Received: 11 February 2024 | Revised: 9 May 2024 | Accepted: 16 May 2024

DOI: 10.1111/ppl.14398

RESOURCE ARTICLE

Physiologia Plantarum

PABLOG: a Primer Analysis tool using a Bee-Like approach on Orthologous Genes

Michele Ferrigno[1,2] | Paola Frazzetto[1] | Andrey Prjibelski[3] | Alexandru I. Tomescu[3] | Giuseppe Diego Puglia[1]

[1]National Research Council of Italy, Institute for Agriculture and Forestry Systems in the Mediterranean, Catania, Italy

[2]Department of Mathematics and Computer Science, University of Catania, Catania, Italy

[3]Department of Computer Science, University of Helsinki, Helsinki, Finland

Correspondence
Giuseppe Diego Puglia,
Email: giuseppediego.puglia@cnr.it

Funding information
Consiglio Nazionale delle Ricerche, Grant/Award Number: FOE-2021 DBA. AD005.225; Agritech National Research Center and European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR), Grant/Award Number: MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022

Edited by J.-.F. Mao

Abstract

RNA-seq data is currently generated in numerous non-model organisms that lack a reference genome. Nevertheless, the confirmation of gene expression levels using RT-qPCR remains necessary, and the existing techniques do not seamlessly interface with the omics pipeline workflow. Developing primers for many targets by utilising orthologous genes can be a laborious, imprecise, and subjective process, particularly for plant species that are not commonly studied and do not have a known genome. We have developed a primer design tool, named PABLOG, that analyses the alignments generated from long or short RNA-seq reads and a reference orthologous gene. PABLOG scans, much like a bee searching several flowers for pollen, and presents a sorted list of potential exon-exon junction locations, ranked according to their reliability. Through computational analysis across the whole genomes of several non-model species, we demonstrate that PABLOG performs more effectively than other methods in identifying exon-exon junctions since it generates significantly fewer false-positive results. Examination of candidate regions at the gene level, in conjunction with laboratory studies, shows that the suggested primers successfully amplified particular targets in non-model plants without any presence of genomic contamination. Our tool includes a consensus sequence feature that enables the complete process of primer design, from aligning with the target gene to determining amplification parameters. The utility can be accessed via the GitHub repository located at: https://github.com/tools4plant-omics/PABLOG.

1 | INTRODUCTION

Plants, being immobile organisms, exhibit plastic responses to environmental changes in order to quickly adjust to new conditions and reproduce (Nicotra et al., 2010). The fast alterations occur due to the extensive diversity of genes and isoforms present in plant genomes. The study of gene expression enables the examination of how genes are transcribed in plants during their responses. However, our comprehension of this mechanism mostly relies on a limited number of species, resulting in the neglect of this regulatory level in non-model plants. Recently, the development of next- and third-generation sequencing technology has allowed researchers to study gene regulation in both model and non-model species (Unamba et al., 2015). Nevertheless, most studies estimate differential expression in different tissues or conditions based on RNA-seq transcript abundance and rarely engage in an appropriate investigation of the gene expression of relevant genes (Tognacca et al., 2023). This could lead to the multiplication of transcriptomic studies aimed at studying only a few target genes instead of exploiting the available datasets. As a part of the validation process, a RT-qPCR of identified mRNAs should be performed to ascertain their presence in the translatable pool and precisely determine the relative concentration (Tognacca et al., 2023). Unfortunately, the design of primers for amplifying a specific mRNA in both model and non-model species is still a time-consuming step, and the present tools do not fully exploit the available information from RNA-seq data. This is particularly true for non-model plants that lack an

