Beverage Plant Research

METHOD

Open Access

https://doi.org/10.48130/bpr-0023-0041

*Beverage Plant Research* **2024**, 4: e008

# GFAnno: integrated method for plant flavonoid biosynthesis pathway gene annotation

Liuxu Du[#], Cui Lu[#], Zhentao Wang[#], LongXiang Zou, Yi Xiong and Qunjie Zhang[*]

*Center of Genomics and Bioinformatics, Guangdong Key Laboratory of Plant Molecular Breeding, College of Agriculture, South China Agricultural University, Guangzhou 510642, China*

[#] These authors contributed equally: Liuxu Du, Cui Lu, Zhentao Wang

[*] Corresponding author, E-mail: s-zqj@163.com

## Abstract

Flavonoids are important secondary metabolites synthesized by tea plants. However, inconsistencies in the variations in gene annotation methods across numerous studies, have hindered the comparison of results from previous studies. In this work, we offer 'GFAnno', an open-source software package which annotates genes and gene families based on sequence features, along with annotated parameters for 18 key genes related to the flavonoid biosynthesis pathway. The package takes a protein sequence file as input, performs gene annotation based on the identity and coverage of pre-prepared known seed protein sequences and the coverage of conserved Hidden Markov Model (HMM) domain. We used 11 dicotyledon, seven monocotyledon, and two basal angiosperm genomes to construct three datasets. We then use the seed species collection to construct seed sequences, use the test species collection to follow strict parameter selection rules, and use the validation species collection to verify the accuracy of the analysis results. The annotation results of validation collection using the filtering parameters by test collection shows that, our parameter selection can effectively exclude various structurally incomplete and abnormal proteins, while correctly distinguishing genes with high sequence similarity, such as Flavonoid 3'-Hydroxylase (F3'H) and Flavonoid 3'5'-Hydroxylase (F3'5'H) in the cytochrome P450 (CYP450). Our work aids ongoing tea plant pan-genome research by offering a convenient software for target gene annotation and sets comparative standards for analyzing the flavonoid biosynthesis pathway and conducting sequence comparison of catalytic enzymes.

**Citation:** Du L, Lu C, Wang Z, Zou L, Xiong Y, et al. 2024. GFAnno: integrated method for plant flavonoid biosynthesis pathway gene annotation. *Beverage Plant Research* 4: e008 https://doi.org/10.48130/bpr-0023-0041

## Introduction

The complexity of plant secondary metabolism poses major challenges for analyses in different species due to the evolution of multiple independent pathways. With the rapid development of third-generation sequencing, the era of pan-genomics has fully arrived, offering us new opportunities to study the diversity of secondary metabolites in plants. However, the classical annotation methods for superfamily, gene family, and multicopy genes are similar, but with different parameter settings. Extensive research is required to manually examine sequence features, such as sequence similarity and coverage, as well as enzyme catalytic domains, in order to determine the identification parameters for each target gene[1,2]. This can create analytical obstacles for researchers outside of the field.

Taking gene annotation in the flavonoid biosynthesis pathway of tea plants as an example. Flavonoids are an important class of biologically active secondary metabolites produced by tea plants[3]. In addition to their myriad of health benefits, flavonoids contribute to the enjoyable flavors and overall sensory experience of tea. The flavonoid pathway is complex, involving a wide array of enzymes, which generates diverse metabolites. In tea plants, the copy number of key genes involved in flavonoid biosynthesis differs significantly among different tea genomes[4–6]. Annotation of the same genome using different tools and parameters often results in different gene numbers. For instance, we used Orthofinder[7] and utilized BLASTP[8] and HMMsearch[9] to annotate the *C. sinensis* genomes 'Shuchazao' (SCZ) and 'Yunkang10' (YK10) in two different studies and received different results[4, 6]. In addition, errors often occur when distinguishing multi-copy genes with very similar sequences within the same gene family. For example, the cytochrome P450 (CYP450, which includes Flavonoid 3'-Hydroxylase (F3'H), Flavonoid 3'5'-Hydroxylase (F3'5'H), Flavone Synthase II (FNSII), and Trans-cinnamate 4-Monooxygenase (C4H)) and the 2-oxoglutarate-dependent dioxygenase (2OGDs, which includes Flavanone 3-Dioxygenase (F3H), Anthocyanidin Synthase (ANS), Leucoanthocyanidin Dioxygenase (LDOX) and Flavonol Synthase (FLS)) superfamilies participate in multiple oxidation and hydroxylation reactions within the flavonoid biosynthesis pathway, leading to the production of various flavonoid compounds[10–13]. F3'H and F3'5'H, which hydroxylate a broad range of flavonoid substrates, are characterized by a high sequence similarity (Supplemental Fig. S1)[10–12]. Therefore, it is necessary to use precise annotation parameters for the accurate identification of different gene copies within these families to avoid false-positive or false-negative annotation results.

Here, we present 'GFAnno,' an open-source software package that integrates the workflow based on sequence similarity and conserved domain analysis into a single package. We used three dicots, two monocots, and a basal angiosperm species to generate the seed sequence, provide strict filtered and doubled

checked parameters, allowing researchers to obtain annotations and classifications for the 18 key genes in the plant flavonoid biosynthesis pathway with just one command. Furthermore, this paper offers a parameter selection process and guidelines for creating parameters for the identification of other target genes, enhancing the applicability and scalability of this workflow.

## Materials and methods

### Workflow of GFAnno

We offer an open-source software package 'GFAnno' to annotate genes and gene families based on sequence features. GFAnno consists of a standalone command line program written in Python which runs on most Linux systems, which are freely available at https://github.com/qunjie-zhang/gfanno. The GFAnno flowchart is summarized in Fig. 1.

### File preparation

Obtain the Hidden Markov Model (HMM), BLASTP seed sequences, and prepare query protein files. Create a config file to include the local of 'seed sequence' file, 'HMM Model file' and the values of parameters (b_iden, b_qcov, b_tcov, and h_cov).

