Genetica (2023) 151:325–338
https://doi.org/10.1007/s10709-023-00196-8

REVIEW

Approaches to increase the validity of gene family identification using manual homology search tools

Benjamin J. Nestor[1,2] · **Philipp E. Bayer**[1,2] · **Cassandria G. Tay Fernandez**[1,2] · **David Edwards**[1,2] · **Patrick M. Finnegan**[1,2]

Received: 7 June 2023 / Accepted: 1 October 2023 / Published online: 10 October 2023

Abstract

Identifying homologs is an important process in the analysis of genetic patterns underlying traits and evolutionary relationships among species. Analysis of gene families is often used to form and support hypotheses on genetic patterns such as gene presence, absence, or functional divergence which underlie traits examined in functional studies. These analyses often require precise identification of all members in a targeted gene family. Manual pipelines where homology search and orthology assignment tools are used separately are the most common approach for identifying small gene families where accurate identification of all members is important. The ability to curate sequences between steps in manual pipelines allows for simple and precise identification of all possible gene family members. However, the validity of such manual pipeline analyses is often decreased by inappropriate approaches to homology searches including too relaxed or stringent statistical thresholds, inappropriate query sequences, homology classification based on sequence similarity alone, and low-quality proteome or genome sequences. In this article, we propose several approaches to mitigate these issues and allow for precise identification of gene family members and support for hypotheses linking genetic patterns to functional traits.

Keywords Gene family identification · Homology · Genome analysis · Sequence similarity · Sequence evolution

Background to gene family identification

Recent increases in the number and quality of sequenced genomes has allowed in-depth comparison of genes between species and individuals through both single reference genomes and multiple species pangenomes (Bayer et al. 2020; Fernandez et al. 2022a). Genes shared between species or closely-related genes in the same species are known as homologs. Homologs known as orthologs originate from a common ancestral gene due to speciation events, while homologs known as paralogs arise from gene duplication in the same species (Fitch 1970; Setubal and Stadler 2018; Glover et al. 2019; Nevers et al. 2020). Identification of the homologs in gene families may take a whole genome approach where many different gene families and homologous members are identified, or a targeted approach where homologs of a specific gene family are identified with high accuracy. In both cases, genes translated from open reading frames to protein sequences are assigned as candidate homologs based on various measures of identity to protein sequences that are already characterised in those families from the same or different species. Sequences are usually classified as candidate homologs if the similarity between translated protein sequences is greater than that expected by chance (Pearson 2013; de Boissier and Habermann 2020). Candidate homologs can also be identified by the presence of conserved sequence regions such as motifs or functionally characterised domains in cases where exon shuffling, sequence rearrangements and modification, or gene fusion events cause low overall sequence identity between homologous sequences (Buljan and Bateman 2009; Forslund et al. 2011; Wu et al. 2012; Gabaldón and Koonin 2013). Homolog identification forms the basis for many downstream analyses in genome exploration such as analysis of trait and gene correlation, gene expression, gene functional mutation, gene ontology-based functional enrichment, phylogenetics, protein structure modelling, and comparative

✉ Benjamin J. Nestor
benjamin.nestor@research.uwa.edu.au

[1] School of Biological Sciences, University of Western Australia, Perth, WA 6009, Australia

[2] Centre for Applied Bioinformatics, University of Western Australia, Perth, WA 6009, Australia

genomics. This diversity of applications reinforces the need for accurate homology searches whether by whole genome or targeted approaches.

An important use of identifying homologs, specifically orthologs, is that the function of an uncharacterised protein sequence can be hypothesised by its relationship to an ortholog that is already functionally characterised. Relationships of orthologs are most usefully derived from model species with many functionally characterised genes such as *Arabidopsis thaliana,* rice (*Oryza sativa*), filamentous fungi (*i.e. Aspergillus nidulans*), mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*), or *Escherichia coli*. However, several studies have challenged the assumption that orthologs always have similar functions, particularly for orthologs in different species with high evolutionary distance. Protein sequence similarity alone does not indicate that sequences will share the same function and expression patterns of this function because high similarity can result from conserved sequence domains or other low-complexity regions (Pearson 2013; Sinha et al. 2018; Stamboulian et al. 2020). Hence, while functional characterisation of genes based on sequence similarity provides a basis for hypothesising gene functions, confirmation of these functions is needed through gene expression or other functional analyses.

The identification of gene families in genomes is often used to form or support hypotheses of functional studies based on the genes present and evolutionary relationships of species. These studies are particularly important where large scale analysis of complex traits is needed or where functional studies such as mutant studies would fail because of the inability to examine mutations of vital or functionally-redundant genes (Favre et al. 2014). An example of the use of homolog identification is in the linking of genetic patterns such as presence or absence of homologs, gene family size, protein structure, or conservation of sequence motifs, functional sites, and residues to functional traits among different species (Khan et al. 2016; Leelananda and Lindert 2016; Glover et al. 2019). Ideally, genetic patterns at specific genomic regions are compared between genomes of a species with the functional trait and a closely-related species lacking the functional trait to minimise genetic differences arising from evolutionary distance. Differences in genetic patterns between these species can then be hypothesised as a potential mechanism that underlies the trait (Huynen et al. 1998; Jim et al. 2004; Nevers et al. 2020). Functional studies such as this have been used to identify genes involved in symbioses with arbuscular mycorrhizae (Delaux et al. 2014; Favre et al. 2014) and symbioses with nitrogen-fixing bacteria (Mergaert et al. 2020; Radhakrishnan et al. 2020). Similar studies have also been performed in prokaryotic microorganisms to predict genes associated with temperature-dependent virulence (Bocsanczy et al. 2017) and gene patterns linked with flagella, pili and thermophily (Jim et al. 2004). However, the validity of analyses that involve homolog identification greatly depend on the accuracy of this identification. A high accuracy of homolog identification is particularly important where a gene is hypothesised to be absent from a genome because gene family members may easily be missed in identification steps.

Automated and manual pipelines for gene family identification

Many automated and manual pipelines for homology searches have been tested and benchmarked in services such as the Quest for Orthologs (Nevers et al. 2022). Well known examples of automated pipelines include Ortho-Markov Cluster Algorithm (OrthoMCL) (Li et al. 2003), Protein Annotation Through Evolutionary Relationship (PANTHER) (Thomas et al. 2003), and OrthoFinder (Emms and Kelly 2019) which have been reviewed extensively (see Glover et al. 2019; de Boissier and Habermann 2020; Nevers et al. 2020). Automated pipelines are generally used to rapidly compare large datasets for whole genome approaches such as genome annotation and the results from using different tools can easily be compared. However, this ability to compare large datasets comes at the cost of requiring a large amount of bioinformatic user-skill, computational power, and in-depth knowledge of tool usage to achieve precise identification of all gene family members without inclusion of members in other gene families (Steinegger et al. 2019; de Boissier and Habermann 2020; Nevers et al. 2020). Furthermore, automated pipelines trade the ability to manually curate homologous sequences between steps of the pipeline in favour of analysis speed and ease of use (Habermann 2016). Automated pipelines for homology search are useful for whole genome analyses, but often fall short for precise identification of gene families where the presence or absence of members must be confirmed with high confidence.

In manual pipelines, major steps are performed separately with simple homology-search tools. These steps often include a homology search tool to identify candidate homologs, usually predicted protein sequences, followed by sequence alignment and phylogenetic analysis. In brief, protein sequences with high similarity to query sequences in the targeted gene family are identified using homology search tools such as the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990; Camacho et al. 2009) or Hidden Markov Modeler (HMMER) (Eddy 2011). Matching sequences that pass a fixed threshold value of significance are extracted (Nevers et al. 2020). These protein sequences are aligned using a multiple sequence aligner such as MUltiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar 2004) or Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh and Standley 2013). Matching

protein sequences are compared to characterised members of the targeted gene family in a phylogenetic tree constructed with tools such as Randomized Axelerated Maximum Likelihood (RAxML) (Stamatakis 2014) or MrBayes (Ronquist et al. 2012). If the matching sequences appear to be part of the targeted gene family based on alignment and phylogenetic grouping, they are then reported as candidate homologs.