accurate reference genome. Poor primer design can result in reduced technical precision and false positive or negative detection of amplification targets. Reliable primers should have absolute specificity, the absence of hairpin structures or cross-dimerization potential, and temperature tolerance (Bustin & Hugget, 2017). These requirements could be addressed with the many primer-designing tools available online (Markham & Zuker, 2005; Untergasser et al., 2012). Recent resources have been developed for improving oligo specificity, taking advantage of publicly available transcriptome (Arvidsson et al., 2008) or genome (Kim et al., 2017; Bae et al., 2021) datasets. However, these tools are intended to be used for organisms with well-annotated genomes, and their application to non-model species makes target sequence identification very cumbersome. In this case, annotated orthologous genes are commonly used for target sequence identification, but this practice requires substantial manual intervention and does not integrate with user-defined elaboration pipelines. Few other tools nowadays make use of orthologous genes, but for different applications, such as assembly quality assessment (Simao et al., 2015) or phylogenomic reconstruction (Manni et al., 2021). Moreover, amplification should always be RNA-specific, as some genomic DNA can be present in RNA preparation even after digestion with DNAse I. This can be achieved by placing the primer binding sites spanning an exon-exon junction, and this feature can be used for isoform-specific amplification as well. Such primers will successfully distinguish between cDNA, genomic DNA, and isoforms. This step can be particularly delicate for non-model species and in de-novo transcriptome assembly, as the lack of a genome reference can make this process arbitrary, inaccurate, and hard to replicate. At present, a few tools, such as SAMtools (Li et al., 2009) and RegTools (Cotto et al., 2023), can identify exon junction regions with different approaches. However, these tools were designed for other purposes, and their performance for primer design over candidate regions has not been assessed. In the present study, we present a novel tool, PABLOG (a Primer Analysis tool using a Bee-Like approach on Orthologous Genes), which has been specifically conceived for selecting candidate regions using orthologous gene alignments at the exon-exon junction and returning high-quality primers for gene expression studies. The proposed tool provides tunable parameters and a ranking score for identified regions in which primers have been designed.

2 | MATERIALS AND METHODS

2.1 | In silico benchmarking

We compared the capability of PABLOG and *find_introns*, a utility from the SAMtools suite (Li et al., 2009), to identify correct exon-exon junctions by running both programmes on homologous alignment at genome-scale with an increasing gradient of phylogenetic distance and comparing the obtained intronic regions with the reference General Feature Format - GFF file (Figure 1). To perform this benchmarking, using the STAR programme (Dobin et al., 2013), we aligned with the default parameters a reference RNA-seq dataset from *Arabidopsis thaliana* (Zuther et al., 2019; Schaarschmidt et al., 2020; http://www.ncbi.nlm.nih.gov/geo, accession number GSE112225) against the genomes of *A. thaliana*, *Camelina sativa*, *Euthrema salsugineum*, *Brassica oleracea*, *Dacus carota*, and *Solanum tuberosum*. The first four genomes fall within the Brassicaceae plant family, while the other two species are in the Apiaceae and Solanaceae families, respectively. The obtained Binary Alignment Map (BAM) files were used as input for PABLOG and *find_introns*, and the obtained intronic regions were compared with the GFF file of each species retrieved from the Ensembl Plants database (https://plants.ensembl.org/index.html). The percentage accuracy value was calculated by dividing the number of introns validated by GFF by the introns detected using the PABLOG or *find_introns* tools.