Perform BLASTP[8] searches with an E-value threshold of $< 1E^{-10}$. Apply the filter parameters of identity ≥ b_iden, query coverages ≥ b_qcov and target coverages within b_qcov range to the search results.

Conduct HMMsearch[9] with an E-value threshold of $< 1E^{-10}$. Apply the filter parameter of domain coverage ≥ h_cov to the search results.

Determine the intersection of the two filtered datasets, which will yield the remaining data representing the candidate genes of the target gene.

### Parameters filtering workflow

The annotation process of GFAnno relies on the protein sequence files from known genes, important structural domain model files, and filtering parameters. Three datasets are used for the parameter selection process: the seed datasets are used to construct BLASTP[8] search the database and optimize the HMM model, the test datasets are used for parameter filtering, and the validation datasets are used to test whether the parameters selected earlier are correct. The production and filtering processes for these files are as shown in Fig. 2.

#### *BLASTP seed sequence generation*

Based on the distribution range of the target gene, select a set of model species that represent various taxonomic positions as starting points for analysis. Obtain the reviewed protein sequence of target genes in selected species from the Swiss-Prot database (www.uniprot.org/uniprotkb?query=reviewed:true)[14]. Remove sequences that are too long, too short, or have incomplete structural domains based on the common features of the target gene, then CD-hit (v 4.8.1)[15] with an identity threshold of 0.9 was used to reduce redundancy to build seed sequence for BLASTP. Subsequently, use an all-by-all BLASTP (v2.2.31+)[8] search to obtain a similarity matrix between seed sequences, find the lowest similarity value, and set it as the initial value for the subsequent 'b_iden' parameter.

#### *HMM model selection*

Important structural domain files for gene families or superfamilies were obtained by downloading from Protein Families database (Pfam)[16] and SMART[17] database . We first use seed sequences to check whether the downloaded HMM domains are suitable for detecting the target gene. If not, we obtained the sequence of the gene family containing the gene from the Conserved Domain Database (CDD)[18] and used HMMbuild to build a custom model.

#### *Parameter selection*

The protein sequences of species in test datasets were first subjected to HMMsearch with an E-value threshold of $< 1E^{-10}$. The protein sequences obtained from the HMMsearch were

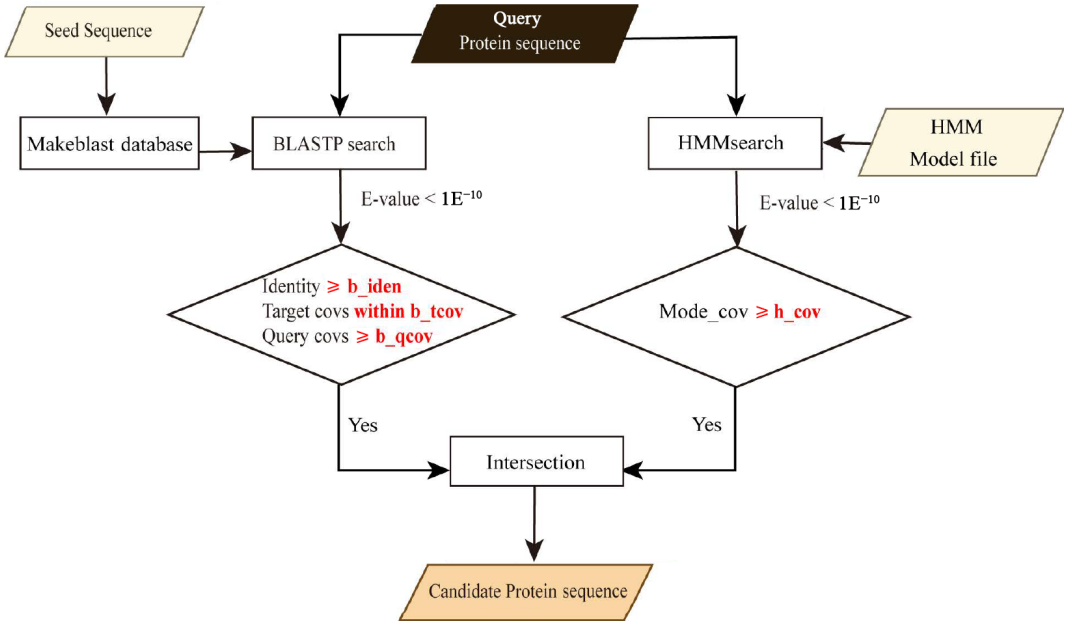


**Fig. 1** Workflow of GFAnno. The local of 'seed sequence' file, 'HMM Model file' and the values of parameters (b_iden, b_qcov, b_tcov, and h_cov) were obtained from the config file.

