Genome Biology

METHOD

Open Access

SynGAP: a synteny-based toolkit for gene structure annotation polishing

Fengqi Wu[1,2,3], Yingxiao Mai[1,2,3], Chengjie Chen[1,2,3*] and Rui Xia[1,2,3*]

*Correspondence: ccj0410@gmail.com; rxia@scau.edu.cn

[1] State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Horticulture, South China Agricultural University, Guangzhou 510640, Guangdong, China
[2] Guangdong Laboratory for Lingnan Modern Agriculture, South China Agricultural University, Guangzhou 510640, Guangdong, China
[3] Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China, Ministry of Agriculture and Rural Affairs, South China Agricultural University, Guangzhou 510640, Guangdong, China

Abstract

Genome sequencing has become a routine task for biologists, but the challenge of gene structure annotation persists, impeding accurate genomic and genetic research. Here, we present a bioinformatics toolkit, SynGAP (Synteny-based Gene structure Annotation Polisher), which uses gene synteny information to accomplish precise and automated polishing of gene structure annotation of genomes. SynGAP offers exceptional capabilities in the improvement of gene structure annotation quality and the profiling of integrative gene synteny between species. Furthermore, an expression variation index is designed for comparative transcriptomics analysis to explore candidate genes responsible for the development of distinct traits observed in phylogenetically related species.

Keywords: SynGAP, Gene structure annotation, Synteny, Comparative transcriptomics

Background

Advances in sequencing and computational technologies, coupled with decreasing costs, have made it possible for researchers to routinely sequence genomes and obtain high-quality assemblies of interest. However, genome annotation, which usually involves three major steps: masking of repetitive DNA sequences, gene structure annotation (GSA), and gene functional annotation, remains a challenging task for biologists, with gene structure annotation the most important and difficult step. Gene structure annotation refers to the determination of gene location in genomic sequences and the accurate defining of genic exons and introns. Given gene transcription is spatial-temporally dependent, GSA could be very complicated. A single gene could be transcribed into multiple transcripts due to alternative splicing and alternative start and termination sites. Accurate GSA is indispensable for genomic and genetic research as substandard GSA can greatly impede downstream research, leading to erroneous bioinformatics analysis and misdirected functional genomics studies [1–3]. Nowadays, a variety of pipelines or workflows have been developed for gene structure annotations, usually integrated with ab initio or homology-based prediction and transcriptome-assisted annotation.

Prominent examples of such pipelines include AUGUSTUS [4], miniprot [5], MAKER [6], and so on. None of them is superior, resulting in substantial variability in GSA quality among different genome assemblies. This fails to keep pace with the significant increase of genome assembly quality, which is attributed to the utilization of the 3rd-generation long-read sequencing technologies. Although manual correction of GSA using toolkits such as Apollo [7] and IGV-GSAman [8] seems to be an effective approach to improve GSA, it relies on comprehensive transcriptome or proteome data and is time-consuming especially when working on pan-genome projects.

During evolution, order of genes on chromosomes is maintained in related species that descend from a common ancestral species. This preserved co-localization of genes on chromosomes of different species, known as gene synteny [9, 10], offers insights into both the inter-species evolutionary relationship of chromosomes and the intra-species genomic changes, such as the quantity and location of genome shuffling events. Generally, the more closely related two species are, the higher the degree of their gene synteny is. Therefore, gene synteny is often used for comparative genomics and transcriptomics analysis to identify homologous genomic blocks and map orthologous genes between species [11].

Since the synteny relationship reflects the conservative arrangement of orthologous genes, it is exceptionally suitable for the comparative analysis of genes in aligned genomic regions across different species. Orthologous gene pairs or unpaired genes within syntenic genomic regions could be easily identified via synteny analysis using tools like MCScanX [11], JCVI [10], and WGDI [12]. These unpaired genes may result from genomic sequence changes, such as gene deletions or insertions, or simply be the consequences of incomplete or inaccurate gene structure annotations. If the latter is the case, synteny analysis could be used for the mutual correction and complement of gene structure annotation in closely related species. Based on this scenario, we developed SynGAP (Synteny-based Gene Structure Annotation Polisher), a toolkit for filling in missing gene structure annotations and correcting inaccurate gene models in related species based on gene synteny. We have also demonstrated the application of SynGAP in comparative transcriptomics analysis by improving the synteny-based identification of orthologous gene pairs.

Results and discussion

Strategy design of SynGAP

Currently, gene structural annotation of genomes is still one of the major challenges for genome sequencing projects, which is evident in a quick assessment of gene structure annotations (GSA) for hundreds of published genomes using BUSCO (Benchmarking Universal Single-Copy Orthologs) measurements [13] (Additional file 1: Fig. S1). It is shown that at least one-fourth of the genomes have less than 90% GSA completeness in sharp contrast to their more than 90% completeness of genome assembly, especially for embryophytes (Additional file 1: Fig. S1a, c) in which a large number of genomes lack well-defined gene structure annotations.

Genomes of phylogenetically related species usually maintain large blocks of conserved genomic regions, and especially the protein-coding genes in these blocks preserve great syntenic relationships, as exemplified in Fig. 1a (Additional file 1: Fig. S2).

