index-e.html (Accessed February 27, 2026).

Quercus Portal (2024). A European genetic and genomic web resources for *Quercus*. Available online at: https://quercusportal.pierroton.inra.fr/ (Accessed February 27, 2026).

Ramos, A. M., Usié, A., Barbosa, P., Barros, P. M., Capote, T., Chaves, I., et al. (2018). The draft genome sequence of cork oak. *Sci. Data* 5, 180069. doi:10.1038/sdata.2018.69

Ren, Y. B., Yu, S. L., Sun, H., and Ma, X. G. (2025). Genomic insights into introgression between *Quercus aquifolioides* and its sympatric relatives across elevational gradients. *Mol. Ecol.* 34, e17747. doi:10.1111/mec.17747

Rey, M. D., Labella-Ortega, M., Guerrero-Sánchez, V. M., Carleial, R., Castillejo, M. Á., Ruggieri, V., et al. (2023). A first draft genome of holm oak (*Quercus ilex* subsp. *ballota*), the most representative species of the Mediterranean forest and the Spanish agrosylvopastoral ecosystem dehesa. *Front. Mol. Biosci.* 10, 1242943. doi:10.3389/fmolb.2023.1242943

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi:10.1016/s0168-9525(00)02024-2

Sayers, E. W., Cavanaugh, M., Frisse, L., Pruitt, K. D., Schneider, V. A., Underwood, B. A., et al. (2025). GenBank 2025 update. *Nucleic Acids Res.* 53, D56–D61. doi:10.1093/nar/gkae1114

Sayers, E. W., Bolton, E. E., Fine, A. M., Kelly, C., Kim, S., Landrum, M., et al. (2026). Database resources of the national center for biotechnology information in 2026. *Nucleic Acids Res.* 54, D20–D27. doi:10.1093/nar/gkae112

Shi, T., Zhang, X., Hou, Y., Jia, C., Dan, X., Zhang, Y., et al. (2024). The super-pangenome of populus unveils genomic facets for its adaptation and diversification in widespread forest trees. *Mol. Plant* 17, 725–746. doi:10.1016/j.molp.2024.03.009

Shumate, A., and Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics* 37, 1639–1643. doi:10.1093/bioinformatics/btaa1016

Shumate, A., and Salzberg, S. (2024). LiftoffTools: a toolkit for comparing gene annotations mapped between genome assemblies. *F1000Res* 11, 1230. doi:10.12688/f1000research.124059.2

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351

Sork, V. L., Fitz-Gibbon, S. T., Puiu, D., Crepeau, M., Gugger, P. F., Sherman, R., et al. (2016). First draft assembly and annotation of the genome of a California endemic oak Quercus lobata née (Fagaceae). *G3 (Bethesda)* 6, 3485–3495. doi:10.1534/g3.116.030411

Sork, V. L., Cokus, S. J., Fitz-Gibbon, S. T., Zink, A. G., Puiu, D., Pellegrini, M., et al. (2022). High-quality genome and methylomes illustrate features underlying evolutionary success of oaks. *Nat. Commun.* 13, 2047. doi:10.1038/s41467-022-29584-y

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi:10.1038/nbt.3988

Szcześniak, M. W., Rosikiewicz, W., and Makałowska, I. (2016). CANTATAdb: a collection of plant long non-coding RNAs. *Plant Cell Physiol.* 57, e8. doi:10.1093/pcp/pcv201

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi:10.1093/nar/gkw1099

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi:10.1101/gr.772403

TreeGenes Database (2026). *Quercus pubescens transcriptome shotgun assembly (qupu_TSA.fasta)*. TreeGenes. Available online at: https://treegenesdb.org/FTP/Transcriptome/TSA/Qupu/ (Accessed February 27, 2026).

Usié, A., Serra, O., Barros, P. M., Barbosa, P., Leão, C., Capote, T., et al. (2023). An improved reference genome and first organelle genomes of *Quercus suber*. *Tree Genet. Genomes* 19, 54. doi:10.1007/s11295-023-01624-8

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., and Vandepoele, K. (2013). TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. *Genome Biol.* 14, R134. doi:10.1186/gb-2013-14-12-r134

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., et al. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* 46, D1190–D1196. doi:10.1093/nar/gkx1002

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., et al. (2023). Welcome to the big leaves: best practices for improving genome annotation in non-model plant genomes. *Appl. Plant Sci.* 11, e11533. doi:10.1002/aps3.11533

Wang, W. B., He, X. F., Yan, X. M., Ma, B., Lu, C. F., Wu, J., et al. (2023). Chromosome-scale genome assembly and insights into the metabolome and gene regulation of leaf color transition in an important oak species, *Quercus dentata*. *New Phytol.* 238, 2016–2032. doi:10.1111/nph.18814

Whittemore, A. T., and Schaal, B. A. (1991). Interspecific gene flow in sympatric oaks. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2540–2544. doi:10.1073/pnas.88.6.2540

Zhang, K. L., Huang, S. Y., Li, Y., Chen, M. X., Zhu, F. Y., and Fang, Y. M. (2025). Genetic and epigenetic views of the adaptive evolution of oaks (*Quercus* L.). *Plant Cell Environ.* 48, 6307–6320. doi:10.1111/pce.15603

Zhou, X., Liu, N., Jiang, X., Qin, Z., Farooq, T. H., Cao, F., et al. (2022). A chromosome-scale genome assembly of *Quercus gilva*: insights into the evolution of *Quercus* section *Cyclobalanopsis* (Fagaceae). *Front. Plant Sci.* 13, 1012277. doi:10.3389/fpls.2022.1012277