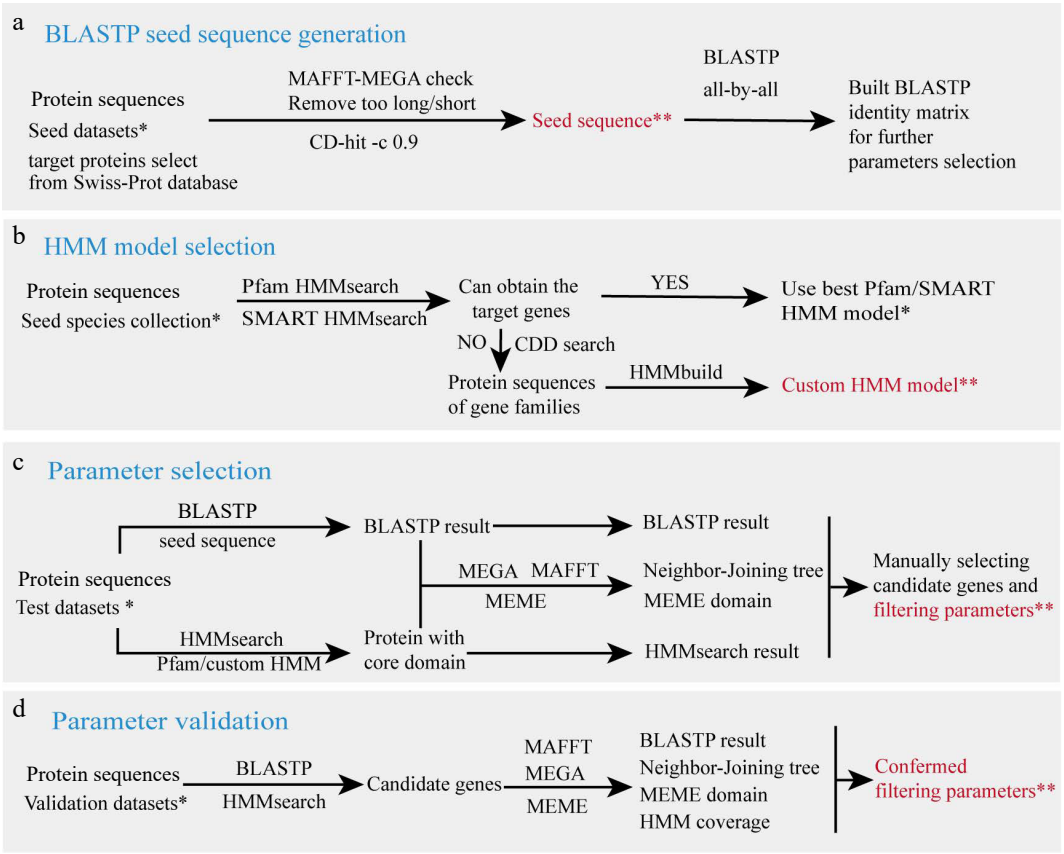


**Fig. 2** Parameter filtering workflow. (a) BLASTP seed sequence generation. (b) HMM model selection. (c) Parameter selection. (d) Parameter validation. Species collections in each step are marked with a single asterisk (*), and output data is marked with double asterisks (**), which are provided in github.

aligned using MAFFT (v7.310) for multiple sequence alignment. The alignment results were subsequently trimmed by TrimAl (v1.4.rev15)[19]. An evolutionary tree was constructed from the trimmed sequences using MEGA (v11.0.10)[20] with the following parameters: Neighbor-Joining method, Poisson correction, and bootstrap value of 1,000. For enhanced visual representation, the generated trees were further refined using iTOL[17]. The protein sequences obtained from the HMMsearch were further analyzed using BLASTP seed sequences with an E-value threshold of $< 1E^{-10}$. The MEME[21] program was used to identify conserved motifs and the Tbtools[22] software was utilized for the visualization of specific domains and conserved motifs. Parameters for further batch searching were established by manually examining the BLASTP identity (b_iden), BLASTP query coverages (b_qcov), BLASTP target coverages (b_tcov), and HMMsearch domain coverage (h_cov) for the target genes.

### *Parameter validation*

We validate the parameters obtained from the above steps using proteins from the test dataset to ensure good filtering results when expanding the dataset.

## Flavonoid biosynthesis pathway used in this work

We conducted a systematic literature search using the PubMed database as primary source, selecting classic studies of these genes in plants or literature focused on functional research of these genes in tea plants (Supplemental Table S1). We incorporated research conducted by Cui et al.[23], which elucidated the involvement of UGT84A enzymes in the production of $\beta$-glucogallin as the acyl donor. Furthermore, we included the work of Yao et al.[24] that demonstrated the participation of Serine Carboxypeptidase-Like (SCPLIA) in the acylation of catechins. To address the challenges posed by the high similarity among members of the CYP450 (Supplemental Fig. S1) and 2OGD superfamilies, we implemented special annotations for these enzymes. Specifically, C4H, FNS II, F3'H, and F3'5'H, which are involved in hydroxylation and oxidation reactions[25], are denoted by an asterisk (*). Similarly, 2OGDs, such as F3H, FLS, ANS, and LDOX, which are involved in oxidation reactions[26, 27], are marked with a double asterisk (**) (Fig. 3).

## Species and protein used in this work

Flavonoids are primarily found in angiosperms, and some basal angiosperm groups have also been found to harbor genes associated with this pathway. In our study, we constructed the seed dataset using three dicots (*Arabidopsis thaliana*[28], *Vitis vinifera*[29] and *C. sinensis* 'Tieguanyin'[30]), two monocots (*Oryza sativa*[31] and *Apostasia shenzhenica*[32]), and a basal angiosperm (*Amborella trichopoda*[33]); the test dataset was constructed using two dicots (*Populus trichocarpa*[34] and *C. sinensis* 'Huangdan'[35]), a monocot (*Zostera marina*[36]), and a basal angiosperm (*Cinnamomum kanehirae*[37]); and the validation dataset was constructed using six dicots (*Actinidia chinensis*[38], *Ficus hispida*[39], *C. sinensis* 'Longjing 43'[40], *C. sinensis* 'Shuchazao'[41] and *C. sinensis* 'DASZ'[4]) and three monocots (*Zea mays*[42], *Asparagus officinalis*[43] and *Musa balbisiana*[44]) (Supplemental Table S2). We selected these species based on