2.2 | Laboratory validation

To test the effectiveness of primers designed by the PABLOG tool, we performed gene-scale benchmarking by simulating a primer design workflow in a non-model plant species. To this end, we used the raw transcriptome data available in the NCBI SRA repository PRJNA897116 from the non-model plant species *Lonicera japonica* Thunb. (Caprifoliaceae), which is phylogenetically distant from species with annotated genomes, to design primers for two distinct genes: *DOF AFFECTING GERMINATION1 (DAG1)* and *TOPLESS (TPL)*. We used the NCBI accession number LOC108211016 from *Daucus carota* L. for *DAG1* and LOC105158348 for *TPL* from *Sesamum indicum* L. as reference orthologous genes to align *L. japonica* RNA-seq reads. We ran the PABLOG tool with default parameters for *DAG1*, and set the size parameter to 50 for *TPL*. The resulting primers are provided in Table S1. The effectiveness of targeting the exon-exon junction was ascertained by amplifying cDNA prepared from *L. japonica* leaves using the designed primers. The reactions were performed on an Eppendorf Mastercycler Nexus gradient using the DreamTaq DNA Polymerase protocol (Thermo Scientific™) with 0.8 µL of 10 mM for each primer and 0.1 µg of cDNA template. The programme run was composed of a first step at 95°C for 3 min and afterwards 30 cycles consisting of 30 s at 95°C, 30 s at a specific T_a (*DAG1* 57°C; *TPL* 59°C) for each pair of primers, and 1 min at 72°C. These cycles were followed by a final extension at 72°C for 5 minutes. For all primers, the fragments obtained were observed on a 2% agarose gel. We purified the PCR products using the MinElute PCR purification kit (Qiagen) and then sequenced them using the corresponding primers with Sanger technology through a sequencing company (Eurofins GmbH). We analysed the obtained DNA sequences using BLASTN, the Basic Local Alignment Search Tool (Camacho et al., 2009), searching in the nucleotide collection set using the BLASTN algorithm to verify their correspondence to target genes.

3 | RESOURCE OVERVIEW

We present PABLOG, a script tool developed in the Python programming language that implements an analysis pipeline for primer design. The PABLOG workflow consists of four main steps (Figure 2):

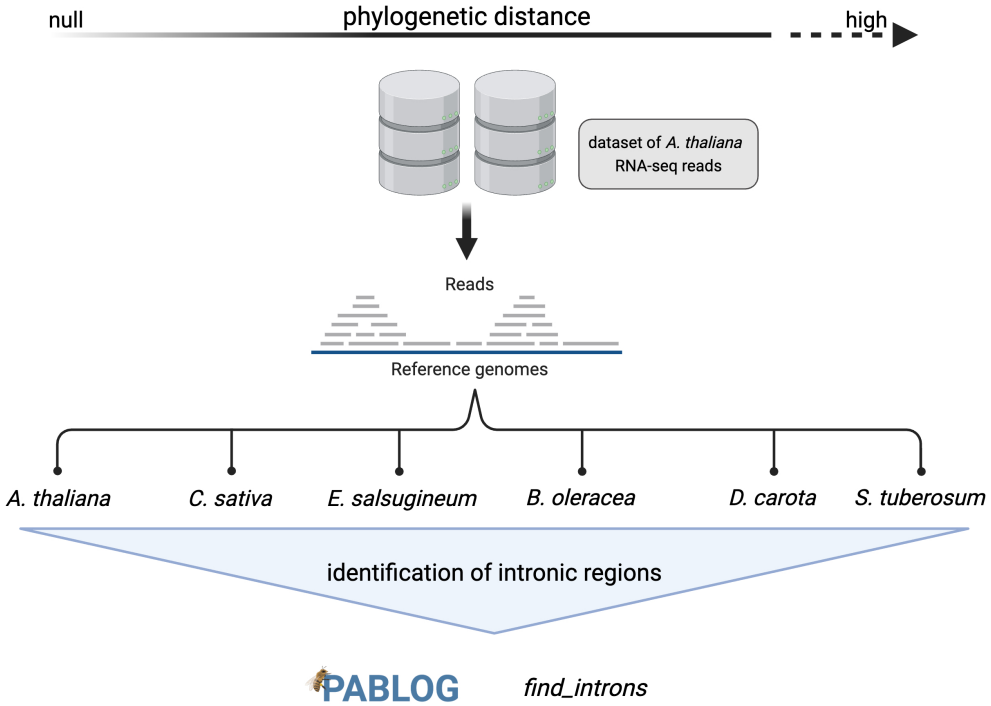

FIGURE 1 *In silico* benchmarking of the PABLOG tool. A reference RNA-seq dataset was aligned with six different annotated genomes following an increasing phylogenetic distance gradient. The alignments were used as input for the identification of intronic regions with PABLOG and *find_introns* (SAMtools). The correctness of the identified regions was validated with the GFF file for each species.