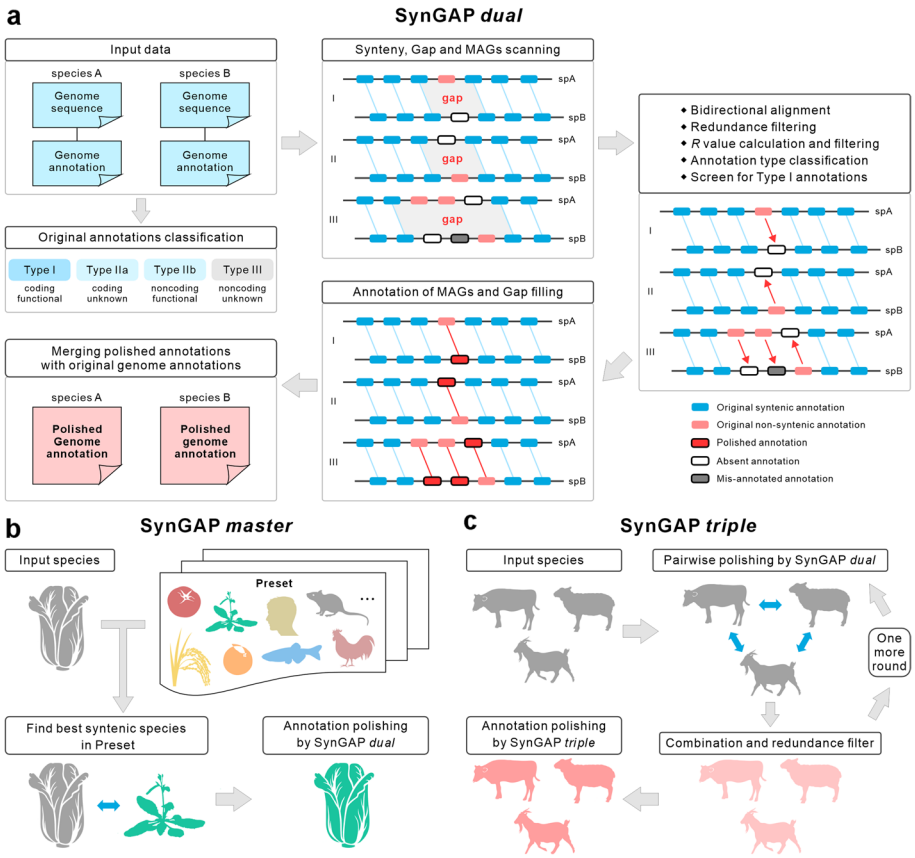

Fig. 1 Strategy design and workflow of SynGAP for GSA polishing. **a** Workflow of SynGAP *dual*. The gaps resulting from mis-annotated or absent gene models are represented by gray diamonds. Synteny gene pairs are depicted by solid blue lines, with the gene possessing synteny indicated by blue boxes. The genes without syntenic counterparts are represented by light red boxes, and potentially absent, or mis-annotated genes are depicted by white boxes and dark boxes, respectively. The red boxes signify the polished annotations, while solid red lines represent synteny gene pairs recovered by SynGAP. **b** Workflow of SynGAP *master*. **c** Workflow of SynGAP *triple*

But their syntenic counterpart of some genes within these blocks are lost because of the failure to detect the correct gene models in genome annotation (Fig. 1a). In other words, the loss of synteny of certain genes is not caused by the change of genomic sequences, but the mis-annotated or absent gene models (MAGs) (Fig. 1a). In these occasions, these MAGs can be recovered by performing a comparison of genomic blocks between close species. SynGAP (Synteny-based Gene Structure Annotation Polisher) was developed, based on this rationale, for the mutual polishing of gene structure annotations for phylogenetically related species. Based on the gene synteny, the potential omissions and errors in the original GSAs of related species can be found and polished. Three independent modules, SynGAP *dual* (Fig. 1a), SynGAP *master* (Fig. 1b), and SynGAP *triple* (Fig. 1c), were developed in the current version of SynGAP.

SynGAP *dual* is a module designed for the mutual gene structure annotation correction of two species. With the genome sequences and genome annotations of two species, synteny blocks are firstly identified using the MCscan pipeline in the JCVI toolkit [10]. The gaps between synteny gene pairs are found and extracted. A gap is

defined as the genomic region between two adjacent syntenic gene pairs within a syntenic block of species A, where annotation of additional genes is found in the corresponding region of the reference species (species B) (Fig. 1a). These gaps imply the presence of MAGs, which is subjected to be annotated (Fig. 1a). Next, for each gap identified, genome sequences within the gap and the corresponding protein sequences of existing gene annotations are extracted to perform bidirectional alignment-based homologous gene prediction using genBlastG [14] or miniprot [5] under default parameters, and preliminary polished annotations are obtained by the integration of homologous prediction results for all gaps and the filter-out of redundant annotations. After that, *R* value, a reliability index that represents the similarity between the polished annotation and its homologous reference annotation, is calculated. The closer the *R* value approaches to 1, the higher the similarity between the polished gene model and its homologous reference. A newly annotated gene with over 70% protein sequence similarity to its homologous reference will gain an *R* value above 0.5, which is a reliable threshold for closely related species, but it is likely too stringent for species of significant evolutionary divergence, such as those from different families. Therefore, a dynamic cutoff mechanism of *R* value was adopted to maximize the capacity of SynGAP to get high-confidence annotations. For any given syntenic block, there are tens or hundreds of syntenic gene pairs identified by synteny analysis before SynGAP polishing. All of these gene pairs can be considered as a set of positive annotations and used to obtain a set of positive *R* values. The lower quantile value of the *R* value set (R_{Q1}) is chosen as the *R* cutoff for the screening of new annotations retrieved by SynGAP, as long as the R_{Q1} is less than 0.5. In other words, when R_{Q1} is smaller than 0.5, R_{cutoff} is set to the R_{Q1}, otherwise, to 0.5.

Moreover, considering some GSAs of reference genomes are not necessarily of high-confidence, we classified the GSAs of a reference genome into four types, type I–IV, according to their protein-coding potential and functional annotation against Swiss-Prot. To ensure the high-confidence of reannotation, only the results generated from original annotations of type I genes (with confident protein-coding potential [15] and functional annotations against Swiss-Prot [16]) were considered as high-confidence polishings. Finally, the high-confidence polished annotations were combined with the original ones to obtain an improved version of GSA of the target species.

In some cases, it may be difficult to quickly find a good reference species with high-quality GSA; therefore, we constructed the module SynGAP *master* for a rapid annotation polishing with several preset reference genomes with high-quality GSAs, including both plant (*Arabidopsis thaliana*, *Oryza sativa*, *Solanum lycopersicum*) and animal species (*Homo sapiens*, *Mus musculus*, *Danio rerio*), and a few other species (Fig. 1b). SynGAP *master* will infer the species that have the largest number of syntenic gene pairs with the input species from the preset genomes, and then use it as the reference to carry out SynGAP *dual* to polish the input species. The *master* mode is an out-of-box solution for quick gene annotation polishing without prior setting of a reference species. Another module, SynGAP *triple*, was designed for three species in combination with two rounds of polishing using SynGAP *dual* (Fig. 1c). Compared to the *dual* mode, SynGAP *triple* could achieve more robust and thorough annotation polishing for each of the three species analyzed (see below).