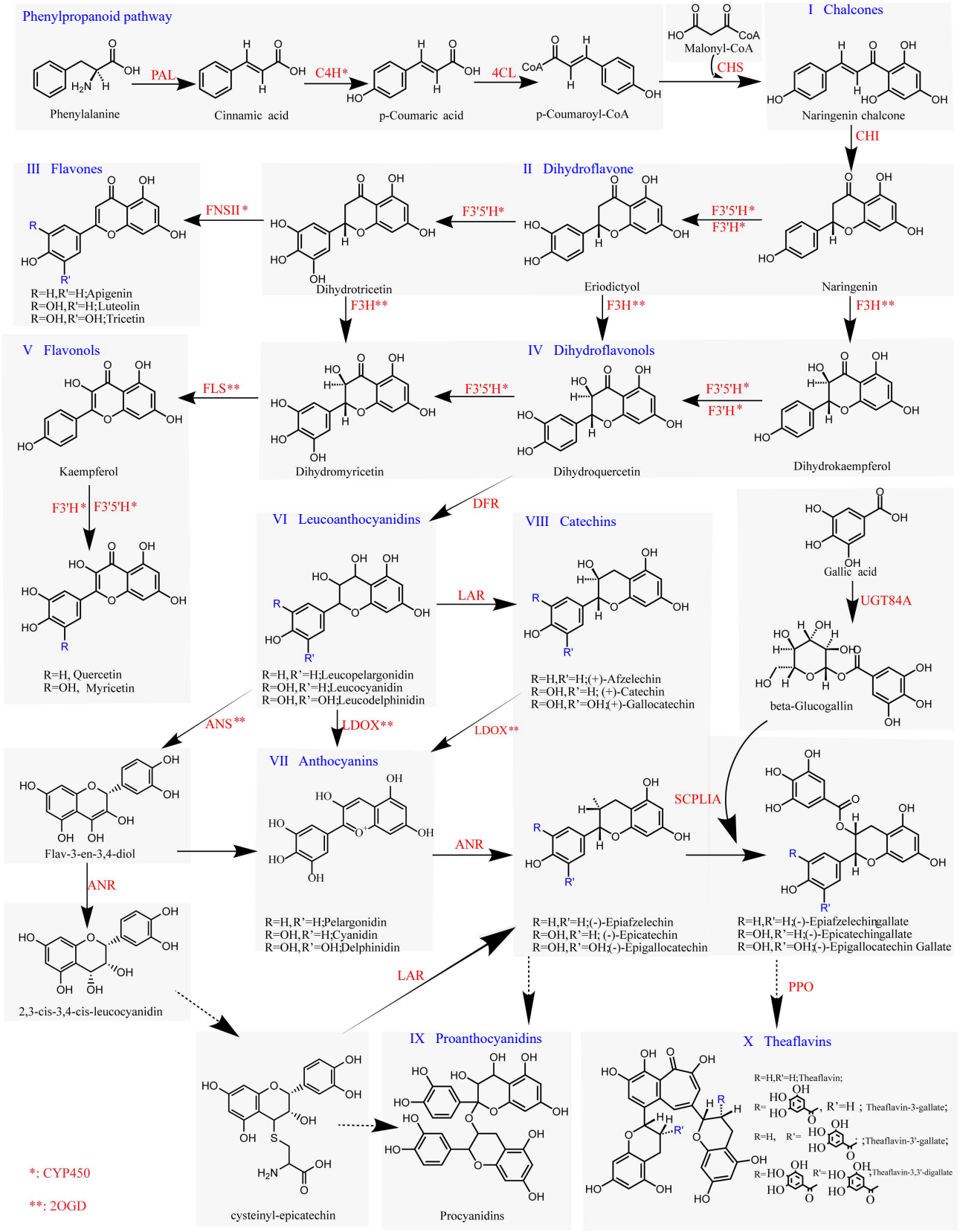


**Fig. 3** Plant flavonoid biosynthesis pathway. Enzymes labeled with '*' indicate members of CYP450s, whereas genes labeled with '**' represent members of 2OGDs. Solid-line arrows indicate well-established mechanisms for the corresponding enzymatic reactions, and dashed-line arrows represent pathways where the mechanistic details are yet to be fully determined.

ensuring their representativeness and the quality of genome annotation for each species, while also including a broad representation of various taxonomic groups within angiosperms. We use The Benchmarking Universal Single-Copy Orthologs (BUSCO)[44] completeness integrity to indicate genome annotation quality. The seed dataset requires a BUSCO integrity score of over 90, while the test and validation datasets require a BUSCO integrity score of over 80 (Supplemental Table S2).

### Filtering parameters for flavonoid biosynthesis-related gene in plants

Following the parameters filtering workflow in Fig. 2, we used proteins of the test species (Supplemental Table S2) downloaded from the Swiss-Prot database which have been validated for gene function (PE ≤ 2) to build the BLASTP seed sequence. In this work, for Uridine Diphosphate Galloyglycosyltransferase (UGT84A), Pfam HMM provided a model for the UGT superfamily (PF00201.21, UDPGT). However, only about 20% coverage was achieved when searching the model with known genes. Therefore, we downloaded the conserved sequence of the UGT superfamily (cl10013) from the CDD and built a custom model named cl10013.hmm for subsequent searches of the UGT84A. The KO IDs corresponding to these 18 genes were extracted from the KEGG database, and the annotation results were further validated using KAAS[45].

To provide a detailed parameter selection process, we provided a detailed presentation of the screening process and parameters for 18 genes (Figs 4–5, Supplemental Figs S2–S10). When selecting parameters, we initially set the b_iden based on the lowest value in the seed sequence identity matrix (Supplemental Table S3). In addition, the initial values of other parameters are as follows: b_qcov ≥ 70, 130 ≥ b_tcov ≥ 70, and h_covc ≥ 90. Subsequently, the process in Fig. 2 is completed using the test dataset to obtain accurate parameters. The final parameters obtained are listed in Table 1, and files and parameters used in this table can all be found on our GitHub page.

To verify the reliability of the parameters, we used the validation dataset to validate the parameters in Table 1 (Supplemental Figs S11 & S21). The results showed that our parameters can effectively identify all the 18 genes in Table 1.

## Input and output

### Input

The input to GFAnno includes the following options:

• The '-f / --fasta' flag specifies the input protein file in FASTA format.

• The '-c / --config' flag specifies the config file.

The config file can contain multiple genes, each with specific settings, including the local of seed file for BLASTP (blastp_seed) and corresponding filtering parameters (b_iden, b_qcov and b_tcov) for the target genes, the local of HMM file for HMMERsearch (hmm_file) and corresponding parameters (h_cov).

