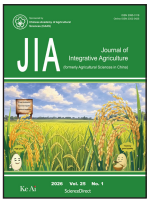
Journal of Integrative Agriculture 2026, 25(1): 150–156

KeAi
CHINESE ROOTS
GLOBAL IMPACT

Available online at ScienceDirect

Journal of Integrative Agriculture



RESEARCH ARTICLE

toGC: A pipeline to correct gene model for functional excavation of dark GPCRs in *Phytophthora sojae*

Min Qiu[1, 2, 3*], Chun Yan[1*], Huaibo Li[1], Haiyang Zhao[1], Siqun Tu[1], Yaru Sun[1], Saijiang Yong[1], Ming Wang[1, 2, 3#], Yuanchao Wang[1, 2, 3#]

[1] *Sanya Institute of Nanjing Agricultural University, Department of Plant Pathology, Nanjing Agricultural University, Nanjing 210095, China*
[2] *The Key Laboratory of Plant Immunity, Nanjing Agricultural University, Nanjing 210095, China*
[3] *Key Laboratory of Soybean Disease and Pest Control (Ministry of Agriculture and Rural Affairs), Nanjing Agricultural University, Nanjing 210095, China*

Highlights

- Developed toGC, a new tool that corrects errors in gene annotations using RNA-seq data.
- Experimentally validated toGC's accuracy, discovering two novel GPCR genes misannotated as one.
- Demonstrated toGC's broad utility by applying it successfully to multiple oomycete species.

Abstract

The accuracy of genomic annotation is crucial for subsequent functional investigations; however, computational protocols used in high-throughput annotation of open reading frames (ORFs) can introduce inconsistencies. These inconsistencies, which lead to non-uniform extension or truncation of sequence ends, pose challenges for downstream analyses. Existing strategies to rectify these inconsistencies are time-consuming and labor-intensive, lacking specific approaches. To address this gap, we developed toGC, a tool that integrates genomic annotation with RNA-seq datasets to rectify annotation inconsistencies. Using toGC, we achieved an accuracy of nearly 100% accuracy in correcting inconsistencies in published *Phytophthora sojae* ORFs. We applied this innovative pipeline to the GPCR-bigrams gene family, which was predicted to have 42 members in the *P. sojae* genome but lacked experimental validation. By employing toGC, we identified 32 GPCR-bigram ORFs with inconsistencies between previous annotations and toGC-corrected sequences. Notably, among these were 5 genes (*GPCR-TKL9*, *GPCR-TKL15*, *GPCR-PDE3*, *GPCR-AC3*, and *GPCR-AC4*) showed substantial inconsistencies. Experimental gene annotation confirmed the effectiveness of toGC, as sequences obtained through cloning matched those annotated by toGC. Importantly, we discovered two novel GPCRs (*GPCR-AC3* and *GPCR-AC4*), which were previously mispredicted as a single gene. CRISPR/Cas9-mediated knockout experiments revealed the involvement of *GPCR-AC4* but not *GPCR-AC3* in oospore production, further confirming their status as two separate genes. In addition to *P. sojae*, the reliability of the toGC pipeline in *Phytophthora capsici* and *Pythium ultimum* further emphasizes the robustness of this pipeline. Our findings highlight the utility of toGC for reliable gene model correction, facilitating investigations into biological functions and offering potential applications in diverse species analyses.

Keywords: gene model correction, transcriptome, open reading frames, G-protein coupled receptors

1. Introduction

A burgeoning number of genomes are undergoing sequencing, unveiling a novel perspective on the distinct attributes of individual organisms and presenting fresh opportunities for functional genomic analysis. Yet, the task of genome annotation still poses challenges, particularly in eukaryotes, where the goal is to establish precise gene models with accurate exon-intron structures for all genes (Danchin *et al.* 2018). However, not all species boast genome annotations of comparable quality to well-studied

Received 1 January 2024; Received in revised form 18 February 2024; Accepted 21 February 2024; Available online 27 March 2024
Min Qiu, E-mail: minqiu@njau.edu.cn; Chun Yan, E-mail: 2022202060@stu.njau.edu.cn; #Correspondence Yuanchao Wang, E-mail: wangyc@njau.edu.cn; Ming Wang, E-mail: mwang@njau.edu.cn
* These authors contributed equally to this study.

© 2026 CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer review under responsibility of Editorial Board of *Journal of Integrative Agriculture*.
doi: 10.1016/j.jia.2024.03.077

organisms such as humans or *Arabidopsis*, which have extensive research and ample experimental evidence for numerous genes (Bult *et al.* 2019), many sequences in non-model organisms are predicted *via* high-throughput computational annotations and may, therefore, be incomplete or erroneous. In reality, even for many model species, such as humans, where extensive research has been conducted on the human genome, annotation errors still exist (Zerbino *et al.* 2020). Similar situations were observed in model fungi and bacteria (Salzberg 2019; Mohanta and Al-Harrasi 2021).