Manual pipelines have the advantage of allowing output sequences to be curated by the user at each step, greatly reducing the errors in the gene family members that are reported and in associated downstream analyses. Manual pipelines that incorporate conserved domain search tools as well as sequence similarity search tools are particularly useful for identifying remote homologs. Remote homologs are sequences that have low protein sequence similarity, making them difficult to detect by automated pipelines that generally use sequence similarity searches to identify homologs (Rost 1999; Habermann 2016; de Boissier and Habermann 2020). One downside to manual pipelines in comparison to automated pipelines is that the separation of steps comes at the cost of increased analysis time and reduced ability to compare results between different tools. However, the advantages of manual pipelines in precise gene family member identification leads them to be widely used for analyses of small targeted gene families.

Manual pipelines are commonly used to identify members of gene families in publications documenting the assembly of newly assembled genomes (Dong et al. 2021; Feng et al. 2021; Huang et al. 2021a; Li et al. 2021; Rai et al. 2021; Wang et al. 2021a; Apablaza et al. 2022). When implementing a manual pipeline to identify a gene family, it is important to consider several issues that affect the confidence in the homologs that are reported. Common issues in manual pipelines include too relaxed or stringent statistical thresholds, inappropriate query sequences, lack of multiple homology search tools, and low-quality proteome sequences (Pearson 2013; Sinha and Lynn 2014; Habermann 2016; Nevers et al. 2020). These issues can lead to exclusion of authentic homologs (false negatives) or the inclusion of inauthentic homologs from different gene families (false positives). False negatives and false positives can lead to poor support for hypotheses being tested and the propagation of errors in analyses of sequences, genetic mechanisms involved in functional traits, and evolutionary relationships among species. The perceived validity and reproducibility of the analyses is also weakened if the homology search process is not thoroughly documented in the research methods. Nevertheless, manual pipelines combined with strong justification of their use and methods are simple and precise tools for identifying gene family members and forming hypotheses. Many different variations of manual pipelines have been developed over the years. Here, we highlight the issues of manual pipelines in accurately identifying candidate homologs and review current approaches in the literature used to overcome these issues and to accurately identify gene family members with high confidence.

Sequence similarity searches using Basic Local Alignment Search Tool (BLAST)

Studies implementing manual pipelines for gene family identification often use the extensively-used homology search tool BLAST, which depends on sequence similarity (Li et al. 2021; Patiranage et al. 2021; Pei et al. 2021; Rai et al. 2021; Wang et al. 2021a; Xu et al. 2021; Zhang et al. 2021; Zhao et al. 2021; Zhong et al. 2021). BLAST uses small local alignments between a query and target sequence to find regions of sequence identity or similarity known as hits scored by pre-defined matrices (Altschul et al. 1990; Pearson 2013). Several statistics are provided in the output of BLAST to help determine if a hit is part of an authentic homolog, including the E-value, alignment length, alignment coverage, and percent identity between the hit and query sequence. The E-value is a statistical score of significance that can be explained as the number of high scoring hits that would be found simply due to random combinations of nucleotides or amino acids in the target sequence that match the query sequence (Wheeler and Bhagwat 2007; Pearson 2013). Hence, a low E-value indicates that a hit is statistically significant and provides evidence that an authentic homolog was identified in the BLAST search. E-value thresholds for statistical significance are usually set in the range of $1e^{-2}$ to $1e^{-20}$ (Wheeler and Bhagwat 2007; Pearson 2013; Setubal and Stadler 2018; Miao et al. 2021; Rai et al. 2021). However, no E-value threshold is applicable for all analyses because E-values are dependent on the size of the target database and the size of query and target sequences. Thresholds of other BLAST output statistics such as alignment coverage above 50–80% and percent identity above 50% can also be used to provide further evidence that hits are authentic homologs (Li et al. 2021; Pei et al. 2021). Where the lengths of typical homologs in the targeted gene family are known, BLAST hits can be filtered by length of alignment or length of the extracted open reading frame (ORF) to avoid shortened pseudogenes or sequences that match only part of the query sequence being included (Niu et al. 2021; Rai et al. 2021; Wang et al. 2021b). By carefully selecting thresholds for E-values, coverage, percent identity, and sequence length to filter BLAST hits, greater confidence can be gained that the selected BLAST hits are authentic homologs.

The foremost issue with BLAST homology searches is that too stringent E-value thresholds in BLAST searches can lead to false negatives, where authentic homologs are

missed, while too relaxed thresholds can lead to false positives, where inauthentic homologs are retained (Pearson 2013; Fujimoto et al. 2016; Habermann 2016; de Boissier and Habermann 2020). Low E-value thresholds are commonly used as the only threshold for determining if BLAST hits are authentic homologs, resulting in a high possibility for false negatives and a low confidence in any genes reported missing from the targeted genome. However, a higher confidence in BLAST searches can be achieved through a step-by-step process to pre-determine an appropriate E-value (Fig. 1). Firstly, a relatively high E-value between 1 and 10 is used to retrieve all potential hits of matching protein sequences with a high likelihood for false positives. The annotations of these protein sequences are then examined to identify a pass and a discard E-value threshold. The pass threshold is the highest E-value of a protein sequence annotated to be in the targeted gene family while the discard threshold is the lowest E-value of a protein

a) Example BLAST output with E-value threshold 1

Hit no.	E-value	Annotation	Classification	
1	0	Targeted gene family	Candidate homolog	✓
2	$1e^{-30}$	Targeted gene family	Candidate homolog (Pass threshold)	✓
3	$1e^{-1}$	Different gene family	Inauthentic homolog (Discard threshold)	x
4	1	Different gene family	Inauthentic homolog	x

→ **Select E-value threshold between $1e^{-1}$ and $1e^{-30}$**

b) Example BLAST output requiring annotation confirmation

Hit no.	E-value	Annotation	Classification	
1	0	Targeted gene family	Candidate homolog	✓
2	$1e^{-30}$	Targeted gene family	Candidate homolog	✓
3	$1e^{-29}$	Different gene family	Potential candidate homolog (Potential pass threshold)	?
4	$1e^{-2}$	Targeted gene family	Potential inauthentic homolog (Potential discard threshold)	?
5	$1e^{-1}$	Different gene family	Inauthentic homolog	x
6	1	Different gene family	Inauthentic homolog	x

→ **Confirm annotations of Hits 3 and 4 through alignment or re-annotation**

Fig. 1 Process for reporting BLAST hits as candidate homologs with high confidence using a pre-selected E-value based on pass and discard thresholds. **a** An E-value threshold between $1e^{-1}$ and $1e^{-30}$ is selected ($1e^{-20}$ for example) based on the pass and discard threshold E-values of BLAST hits. **b** Confirmation of sequence annotations is needed before selection of an E-value threshold. In this case, the potential pass threshold sequence is annotated in a different gene family, while the potential discard threshold sequence is annotated in the targeted gene family

sequence annotated as not in the targeted gene family. If the pass threshold E-value is relatively high compared to the E-values of other genes in the targeted gene family, or the discard threshold E-value is relatively low compared to E-values of other discarded genes, then the annotations of these sequences should be re-confirmed by BLAST searches to the National Center for Biotechnology Information (NCBI) non-redundant (NR) database (Sayers et al. 2022) or alignment with other protein sequences in the targeted gene family. A final E-value threshold between the pass and discard thresholds and if applicable also between the typical E-value range of $1e^{-2}$ to $1e^{-20}$ can then be chosen to filter the BLAST output and ensure that there are no false negatives or false positive homologs. In literature reports, the chosen E-value and the pass and discard threshold E-values should be reported as well as the annotations of the pass and discard threshold sequences to increase reader confidence in reported candidate homologs.