For first-time users of the software, you can use the '-g / --generate' parameter to generate a config file (flavno.ini) for the 18 genes involved in the plant catechin biosynthesis pathway, along with the required blast seed files and hmm_file files. If users need to annotate and filter other genes, they can refer to the format of the initial config file to create new filtering schemes.

### Output

The output of GFAnno includes annotated target gene sequences along with their alignment parameters. A folder is created with the prefix specified by '-o', and within this folder, a series of FASTA files are generated, named using both the specified prefix and the names of the target genes. The 'stat' file contains at least four columns of information: gene ID, blastp_identity with the seed file, blastp_qcovs, blastp_tcovs with the seed file, and hmm_coverage. If multiple HMM files are provided, additional columns will be output for the hmm_coverage of the second and third HMM models.

For example, running it with '-o flavno -c flavno.ini' will create a folder named 'flavno. In this folder, you can find a series of FASTA files named after gene names. These files contain candidate genes that have been annotated by GFAnno using the input file. Additionally, there is a 'flavno.stat' file that provides alignment parameters for the corresponding genes.

## Discussion

Annotation methods for superfamily, gene family, and multi-copy genes are similar, but with different parameter settings. Superfamily and family analysis focus on identifying characteristic structural domains and each gene or gene family has its own sequence characteristics. In the flavonoid metabolic pathway, DFR and ANR belong to the Short-chain dehydrogenases/reductases (SDRs) family (Supplemental Fig. S10), while F3H, FLS, ANS, and LDOX belong to the 2OGD superfamily (Fig. 4), C4H, FNS II, F3'H, and F3'5'H belong to the CYP450 superfamily (Fig. 5 & Supplemental Fig. S1). Species within the same superfamily share conserved domains, so they can use the same HMM model. Furthermore, conserved motif analysis showed that the candidate genes classified as CYP450 superfamily members shared four common regions, namely the heme-binding domain Phe-X-X-Glu-Arg-Arg-X-Cys-X-Gly, which is responsible for binding carbon monoxide, the Glu-X-X-Arg and the Ala-Gly-X-(Asp/Glu)-Thr-(Thr/Ser) motifs, which stabilizes the core structure and facilitates oxygen or substrate binding, respectively, and the hinge region Pro-Pro-Gly-Pro-Thr-(Gly/Pro)-Trp-Pro, which determines the correct orientation of CYP450 enzymes[46]. Characteristic iron-binding sites His-X-Asp-X(n)-His and the 2-oxoglutarate binding domain Arg-X-Ser, which are specific to 2OGDs[47]. The sequence similarity of LDOX and FLS, F3'H and F3'5'H is very high, and their phylogenetic relationship is also very close (Figs 4b & 5b). In our work, by using GFAnno, our parameters can accurately distinguish LDOX and FLS in the 2OGDs superfamily (Fig. 4a) and distinguish F3'H and F3'5'H in the CYP450 superfamily (Fig. 5a).

To assess the accuracy of the GFAnno flavonoid pathway gene annotation parameters, we utilized all proteins from *Ficus hispida* in a validation dataset as input and compared the results with KAAS annotations result (Supplemental Table S4). Supplemental Fig. S22 illustrates the phylogenetic relationships and comparison of conserved domains between the annotation results of GFAnno and KAAS. Red stars represent false positives from KAAS machine annotation, while green stars represent false negatives from KAAS annotation. Among the 18 genes, GFAnno effectively filtered out structurally incomplete sequences and some fused genes, resulting in an increase of 29.41% in accurate annotation results and the exclusion of 52.94% annotation errors. Additionally, other functional annotation software, such as InterPro[48] and blast2GO[49],