SynGAP effectively improves gene structure annotations

To verify the effectiveness of synteny-based genome structure annotation polishing, we tested SynGAP for a large number of species including both plants and animals (Fig. 2, Additional files 2–6: Tables S1–5). For SynGAP *dual*, the first test was set to use the gene structure annotation of *A. thaliana*, a model plant most popularly used in the plant community, comparing to a modified version of *A. thaliana* GSA with 100 gene annotations randomly deleted and three other species with different evolutional distance

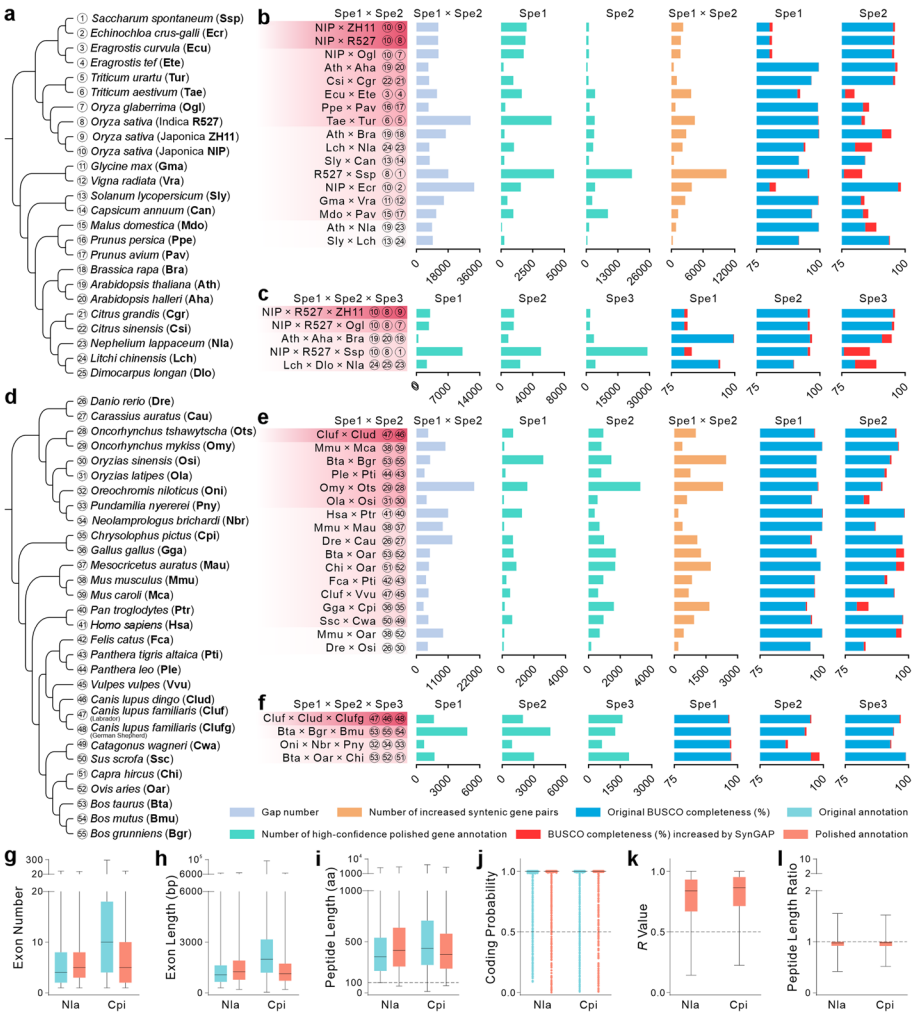

Fig. 2 Evaluation of SynGAP performance in GSA polishing. **a**, **d** Phylogenetic trees of species used for SynGAP performance evaluation (**a** plants; **d** animals). Species are assigned numerical identifications for ease of reference. **b**, **e** Number of gene annotations and synteny gene pairs, and BUSCO completeness of GSAs polished by SynGAP *dual* (**b** plants; **e** animals) in different species combinations. The red background varying in shades behind species pairs indicate the evolutionary distance of species compared, with darker colors indicating short distance. **c**, **f** Number of gene annotations and BUSCO completeness of GSAs polished by SynGAP *triple* (**c** plants; **f** animals) in different species combinations. **g**–**l** Comparison of sequence characteristics between the polished GSAs and the original whole-genome GSAs (**g** exon number per annotation; **h** exon length; **i** peptide length; **j** protein-coding potential of polished annotations predicted by CPC2 [15]; **k** *R* value of polished annotations; **l** peptide length ratio between the polished annotations and their homologous reference annotations). Note: The polishing performance is evaluated through the utilization of a benchmark dataset comprised predominantly of genomes sourced from the public Ensembl database. The polishing results exclusively reflect improvements made to the benchmark genomes

to *A. thaliana*. We found that randomly deleted gene annotations were almost completely retrieved by SynGAP *dual* (95% at average in Ath-to-Ath polishing) with low false positive rate (0.1% at average in Ath-to-Ath polishing). The retrieved gene models kept high CDS structure concordance to the original ones (> 85% at average in Ath-to-Ath polishing) and had a high rate of gene structure accuracy (> 93% for GT-AG intron splice, > 89% for start codon, and nearly 100% for stop codon). Although its annotation performance decreased along with the increase of evolutionary distance, SynGAP can still retrieve considerable annotations with high quality (Additional file 1: Fig. S3a–f, Additional file 2: Table S1, Additional file 3: Table S2). The decrease in gene synteny and sequence similarity across tested species, resulting from the increase in evolutionary distance, may impact or restrict the efficacy of SynGAP. And variations in the quality of their genome assemblies and original GSA annotations among the species used for comparative polishing may also contribute as confounding factors. Collectively, all these tests demonstrated the applicability of SynGAP in GSA polishing. With the application of the dynamic *R* threshold filtering mechanism, SynGAP can effectively remove the low-quality polished annotations while maintaining strong applicability for MAG polishing with most newly annotated gene models belonging to the type I annotations (Additional file 1: Figs. S3h, S4). With the exception of the retrieved annotations, few gene models were found in both Ath0 and Ath1 that were suspected to be non-coding genes or pseudogenes (Additional file 3: Table S2, Additional file 1: Fig. S3i).

We then tested SynGAP for different combinations of plant genomes. All genomes obtained an increasing number of high-confidence gene annotations after being polished by SynGAP *dual*. Not only the number of synteny gene pairs increased, but also the completeness of GSA evaluated by BUSCO assessment was greatly improved for some species (Fig. 2a and b, Additional file 1: Fig. S5, Additional files 3–6: Tables S2–5). These improvements of GSA could be further enhanced by using the SynGAP *triple* module (Fig. 2c). The polishing effect was more dramatic for species with poor-quality GSA, for instance, *Nephelium lappaceum*, whose original GSA was of 80.3% of BUSCO completeness, was improved with an increase of 2864 annotations and 6.6% BUSCO completeness in comparison to *Litchi chinensis*. GSA of *N. lappaceum* was even further improved with an increase of 4364 annotations and the BUSCO completeness was increased by 8.4% in comparison to *L. chinensis* and *Dimocarpus longan* using the *triple* mode (Fig. 2c). In some cases, the improvement of BUSCO completeness was not associated with the increase of the number of high-confidence annotations or syntenic gene pairs, which might be because there are few universal single-copy orthologs that were mis-annotated or absent in those species. For instance, *Echinochloa crus-galli*, whose original GSA was of 97.2% of BUSCO completeness, was improved with an increase of 3678 annotations in comparison to *O. sativa* (Japonica group c.v. NIP), while the BUSCO completeness only increased by 1.2% (Additional file 3: Table S2).