Frequent errors encountered in gene models encompass erroneous exon and gene boundaries (Guigo *et al.* 2000), retention of non-coding nucleotide sequences within coding exons (Dragan *et al.* 2016), missed genes (Denton *et al.* 2014), and fragmentation or fusion of gene models (Guigo *et al.* 2006). These errors can propagate into downstream applications, resulting in incongruous or erroneous findings (Hadley 2003). Such errors can significantly affect the study of individual proteins, including transcript quantification, phylogenetic tree construction, and predictions related to protein structure or function (Deutekom *et al.* 2019). Thus, gene sequence information must be accurate, dependable, and presented in a clear and consistent manner. It is imperative to rectify gene models for our target genes before embarking on any functional analyses. However, in many cases that focused on only a small subset of genes and without the necessity for a comprehensive reannotation of the entire genome, a need arises for a swift and convenient method to promptly assess the accuracy of annotations for a small set of genes.

In this study, we introduce a comprehensive pipeline, toGC (Transcriptome-based Gene Model Correction), designed to rectify gene models by integrating genomic annotation with a diverse range of RNA-seq datasets. An illustrative case study in *Phytophthora sojae* underscores the efficacy of this approach. The concept we propose, utilizing consensus sequences of transcripts for gene model correction, can be extended to various analyses of non-model organisms.

2. Materials and methods

2.1. Simplified instructions for toGC workflow

Firstly, we acquired transcriptomic raw data from public databases and converted them into FASTQ files. Subsequently, Cutadapt and FastQC were employed to filter and eliminate Illumina sequencing data adapters. The resultant quality-controlled FASTQ files were aligned to a reference genome using read mappers such as Bowtie2, BWA, or HISAT2, generating SAM files. These SAM files were then converted to BAM files and sorted. Further, the sorted BAM files were merged using samtools. These merged BAM files include index information of the reference genome and are ready for importation into visual exploration software, such as IGV, along with the reference genome and annotated GFF files. The subsequent step involves visually assessing the target genes and their gene models by comparison with transcriptome information. For more detailed instructions on toGC, please refer to Appendix A.

2.2. Extraction of gDNA and cDNA of *P. sojae*

Genomic DNA was extracted from fresh *P. sojae* mycelial samples. Total RNAs were extracted from fresh *P. sojae* mycelial samples using Trizol extraction method. Complementary DNA (cDNA) was synthesized from the extracted total RNA. The synthesis of cDNA was performed using HiScript reverse transcriptase (R101-02, Vazyma, Nanjing, China), followed by the addition of gDNA eraser enzyme (Vazyme, Nanjing, China) to eliminate any genomic DNA.

2.3. Targeted gene deletion and screening of knockout mutants

Gene deletion mutants were generated using the CRISPR-mediated gene replacement strategy (Fang and Tyler 2016). The *hygromycin* gene, ligated with two 1.0-kb fragments flanking the target gene, was used as donor DNA in homology-directed repair (HDR) (Appendix B-a). We transformed *P. sojae* using PEG-mediated protoplast transformation (Hua *et al.* 2008). Putative resistant transformants were selected by growth on 10% (v/v) V8 medium containing 50 μg mL^{-1} geneticin and screened by PCR using the primer combinations as shown in Appendix B-a.

The primer set F1/R1 (Appendix C), located within the *GPCR-AC3* or *GPCR-AC4* ORF, was used to screen for the deletion of *GPCR-AC3* or *GPCR-AC4* in the genome of resistant transformants (Appendix B-b). The primer sets F2/R2 and F3/R3 (Appendix C) were used to detect homologous recombination events (Appendix B-b). Moreover, Sanger sequencing across the junctions of the flanking sequences consistently supported the replacement of the *GPCR-AC3* or *GPCR-AC4* with the *hygromycin* gene (Appendix B-c and d).

2.4. Growth and oospore production

To determine growth rate, all strains were grown on V8 medium at 25°C in the dark. Colony diameters on V8 medium during 4 days and the relative growth rates were calculated.

To monitor and quantify oospore production, strains were grown on lima bean agar (LBA) medium at 25°C in the dark. After 10 days, the cultures were examined microscopically. Three blocks of 1 cm^2 were excised from the medium, and within each block, three random fields at 40× magnification were selected for counting the number of oospores. All assays were repeated at least three times.

3. Results

3.1. Computed gene model may not in line with cloned authentic sequences

Oomycetes are a widespread group of filamentous microorganisms that pose a significant threat to global food security and natural ecosystems (Thines 2018). Genomic sequences of oomycetes are sprouting in recent years, providing valuable insights into oomycete research field (McGowan and Fitzpatrick 2020). We are wondering whether the released gene models are correct, so we have collected 20 published genes with available sequences by cloning in *P. sojae* (Appendix D). Due to the genome of *P. sojae* had