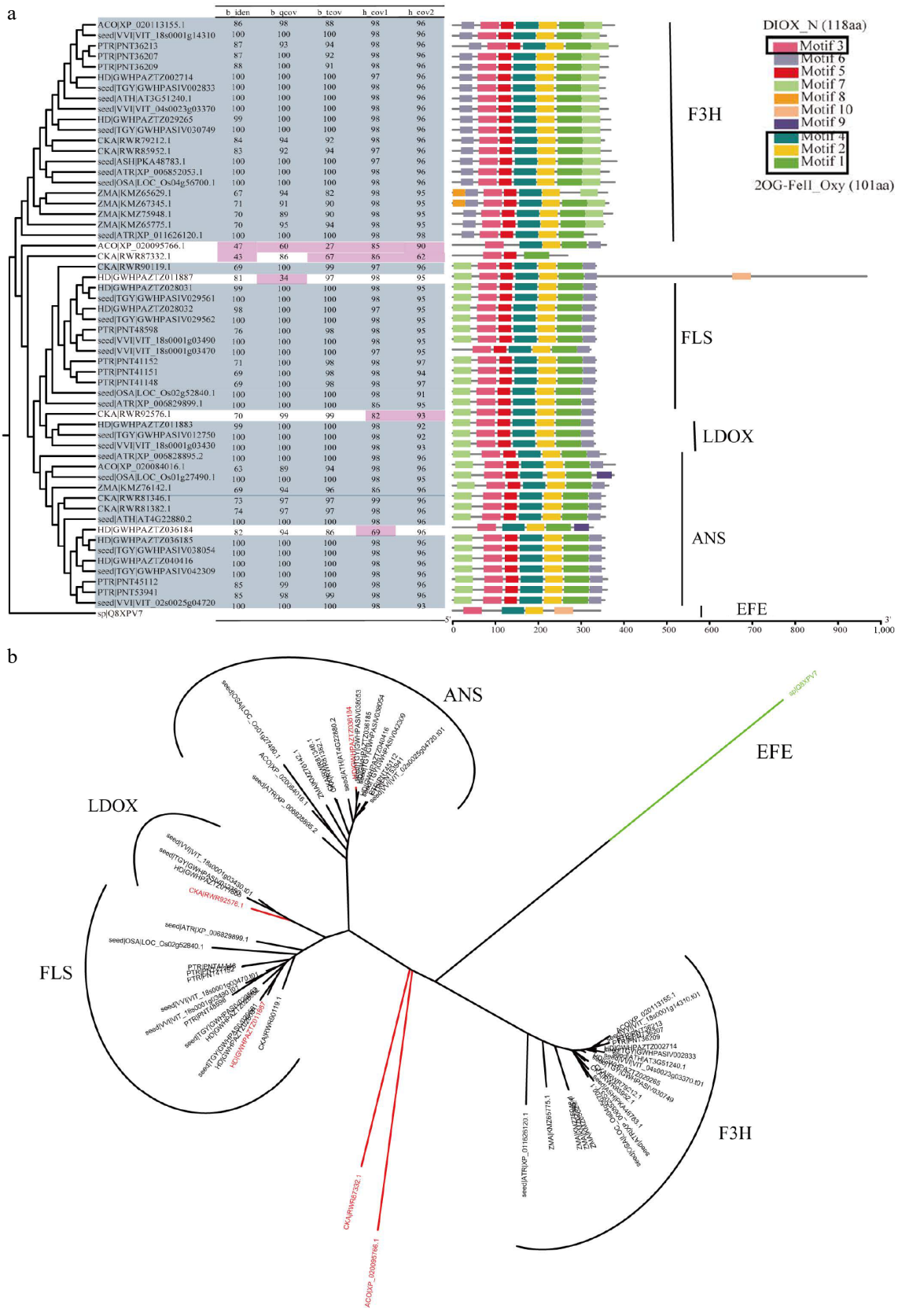


**Fig. 4** Overview of F3H, ANS, LDOX, and FLS in 2OGD superfamily. (a) The parameter selection for candidate genes in F3H, ANS, LDOX, and FLS. The phylogenetic tree, four parameters, and conservation module diagram to illustrate the selection process. In the phylogenetic tree, genes selected for are represented in blue-gray, while parameters that have been excluded are marked in red. The conserved domains were created using MEME[21], and the HMM models used in HMMsearch are outlined in black boxes, with the model length displayed in parentheses following the model's name. (b) Neighbor join tree shows the distance relationships between four genes.

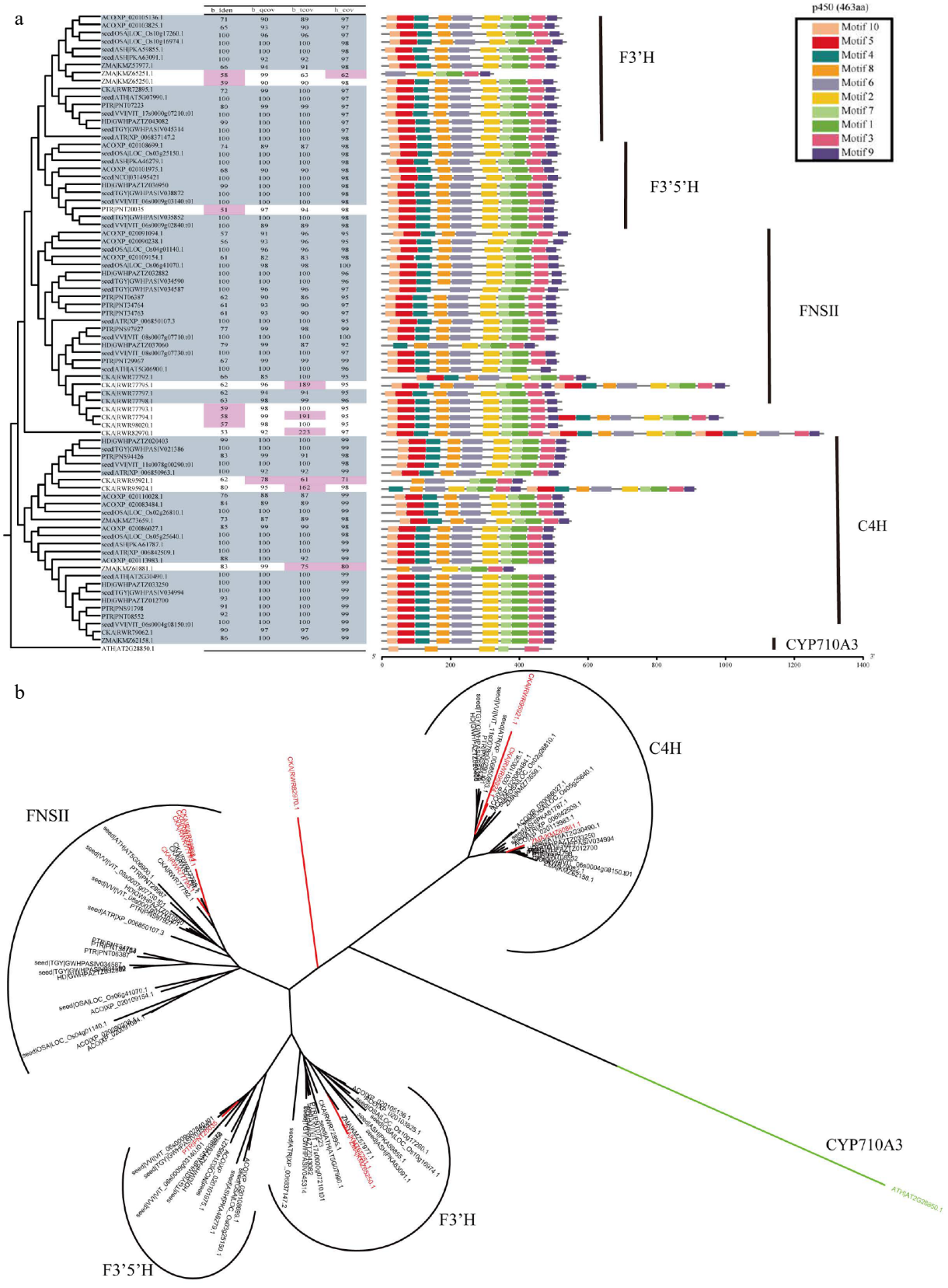


**Fig. 5** Overview of F3'H, F3'5'H, FNSII and C4H in CYP450 superfamily. (a) The parameter selection for candidate genes. The phylogenetic tree, four parameters, and conservation module diagram to illustrate the selection process. In the phylogenetic tree, genes selected for are represented in blue-gray, while parameters that have been excluded are marked in red. The conserved domains were created using MEME[21], and the HMM models used in HMMsearch are outlined in black boxes, with the model length displayed in parentheses following the model's name. (b) Neighbor join tree shows the distance relationships between four genes.