Similar tests were conducted for animal genomes. Generally, similar polishing effect could be achieved for all the different combinations (Fig. 2d–f). For example, the genome annotation of *Chrysolophus pictus* was improved with an increase of 1607 annotations and 4.6% BUSCO completeness in reference to *Gallus gallus* (Fig. 2e). And the number of synteny gene pairs between *C. pictus* and *G. gallus* was increased by 1670. Therefore, SynGAP can be applied to both plant and animal genomes for GSA polishing.

To evaluate the quality of the annotations retrieved by SynGAP, a comparison of sequence characteristics between new annotations retried by SynGAP and the original whole-genome annotations was performed. Results from *A. thaliana*, *Oryza sativa* (Japonica group c.v. Nipponbare), *S. lycopersicum*, *N. lappaceum*, *Bos taurus*, and *C. pictus* (polished by *Arabidopsis halleri*, *O. sativa* (Indica group c.v. R527), *Capsicum annuum*, *L. chinensis*, *Ovis aries*, and *G. gallus*, respectively) were used for this comparison. Exon number per annotation, exon length, and protein length of the polished annotations were overall less or shorter than that of the original whole-genome GSAs (Fig. 2g–i, Additional file 1: Fig. S6a–c), indicating most genes polished by SynGAP are short in length and of less exons. This observation aligns well with the general fact that genes of shorter length are more susceptible to being incorrectly or incompletely annotated. The majority of SynGAP-retrieved genes possessed a good level of protein-coding potential, with coding probability over 0.5, and could encode protein sequences longer than 100 aa (Fig. 2i–j, Additional file 1: Fig. S6c–d). Meanwhile, over 50% of the polished annotations keep high sequence similarity with their homologous reference annotations, with an R value over 0.5 and a peptide length ratio near 1 (Fig. 2k–l, Additional file 1: Fig. S6e–f). Certain annotations display sequence shortening in comparison to their corresponding homologous reference annotations. This may be attributed to the premature termination of the gene caused by the revised annotation, or it is plausible that the gene has been turned into a pseudogene, thereby retaining only a portion of gene sequence compared to the reference. All these results indicate that polished annotations obtained by SynGAP are principally credible.

Genes obtained by SynGAP are of critical biological function

To assess the potential biological function of new annotations obtained by SynGAP, we conducted GO enrichment analysis for these newly annotated genes from four species and found that they were enriched in all kinds of different biological processes like photosynthesis, cell differentiation, secondary metabolism, reproduction, and immune response (Additional file 1: Fig. S7). This suggests that the MAGs could be functionally important genes which may be involved in a wide range of biological processes.

To further validate the accuracy of annotation polished by SynGAP, transcriptome data were used to verify the presence of these newly annotated genes. Although the genome structure annotations of broadly studied species such as *A. thaliana*, *O. sativa*, and *S. lycopersicum* are supposed to be relatively complete, MAGs could still be found by SynGAP. We analyzed the GSA of *A. thaliana* in comparison to *A. halleri* and obtained 211 polished annotations. Among them, we discovered *AtBZO2*, the homolog of *AtBZO1* (AT1G65880), was omitted from the original GSA in *A. thaliana* (Fig. 3a). This gene encodes benzoate-CoA ligase, which is involved in the biosynthesis of benzoyloxyglucosinolate in *A. thaliana* seeds [17]. Similarly, for the GSA of *O. sativa* (c.v. Nipponbare) and *S. lycopersicum*, after being polished against *O. sativa* (c.v. R527) and *C. annuum*, new annotations of 1938 and 272 genes were obtained for *O. sativa* and *S. lycopersicum*, respectively. For instance, *NIP_BRR2*, the homolog of *AtBRR2* (AT1G20960), was found to be mistakenly annotated as two genes (Fig. 3b). This gene encodes a DEAD/DExH box ATP-dependent RNA helicase, which is required for proper splicing of *FLC* and its mutation leads to the decrease of *FLC* expression level and the early flowering