We have provided example data for choosing an E-value threshold during a BLAST homology search (Table 1). Here, BLAST searches were performed against the NCBI RefSeq predicted protein sequence database for *Nelumbo nucifera* (lotus) to identify members of the *PHOSPHATE TRANSPORTER 1* (*PHT1*) and *PHOSPHATE1* (*PHO1*) gene families. These gene families are important for phosphate transport in plants and are frequently the subject of homology searches involving plant genomes. In this example, all sequences annotated as PHT1 that were retrieved from the *N. nucifera* database by BLAST using 20 PHT1 protein query sequences from the model plant species *Arabidopsis thaliana* and *Oryza sativa* (rice) had an E-value of 0. The next best match, which belonged to another gene family, had an E-value of $6.05e^{-14}$. Hence, an E-value threshold below $6.05e^{-14}$, such as $1e^{-20}$, would be appropriate to retrieve homologs in this homology search. However, the situation was different for the identification of PHO1 protein sequences. Almost all sequences annotated as PHO1 that were retrieved from the *N. nucifera* database by BLAST using 14 PHO1 protein query sequences from *A. thaliana* and *O. sativa* had an E-value of 0, indicating that they are likely true homologs. In contrast, the next best match was to a non-PHO1 family member which had an E-value of $9.74e^{-34}$. Therefore, using the E-value threshold of $1e^{-20}$ as in the PHT1 search would lead to false positives and be inappropriate for this homology search. In addition, a short protein sequence 90 amino acids in length with a PHO1 annotation was recovered that had an E-value of $9.34e^{-10}$. This sequence would have been discarded if an E-value threshold above $9.74e^{-34}$ was used. In order to demonstrate that no *N. nucifera* PHO1 protein sequence homologs have been missed in the analysis, this short protein sequence annotated as PHO1 must be examined further to determine whether it can be filtered out due to its low sequence length or low alignment coverage, or whether the sequence is a potentially important PHO1 ortholog to be considered further in the study at hand. Once the outlier sequence has been kept or discarded, the E-value threshold can be set below $9.74e^{-34}$, such as $1e^{-40}$, to retrieve homologs for further analysis. An E-value threshold of $1e^{-40}$ is seemingly quite low, but is still appropriate for the PHO1 analysis since sequence hits with lower E-values have been checked to increase the confidence that there are no false negatives or positives.

In cases where homologs are short or have little similarity to a query sequence, known as remote homologs (Habermann 2016; Yang et al. 2021), a more diverse group of query sequences can increase the accuracy of BLAST searches. Remote homologs often still share sequence similarity due to conserved protein structure or conserved domains of the targeted gene family, meaning they will likely be missed by strict E-value cut-offs in BLAST, which matches protein sequences based only on high sequence similarity (Pearson 2013; Sinha and Lynn 2014; Habermann 2016). These remote homologs are more likely to be captured if using a diverse range of query sequences as this will represent the variation present in the targeted gene family. In recent studies on gene family identification in plants (Liu et al. 2021; Rai et al. 2021), sequences from up to 15 species have been used as BLAST query sequences. If using a single or very small set of homologs, large collections from NCBI NR, the Universal Protein Resource (UniProt; https://www.uniprot.org/) (The Uniprot Consortium 2015), or Ensembl (Cunningham et al. 2021) could also be used (Liu et al. 2021). Clustered protein sequence databases such as UniRef (Suzek et al. 2015) provide a smaller sequence subset for selecting queries or confirming identified homologs that still maintains a high level of sequence diversity for remote homolog detection. UniRef contains clustered protein sequences from UniProt based on sequence identities ranging from 50% (UniRef50) to 100% (UniRef100). These clusters can be used to select a diverse range of protein sequences as query sequences, or as a database to investigate the annotations of sequences that cluster with identified sequences. The use of more diverse query sequences selected from a wide range of species or clustered databases may increase computational requirements for homology searches, but will greatly increase the confidence that can be given to homology searches based on sequence similarity.

In many cases, a targeted gene family will be too large to use queries from multiple species. Here, queries specific to the targeted species can be generated using Position-Specific Iterative BLAST (PSI-BLAST) (Altschul et al. 1997). In PSI-BLAST searches, an initial BLAST search retrieves a list of high-scoring hits to sequences based on an E-value threshold, and these sequences are then used to produce an alignment and a position-specific scoring matrix (PSSM) (Altschul et al. 1997; Sinha et al. 2018). Residue scores in

Table 1 Example data for choosing an E-value threshold in the NCBI RefSeq predicted protein sequence database of *Nelumbo nucifera* (GenBank accession: GCF_000365185.1). Members of the *PHOSPHATE TRANSPORTER 1* (*PHT1*) and *PHOSPHATE1* (*PHO1*) gene families were identified using BLAST (v. 12.2.0) with query protein sequences from *Arabidopsis thaliana* and *Oryza sativa*

PHOSPHATE TRANSPORTER 1 (PHT1)					
Subject ID	E-value	Query coverage per subject (%)	Subject protein sequence length (aa)	Annotation	Note
XP_010262309.1	**0**	**89**	**541**	**Inorganic phosphate transporter 1-4 like**	**Likely in targeted gene family**
XP_010262310.1	**0**	**89**	**541**	**Inorganic phosphate transporter 1-4 like**	**Likely in targeted gene family**
XP_010278453.1	$6.05e^{-14}$	90	496	Organic cation/carnitine transporter 7	Not likely in targeted gene family
XP_010278454.1	$6.05e^{-14}$	90	496	Organic cation/carnitine transporter 7	Not likely in targeted gene family
PHO1 (PHOSPHATE1)					
Subject ID	E-value	Query coverage per subject (%)	Protein sequence length (aa)	Annotation	Note
XP_010258273.1	**0**	**100**	**775**	**Phosphate transporter PHO1 homolog 3-like isoform X1**	**Likely in targeted gene family**
XP_010258274.1	**0**	**89**	**743**	**Phosphate transporter PHO1 homolog 3-like isoform X2**	**Likely in targeted gene family**
XP_010246611.1	$9.74e^{-34}$	40	471	SPX and EXS domain-containing protein 1-like isoform X3	Not likely in targeted gene family
XP_010246612.1	$9.74e^{-34}$	40	471	SPX and EXS domain-containing protein 1-like isoform X3	Not likely in targeted gene family
XP_010246614.1	$9.74e^{-34}$	40	471	SPX and EXS domain-containing protein 1-like isoform X3	Not likely in targeted gene family
XP_010246615.1	$9.74e^{-34}$	40	471	SPX and EXS domain-containing protein 1-like isoform X3	Not likely in targeted gene family
XP_010246609.1	$1.29e^{-33}$	40	497	SPX and EXS domain-containing protein 1-like isoform X1	Not likely in targeted gene family
XP_010246610.1	$1.78e^{-33}$	40	492	SPX and EXS domain-containing protein 1-like isoform X2	Not likely in targeted gene family
XP_019055870.1	$6.52e^{-31}$	48	500	SPX and EXS domain-containing protein 1-like isoform X1	Not likely in targeted gene family
XP_019055871.1	$1.21e^{-29}$	39	484	SPX and EXS domain-containing protein 5-like isoform X2	Not likely in targeted gene family
XP_010249411.1	$2.36e^{-29}$	37	331	SPX and EXS domain-containing protein 3-like	Not likely in targeted gene family
XP_010246616.1	$3.71e^{-28}$	40	463	SPX and EXS domain-containing protein 3-like isoform X4	Not likely in targeted gene family
XP_010265690.1	$9.34e^{-10}$	8	90	Phosphate transporter PHO1 homolog 1-like	Targeted gene family, but short protein sequence. Needs inspection.
XP_010250106.1	$9.75e^{-05}$	10	288	Domain-containing protein 1-like	Not likely in targeted gene family

Table 1 (continued)

Query sequences from *Arabidopsis thaliana* and *Oryza sativa* were retrieved from the Araport11 protein sequence database through The Arabidopsis Information Resource (Berardini et al. 2015) and the Swiss-Prot database, UniProtKB (The Uniprot Consortium 2015). PHT1: AT5G43350, AT5G43370, AT5G43360, AT2G38940, AT2G32830, AT5G43340, AT3G54700, AT1G20860, AT1G76430, Q7XDZ7, Q01MW8, Q7X7V2, Q8H6H0, Q8H6G9, Q8H6G8, Q8H6G7, Q69T94, Q8H6H4, Q8GSD9, Q7XDZ7. PHO1: Q8S403, Q93ZF5, Q6R8G8, Q6R8G7, Q6R8G6, Q6R8G5, Q6R8G4, Q6R8G3, Q6R8G2, Q9LJW0, Q6R8G0, Q657S5, Q6K991, Q651J5