been annotated in two versions, the initial *P. sojae* V1.1 (GenBank assembly accession: GCA_000149755.1) and the latest *P. sojae* V3.0 (GenBank assembly accession: GCA_000149755.2) (Tyler *et al.* 2006), we conducted a multiple alignment of published genes among V1.1, V3.0, and authentic cloned sequences. We found that, in *P. sojae*, approximately 40% of the ORFs of published genes were initially annotated incorrectly, leading to the sequence differences in coding length or even completely incorrect reading frames (Fig. 1-A and B). For example, in Fig. 1-C, *PsHint1*, the histidine triad nucleotide-binding protein 1, was predicted the same start codon in both genomic versions, whereas the cloned sequence inferred that the coding actually starts with an earlier start codon (Zhang *et al.* 2016). *PsABCF1*, a putative ATP-binding cassette protein, also has an error in the start codon in the V3.0 annotation (Liu 2014), while the protein phosphatase *PsPP2C* has errors in intron prediction in V1.1 and the start codon in V3.0, respectively (Qiu *et al.* 2023). Another example is the SET domain protein PsKAMT3 regulating histone H3K36 trimethylation, which should have been coded by a shorter sequence due to an early occurrence of stop codon relative to the V3.0 genome (Chen *et al.* 2023). In addition, *PsMPK7* has errors in start codon and intron in V1.1 and the start codon in V3.0 separately (Gao *et al.* 2015). Notably, the *P. sojae* genome encodes three 6mA methyltransferases as core regulators involved in epigenetic modification, but all three have respective errors in their predicted ORFs, with *PsDMAT1* had an error in start codon that resulting in extending length, *PsDMAT2* having an error in intron that results in wrong stop codon, and *PsDMAT3* having a wrong start codon in both genomic versions (Chen *et al.* 2018). These errors underscore the importance of manual curation to ensure accuracy of gene model, thereby enhancing the success in downstream functional studies.

3.2. toGC: Correction of gene models through consensus transcribed sequences

The lack of congruency between predicted gene models and authentic ones infers the necessity to reanalyze gene models of interest prior to conducting any phylogenetic analysis or experiments. Although amplifying DNA products from cDNA and using Sanger sequencing to confirm sequence authenticity is a practical strategy for a handful of genes, it becomes time-consuming and expensive when dealing with dozens or hundreds of genes. As well known transcriptomic sequencing reveals transcripts' abundance, which generally draws more attention. However, it is also good at qualitatively uncovering the information of transcribed sequences, particularly displaying full profile of an authentic model, such as correct start and stop codon sites, no intron disturbance, even displaying alternatively spliced transcripts in different samples. By means of this context, we developed a user-friendly workflow, toGC, which incorporates multiple transcriptome data together (Ye *et al.* 2011; Wang *et al.* 2018; Qiu *et al.* 2020) to capture consensus transcribed sequences, and combine with visual tool (i.e., IGV) to vividly correct gene models (Fig. 1-D). Briefly, transcriptomic raw data from multiple stages of *P. sojae* were gathered, as gene expression may vary at different stages. After a standard RNA-seq data analysis workflow, including quality control, mapping, alignment along with conversion of SAM files into BAM files, these data were cataloged and merged into specific BAM files according to the developmental stages, such as mycelia, sporangia, zoospore, germinated cysts and infection. These combined BAM files contain index information of *P. sojae* genome V3.0 and ready for next importing to visual exploration software, such as IGV, along with *P. sojae* genome V3.0 file and annotated GFF files (Fig. 1-D), the details of toGC are available in Appendix A.

To assess the effectiveness of the toGC pipeline in correcting gene models, we applied it to the published genes mentioned in Fig. 1-C. We identified two types of errors in predicted gene models, including errors in start codons (e.g., *PsHint1*, *PsABCF1*, *PsPP2C*, *PsMPK7*, *PsDMAT1* and *PsDMAT3*), and introns within genes (e.g., *PsMPK7*, *PsPP2C*, *PsKAMT3*, *PsDMAT2*, and *PsDMAT3*). We observed that all errors related to reading frames (start codon and intron) got corrected (Fig. 1-C; Appendix D). It is noteworthy that the toGC pipeline employs the genomic information as a blueprint and incorporates transcriptome data; hence, it cannot correct errors resulted from genome sequence or genes with no transcription. Overall, the toGC pipeline surprisingly yielded an accuracy of nearly 100% in ORF prediction for the published genes (Fig. 1-C; Appendix D).

3.3. Practical performance of toGC in correcting predicted gene models in GPCR family

To double confirm the effectiveness of the toGC pipeline, we turned our attention to genes within a large family, GPCR-bigrams (*G-protein coupled receptor domain and a catalytic accessory domain*). G protein-coupled receptors (GPCRs), commonly referred to as seven transmembrane receptors or 7TM receptors, constitute a substantial group of proteins functioning as cell surface receptors. They are responsible for detecting molecules external to the cell and triggering cellular responses (Weis and Kobilka 2018). Notably, oomycetes feature a distinctive class of GPCRs known as GPCR-bigrams, characterized by the presence of an accessory catalytic domain alongside the GPCR domain (Van den Hoogen *et al.* 2018). It has been reported that the *P. sojae* genome contains 42 GPCR-bigrams, which can be further classified into 1 AP-GPCR (aspartic protease-GPCR), 7 GPCR-ACs (GPCR-adenylyl cyclase), 5 GPCR-INPPs (GPCR-inositol polyphosphate phosphatase), 2 GPCR-PDEs (GPCR-phosphodiesterase), 12 GPCR-PIPKs (GPCR-phosphatidylinositol-4-phosphate 5-kinase) and 15 GPCR-TKLs (GPCR-tyrosine kinase-like), however, whose models were not certificated yet by cloning (Van den Hoogen *et al.* 2018). Using the toGC pipeline, we corrected all the ORFs associated with the GPCR-bigrams. We compared toGC-corrected gene sequences with the original gene models and found that 32 ORFs of GPCR-bigrams contained inconsistencies among V1.1, V3.0 and toGC-corrected sequences (Appendix E). Specifically, when comparing with corrected sequences, 24 genes had inconsistencies in the