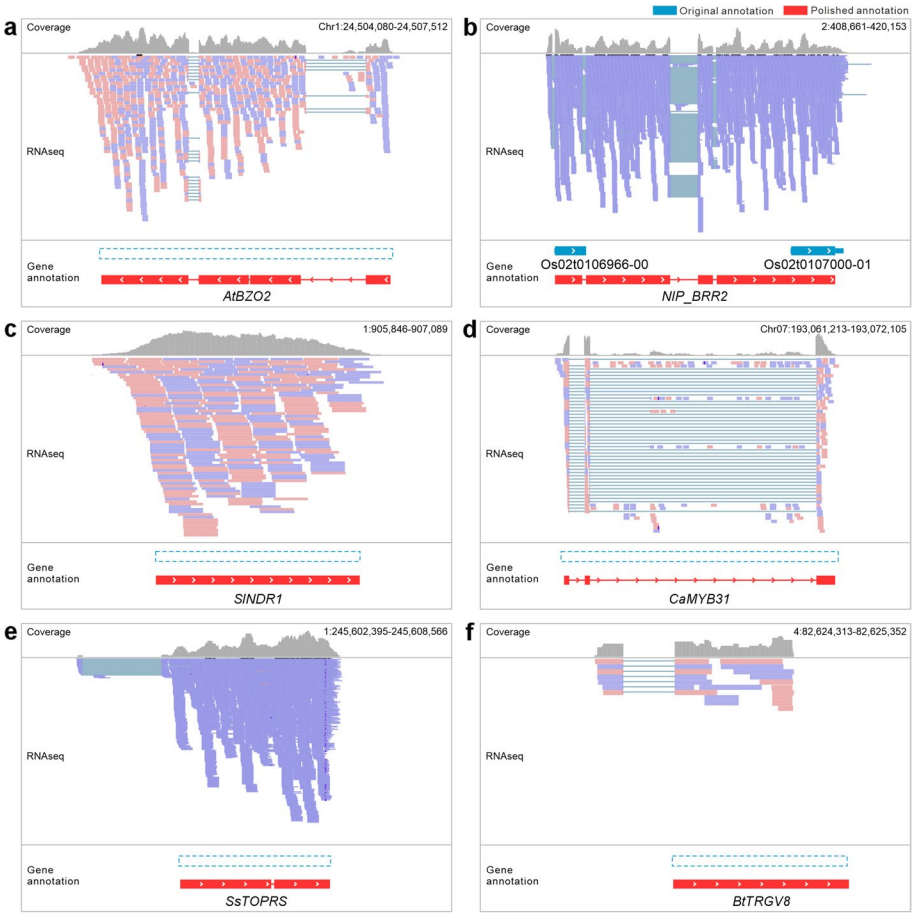

Fig. 3 Demonstrating examples of gene models polished by SynGAP in various species. **a** *AtBZO2* in *A. thaliana* in reference to *A. halleri*. **b** *NIP_BBR2* in *O. sativa* (c.v. Nipponbare) in reference to *O. sativa* (c.v. R527). **c** *SlNDR1* in *S. lycopersicum* in reference to *C. annuum*. **d** *CaMYB31* in *C. annuum* in reference to *S. lycopersicum*. **e** *SsTOPORS* in *S. scrofa* in reference to *C. wagneri*. And *SsTOPORS* is likely a processed pseudogene copy that is expressed. **f** *BtTRGV8* in *B. taurus* in reference to *O. aries*. Red gene models are these polished by SynGAP. Blue ones and blue boxes of dotted lines denote the original incorrect gene models and absent genes, respectively

phenotype [18]. *SlNDR1*, the homolog of *AtNDR1* (AT3G20600) which may be required for non-race specific resistance to bacterial and fungal pathogens, mediating systemic acquired resistance (SAR) response [19], was absent in *S. lycopersicum* (Fig. 3c). Meanwhile, the annotation of *CaMYB31*, the key regulatory transcription factor for capsaicin synthesis [1], was missing in the original annotation, but retrieved in mutual polishing (Fig. 3d).

Similar cases could also be found in animals. After being polished by *Catagonus wagneri*, 633 new annotations were obtained for *Sus scrofa*. For example, *SsTOPORS*, the homolog of *MmTOPORS* which functions as an E3 ubiquitin-protein ligase and a probable tumor suppressor involved in cell growth, cell proliferation, and apoptosis [20], was found to be missing in the original annotation (Fig. 3e). And *BtTRGV8*, the homolog of *HsTRGV8*, was retrieved in the GSA polishing of *B. taurus* against *O. aries* (Fig. 3f). It encodes V region of the variable domain of T cell receptor (TR) gamma chain that participates in the antigen recognition. It recognizes a variety of self and foreign

non-peptide antigens and takes innate-like immune responses involved in pathogen clearance and tissue repair [21, 22]. All the above results demonstrated the effectiveness of SynGAP in the retrieval of MAGs of critical biological function via mutual GSA comparative analysis.

Employment of SynGAP in comparative transcriptome analysis across species

With the increasing popularity of comparative genomic analysis, comparative transcriptome analysis is becoming more and more popular. Comparative transcriptome analysis is the comparison of expression patterns between homologous genes in different species. One of the main challenges of this approach is to establish the best one-to-one relationship of homologous genes between two species. SynGAP incorporates another function module, *genepair*, to generate high-confidence cross-species homologous gene pairs by combining the improved synteny (from SynGAP *dual* or *triple*) and best two-way BLAST (Fig. 4a). Paired genes found by SynGAP accounted for the majority of the total number of genes in each species, reaching over 65% for gene pairing within species and between species from the same genus but decreasing with the increase of evolutionary distance (Fig. 4b, Additional file 1: Fig. S8a–b), which is similar to the performance of other strategies such as OrthoFinder [23] (Additional file 1: Fig. S8c–f). For *O. sativa* (c.v. Nipponbare), in reference to *O. sativa* (c.v. R527), unpaired genes tended to have shorter exon, less intron numbers, weaker protein-coding potential, and higher rate of non-ATG start codons (Additional file 1: Fig. S9), suggesting that these unpaired genes are more likely to be noncoding genes or short peptides with unknown functions [24]. Therefore, the high-confidence gene pairs obtained by SynGAP *genepair* are more suitable to be used as the reference genes for comparative transcriptome analysis.

Another challenge of comparative transcriptome analysis is the alignment of developmental stages between two species, as it is always difficult to collect samples corresponding to each other in two species although the whole developmental process is largely conserved. To solve this issue, we adopted another parameter, expression variation index (*EVI*), which is calculated based on gene expression level, expression change, and the difference of expression trend in a time-course transcriptome data (Fig. 4c). In a conserved biological process in closely related species, the expression pattern of homologous genes, reflected by the difference in expression levels and trend of expression changes, should be largely similar. This is the basis for the design of *EVI*, unlike doing a stage-to-stage comparison as most of the studies do [25–27]. A functional module, *evi*, was designed in SynGAP to calculate the *EVI* value for each high-confidence gene pair (Fig. 4a). The higher the *EVI* value of a gene pair is, the more dramatically the two homologous genes are differentially expressed.

To evaluate the reliability of *EVI*, we tested it using a couple of publicly available RNA-seq datasets from published studies. Firstly, two apple cultivars with different peel coloring phenotypes were used for testing [28]. For all the genes ordered by the *EVI* values calculated, two genes, *MdGST* and *MdMYB10*, ranked high (Fig. 4d), suggesting a high level of differential expression. And indeed, these two genes were identified as candidate genes that cause the apple cultivar "Blondee" to turn yellow in peel [28]. Meanwhile, GSEA analysis using *EVI* as a weight indicator could significantly