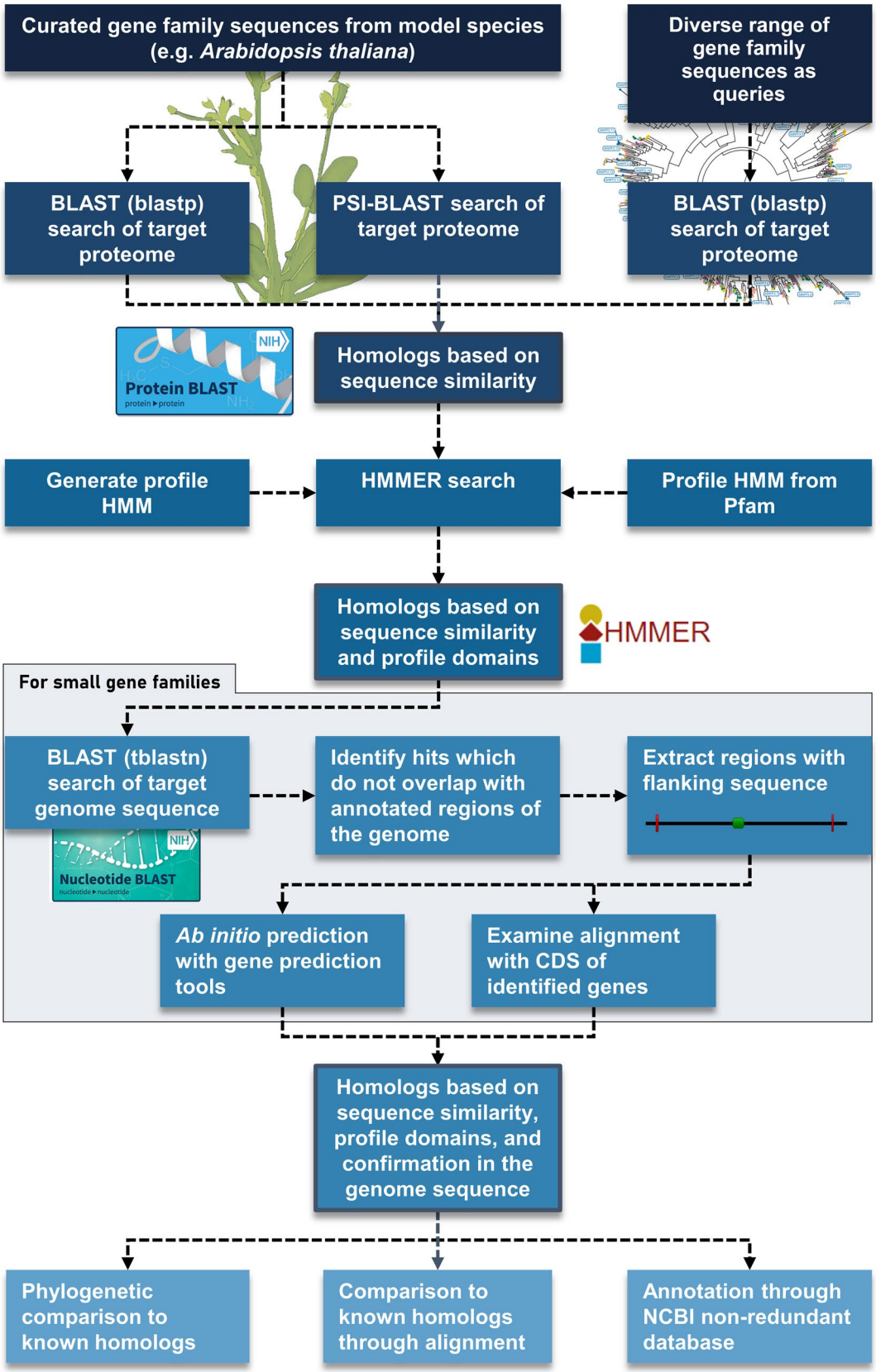
Sequences likely to be authentic homologs are highlighted in bold, and potential homologs that require further examination are underlined. A dotted line is placed between the pass and discard E-values for both homology searches and indicates the range where an appropriate E-value threshold can be set. For simplicity, the results have been trimmed to show only sequence hits with E-values surrounding the pass and discard thresholds and potential homologs requiring further investigation

the PSSM are used for iterative similarity searches to the target sequences with the scores modified after each iteration based on alignment hits that pass a threshold E-value. PSI-BLAST can be useful for identifying divergent sequences because the generated PSSM is designed specifically for the target proteome and gene family (Altschul et al. 1997; Andolfo et al. 2021). The PSI-BLAST approach can be used as an alternative to a large set of query species for greater confidence that all members of the targeted gene family have been identified.

Profile domain searches using Hidden Markov Modeler (HMMER)

A second commonly used method for homology searches is HMMER (Andolfo et al. 2021; Dong et al. 2021; Feng et al. 2021; Guérin et al. 2021; Huang et al. 2021a; Qin et al. 2021; Wu et al. 2021; Apablaza et al. 2022). HMMER is used to search for homologs based on conserved sequence domains rather than sequence similarity, which allows identification of remote homologs with low overall sequence similarity (Rost 1999; Habermann 2016; de Boissier and Habermann 2020). Sequence domains in a gene family that are conserved across many species often have functional importance and so these domains are expected to be detectable in the majority of homologs in the gene family (Richardson 1981; Ghouila et al. 2014; Lees et al. 2016). HMMER identifies conserved domains based on probabilistic models of sequences known as profile Hidden Markov Models (profile HMMs) (Eddy 1998). Like BLAST, HMMER outputs an E-value statistic to aid in determining if target sequences are authentic homologs of the targeted gene family. The E-value in this case refers to the expected number of false positive sequences being included, with the E-value increasing as dataset size increases. An appropriate range for the threshold of this E-value can be determined through the method used for BLAST (Fig. 1). HMMER is often used in combination or as an alternative to BLAST and both programs can be highly adept at detecting authentic homologs if used with appropriate thresholds.

HMMER can be combined with BLAST by using the candidate homologs of one program as the target sequences for the next program (Liu et al. 2021; Yan et al. 2021), or by using both tools simultaneously and then filtering for common sequences (Pareek et al. 2021; Wang et al. 2021a). Similar to BLAST searches, it is important to generate profiles from several phylogenetically-related species or use family-specific profile HMMs from the target species or the Pfam database (Mistry et al. 2021) through InterPro (Paysan-Lafosse et al. 2022). Using suitable profiles will reduce false negatives where homologs are evolutionarily distant or have divergent domain structures (Ghouila et al. 2014) and also reduce false positives where domains in the profile HMM are not specific to the target gene family (Sinha et al. 2018). Extensive profile HMMs for conserved domains of diverse groups of sequences can be downloaded from Pfam or new profiles can be generated from a user's own sequence sets to identify matching sequences with the same sequence domains. Homologs can be further confirmed by searching other conserved domain and signature databases such as Simple Modular Architecture Research Tool (SMART) (Schultz et al. 1998), the Conserved Domain Database (CDD) using CD-Search (Marchler-Bauer and Bryant 2004), or PROSITE (Sigrist et al. 2012). Greater control of conserved domain detection and use of user-made motifs can be achieved following a similar method to Andolfo et al. (2021) where Multiple EM for Motif Elicitation (MEME) was used to extract motifs from Pfam domains and these were then searched in target sequences using Motif Alignment and Search Tool (MAST) (Bailey et al. 2015). Several protein family databases including PANTHER, CDD, Pfam, SMART, and PROSITE, among other useful protein motif, domain, signature and site databases, can also be searched simultaneously using InterPro through InterProScan (Paysan-Lafosse et al. 2022) to provide a comprehensive analysis of potential orthology for identified protein sequences. In summary, protein family database searches available through tools such as HMMER or InterProScan are powerful tools alone or in combination with BLAST for accurate identification of homologs in gene families.