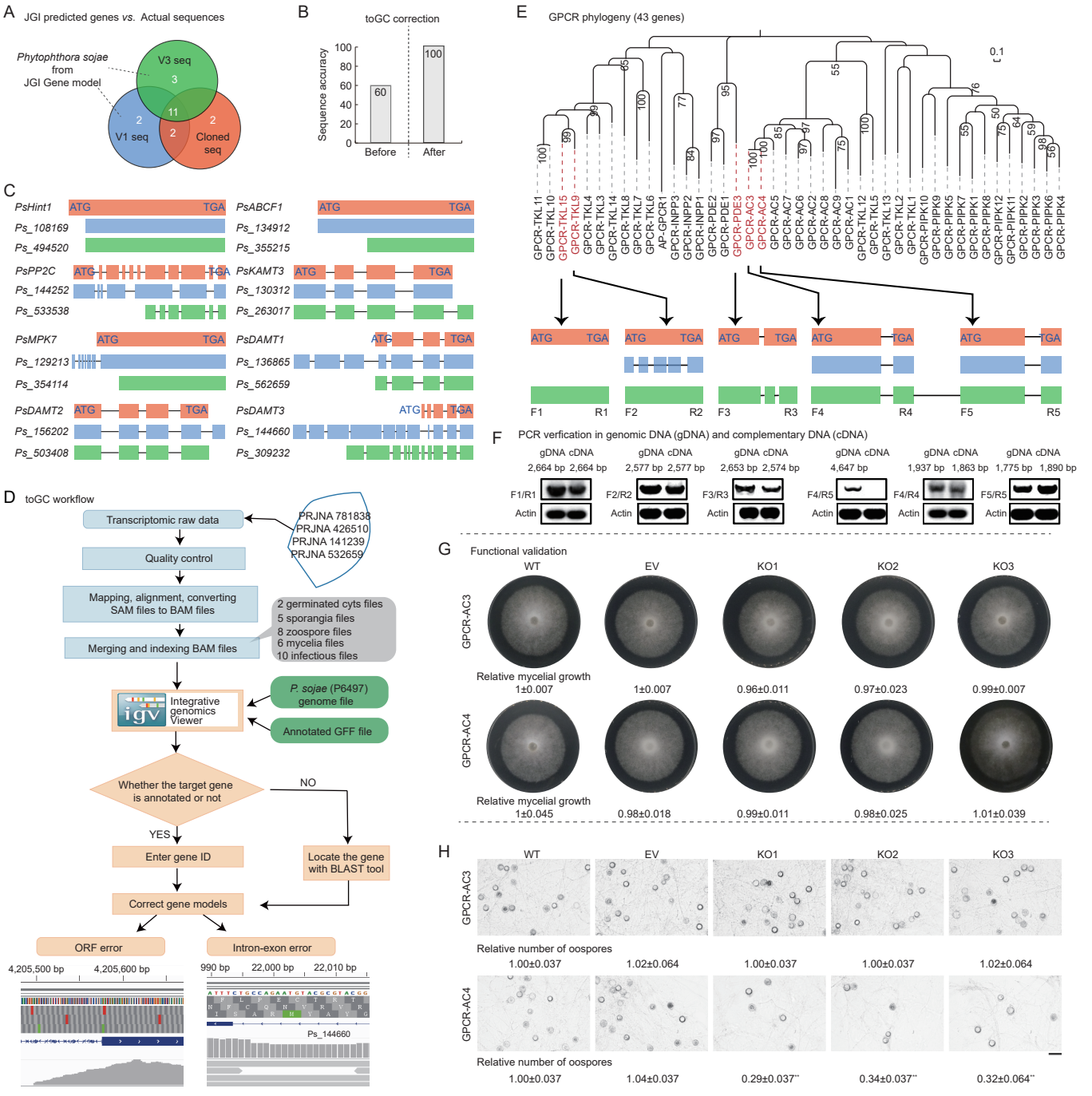


Fig. 1 Development and application of toGC on corrections of gene models. A, Venn diagram showed the comparison of actual gene sequences (cloned seq, orange) with predicted sequence from *Phytophthora sojae* V1.1 (V1 seq, blue) and V3.0 (V3 seq, green). The V1 seq and V3 seq represent that the predicted sequences of genes from two versions of annotated genomes, the initial V1.1 and the updated V3.0 (Tyler *et al.* 2006). These genes were collected from published works in *P. sojae* and they were summarized in Appendix D. B, accuracy of the published 20 gene sequences by predicted method (before) or by toGC pipeline (after). C, gene models of published 20 genes from predicted database or cloned actual sequences. The cloned sequences, the predicted sequences from *P. sojae* V1.1 and V3.0, were respectively filled with orange, blue and green. D, general workflow of toGC pipeline. E, correction of GPCR gene family by toGC pipeline. The corrected sequence, and the predicted sequences from *P. sojae* V1.1 and V3.0, were respectively filled with orange, blue and green. These sequences were summarized in Appendix F. F, PCR verification of 5 genes (*GPCR-TKL9*, *GPCR-TKL15*, *GPCR-PDE3*, *GPCR-AC3*, *GPCR-AC4*) by genomic DNA (gDNA) and complementary DNA (cDNA), those genes had huge variations among V1.1, V3.0 and toGC-corrected sequences. G, growth characteristics after 4 days of the wild-type (WT), empty vector control lines of *GPCR-AC3/AC4* (EV), *GPCR-AC3/AC4*-knockout lines (KO1, KO2 and KO3). Relative growth rates were displayed under each picture, with the number in wild-type strain P6497 set at 1. H, microscopic visualization of oospores of WT, empty vector control lines of *GPCR-AC3/AC4* (EV), *GPCR-AC3/AC4*-knockout lines (KO1, KO2 and KO3). The relative numbers of oospores were displayed under each picture, with the number in wild-type strain P6497 set at 1. Scale bar, 50 μm. Asterisk indicates significant difference at $P<0.01$ (**).

V3.0 annotation, while 14 genes exhibited inconsistencies in V1.1 (Appendix E). Among these inconsistences, five of them displayed huge variations among V1.1, V3.0 and toGC-corrected sequences (*GPCR-TKL9*, *GPCR-TKL15*, *GPCR-PDE3*, *GPCR-AC3*, *GPCR-AC4*) (Fig. 1-E), we further cloned these 5 genes individually for verification.

GPCR-TKL9 was found to have particularly different ORFs in *P. sojae* V3.0 (*Ps313892* without introns, 1,677 bp) and V1.1 (*Ps129200* with four introns, 2,577 bp) (Fig. 1-E). However, *GPCR-TKL9* has a very low transcriptional abundance in all transcriptome files (mycelia, sporangia, germinated cyst, zoospore and infection). Thus, we were unable to reanalyze this gene by incorporating transcriptome data. To address this issue, we cloned the *GPCR-TKL9* sequence from genomic DNA (gDNA) and complementary DNA (cDNA) of *P. sojae*. Through gel electrophoresis and Sanger sequencing, we confirmed that real sequence of *GPCR-TKL9* is 2,577 bp without introns (Fig. 1-F; Appendix F).