(1) alignment analysis; (2) coverage analysis; (3) generation of consensus sequence; and (4) primer design.

The pipeline starts by analysing the reference orthologous gene Fasta file (containing both introns and exons) and the BAM file obtained by aligning RNA-seq reads to the reference Fasta. At this step, the tool uses the HTSeq Python module (Putri et al., 2022) to parse alignment files and relative data. The goal of this step is to extract exon-exon junctions as candidate regions for priming sites. This is achieved by scanning the Concise Idiosyncratic Gapped Alignment Report (CIGAR) strings and searching for a skip or clip operation (N, H, or S cigar codes) between two matches (M). Every time this pattern occurs, the corresponding region is noted as a candidate region. The list of candidate regions proceeds into the coverage analysis step, in which the Average Coverage and the Goodness Score are computed from the coverage variation in that region as follows:

$$AverageCoverage = \frac{cov_{start} + cov_{end}}{2}$$

$$GoodnessScore = \frac{AverageCoverage - cov_{max}}{AverageCoverage} * 100$$

, where cov_{end} and cov_{start} are the alignment coverage values at the *end* and *start* positions of exon 1 and 2, respectively, and cov_{max} is the highest coverage value observed within the intronic region (Figure 3). Such value quantifies the variation of the coverage in the region with respect to *start* and *end*: if the region in between shows coverage values close to *start* and *end* values, Goodness Score will be low, suggesting the region is not very good to be used as a candidate site to develop primers. On the contrary, an optimal case would show cov_{max} to be close to zero, resulting in a 100% Goodness Score region. In addition, this computation is influenced by the alignment and read-calling quality. For instance, in an area with a high degree of mismatch, the cov_{end}, cov_{start}, and cov_{max} values will be similar to each other, resulting in a very low Goodness Score.

Once the candidate regions have been identified, PABLOG generates a consensus sequence using the pileup engine (pysam.bcftools utilities), provided in the same Python module of pysam. From this consensus sequence, the regions found in the previous steps are removed since they represent introns and we are interested only in exon-exon junctions for RT-qPCR amplification.

In the last step of the pipeline, the consensus sequence is fed into the Primer3 integrated tool for primer design using the Python module primer3-py, where the consensus sequence is provided as an input sequence along with the specific position (exon-exon identified site) where the primers have to be designed. All other Primer3 parameters are left by default, as the authors intended. The user can change these by editing the text file "primer3_settings.txt".

PABLOG results are returned to the user as a single text file that contains all regions found and, for each of them, the primer designed with the relative Goodness Score of the identified region. A step-by-step guide for tool installation and running is provided in the online repository and as Supplementary Material in this article.

4 | RESULTS AND MINOR DISCUSSION

Figure 4 illustrates the efficacy of the PABLOG and *find_introns* tools in detecting potential locations suitable for primer development. In general, PABLOG demonstrated superior accuracy in identifying