**Table 1.** Files and parameters used for the annotation pipeline.

| Enzyme | HMM/CDD ID | Parameter setting | | | |
|---|---|---|---|---|---|
| | | b_iden | b_qcov | b_tcov | h_cov |
| 4CL | PF13193.6 (AMP-binding_C) PF00501(AMP-binding) | 40 | 70 | 60–120 | 90/90 |
| CHI | PF02431.15 (Chalcone) | 35 | 70 | 70–130 | 90 |
| CHS | PF02797.15 (Chal_sti_synt_C) PF00195.19 (Chal_sti_synt_N) | 50 | 80 | 80–120 | 90/90 |
| LAR | PF05368.13 (NmrA) | 30 | 65 | 80–120 | 90 |
| PAL | PF00221.19 (Lyase_aromatic) | 60 | 80 | 80–120 | 90 |
| PPO | PF12142.8 (PPO1_DWL) PF12143.8 (PPO1_KFDV) | 35 | 70 | 70–130 | 90/90 |
| SCPL4* | PF00450.22 (Peptidase_S10) | 70 | 80 | 80–120 | 90 |
| SCPL5* | PF00450.22 (Peptidase_S10) | 70 | 80 | 80–120 | 90 |
| UGT84A | cl10013 | 40 | 80 | 80–120 | 90 |
| DFR | PF01370.21 (Epimerase) | 60 | 80 | 80–120 | 90 |
| ANR | | 55 | 80 | 80–120 | 90 |
| F3H | PF03171.23 (2OG-FeII_Oxy) PF14226.9 (DIOX_N) | 60 | 80 | 80–120 | 90/90 |
| FLS | | 60 | 80 | 80–120 | 90/90 |
| LDOX | | 50 | 80 | 80–120 | 90/90 |
| ANS | | 60 | 80 | 80–120 | 90/90 |
| F3'H | PF00067.25 (p450) | 50 | 80 | 80–120 | 90 |
| F3'5'H | | 60 | 80 | 80–120 | 90 |
| C4H | | 70 | 80 | 80–120 | 90 |
| FNSII | | 50 | 80 | 80–120 | 90 |

* SCPL4, SCPL5 in SCPLIA[24]. 'b_iden' denotes the identity of BLASTP; 'b_qcov' represents the 'query coverage per HSP' of BLASTP, 'b_tcov' represents the 'target coverage per HSP' of BLAST, and 'h_covc' indicates the HMM coverage of HMMsearch.

employs a similar approach using BLASTP and HMMsearch separately, which tends to yield a large number of sequences from the same gene family. Therefore, we also compiled the results obtained solely using BLASTP and HMMsearch (Supplemental Table S4). It can be observed that such annotations effectively identify the source of sequences but face challenges in determining the completeness of the target sequences (Figs 4 & 5; Supplemental Figs S2–S22).

The above results indicate that, when conducting batch analyses for pan-genomic studies, accuracy is crucial. Therefore, our study serves as a valuable addition to genes and gene families annotation based on sequence features, it also provides a firm foundation for future flavonoid studies.

## Conclusions

In conclusion, our study addresses the challenges in annotating genes related to the flavonoid biosynthesis pathway, offering 'GFAnno,' an open-source software package for gene and gene family annotation. We provide accurate parameters for precise identification of plant flavonoid biosynthesis-related genes. Our contribution lies in streamlining target gene annotation and establishing comparative benchmarks for analyzing the flavonoid biosynthesis pathway and comparing catalytic enzyme sequences, benefiting ongoing tea plant pan-genome research.

## Author contributions

The authors confirm contribution to the paper as follows: study conception & design and project management: Zhang Q; data analyses: Lu C, Du L, Xiong Y, Zou L; Python code: Du L, Wang Z; draft manuscritp preparation: Lu C, Du L; manuscript revision: Zhang Q. All authors reviewed the results and approved the final version of the manuscript..

## Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The GFAnno software is accessible at https://github.com/qunjie-zhang/gfanno.

## Acknowledgments

This study was supported by the Natural Science Foundation of China (32170625) (to Qunjie Zhang), Double first-class discipline promotion project (2021B10564001) (to Qunjie Zhang) and the Guangdong Special Support 318 Program (to Qunjie Zhang).

## Conflict of interest

The authors declare that they have no conflict of interest.