GPCR-TKL15 and *GPCR-PDE3* were predicted from *P. sojae* V3.0 genome, but *GPCR-TKL15* was not annotated in *P. sojae* V1.1 genome, and *GPCR-PDE3* was entirely absent from *P. sojae* V1.1 genome (Fig. 1-E). To verify the existence of the two genes, we cloned their real sequences from gDNA and cDNA of *P. sojae*. The experimental gene annotation further confirmed the correctness of this tool, as evidenced by the congruence of sequences between practical cloning and toGC annotation (Fig. 1-F; Appendix F).

Interestingly, we found two GPCR-ACs (*GPCR-AC3* and *GPCR-AC4*) were incorrectly predicted as a single gene (*Ps341867*) in *P. sojae* V3.0 genome (Fig. 1-E). As shown in Fig. 1-F; the full length of *Ps341867* can be amplified in gDNA but not in cDNA (F4/R5), indicating that the introns may have been wrongly predicted in this sequence. Furthermore, two DNA fragments were separately amplified from cDNA using two primer pairs (F4/R4 and F5/R5) (Fig. 1-F; Appendix C), confirming that indeed they are two separate genes. Most of the errors found in the GPCR-bigrams resulted in partial changes in protein length but they still belong to GPCR-bigrams. However, two errors found in introns (*Ps477243* and *Ps390167*) resulted in a premature termination, causing the corrected sequences to be excluded from the GPCR-bigrams (Appendix E). Thus in *P. sojae*, after correction, there are a total of 1 AP-GPCR, 9 GPCR-ACs, 3 GPCR-INPPs, 3 GPCR-PDEs, 12 GPCR-PIPKs and 15 GPCR-TKLs. Of particular interest is the discovery of two genes (*GPCR-AC3* and *GPCR-AC4*) that were erroneously predicted as a single gene, potentially relegating one of them to remain an untapped dark gene forever. To further investigate the function of these newly discovered genes, CRISPR/Cas9-mediated technology was utilized to generate three independent knockout mutants for each gene (Appendix B). Results showed that there were no obvious changes in oospore production and mycelium growth in the *GPCR-AC3* knockout mutants. However, in the *GPCR-AC4* knockout mutants, mycelium growth remained unaffected while a significant reduction was observed in oospore production (Fig. 1-G and H). Therefore, the toGC pipeline developed herein not only display strength in identifying previously overlooked “dark genes” but also demonstrates its potential in an applied perspective for functional investigations.