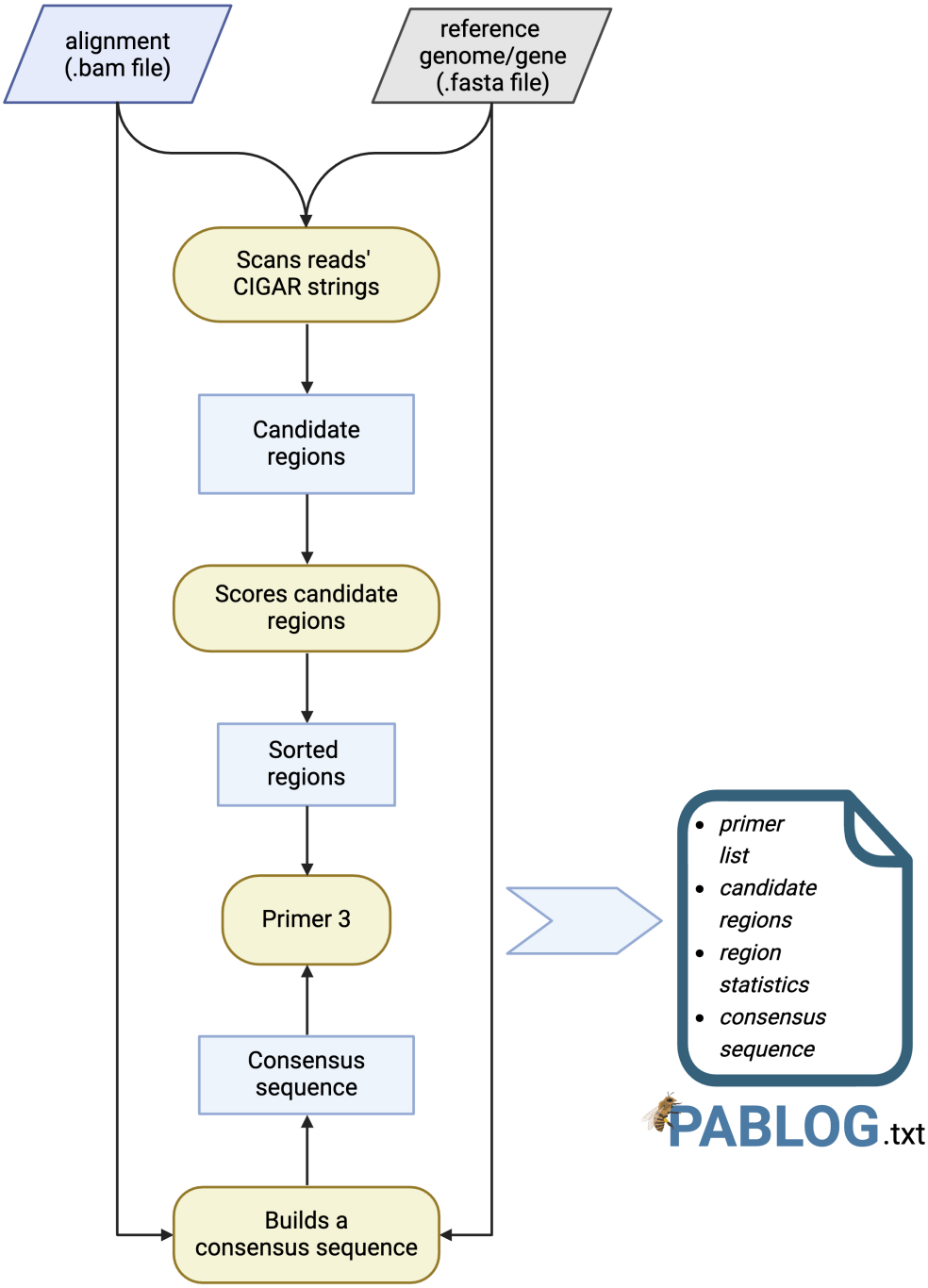

FIGURE 2 A schematic of the PABLOG tool's workflow. The parallelograms represent input files, rounded rectangles for elaboration steps, and squared rectangles for intermediate data. PABLOG outputs a text file with the primer list, candidate regions, region statistics, and consensus sequence used to design primers.

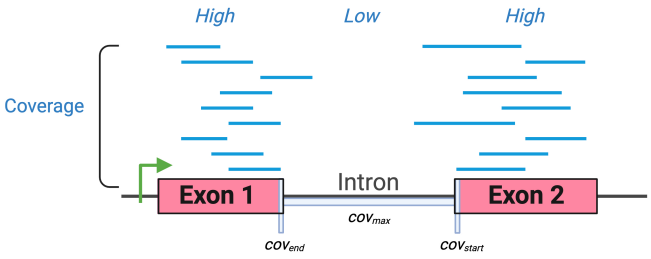

FIGURE 3 Schematic representation of coverage evaluation in exon 1 (cov_{end}), exon 2 (cov_{start}), and the intron (cov_{start}) used for the computation of the Goodness Score.