Curated gene family sequences from model species (e.g. *Arabidopsis thaliana*)
Diverse range of gene family sequences as queries
BLAST (blastp) search of target proteome
PSI-BLAST search of target proteome
BLAST (blastp) search of target proteome
Protein BLAST
protein ▶ protein
Homologs based on sequence similarity
Generate profile HMM
HMMER search
Profile HMM from Pfam
Homologs based on sequence similarity and profile domains
HMMER
For small gene families
BLAST (tblastn) search of target genome sequence
Nucleotide BLAST
nucleotide ▶ nucleotide
Identify hits which do not overlap with annotated regions of the genome
Extract regions with flanking sequence
Ab initio prediction with gene prediction tools
Examine alignment with CDS of identified genes
Homologs based on sequence similarity, profile domains, and confirmation in the genome sequence
Phylogenetic comparison to known homologs
Comparison to known homologs through alignment
Annotation through NCBI non-redundant database

◂**Fig. 2** Suggested approach for homolog identification based on sequence identity and profile domains. Input sequences are either from model species or a diverse range of protein sequences using a phylogenetically diverse species range or clustered protein sequence database. These input sequences are compared with the target predicted proteome using Basic Local Alignment Search Tool (BLAST) protein (blastp) or Position-Specific Iterative BLAST (PSI-BLAST). Hidden Markov Modeler (HMMER) is then applied with a user-generated profile or downloaded Pfam profile to identify candidate homologs based on both sequence identity and profile domains. If the target gene family is relatively small, the genome sequence can be checked to ensure that no unannotated genes have been missed. In the genome sequence check, a translated BLAST nucleotide (tblastn) search of the genome sequence will identify potential unannotated genes that can be extracted along with flanking sequences. The potential genes can be validated by using gene prediction tools or alignment with the coding DNA sequence (CDS) of previously identified genes in the target gene family. After all candidate homologs have been identified, the protein translations are further confirmed by phylogenetic comparison and alignment to other known proteins in the same gene family. Confirmation can also be performed using a BLAST search of the National Center for Biotechnology Information (NCBI) non-redundant (NR) database and examination of the top hit annotations

Confirming genes in the genome sequence

A common source of false negatives and false positives in homology searches is from low-quality predicted proteome databases derived from genome sequences. Low-quality predicted proteomes often contain a high proportion of fragmented, chimeric, or contaminant protein sequences that arise due to errors in genome sequencing or in the assembly and gene prediction stages (Li et al. 2006; El-Metwally et al. 2013; Richards 2018). False negatives will occur if fragments or chimeric sequences belonging to authentic homologs contain enough errors that they are undetectable by query sequences or profile HMMs (Nevers et al. 2020). On the other hand, false positives will occur if inauthentic homologs are represented as fragmented or chimeric sequences, of which a large proportion of the sequence is a low complexity or conserved domain region and has a strong match in the homology search. To alleviate both these issues, it is important to use high-quality genome and predicted proteome databases such as reference proteomes that are available through UniProt. Recent advances in genome sequencing such as long read sequencing and improvements in genome assembly and annotation tools have made high-quality genome and predicted proteome databases readily available for many species (Angel et al. 2018; Rice and Green 2019; Fernandez et al. 2022a). However, high-quality proteomes are still lacking for many non-model species, meaning homolog identification in these species must be accompanied by high-quality genome assembly and proteome prediction. Using these sources of high-quality proteomes will greatly reduce the chance of false negatives and false positives in reported gene family members.

In some cases, high-quality proteome databases are not available and additional verification is needed to demonstrate the presence or absence of gene family members in the genome. Even when an in-depth homology search is performed, authentic homologs may still be missed if the proteome being searched has a relatively low completeness score due to poor-quality gene prediction (Dohmen et al. 2016). Although the quality of sequencing and genome assembly methods are rapidly improving and associated costs are decreasing, most genome assemblies will still likely contain misassemblies due to base changes or larger insertions/deletions (indels), which often prevent annotation tools from correctly predicting genes (Watson and Warr 2019; Huang et al. 2021b; Fernandez et al. 2022b). However, these missed genes can be detected by examining the genome sequence of these species. Regions of the genome sequence that have sequence identity with the targeted gene family but lack a predicted gene can be extracted and ORFs predicted using gene prediction tools. The BLAST tblastn tool can be used to search all frame translations of the genome sequence for regions of sequence identity using protein query sequences. This method was used in Fernandez-Pozo et al. (2021) and in Marsh et al. (2023) to extract gene hits within over 1 kbp flanking regions and predict ORFs using Augustus (Stanke et al. 2006). Similar methods to predict ORFs were followed in Chen et al. (2021) using the BLAST-Like Alignment Tool (BLAT) (Ward and Moreno-Hagelsieb 2014) to detect potential homologs, and in Ji et al. (2021) using GeneWise (Birney et al. 2004). Other tools that can be used to predict genes in extracted nucleotide sequences include SNAP (Korf 2004) and Fgenesh (Salamov and Solovyev 2000). Alternatively, extracted regions with potential genes can be examined by alignment with the coding DNA sequence (CDS) of genes from the targeted gene family. Comparing tools for different gene predictions and genome sequence homology searches is difficult, because the results will largely differ depending on the target species and gene family, but implementing any form of the genome search and ORF prediction or alignment approach will lead to a higher confidence that all authentic homologs of a gene family have been identified.

Final confirmation of candidate homologs

Once candidate homologs are extracted in a homology search, it is important to further confirm them to ensure that they are part of the targeted gene family. The presence of false positives may be less detrimental than false negatives in this case because the absence of false negatives cannot be verified while false positives can often be detected and removed by several methods. Methods to detect false positives include aligning sequences and then removing those

sequences that are inconsistent with known homologs (Cao et al. 2021; Niu et al. 2021; Zhang et al. 2021) or building phylogenetic trees and removing sequences that occur as single-member deeply-rooted clades or highly divergent branches lacking orthologs of the targeted gene family (Li et al. 2006; Thanki et al. 2018). A BLAST search of candidate homologs against a database such as NCBI NR or UniProt is also a useful method to confirm if the sequence is part of the target gene family based on the annotations of top hits (Fernandez-Pozo et al. 2021). For validating a large number of genes, CD-HIT (Fu et al. 2012) can be used to select candidate homologs that cluster with known proteins (Rai et al. 2021). A summarised pipeline of the approaches and options for mitigating the manual pipeline issues we have discussed in this article is provided in Fig. 2. Although the approaches will likely extend the time and complexity of homolog searches using manual pipelines, their use will greatly increase the validity and thoroughness of gene family member identification allowing greater confidence in gene family analyses and reporting of non-functional or absent genes.

Summary

Manual pipelines are widely used to identify gene family members in targeted gene family studies with the goal of linking gene patterns to functional traits, but several issues often hinder the validity of reported gene families. We suggest several approaches to mitigate issues with manual pipelines and minimise the number of false negatives and false positives in analyses. The foremost issue is that false negatives and false positives often result from the use of strict thresholds such as E-values, without these threshold values being validated for the specific analysis. An appropriate E-value can be selected from a pass and discard threshold based on the annotations of matching sequences. Furthermore, inappropriate query sequences are often used in homology search tools, which can result in false negatives by excluding authentic homologs or the inclusion of false positive genes. Among the options for query sequence selection is the use of divergent query sequences from a wide range of phylogenetically diverse species or clustered protein sequence databases. In addition, combining similarity and conserved domain search tools can increase the ability to identify and validate all members of a gene family. False negatives and false positives also result when using low-quality predicted proteomes that may not include some protein sequences or contain fragmented and chimeric protein sequences. In these cases, missing genes can be confirmed by alignment or gene prediction of regions containing potential genes in the genome sequence. We believe that the issues and approaches detailed in this article are important to consider for analyses requiring precise identification of all members of a targeted gene family. Implementation of these approaches in manual homology searches will greatly increase the confidence in gene family identification and the ability for accurate down-stream analyses on relating gene presence and absence to traits in model and non-model species.

Acknowledgements The writing of this article was supported by the Australian Research Council (DP200101013) grant to Patrick M. Finnegan. Benjamin J. Nestor and Cassandria G. Tay Fernandez are supported by Research Training Program scholarships. Benjamin J. Nestor is further supported by a University Postgraduate Award from The University of Western Australia.