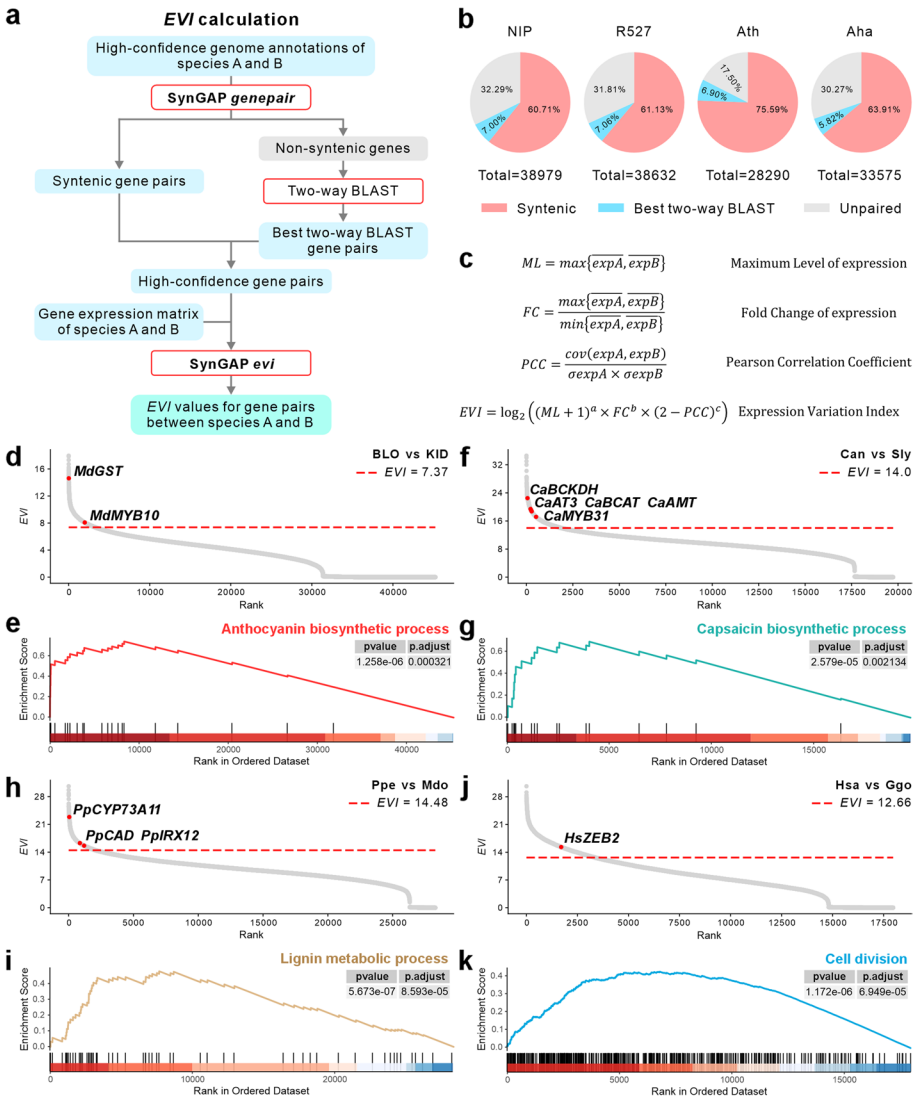

Fig. 4 Cross-species gene differential expression analysis employing SynGAP *evi*. **a** Workflow for SynGAP *genepair* and *evi*. **b** Statistics of gene pairs identified by SynGAP of "NIP (*O. sativa* Japonica c.v. Nipponbare) vs R527 (*O. sativa* Indica c.v. R527)" and "Ath (*A. thaliana*) vs Aha (*A. halleri*)." **c** Formula of *EVI* calculation. *expA* and *expB* represent the temporal expression levels of cross-species gene pairs (gene A and gene B). $\overline{expA}$ and $\overline{expB}$ are the average expression of gene A and gene B in a temporal process, and expression values below 0.1 are set to 0. *ML*, *FC*, and *PCC* represent the maximum expression level, fold change of expression, and expression pattern correlation of gene pairs, respectively. The indexes *a*, *b*, and *c* are set to 1, 1, and 4 by default, which are derived from broad test. **d**, **f**, **h**, **j** Ranked *EVI* of gene pairs. The red dashed line represents the threshold automatically generated by SynGAP, and the gene pairs with *EVI* exceeding the threshold are considered to show remarkable differential expression (**d** KID (*Malus domestica* c.v. Kidd's D-8) vs BLO (*M. domestica* c.v. Blondee); **f** Can (*C. annuum*) vs Sly (*S. lycopersicum*); **h** Ppe (*Prunus persica*) vs Mdo (*M. domestica*); **j** Hsa (*Homo sapiens*) vs Ggo (*Gorilla gorilla*)). **e**, **g**, **i**, **k** GSEA enrichment analysis based on *EVI* (**e** KID vs BLO; **g** Can vs Sly; **i** Ppe vs Mdo; **k** Hsa vs Ggo)

enrich the genes to the GO term of anthocyanin biosynthetic process, which is the main biological process responsible for apple peel coloration (Fig. 4f).

Although both tomato and pepper belong to Solanaceae, pepper specifically synthesizes capsaicin in the placenta [29]. We used SynGAP to conduct comparative

transcriptome analysis using RNA-seq data sampled from the placenta of pepper and tomato fruit at different development stages [30, 31]. A list of homologous genes, enriched into the GO term of capsaicin biosynthetic process, were identified by the ranking of calculated *EVI* (Fig. 4g). Among them, *CaBCKDH*, *CaAT3*, *CaBCAT*, *CaAMT*, and *CaMYB31*, as key enzymes and regulators of the capsaicin metabolism pathway [29, 30, 32, 33], showed strong differential expression signals (high *EVI* value, Fig. 4e). Similarly, in Rosaceae, peach is a drupe fruit with ovary wall hardening into a stone-like shell, while apple is a pome fruit with no lignifying of the ovary wall. Using SynGAP in combination with temporal RNA-seq data from these two fruits [27], we could profile a list of genes enriched in the lignin metabolic process by the GSEA analysis (Fig. 4j). Among them, *PpCYP73A11*, *PpCAD*, and *PpIRX12*, which encode key enzymes in the lignin metabolism pathway, ranked high in *EVI* value (Fig. 4h), suggesting that they may be important factors contributing to different fruit type formation of peach and apple.

We also tested the performance of *EVI* in organisms other than plants. For instance, compared to gorillas, the human brain undergoes rapid expansion during evolution, which is one of the main reasons for the intellectual differences between humans and gorillas. In the *EVI*-based comparative transcriptome analysis using RNA-seq data of different development processes of human and gorilla brains, it was found that *ZEB2*, one of the key regulatory factors of brain expansion [34], had a strong differential expression signal, as reflected by a high *EVI* value (Fig. 4i), and GSEA analysis enriched a large number of genes related to cell division which likely contribute to the brain expansion (Fig. 4k).

Taken together, all the results above demonstrated that SynGAP is a reliable tool for comparative transcriptome analysis, and the *EVI* could serve as an effective index value to identify candidate key genes that are responsible for certain trait development.

Conclusions

Although the rapid advancements in sequencing and computational technologies significantly lower the barriers of genome sequencing, accurate gene structure annotation persists to be a challenge for biologists. Gene synteny, the preservation of gene order in aligned genomic blocks across species, is generally conserved to a certain extent based on their phylogenetic relationship. Here, as a proof-of-concept demonstration, SynGAP is developed for inter-species polishing of gene structure annotation, utilizing gene synteny relationships across species. Our results showed that SynGAP is a powerful toolkit with exceptional capabilities in the improvement of GSA quality and the profiling of integrative gene synteny between species. It can be broadly applied in comparative genomics and transcriptomics analyses to facilitate the exploration of evolutionarily genomics changes and the identification of candidate genes responsible for the development of distinct traits observed in related species.