effective intronic regions compared to *find_introns*, even when handling phylogenetically distant species. Through genome-wide benchmarking for PABLOG, we found that as the phylogenetic distance increased, there was a significant decrease in both the accuracy and quantity of candidate regions. Within the Brassicaceae family, the PABLOG average accuracy was 90.63% ± 0.05, while for *D. carota* it was 29.72% and for *S. tuberosum* it was 16.66% (Figure 4A). In contrast, the *find_introns* analysis exhibits a far lower accuracy (30.39% within Brassicaceae, 11.98% for *D. carota*, and 9.70% for *S. tuberosum*), mostly because there is limited concordance between the candidate regions identified in the study and the corresponding

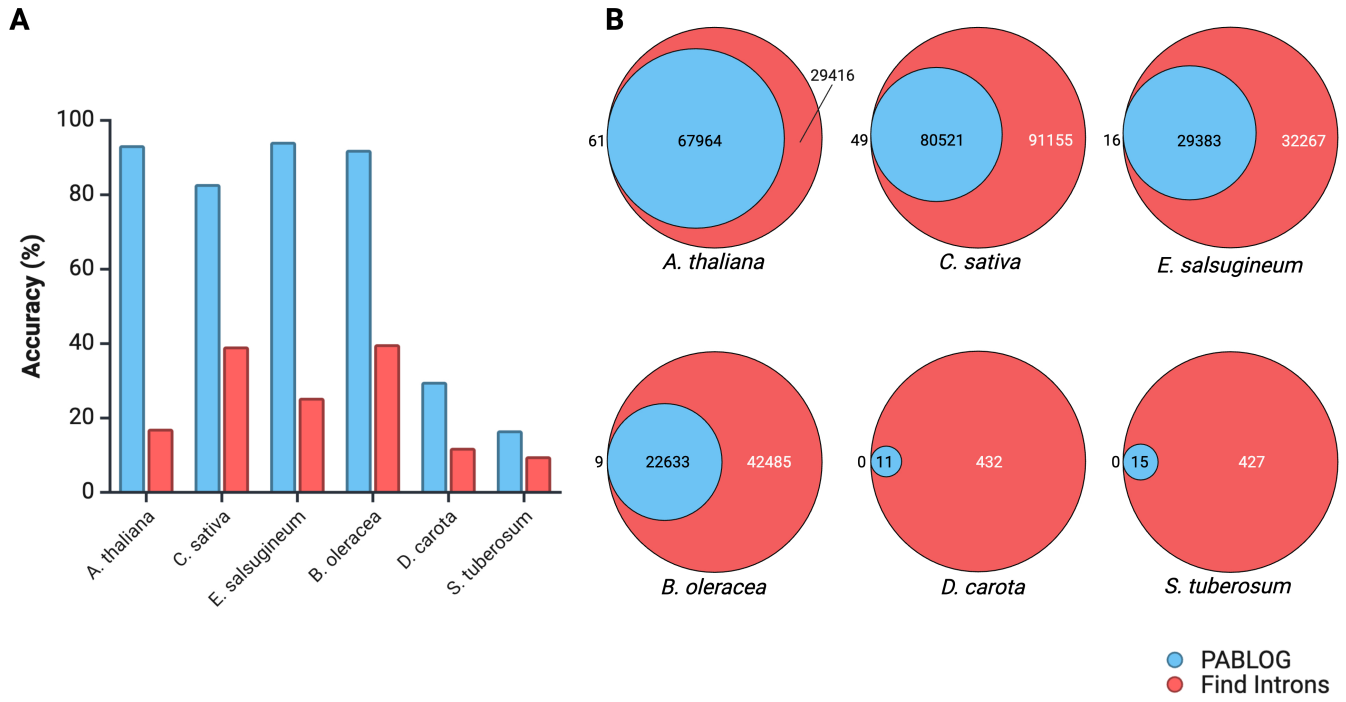

FIGURE 4 Comparison of the genome-scale identification of exon-exon junction regions in PABLOG and in *find_introns* (SAMtools) tools. A) The accuracy percentage was calculated by dividing the number of introns validated by GFF by the total number of introns detected by either the PABLOG or *find_introns* tools. B) Venn diagrams showing an overlap between the GFF-validated regions identified by *find_introns* (green circle) and PABLOG (brown circle).