Author contributions BJN: conceived and wrote the manuscript. PB, CGTF, DE and PMF: reviewed and edited the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. The writing of this article was supported by the Australian Research Council (DP200101013) grant to Patrick M. Finnegan. Benjamin J. Nestor and Cassandria G. Tay Fernandez are supported by Research Training Program scholarships. Benjamin J. Nestor is further supported by a University Postgraduate Award from The University of Western Australia.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval The authors declare that no ethical approval was required for the writing of this article.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

Andolfo G, Sánchez CS, Cañizares J, Pico MB, Ercolano MR (2021) Large-scale gene gains and losses molded the NLR defense arsenal during the *Cucurbita* evolution. Planta 254:1–14. https://doi.org/10.1007/s00425-021-03717-x

Angel VDD, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Pettersson OV, Amselem J, Bouri L, Bocs S, Klopp C, Gibrat J-F, Vlasova A, Leskosek BL, Soler L, Binzer-Panchal M, Lantz H (2018) Ten steps to get started in genome assembly and annotation. F1000Research. https://doi.org/10.12688/f1000research.13598.1

Apablaza H, Solís M, Conejera D, Fonseca A, Cid J, Tarifeño-Saldivia E, Valenzuela S, Emhart V, Fernández M (2022) bHLH transcription factors undergo alternative splicing during cold acclimation in a *Eucalyptus* hybrid. Plant Mol Biol Rep 40:310–326. https://doi.org/10.1007/s11105-021-01313-7

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME suite. Nucleic Acids Res 43:W39–W49. https://doi.org/10.1093/nar/gkv416

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020) Plant pan-genomes are the new reference. Nature Plants 6:914–920. https://doi.org/10.1038/s41477-020-0733-0

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis 53:474–485. https://doi.org/10.1002/dvg.22877

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995. https://doi.org/10.1101/gr.1865504

Bocsanczy AM, Huguet-Tapia JC, Norman DJ (2017) Comparative genomics of *Ralstonia solanacearum* identifies candidate genes associated with cool virulence. Front Plant Sci 8:1565–1565. https://doi.org/10.3389/fpls.2017.01565

Buljan M, Bateman A (2009) The evolution of protein domain families. Biochem Soc Trans 37:751–755. https://doi.org/10.1042/BST0370751

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:1–9. https://doi.org/10.1186/1471-2105-10-421

Cao Y-L, Li Y-l, Fan Y-F, Li Z, Yoshida K, Wang J-Y, Ma X-K, Wang N, Mitsuda N, Kotake T, Ishimizu T, Tsai K-C, Niu S-C, Zhang D, Sun W-H, Luo Q, Zhao J-H, Yin Y, Zhang B, Wang J-Y, Qin K, An W, He J, Dai G-L, Wang Y-J, Shi Z-G, Jiao E-N, Wu P-J, Liu X, Liu B, Liao X-Y, Jiang Y-T, Yu X, Hao Y, Xu X-Y, Zou S-Q, Li M-H, Hsiao Y-Y, Lin Y-F, Liang C-K, Chen Y-Y, Wu W-L, Lu H-C, Lan S-R, Wang Z-W, Zhao X, Zhong W-Y, Yeh C-M, Tsai W-C, Van de Peer Y, Liu Z-J (2021) Wolfberry genomes and the evolution of *Lycium* (Solanaceae). Commun Biol. https://doi.org/10.1038/s42003-021-02152-8

Chen Z, Vining KJ, Qi X, Yu X, Zheng Y, Liu Z, Fang H, Li L, Bai Y, Liang C, Li W, Lange BM (2021) Genome-wide analysis of terpene synthase gene family in *Mentha longifolia* and catalytic activity analysis of a single terpene synthase. Genes 12:518. https://doi.org/10.3390/genes12040518

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean Irina M, Austine-Orimoloye O, Azov Andrey G, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin FL, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez Jose G, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh Denye N, Parker A, Parton A, Perry M, Piližota I, Prosovetskaia I, Sakthivel Manoj P, Salam Ahamed Imran A, Schmitt Bianca M, Schuilenburg H, Sheppard D, Pérez-Silva José G, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner M-M, Szpak M, Thormann A, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh Thomas A, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt Sarah E, IIsley Garth R, Loveland Jane E, Martin Fergal J, Moore B, Mudge Jonathan M, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion Stephen J, Dyer S, Harrison Peter W, Howe Kevin L, Yates Andrew D, Zerbino Daniel R, Flicek P, (2021) Ensembl 2022. Nucleic Acids Res 50:D988–D995. https://doi.org/10.1093/nar/gkab1049

de Boissier P, Habermann BH (2020) A practical guide to orthology resources. Evolutionary Biology—A Transdisciplinary Approach. Springer, Cham, pp 41–77

Delaux PM, Varala K, Edger PP, Coruzzi GM, Pires JC, Ané JM (2014) Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. PLoS Genet. https://doi.org/10.1371/journal.pgen.1004487

Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C (2016) DOGMA: Domain-based transcriptome and proteome quality assessment. Bioinformatics 32:2577–2581. https://doi.org/10.1093/bioinformatics/btw231

Dong S, Liu M, Liu Y, Chen F, Yang T, Chen L, Zhang X, Guo X, Fang D, Li L, Deng T, Yao Z, Lang X, Gong Y, Wu E, Wang Y, Shen Y, Gong X, Liu H, Zhang S (2021) The genome of Magnolia biondii Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids. Horticulture Res. https://doi.org/10.1038/s41438-021-00471-9

Eddy SR (1998) Profile hidden markov models. Bioinformatics 14:755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:1002195–1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

El-Metwally S, Hamza T, Zakaria M, Helmy M (2013) Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol 9:e1003345–e1003345. https://doi.org/10.1371/journal.pcbi.1003345

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–14. https://doi.org/10.1186/S13059-019-1832-Y

Favre P, Bapaume L, Bossolini E, Delorenzi M, Falquet L, Reinhardt D (2014) A novel bioinformatics pipeline to discover genes related to arbuscular mycorrhizal symbiosis based on their evolutionary conservation pattern among higher plants. BMC Plant Biol 14:333–333. https://doi.org/10.1186/s12870-014-0333-0

Feng S, Liu Z, Cheng J, Li Z, Tian L, Liu M, Yang T, Liu Y, Liu Y, Dai H, Yang Z, Zhang Q, Wang G, Zhang J, Jiang H, Wei A (2021) Zanthoxylum-specific whole genome duplication and recent activity of transposable elements in the highly repetitive paleotetraploid *Z. bungeanum* genome. Horticulture Res. https://doi.org/10.1038/s41438-021-00665-1

Fernandez CGT, Nestor BJ, Danilevicz MF, Gill M, Petereit J, Bayer PE, Finnegan PM, Batley J, Edwards D (2022a) Pangenomes as a resource to accelerate breeding of under-utilised crop species. Int J Mol Sci 23:2671. https://doi.org/10.3390/ijms23052671

Fernandez CGT, Nestor BJ, Danilevicz MF, Marsh JI, Petereit J, Bayer PE, Batley J, Edwards D (2022b) Expanding gene-editing potential in crop improvement with pangenomes. Int J Mol Sci. https://doi.org/10.3390/IJMS23042276