**Supplementary Information** accompanies this paper at (https://www.maxapress.com/article/doi/10.48130/bpr-0023-0041)

## Dates

Received 11 July 2023; Revised 9 December 2023; Accepted 14 December 2023; Published online 1 March 2024

## References

1. Zhang L, Chen F, Zhang X, Li Z, Zhao Y, et al. 2020. The water lily genome and the early evolution of flowering plants. *Nature* 577:79–84
2. Lozano R, Hamblin MT, Prochnik S, Jannink JL. 2015. Identification and distribution of the NBS-LRR gene family in the Cassava genome. *Bmc Genomics* 16:360
3. Ashihara H, Deng WW, Mullen W, Crozier A. 2010. Distribution and biosynthesis of flavan-3-ols in *Camellia sinensis* seedlings and expression of genes encoding biosynthetic enzymes. *Phytochemistry* 71:559–566
4. Zhang W, Zhang Y, Qiu H, Guo Y, Wan H, et al. 2020. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications* 11:3719
5. Zhang QJ, Li W, Li K, Nan H, Shi C, et al. 2020. The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Molecular Plant* 13:935–938
6. Lin P, Wang K, Wang Y, Hu Z, Yan C, et al. 2022. The genome of oil-Camellia and population genomics analysis provide insights into seed oil domestication. *Genome Biology* 23:14
7. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–10
9. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, et al. 2018. HMMER web server: 2018 update. *Nucleic Acids Research* 46:W200–W204
10. Guo L, Gao L, Ma X, Guo F, Ruan H, et al. 2019. Functional analysis of flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylases from

tea plant (*Camellia sinensis*), involved in the B-ring hydroxylation of flavonoids. *Gene* 717:144046
11. Wang YS, Xu YJ, Gao LP, Yu O, Wang XZ, et al. 2014. Functional analysis of flavonoid 3',5'-hydroxylase from tea plant (*Camellia sinensis*): critical role in the accumulation of catechins. *BMC Plant Biology* 14:347
12. Wei K, Wang L, Zhang C, Wu L, Li H, et al. 2015. Transcriptome analysis reveals key flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase genes in affecting the ratio of dihydroxylated to trihydroxylated catechins in *Camellia sinensis*. *Plos ONE* 10:e137925
13. Xiong S, Tian N, Long J, Chen Y, Qin Y, et al. 2016. Molecular cloning and characterization of a flavanone 3-Hydroxylase gene from *Artemisia annua* L.. *Plant Physiology and Biochemistry* 105:29–36
14. The UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* 51:D523–D531
15. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–52
16. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research* 49:D412–D419
17. Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Research* 49:D458–D460
18. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* 45:D200–D203
19. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–73
20. Tamura K, Stecher G, Kumar S. 2021. MEGA11: molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution* 38:3022–27
21. Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Research* 43:W39–W49
22. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, et al. 2020. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant* 13:1194–202
23. Cui L, Yao S, Dai X, Yin Q, Liu Y, et al. 2016. Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (*Camellia sinensis*). *Journal of Experimental Botany* 67:2285–97
24. Yao S, Liu Y, Zhuang J, Zhao Y, Dai X, et al. 2022. Insights into acylation mechanisms: co-expression of serine carboxypeptidase-like acyltransferases and their non-catalytic companion paralogs. *The Plant Journal* 111:117–33
25. Singh K, Kumar S, Rani A, Gulati A, Ahuja PS. 2009. *Phenylalanine ammonia-lyase* (*PAL*) and *cinnamate 4-hydroxylase* (*C4H*) and catechins (flavan-3-ols) accumulation in tea. *Functional & Integrative Genomics* 9:125–34
26. Singh K, Rani A, Kumar S, Sood P, Mahajan M, et al. 2008. An early gene of the flavonoid pathway, flavanone 3-hydroxylase, exhibits a positive relationship with the concentration of catechins in tea (*Camellia sinensis*). *Tree Physiology* 28:1349–56
27. Lin GZ, Lian YJ, Ryu JH, Sung MK, Park JS, et al. 2007. Expression and purification of His-tagged flavonol synthase of *Camellia sinensis* from *Escherichia coli*. *Protein Expression and Purification* 55:287–92
28. Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
29. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *Plos ONE* 2:e1326
30. Zhang X, Chen S, Shi L, Gong D, Zhang S, et al. 2021. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nature Genetics* 53:1250–59
31. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, et al. 2013. Improvement of the *Oryza sativa Nipponbare* reference genome using next generation sequence and optical map data. *Rice* 6:4
32. Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, et al. 2017. The *Apostasia* genome and the evolution of orchids. *Nature* 549:379–83
33. Amborella Genome Project , Albert VA, Barbazuk WB, Depamphilis CW, DER JP, et al. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* 342:1241089
34. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–604
35. Wang P, Yu J, Jin S, Chen S, Yue C, et al. 2021. Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Horticulture Research* 8:107
36. Olsen JL, Rouzé P, Verhelst B, Lin YC, Bayer T, et al. 2016. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature* 530:331–35
37. Chaw SM, Liu YC, Wu YW, Wang HY, Lin CYI, et al. 2019. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nature Plants* 5:63–73
38. Wu H, Ma T, Kang M, Ai F, Zhang J, et al. 2019. A high-quality *Actinidia chinensis* (kiwifruit) genome. *Horticulture Research* 6:117
39. Zhang X, Wang G, Zhang S, Chen S, Wang Y, et al. 2020. Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. *Cell* 183:875–889.E17
40. Wang X, Feng H, Chang Y, Ma C, Wang L, et al. 2020. Population sequencing enhances understanding of tea plant evolution. *Nature Communications* 11:4447
41. Xia E, Tong W, Hou Y, An Y, Chen L, et al. 2020. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Molecular Plant* 13:1013–1026
42. Eichten SR, Foerster JM, de Leon N, Kai Y, Yeh CT, et al. 2011. B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiology* 156:1679–90
43. Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, et al. 2017. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications* 8:1279
44. Wang Z, Miao H, Liu J, Xu B, Yao X, et al. 2019. *Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nature Plants* 5:810–21
45. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* 35:W182–W185
46. Werck-Reichhart D, Feyereisen R. 2000. Cytochromes P450: a success story. *Genome Biology* 1:reviews3003.1
47. Ding Q, Wang F, Xue J, Yang X, Fan J, et al. 2020. Identification and expression analysis of hormone biosynthetic and metabolism genes in the 2OGD family for identifying genes that may be involved in tomato fruit ripening. *International Journal of Molecular Sciences* 21:5344
48. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, et al. 2023. InterPro in 2022. *Nucleic Acids Research* 51:D418–D427
49. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–76