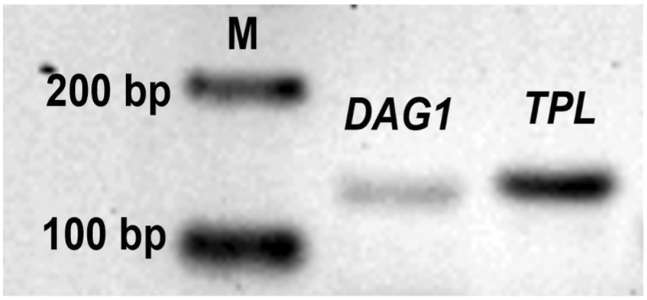

FIGURE 5 Electrophoresis gel of the fragments obtained using the two pairs of primers designed with PABLOG. M: 100 bp ladder; amplicon of *DAG1* (expected length 134 bp) and *TPL* (expected length 135 bp) target fragments, respectively.

regions that were confirmed using the GFF file. Conversely, the PABLOG tool detected a smaller number of possible locations, but nearly all of them were verified in the GFF file (Figure 4B). Notwithstanding the increased phylogenetic distance, both *find_introns* and PABLOG detected a higher amount of introns for *C. sativa* with respect to the *A. thaliana* genome. This is likely due to the bigger hexaploid genome of *C. sativa*, which arose from a whole-genome triplication event.

These findings emphasize the significance of utilissing, where feasible, a reference gene from a phylogenetically proximate species for precise primer design. Our tool, PABLOG, offers a practical solution for implementing this technique. It has a low false-positive rate and great flexibility, enabling users to effortlessly substitute the reference gene in cases when no candidate areas are detected.

Regarding the laboratory validation, we noticed a distinct amplification of each primer pair that was constructed using PABLOG. Additionally, the *L. japonica* amplicons displayed the expected length, as depicted in Figure 5. The BLASTN search with the DNA sequences of the *L. japonica* DAG1 and TPL amplicons returned the corresponding genes as the top results, thus verifying the accuracy of the analysis (Table S2). The sequencing of the amplicons did not reveal any genomic areas. In comparison to the genome-wide benchmarking analysis, the laboratory tests demonstrated the ability of PABLOG to design effective primers for non-model species, specifically using *D. carota* or *S. indicum* for reference orthologous genes and *L. japonica* for RNA-seq readings. The selection of understudied genes further demonstrates that this approach can be used for multiple research areas.

5 | CONCLUSIONS

Our tool regards a crucial step in gene expression studies that is often overlooked and error-prone, particularly for non-model species. PABLOG provides unprecedented integrability with RNA-seq pipelines without replacing any well-established tools but combining them into a rapid and consistent analysis workflow. It takes as input aligned RNA-seq data and an orthologous gene sequence without the need for a GFF file. Then PABLOG scans it to identify candidate regions in which to design primers, and returns oligo sequences when reliable

regions are found. PABLOG is an open system that can be finely tuned to any species both at the candidate region identification and in the primer design step through the change of several intuitive parameters. This characteristic could allow PABLOG usage to be expanded to other organisms than plants.

AUTHOR CONTRIBUTIONS

MF: assisted in the conceptualisation of the *in silico* experiments, wrote the code, performed and analysed the *in silico* experiments; PF: performed the laboratory experiments; AP: performed the genome-scale *in silico* experiments; AT: supervised the genome-scale *in silico* experiments; GDP: conceived the idea, led the *in silico* and laboratory experiments, and wrote the paper.

ACKNOWLEDGMENTS

Figures in this article were created with BioRender.com.

FUNDING INFORMATION

This study was carried out within the CNR project FOE-2021 DBA. AD005.225 and the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

DATA AVAILABILITY AND FAIR (FINDABLE ACCESSIBLE INTEROPERABLE REUSABLE) COMPLIANCE STATEMENT

We have utilised RNA-seq data from NCBI SRA repository PRJNA897116.