3.4. toGC pipeline corrects gene models in *Phytophthora capsici* and *Pythium ultimum*

To validate the robustness of the toGC pipeline in other oomycete species, we gathered transcriptomic raw data from *P. capsici* (Wang *et al.* 2023). Merged BAM files were generated for gene model correction with toGC, focusing on published genes with cloned sequences. *PcPdeH* (*Pc511692*) is a putative phosphodiesterase crucial for polarized growth and pathogenicity in *P. capsici*. The actual PcPdeH sequence exhibited an earlier start codon than initially predicted (Li *et al.* 2020). Employing the same steps as used for *P. sojae*, the toGC pipeline successfully corrected the gene model of *PcPdeH*, aligning with the cloned sequence (Appendix G-a).

To further confirm the robustness of the toGC pipeline, we obtained transcriptomic raw data from *P. ultimum* (Feng *et al.* 2021) and generated merged BAM files for gene model correction. Our focus shifted to an RNA-binding protein (*PuM90*), essential for oospore production. Previous confirmation indicated its automated gene model incorrectly predicted the start codon and introns, with the predicted gene model spanning 1,233 bp, while the actual sequence is 2,631 bp with an earlier start codon and two introns (Feng *et al.* 2021). Using the same approach, we observed that the toGC pipeline successfully corrected gene model of PuM90, aligning with the cloned sequence (Appendix G-b). These results affirm the applicability of the toGC pipeline in multiple oomycete pathogens.

4. Discussion

In this study, we introduce toGC as an innovative method aimed at rectifying inaccuracies in gene models resulting from erroneous annotations of ORF start and stop sites, as well as introns. We observed these discrepancies in published *P. sojae* genes, where they occurred at a surprisingly high frequency of up to 40% (Appendix D). These discrepancies necessitated additional efforts from researchers who had to invest considerable time in cloning the true sequences of their candidate genes, as documented in previous studies (Zhang *et al.* 2016; Chen *et al.* 2023). It is crucial to acknowledge that this rate of ORF annotation inconsistencies is likely an underestimation. This is particularly true for genes with extremely low transcriptional abundance, as exemplified by the case of *GPCR-TKL9* in our study. Furthermore, if draft-quality genomes were included, the rate would likely be even higher.

Genuine errors in start and stop site predictions can arise from several sources, including the inherent imprecision of automated computational methods, sequencing errors introducing premature stop codons or masking start site signals, and draft-quality genome sequencing causing truncation of ORFs that span contig edges. While some ORF annotation differences are undoubtedly legitimate, distinguishing them from sequencing errors is a daunting task without additional validation strategies. Even state-of-the-art eukaryotic gene prediction algorithms can make significant errors that have a profound impact on subsequent analyses.

A recent study has revealed that gene prediction errors in ORFs can affect up to 50% of subsequent coding protein sequences (Meyer *et al.* 2020).

In our dataset, we observed that inconsistencies in start codon annotations were more prevalent than those at stop codon ends (Fig. 1-C). This is likely due to the relative ease of identifying the first stop codon along an ORF compared to the complexity of discriminating between multiple potential start codons at the 5′ sequence end. Abnormally short 3′ sequence lengths may be linked to sequencing errors introducing premature stop codons in the wrong introns or programmed frameshifts within these ORFs. Additionally, inconsistencies in intron annotations were the most frequent among all types of annotation errors (Fig. 1-C and E). This could be attributed to limitations in previous gene prediction algorithms or insufficient experimental data.

To address these challenges and improve the accuracy of gene structure prediction, we integrated 31 transcriptome datasets encompassing five developmental or infectious stages of *P. sojae*. This integration led to the development of the toGC workflow, which leverages transcriptome data to achieve a 100% accuracy rate in ORF prediction for *P. sojae* genes (Fig. 1-C and E). The versatility of this tool is demonstrated across various gene types, including G protein-coupled receptors (*GPCR-AC3* and *GPCR-AC4*), mitogen-activated protein kinase (*PsMPK7*), methyltransferases (*PsDAMT1*, *PsDAMT2*, and *PsDAMT3*), protein phosphatase (*PsPP2C*), ATP-binding cassette protein (*PsABCF1*), among others. It is notable that *GPCR-AC3* and *GPCR-AC4* were initially mispredicted as a single gene (Fig. 1-E and F), potentially leaving one of them undiscovered as a dark gene indefinitely. Importantly, these two genes are closely situated in the genome, sharing identical protein domains with nearly 80% sequence similarity (Appendix F). This raises the question of whether they are the same gene with similar biological functions. Subsequently, functional validation of these two genes was conducted, revealing distinct biological roles (Fig. 1-G) and confirming their status as two separate genes.