Methods

Overview of the SynGAP

SynGAP (Synteny-based Gene Structure Annotation Polisher) is a command-line software written in Python 3, suitable for Linux operating systems. And we provide images

that can be used for MacOS and Windows. The source code and tutorial for this toolkit can be obtained for free from a GitHub site (https://github.com/yanyew/SynGAP). SynGAP comprises five subroutines: *dual*, *master*, *triple*, *genepair*, and *evi*. It supports two main workflows: (1) GSA polishing for related species (*dual*, *master*, and *triple*) and (2) comparative transcriptome analysis of related species (*genepair* and *evi*).

Data source

Genomic data, including genomic data from 118 embryophytes and 307 vertebrates, were collected from Ensembl Plants, Ensembl, Rice RC, Sol Genomics Network, and SapBase databases [35–39]. Transcriptome data was downloaded from the SRA database [27, 28, 30, 31, 34, 40–42].

Classification of gene annotations

DIAMOND [43] was used to align the protein sequences of the original gene annotations to the Swiss-Prot database [16] for functional annotation. And the *E* value cutoff is set to $1e-8$. Protein-coding potential calculation was carried out using CPC2 [15] under default parameters. Based on the gene functional annotation and protein-coding potential, the original gene annotations are classified into four categories: type I genes are protein-coding potential (coding potential ≥ 0.5) and could be functionally annotated against Swiss-Prot; type IIa ones are protein-coding, but fail to be annotated functionally; type IIb genes do not have good protein-coding potential, but are of functional annotation; type III ones are of neither protein-encoding nor functional annotation.

Calculation of *R* value

To evaluate the reliability of each polished annotation, pairwise sequence global alignment against its homologous reference annotation was performed using EMBOSS NEEDLE [44] under default parameters. An evaluation parameter, *R* value, was designed based on the similarity and gap information from the alignment results. The calculation formula for *R* value is defined as:

$$R = \frac{Similarity \times \left(Alignment\ Length - Gaps\right)}{Alignment\ Length^2}$$

The closer the *R* value approaches 1, the higher the similarity between the polished annotation and its reference homolog.

A dynamic *R* cutoff threshold (R_{cutoff}) was adopted to filter out low-quality predicted annotations given the fact that the *R* value varies according to the distance of compared species and the evolutionary time of genomic duplications (WGD events or segmental duplications) giving rise to syntenic blocks.

Therefore, we defined the cutoff threshold for *R* value (R_{cutoff}) as:

$$R_{cutoff} = \min\left(R_{Q1}, 0.5\right)$$

R_{Q1} is the lower quantile *R* value for a given syntenic block; min() means the selection of the smaller value of R_{Q1} and 0.5. When R_{Q1} is smaller than 0.5, R_{cutoff} is set to the R_{Q1}, otherwise, to 0.5.

BUSCO assessment

BUSCO [13] was used to evaluate the completeness of GSA for various species under default parameters, using embryophyta_odb10 dataset for embryophytes and vertebrata_odb10 for vertebrates.

Phylogenetic analysis

OrthoFinder [23] was used to identify the single-copy orthologs among 25 plant species and 30 animal species and estimated the species tree respectively under default parameters.

Transcriptome data analysis

Quality control of raw data was conducted using FastQC [45] to confirm acceptable quality for downstream analysis. Trimmomatic [46] was invoked to remove the low-quality bases present in the sequencing data at the 3′ end of the splice sequence and read segment. All sequence data were compared to the reference genome using STAR [47]. The expression of genes was calculated using StringTie software [48] and normalized to TPM.

GSEA analysis

GSEA enrichment analysis was conducted using R package clusterProfiler [49] using *EVI* as the weight indicator.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03359-8.

Additional file 1: Supplementary Figs. S1 to S9.

Additional file 2: Supplementary Table S1 Estimation and data statistics for positive tests of SynGAP *dual* using plant species.

Additional file 3: Supplementary Table S2 Estimation and data statistics for SynGAP *dual* using plant species.

Additional file 4: Supplementary Table S3 Estimation and data statistics for SynGAP *dual* using animal species.

Additional file 5: Supplementary Table S4 Estimation and data statistics for SynGAP *triple* using plant species.

Additional file 6: Supplementary Table S5 Estimation and data statistics for SynGAP *triple* using animal species.

Additional file 7: Supplementary Table S6 List of accession links for the whole genome assemblies and genome annotations.

Additional file 8: Supplementary Table S7 List of accession numbers for the RNA-seq data.

Additional file 9: Review history. The peer review history.

Acknowledgements

We thank all labmates in the Xia lab for their generous help. We also appreciated the insightful suggestions from Dr. Yi Liao at South China Agricultural University.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 9.

Authors' contributions
F.W., C.C., and R.X. conceived the project; F.W., C.C., R.X., and Y.M. designed the functions of the toolkit. F.W. performed all the Python coding. F.W. tested the functions and prepared the tutorial manual. F.W., C.C., and R.X. prepared the figures and wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work is supported by the open competition program of top ten critical priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan of Guangdong Province (2022SDZG05) and the Key Area Research and Development Program of Guangdong Province (2022B0202070003). This work is also supported by the National Science Foundation of China (#32072547, #32372665, and #32102320).

Availability of data and materials
SynGAP is written in Python3, and it is released as open-source software under the GPL 3.0 license. The source code and documentation of SynGAP is available on GitHub (https://github.com/yanyew/SynGAP) [50] and Zenodo (https://doi.org/10.5281/zenodo.12771740) [51]. All described datasets are publicly available through the corresponding repositories. Genome assemblies and genome annotations are retrieved from Ensembl Plants, Ensembl, Rice RC, Sol Genomics Network, and SapBase databases with accession links in supplementary tables (Additional file 7: Table S6). RNA-seq data are available at SRA database with accession numbers in supplementary tables (Additional file 8: Table S7). The datasets generated and analyzed during the current study are available at figshare (https://doi.org/10.6084/m9.figshare.24657396) [52].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 December 2023 Accepted: 29 July 2024
Published online: 13 August 2024