The PABLOG tool is available under a GNU GENERAL PUBLIC LICENSE (GPL 3.0), along with the data used for the evaluation, at the following GitHub repository link: https://github.com/tools4plant-omics/PABLOG.

ORCID

Michele Ferrigno https://orcid.org/0009-0006-3500-2163
Paola Frazzetto https://orcid.org/0009-0007-1449-5904
Andrey Prjibelski https://orcid.org/0000-0003-2816-4608
Alexandru I. Tomescu https://orcid.org/0000-0002-5747-8350
Giuseppe Diego Puglia https://orcid.org/0000-0002-2327-3613

REFERENCES

Arvidsson, S., Kwasniewski, M., Riaño-Pachón, D.M., Mueller-Roeber, B. (2008) QuantPrime - A flexible tool for reliable high-throughput primer design for quantitative PCR. *BMC Bioinformatics* 9, 1–15

Bae, J., Jeon, H., Kim, M.S. (2021) GPrimer: a fast GPU-based pipeline for primer design for qPCR experiments. *BMC Bioinformatics* 22, 220

Bustin, S., Huggett, J. (2017) qPCR primer design revisited. *Biomolecular Detection and Quantification*, 14, 19–28.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421

Cotto, K.C., Feng, Y.Y., Ramu, A., Richters, M., Freshour, S.L., Skidmore, Z.L., Xia H,, McMichael, J.F., Kunisaki, J., Campbell, K.M., Chen, T.H.P., Rozycki, E.B., Adkins, D., Devarakonda, S., Sankararaman, S., Lin, Y., Chapman, W.C., Maher, C.A., Arora, V., Dunn, G.P., Uppaluri, R., Govindan, R., Griffith, O.L., Griffith, M. (2023) Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nature Communication*, 14.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.

Kim, H., Kang, N.N., An, K.H., Kim, D., Koo, J.H., Kim, M.S. (2017) MRPrimerV: A database of PCR primers for RNA virus detection. *Nucleic Acids Research*, 45, D475–D481.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M. (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38, 4647–4654.

Markham, N.R., Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Research*, 33, W577-81.

Nicotra, A.B., Atkin, O.K., Bonser, S.P., Davidson, A.M., Finnegan, E.J., Mathesius, U., Poot, P., Purugganan, M.D., Richards, C.L., Valladares, F., van Kleunen M. (2010) Plant phenotypic plasticity in a changing climate. *Trends in Plant Science*, 15, 684–692.

Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E., Zanini, F. (2022) Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* 38, 2943–2945.

Schaarschmidt, S., Fischer, A., Zuther, E., Hincha, D.K. (2020) Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant *Arabidopsis thaliana*. *International Journal of Molecular Science*, 21.

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

Tognacca, R.S., Rodríguez, F.S., Aballay, F.E., Cartagena, C.M., Servi, L., Petrillo, E. (2023) Alternative splicing in plants: current knowledge and future directions for assessing the biological relevance of splice variants. *Journal of Experimental Botany*, 74, 2251–2272.

Unamba, C.I.N., Nag, A., Sharma, R.K. (2015) Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. *Frontiers in Plant Science*, 6.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G. (2012) Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40, e115–e115.

Zuther, E., Schaarschmidt, S., Fischer, A., Erban, A., Pagter, M., Mubeen, U., Giavalisco, P., Kopka, J., Sprenger, H., Hincha, D.K. (2019) Molecular signatures associated with increased freezing tolerance due to low temperature memory in Arabidopsis. *Plant Cell Environment*, 42, 854–873.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ferrigno, M., Frazzetto, P., Prjibelski, A., Tomescu, A.I. & Puglia, G.D. (2024) PABLOG: a Primer Analysis tool using a Bee-Like approach on Orthologous Genes. *Physiologia Plantarum*, 176(3), e14398. Available from: https://doi.org/10.1111/ppl.14398