It is essential to highlight that the toGC workflow depends on gene candidates with adequate transcriptional abundance. Therefore, it is advisable to consider multiple stages, especially if a candidate gene exhibits low expression in a particular stage. This rationale underlies our decision to incorporate data from five distinct stages of *P. sojae* in our study. If a gene demonstrates low expression across all analyzed transcriptional data, this method may not effectively correct the gene models for such genes. Another noteworthy limitation of this method is that it cannot rectify errors arising from genome sequencing or assembly issues. Since the transcriptional data utilized in the workflow are derived from alignments to sequenced reference genomes, any inaccuracies in the underlying genome sequences may persist in the corrected gene models.

Our method was chosen for its time efficiency and convenience, enabling the identification of gene sequences in species with sufficient RNA-seq data or expressed sequence tag (EST) data. While more accurate strategies for verifying gene structure predictions exist, such as higher-quality genome sequencing or assembly, these methods are computationally intensive and not always updated with each release of gene databases. The primary objective of our pipeline is to enhance gene sequence quality using existing genome sequences with minimal computational cost, time consumption, and labor intensity. For users working with *P. sojae*, *P. capsici and P. ultimum*, we have integrated transcriptome data and provided user-friendly software tools (e.g., IGV) for accomplishing the subsequent steps. For users working with other species, we have made it straightforward to integrate transcriptome data, providing all the necessary scripts in the Appendix files. This user-friendly workflow allows for the correction of gene models using genomic and transcriptome data, even if they may not be of the highest quality. In the future, we intend to expand this method to allow for gene model correction in additional species.

5. Conclusion

In summary, our work has demonstrated a notable enhancement in the precision of gene model correction through the utilization of the toGC tool. Furthermore, we have showcased its remarkable performance in elucidating the intricacies of ‘dark genes’ that have hitherto remained unannotated. Our study serves as a beacon, underscoring the efficacy of the integrated transcriptomic methodologies as a means to elevate the caliber of oomycete genomes, while requiring only modest financial and labor investments.

Acknowledgements

This work was supported by the grants to Min Qiu and Ming Wang from the National Natural Science Foundation of China (32100160 and 32100044), the grants to Ming Wang from the Jiangsu “Innovative and Entrepreneurial Talent” Program, China (JSSCRC2021510), and the grants to Yuanchao Wang from the Chinese Modern Agricultural Industry Technology System (CARS-004-PS14).

Declaration of competing interest

The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used DeepSeek in order to refine the language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published artice.

Appendices associated with this paper are available at https://doi.org/10.1016/j.jia.2024.03.077

References

Bult C J, Blake J A, Calvi B R, Cherry J M, DiFrancesco V, Fullem R, Howe K L, Kaufman T, Mungall C, Perrimon N, Shimoyama M, Sternberg P W, Thomas P, Westerfield M, Consorti A G R. 2019. The alliance of genome resources: building a modern data ecosystem for model organism databases. *Genetics*, **213**, 1189–1196.

Chen H, Fang Y, Song W, Shu H, Li X, Ye W, Wang Y, Dong S. 2023. The SET domain protein PsKMT3 regulates histone H3K36 trimethylation and modulates effector gene expression in the soybean pathogen *Phytophthora sojae*. *Molecular Plant Pathology*, **24**, 346–358.

Chen H, Shu H, Wang L, Zhang F, Li X, Ochola S O, Mao F, Ma H, Ye W, Gu T, Jiang L, Wu Y, Wang Y, Kamoun S, Dong S. 2018. Phytophthora methylomes are modulated by 6mA methyltransferases and associated with adaptive genome regions. *Genome Biology*, **19**, 181.

Danchin A, Ouzounis C, Tokuyasu T, Zucker J D. 2018. No wisdom in the crowd: Genome annotation in the era of big data-current status and future prospects. *Microbial Biotechnology*, **11**, 588–605.

Denton J F, Lugo-Martinez J, Tucker A E, Schrider D R, Warren W C, Hahn M W. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology*, **10**, e1003998.

Deutekom E S, Vosseberg J, van Dam T J P, Snel B. 2019. Measuring the impact of gene prediction on gene loss estimates in eukaryotes by quantifying falsely inferred absences. *PLoS Computational Biology*, **15**, e1007301.

Dragan M A, Moghul I, Priyam A, Bustos C, Wurm Y. 2016. GeneValidator: Identify problems with protein-coding gene predictions. *Bioinformatics*, **32**, 1559–1561.

Fang Y, Tyler B M. 2016. Efficient disruption and replacement of an effector gene in the oomycete *Phytophthora sojae* using CRISPR/Cas9. *Molecular Plant Pathology*, **17**, 127–139.

Feng H, Wan C, Zhang Z, Chen H, Li Z, Jiang H, Yin M, Dong S, Dou D, Wang Y, Zheng X, Ye W. 2021. Specific interaction of an RNA-binding protein with the 3´-UTR of its target mRNA is critical to oomycete sexual reproduction. *PLoS Pathogens*, **17**, e1010001.