Fernandez-Pozo N, Metz T, Chandler JO, Gramzow L, Mérai Z, Maumus F, Mittelsten Scheid O, Theißen G, Schranz ME, Leubner-Metzger G, Rensing SA (2021) *Aethionema arabicum* genome annotation using PacBio full-length transcripts

provides a valuable resource for seed dormancy and Brassicaceae evolution research. Plant J 106:275–293. https://doi.org/10.1111/tpj.15161
Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–99. https://doi.org/10.2307/2412448
Forslund K, Pekkari I, Sonnhammer ELL (2011) Domain architecture conservation in orthologs. BMC Bioinform 12:326–326. https://doi.org/10.1186/1471-2105-12-326
Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565
Fujimoto MS, Suvorov A, Jensen NO, Clement MJ, Bybee SM (2016) Detecting false positive sequence homology: a machine learning approach. BMC Bioinform 17:101–101. https://doi.org/10.1186/s12859-016-0955-3
Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nat Rev Genet 14:360–366. https://doi.org/10.1038/nrg3456
Ghouila A, Florent I, Guerfali FZ, Terrapon N, Laouini D, Ben Yahia S, Gascuel O, Bréhélin L (2014) Identification of divergent protein domains by combining HMM-HMM comparisons and co-occurrence detection. PLoS ONE 9:95275–95275. https://doi.org/10.1371/journal.pone.0095275
Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, Martin M-J, Muffato M, Patricio M, Pereira C (2019) Advances and applications in the quest for orthologs. Mol Biol Evol 36:2157–2164. https://doi.org/10.1093/molbev/msz150
Guérin C, Mouzeyar S, Roche J (2021) The landscape of the genomic distribution and the expression of the F-box genes unveil genome plasticity in hexaploid wheat during grain development and in response to heat and drought stress. Int J Mol Sci 22:3111. https://doi.org/10.3390/ijms22063111
Habermann BH (2016) Oh brother, where art thou? Finding orthologs in the twilight and midnight zones of sequence similarity. In: Pontarotti P (ed) Evolutionary Biology. Springer, Cham, pp 393–419
Huang H, Liang J, Tan Q, Ou L, Li X, Zhong C, Huang H, Møller IM, Wu X, Song S (2021a) Insights into triterpene synthesis and unsaturated fatty-acid accumulation provided by chromosomal-level genome analysis of *Akebia trifoliata* subsp. australis. Horticulture Res. https://doi.org/10.1038/s41438-020-00458-y
Huang Y-T, Liu P-Y, Shih P-W (2021b) Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biol. https://doi.org/10.1186/s13059-021-02282-6
Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5. https://doi.org/10.1016/S0014-5793(98)00276-2
Ji Y-T, Xiu Z, Chen C-H, Wang Y, Yang J-X, Sui J-J, Jiang S-J, Wang P, Yue S-Y, Zhang Q-Q, Jin J-l, Wang G-S, Wei Q-Q, Wei B, Wang J, Zhang H-L, Zhang Q-Y, Liu J, Liu C-J, Jian J-B, Qu C-Q (2021) Long read sequencing of *Toona sinensis* (A. Juss) Roem: a chromosome-level reference genome for the family Meliaceae. Mol Ecol Res 21:1243–1255. https://doi.org/10.1111/1755-0998.13318
Jim K, Parmar K, Singh M, Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. Genome Res 14:109–115. https://doi.org/10.1101/gr.1586704
Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
Khan FI, Wei DQ, Gu KR, Hassan MI, Tabrez S (2016) Current updates on computer aided protein modeling and designing. Int J Biol Macromol 85:48–62. https://doi.org/10.1016/j.ijbiomac.2015.12.072
Korf I (2004) Gene finding in novel genomes. BMC Bioinform 5:59–59. https://doi.org/10.1186/1471-2105-5-59
Leelananda SP, Lindert S (2016) Computational methods in drug discovery. Beilstein J Org Chem 12:2694–2718. https://doi.org/10.3762/bjoc.12.267
Lees JG, Dawson NL, Sillitoe I, Orengo CA (2016) Functional innovation from changes in protein domains and their combinations. Curr Opin Struct Biol 38:44–52. https://doi.org/10.1016/j.sbi.2016.05.016
Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503
Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK-S, Zheng W, Dehal P, Wang J, Durbin R (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:D572–D580. https://doi.org/10.1093/nar/gkj118
Li H-L, Wu L, Dong Z, Jiang Y, Jiang S, Xing H, Li Q, Liu G, Tian S, Wu Z, Wu B, Li Z, Zhao P, Zhang Y, Tang J, Xu J, Huang K, Liu X, Zhang W, Liao Q, Ren Y, Huang X, Li Q, Li C, Wang Y, Xavier-Ravi B, Li H, Liu Y, Wan T, Liu Q, Zou Y, Jian J, Xia Q, Liu Y (2021) Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. Horticulture Res. https://doi.org/10.1038/s41438-021-00627-7
Liu H, Wang X, Wang G, Cui P, Wu S, Ai C, Hu N, Li A, He B, Shao X, Wu Z, Feng H, Chang Y, Mu D, Hou J, Dai X, Yin T, Ruan J, Cao F (2021) The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. Nature Plants 7:748–756. https://doi.org/10.1038/s41477-021-00933-x
Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh454
Marsh JI, Nestor BJ, Petereit J, Fernandez CGT, Bayer PE, Batley J, Edwards D (2023) Legume-wide comparative analysis of pod shatter locus *PDH1* reveals phaseoloid specificity, high cowpea expression and stress responsive genomic context. The Plant J Press. https://doi.org/10.1111/tpj.16209
Mergaert P, Kereszt A, Kondorosi E (2020) Gene expression in nitrogen-fixing symbiotic nodule cells in *Medicago truncatula* and other nodulating plants. Plant Cell 32:42–68. https://doi.org/10.1105/tpc.19.00494
Miao J, Feng Q, Li Y, Zhao Q, Zhou C, Lu H, Fan D, Yan J, Lu Y, Tian Q, Li W, Weng Q, Zhang L, Zhao Y, Huang T, Li L, Huang X, Sang T, Han B (2021) Chromosome-scale assembly and analysis of biomass crop *Miscanthus lutarioriparius* genome. Nature Commun. https://doi.org/10.1038/s41467-021-22738-4
Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SC, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913
Nevers Y, Defosset A, Lecompte O (2020) Orthology: Promises and challenges. In: Pontarotti P (ed) Evolutionary Biology—A Transdisciplinary Approach. Springer, Cham, pp 203–228
Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, Codo L, Cosentino S, Marcet-Houben M, Vlasova A, Poidevin L, Kress A, Hickman M, Persson E, Piližota I, Guijarro-Clarke C, OpenEBench team, Quest for Orthologs Consortium, Iwasaki W, Lecompte O, Sonnhammer E, Roos DS, Gabaldón T, Thybert D, Thomas PD, Hu Y, Emms DM, Bruford E, Capella-Gutierrez S, Martin MJ, Dessimoz C, Altenhoff A (2022) The quest for