References

1. Zhu Z, Sun B, Cai W, Zhou X, Mao Y, Chen C, et al. Natural variations in the MYB transcription factor *MYB31* determine the evolution of extremely pungent peppers. New Phytol. 2019;223(2):922–38.
2. Jiang S, Lv F, Gao L, Gu J, Yang R, Li S, et al. Novel R2R3-MYB transcription factor LiMYB75 enhances leaf callus regeneration efficiency in *Lagerstroemia indica*. Forests. 2023;14(3):517.
3. Nie B, Chen X, Hou Z, Li C, Sun W, Ji J, et al. Haplotype-phased genome revealed the butylphthalide biosynthesis and hybrid origin of Ligusticum chuanxiong. bioRxiv. 2023:2023.06. 13.544868.
4. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(suppl_2):W435–9.
5. Li H. Protein-to-genome alignment with miniprot. Bioinformatics. 2023;39(1):btad014.
6. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinform. 2014;48(1):4–11.
7. Lewis SE, Searle S, Harris N, Gibson M, Iyer V, Richter J, et al. Apollo: a sequence annotation editor. Genome Biol. 2002;3:1–14.
8. Chen C, Li J, Feng J, Liu B, Feng L, Yu X, et al. sRNAanno—a database repository of uniformly annotated small RNAs in plants. Hort Res. 2021;8:45.
9. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. 2005;21(12):673–82.
10. Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, Lomas JS, et al. JCVI: a versatile toolkit for comparative genomics analysis. iMeta. 2024:e211.
11. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49-e.
12. Sun P, Jiao B, Yang Y, Shan L, Li T, Li X, et al. WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant. 2022;15(12):1841–51.
13. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
14. She R, Chu JS-C, Uyar B, Wang J, Wang K, Chen N. genBlastG: using BLAST searches to build homologous gene models. Bioinformatics. 2011;27(15):2141–3.
15. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45(W1):W12–6.
16. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15.
17. Kliebenstein DJ, D'Auria JC, Behere AS, Kim JH, Gunderson KL, Breen JN, et al. Characterization of seed-specific benzoyloxyglucosinolate mutations in *Arabidopsis thaliana*. Plant J. 2007;51(6):1062–76.

18. Mahrez W, Shin J, Munoz-Viana R, Figueiredo DD, Trejo-Arellano MS, Exner V, et al. BRR2a affects flowering time via *FLC* splicing. PLoS Genet. 2016;12(4):e1005924.
19. McNeece BT, Pant SR, Sharma K, Niruala P, Lawrence GW, Klink VP. A Glycine max homolog of NON-RACE SPECIFIC DISEASE RESISTANCE 1 (NDR1) alters defense gene expression while functioning during a resistance response to different root pathogens in different genetic backgrounds. Plant Physiol Biochem. 2017;114:60–71.
20. Lin L, Ozaki T, Takada Y, Kageyama H, Nakamura Y, Hata A, et al. topors, a p53 and topoisomerase I-binding RING finger protein, is a coactivator of p53 in growth suppression induced by DNA damage. Oncogene. 2005;24(21):3385–96.
21. Vantourout P, Hayday A. Six-of-the-best: unique contributions of γδ T cells to immunology. Nat Rev Immunol. 2013;13(2):88–100.
22. Lefranc M-P. Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014;5:78999.
23. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:1–14.
24. Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Mol Cell. 2022;82(15):2885-99.e8.
25. Külahoglu C, Denton AK, Sommer M, Maß J, Schliesky S, Wrobel TJ, et al. Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. Plant Cell. 2014;26(8):3243–60.
26. Yu Y, Hu H, Doust AN, Kellogg EA. Divergent gene expression networks underlie morphological diversity of abscission zones in grasses. New Phytol. 2020;225(4):1799–815.
27. Li M, Galimba K, Xiao Y, Dardick C, Mount SM, Callahan A, et al. Comparative transcriptomic analysis of apple and peach fruits: insights into fruit type specification. Plant J. 2022;109(6):1614–29.
28. El-Sharkawy I, Liang D, Xu K. Transcriptome analysis of an apple (*Malus* × *domestica*) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. J Exp Bot. 2015;66(22):7359–76.
29. Sun B, Chen C, Song J, Zheng P, Wang J, Wei J, et al. The *Capsicum* MYB31 regulates capsaicinoid biosynthesis in the pepper pericarp. Plant Physiol Biochem. 2022;176:21–30.
30. Kim S, Park M, Yeom S-I, Kim Y-M, Lee JM, Lee H-A, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet. 2014;46(3):270–8.
31. Shinozaki Y, Nicolas P, Fernandez-Pozo N, Ma Q, Evanich DJ, Shi Y, et al. High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. Nat Commun. 2018;9(1):364.
32. del Rosario A-J, del Carmen R-G, López MG, Rivera-Bustamante RF, Ochoa-Alejo N. Virus-induced silencing of Comt, p Amt and Kas genes results in a reduction of capsaicinoid accumulation in chili pepper fruits. Planta. 2008;227:681–95.
33. Mazourek M, Pujar A, Borovsky Y, Paran I, Mueller L, Jahn MM. A dynamic interface for capsaicinoid systems biology. Plant Physiol. 2009;150(4):1806–21.
34. Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, et al. An early cell shape transition drives evolutionary expansion of the human forebrain. Cell. 2021;184(8):2084-102.e19.
35. Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, et al. The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. Nucleic Acids Res. 2015;43(D1):D1036–41.
36. Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021;184(13):3542-58.e16.
37. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50(D1):D988–95.
38. Li J, Chen C, Zeng Z, Wu F, Feng J, Liu B, et al. SapBase: a central portal for functional and comparative genomics of Sapindaceae species. J Integr Plant Biol. 2024.
39. Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022;50(D1):D996–1003.
40. Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. Plant J. 2016;88(6):1058–70.
41. Tao Y, An L, Xiao F, Li G, Ding Y, Paul MJ, et al. Integration of embryo–endosperm interaction into a holistic and dynamic picture of seed development using a rice mutant with notched-belly kernels. Crop J. 2022;10(3):729–42.
42. Li C, Li S, Yang C, Ding Y, Zhang Y, Wang X, et al. Blood transcriptome reveals immune and metabolic-related genes involved in growth of pasteurized colostrum-fed calves. Front Genet. 2023;14:1075950.
43. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18(4):366–8.
44. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.
45. Wingett SW, Andrews S. FastQ Screen: a tool for multi-genome mapping and quality control. F1000Research. 2018;7:1338.
46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
48. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
49. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. The Innovation. 2021;2(3):100141.

50. Wu F, Mai Y, Chen C, Xia R. SynGAP: a synteny-based toolkit for gene structure annotation polishing. Github. 2023. https://github.com/yanyew/SynGAP.
51. Wu F, Mai Y, Chen C, Xia R. SynGAP: a synteny-based toolkit for gene structure annotation polishing. Zenodo 2024.https://doi.org/10.5281/zenodo.12771740.
52. Wu F, Mai Y, Chen C, Xia R. SynGAP: a synteny-based toolkit for gene structure annotation polishing. Datasets. Figshare 2023. https://doi.org/10.6084/m9.figshare.24657396.

Publisher's Note