Gao J, Cao M, Ye W, Li H, Kong L, Zheng X, Wang Y. 2015. PsMPK7, a stress-associated mitogen-activated protein kinase (MAPK) in *Phytophthora sojae*, is required for stress tolerance, reactive oxygenated species detoxification, cyst germination, sexual reproduction and infection of soybean. *Molecular Plant Pathology*, **16**, 61–70.

Guigo R, Agarwal P, Abril J F, Burset M, Fickett J W. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*, **10**, 1631–1642.

Guigo R, Flicek P, Abril J F, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic V B, Birney E, Castelo R, Eyras E, Ucla C, Gingeras T R, Harrow J, Hubbard T, Lewis S E, Reese M G. 2006. EGASP: The human ENCODE genome annotation assessment project. *Genome Biology*, **7**, S2.1–31.

Hadley C. 2003. Righting the wrongs-DNA and protein sequence databases are increasingly useful research tools. But to maximize their potential, the errors in them need to be addressed. *EMBO Reports*, **4**, 829–831.

Van den Hoogen D J, Meijer H J G, Seidl M F, Govers F. 2018. The ancient link between G-protein-coupled receptors and C-terminal phospholipid kinase domains. *mBio*, **9**, e02119-17.

Hua C L, Wang Y L, Zheng X B, Dou D L, Zhang Z G, Govers F, Wang Y C. 2008. A *Phytophthora sojae* G-protein alpha subunit is involved in chemotaxis to soybean isoflavones. *Eukaryotic Cell*, **7**, 2133–2140.

Li X, Liu Y, Tan X Q, Li D L, Yang X Y, Zhang X, Zhang D Y. 2020. The high-affinity phosphodiesterase is involved in the polarized growth and pathogenicity of. *Fungal Biology*, **124**, 164–173.

McGowan J, Fitzpatrick D A. 2020. Recent advances in oomycete genomics. *Advances in Genetics*, **105**, 175–228.

Meyer C, Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson J D. 2020. Understanding the causes of errors in eukaryotic protein-coding gene prediction: A case study of primate proteomes. *BMC Bioinformatics*, **21**, 513.

Mohanta T K, Al-Harrasi A. 2021. Fungal genomes: Suffering with functional annotation errors. *IMA Fungus*, **12**, 32.

Qiu M, Li Y, Zhang X, Xuan M, Zhang B, Ye W, Zheng X, Govers F, Wang Y. 2020. G protein alpha subunit suppresses sporangium formation through a serine/threonine protein kinase in *Phytophthora sojae*. *PLoS Pathogens*, **16**, e1008138.

Qiu M, Tian M, Yong S, Sun Y, Cao J, Li Y, Zhang X, Zhai C, Ye W, Wang M, Wang Y. 2023. Phase-specific transcriptional patterns of the oomycete pathogen *Phytophthora sojae* unravel genes essential for asexual development and pathogenic processes. *PLoS Pathogens*, **19**, e1011256.

Salzberg S L. 2019. Next-generation genome annotation: We still struggle to get it right. *Genome Biology*, **20**, 92.

Thines M. 2018. Oomycetes. *Current Biology*, **28**, R812-R813.

Tyler B M, Tripathy S, Zhang X, Dehal P, Jiang R H, Aerts A, Arredondo F D, Baxter L, Bensasson D, Beynon J L, Chapman J, Damasceno C M, Dorrance A E, Dou D, Dickerman A W, Dubchak I L, Garbelotto M, Gijzen M, Gordon S G, Govers F, *et al.* 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, **313**, 1261–1266.

Wang W Z, Xue Z L, Xie L F, Zhou X, Zhang F, Zhang S C, Govers F, Liu X L. 2023. Sterol-sensing domain (SSD)-containing proteins in sterol auxotrophic mediate sterol signaling and play a role in asexual reproduction and pathogenicity. *Microbiology Spectrum*, **11**, e0379722.

Wang Y, Ye W, Wang Y. 2018. Genome-wide identification of long non-coding RNAs suggests a potential association with effector gene transcription in *Phytophthora sojae*. *Molecular Plant Pathology*, **19**, 2177–2186.

Weis W I, Kobilka B K. 2018. The molecular basis of G protein-coupled receptor activation. *Annual Review of Biochemistry*, **87**, 897–919.

Ye W, Wang X, Tao K, Lu Y, Dai T, Dong S, Dou D, Gijzen M, Wang Y. 2011. Digital gene expression profiling of the *Phytophthora sojae* transcriptome. *Molecular Plant-Microbe Interactions*, **24**, 1530–1539.

Zerbino D R, Frankish A, Flicek P. 2020. Progress, challenges, and surprises in annotating the human genome. *Annual Review of Genomics and Human Genetics*, **21**, 55–79.

Zhang X, Zhai C, Hua C, Qiu M, Hao Y, Nie P, Ye W, Wang Y. 2016. PsHint1, associated with the G-protein α subunit PsGPA1, is required for the chemotaxis and pathogenicity of *Phytophthora sojae*. *Molecular Plant Pathology*, **17**, 272–285.

Executive Editor-in-Chief Fanghao Wan
Managing Editor Juan Zhang