orthologs orthology benchmark service in 2022. Nucleic Acids Res 50:W623–W632. https://doi.org/10.1093/nar/gkac330
Niu Z, Zhu F, Fan Y, Li C, Zhang B, Zhu S, Hou Z, Wang M, Yang J, Xue Q, Liu W, Ding X (2021) The chromosome-level reference genome assembly for *Dendrobium officinale* and its utility of functional genomics research and molecular breeding study. Acta Pharmaceutica Sinica B 11:2080–2092. https://doi.org/10.1016/j.apsb.2021.01.019
Pareek A, Mishra D, Rathi D, Verma JK, Chakraborty S, Chakraborty N (2021) The small heat shock proteins, chaperonin 10, in plants: an evolutionary view and emerging functional diversity. Environ Exp Bot 182:104323. https://doi.org/10.1016/j.envexpbot.2020.104323
Patiranage DS, Asare E, Maldonado-Taipe N, Rey E, Emrani N, Tester M, Jung C (2021) Haplotype variations of major flowering time genes in quinoa unveil their role in the adaptation to different environmental conditions. Plant, Cell Environ 44:2565–2579. https://doi.org/10.1111/pce.14071
Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar Gustavo A, Bileschi Maxwell L, Bork P, Bridge A, Colwell L, Gough J, Haft Daniel H, Letunić I, Marchler-Bauer A, Mi H, Natale Darren A, Orengo Christine A, Pandurangan Arun P, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu Cathy H, Bateman A (2022) InterPro in 2022. Nucleic Acids Res 51:D418–D427. https://doi.org/10.1093/nar/gkac993
Pearson WR (2013) An introduction to sequence similarity ("homology") searching. Curr Protocols Bioinform 42:3.1.1-3.1.8. https://doi.org/10.1002/0471250953.bi0301s42
Pei L, Wang B, Ye J, Hu X, Fu L, Li K, Ni Z, Wang Z, Wei Y, Shi L, Zhang Y, Bai X, Jiang M, Wang S, Ma C, Li S, Liu K, Li W, Cong B (2021) Genome and transcriptome of *Papaver somniferum* Chinese landrace CHM indicates that massive genome expansion contributes to high benzylisoquinoline alkaloid biosynthesis. Horticulture Res. https://doi.org/10.1038/s41438-020-00435-5
Qin N, Gao Y, Cheng X, Yang Y, Wu J, Wang J, Li S, Xing G (2021) Genome-wide identification of CLE gene family and their potential roles in bolting and fruit bearing in cucumber (Cucumis sativus L.). BMC Plant Biol. https://doi.org/10.1186/s12870-021-02900-2
Radhakrishnan GV, Keller J, Rich MK, Vernié T, Mbadinga Mbadinga DL, Vigneron N, Cottret L, Clemente HS, Libourel C, Cheema J, Linde A-M, Eklund DM, Cheng S, Wong GKS, Lagercrantz U, Li F-W, Oldroyd GED, Delaux P-M (2020) An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. Nat Plants 6:280–289. https://doi.org/10.1038/s41477-020-0613-7
Rai A, Hirakawa H, Nakabayashi R, Kikuchi S, Hayashi K, Rai M, Tsugawa H, Nakaya T, Mori T, Nagasaki H, Fukushi R, Kusuya Y, Takahashi H, Uchiyama H, Toyoda A, Hikosaka S, Goto E, Saito K, Yamazaki M (2021) Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis. Nat Commun. https://doi.org/10.1038/s41467-020-20508-2
Rice ES, Green RE (2019) New approaches for genome assembly and scaffolding. Annual Rev Animal Biosci 7:17–40. https://doi.org/10.1146/annurev-animal-020518-115344
Richards S (2018) Full disclosure: genome assembly is still hard. PLoS Biol 16:1–5. https://doi.org/10.1371/journal.pbio.2005894
Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339. https://doi.org/10.1016/S0065-3233(08)60520-3
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. https://doi.org/10.1093/sysbio/sys029
Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng Des Sel 12:85–94. https://doi.org/10.1093/protein/12.2.85
Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res 10:516–522. https://doi.org/10.1101/gr.10.4.516
Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau Donald C, Farrell Catherine M, Feldgarden M, Fine AM, Funk K, Hatcher E, Kannan S, Kelly C, Kim S, Klimke W, Landrum Melissa J, Lathrop S, Lu Z, Madden Thomas L, Malheiro A, Marchler-Bauer A, Murphy Terence D, Phan L, Pujar S, Rangwala Sanjida H, Schneider Valerie A, Tse T, Wang J, Ye J, Trawick Barton W, Pruitt Kim D, Sherry Stephen T (2022) Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res 51:D29–D38. https://doi.org/10.1093/nar/gkac1032
Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci 95:5857–5864. https://doi.org/10.1073/pnas.95.11.5857
Setubal JC, Stadler PF (2018) Gene phylogenies and orthologous groups. Comparative Genomics: Methods and Protocols. Humana Press Inc., New York, pp 1–28
Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2012) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347. https://doi.org/10.1093/nar/gks1067
Sinha S, Lynn AM (2014) HMM-ModE: Implementation, benchmarking and validation with HMMER3. BMC Res Notes 7:1–11. https://doi.org/10.1186/1756-0500-7-483
Sinha S, Eisenhaber B, Lynn AM (2018) Predicting protein function using homology-based methods. Bioinformatics: Sequences, Structures, Phylogeny. Springer
Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033
Stamboulian M, Guerrero RF, Hahn MW, Radivojac P (2020) The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. Bioinformatics 36:i219–i226. https://doi.org/10.1093/bioinformatics/btaa468
Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: *Ab initio* prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439. https://doi.org/10.1093/nar/gkl200
Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J (2019) HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473. https://doi.org/10.1186/s12859-019-3019-7
Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, The Uniprot Consortium (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932. https://doi.org/10.1093/bioinformatics/btu739
Thanki AS, Soranzo N, Haerty W, Davey RP (2018) GeneSeqToFamily: a Galaxy workflow to find gene families based on the Ensembl Compara GeneTrees pipeline. GigaScience 7:giy005. https://doi.org/10.1093/gigascience/giy005
The Uniprot Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212. https://doi.org/10.1093/nar/gku989
Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141. https://doi.org/10.1101/gr.772403

Wang L, Lei T, Han G, Yue J, Zhang X, Yang Q, Ruan H, Gu C, Zhang Q, Qian T, Zhang N, Qian W, Wang Q, Pang X, Shu Y, Gao L, Wang Y (2021a) The chromosome-scale reference genome of *Rubus chingii* Hu provides insight into the biosynthetic pathway of hydrolyzable tannins. Plant J 107:1466–1477. https://doi.org/10.1111/tpj.15394

Wang X, Cai X, Xu C, Wang Q (2021b) Identification and characterization of the *NPF*, *NRT2* and *NRT3* in spinach. Plant Physiol Biochem 158:297–307. https://doi.org/10.1016/j.plaphy.2020.11.017

Ward N, Moreno-Hagelsieb G (2014) Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS ONE 9:e101850. https://doi.org/10.1371/journal.pone.0101850

Watson M, Warr A (2019) Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol 37:124–126. https://doi.org/10.1038/s41587-018-0004-z

Wheeler D, Bhagwat M (2007) BLAST QuickStart. In: Bergman NH (ed) Comparative Genomics. Humana Press, Totowa

Wu YC, Rasmussen MD, Kellis M (2012) Evolution at the subgene level: domain rearrangements in the *Drosophila* phylogeny. Mol Biol Evol 29:689–705. https://doi.org/10.1093/molbev/msr222

Wu D, He G, Tian W, Saleem M, Li D, Huang Y, Meng L, He Y, Liu Y, He T (2021) OPT gene family analysis of potato (*Solanum tuberosum*) responding to heavy metal stress: comparative omics and co-expression networks revealed the underlying core templates and specific response patterns. Int J Biol Macromol 188:892–903. https://doi.org/10.1016/j.ijbiomac.2021.07.183

Xu P, Wang Y, Sun F, Wu R, Du H, Wang Y, Jiang L, Wu X, Wu X, Yang L, Xing N, Hu Y, Wang B, Huang Y, Tao Y, Gao Q, Liang C, Li Y, Lu Z, Li G (2021) Long-read genome assembly and genetic architecture of fruit shape in the bottle gourd. Plant J 107:956–968. https://doi.org/10.1111/tpj.15358

Yan L, Zhang J, Chen H, Luo H (2021) Genome-wide analysis of ATP-binding cassette transporter provides insight to genes related to bioactive metabolite transportation in *Salvia miltiorrhiza*. BMC Genomics. https://doi.org/10.1186/s12864-021-07623-0

Yang F-X, Gao J, Wei Y-L, Ren R, Zhang G-Q, Lu C-Q, Jin J-P, Ai Y, Wang Y-Q, Chen L-J, Ahmad S, Zhang D-Y, Sun W-H, Tsai W-C, Liu Z-J, Zhu G-F (2021) The genome of *Cymbidium sinense* revealed the evolution of orchid traits. Plant Biotechnol J 19:2501–2516. https://doi.org/10.1111/pbi.13676

Zhang Y, Zhang G-Q, Zhang D, Liu X-D, Xu X-Y, Sun W-H, Yu X, Zhu X, Wang Z-W, Zhao X, Zhong W-Y, Chen H, Yin W-L, Huang T, Niu S-C, Liu Z-J (2021) Chromosome-scale assembly of the *Dendrobium chrysotoxum* genome enhances the understanding of orchid evolution. Horticulture Res. https://doi.org/10.1038/s41438-021-00621-z

Zhao L, Chen P, Liu P, Song Y, Zhang D (2021) Genetic effects and expression patterns of the Nitrate Transporter (*NRT*) gene family in *Populus tomentos*. Front Plant Sci 12:661635. https://doi.org/10.3389/fpls.2021.661635

Zhong M-C, Jiang X-D, Cui W-H, Hu J-Y (2021) Expansion and expression diversity of *FAR1/FRS-like* genes provides insights into flowering time regulation in roses. Plant Divers 43:173–179. https://doi.org/10.1016/j.pld.2020.11.002