

Feb 16, 2025

A PCR primer design method for identifying spider mite species using k-mer counting

DOI

<https://dx.doi.org/10.17504/protocols.io.q26g7m78qgwz/v1>

Tomoko Matsuda¹

¹Nihon BioData Corporation



Tomoko Matsuda

Nihon BioData Corporation

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.q26g7m78qgwz/v1>

Protocol Citation: Tomoko Matsuda 2025. A PCR primer design method for identifying spider mite species using k-mer counting. protocols.io <https://dx.doi.org/10.17504/protocols.io.q26g7m78qgwz/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: February 16, 2025

Last Modified: February 16, 2025

Protocol Integer ID: 120479

Keywords: seq datasets of spider mite species, identifying spider mite species, pcr primer design method, spider mite species, primer design method, specific primer candidate sequence, specific primer, specific species, practical species discrimination, single primer, primer, bioinformatics approach, species, multiplex pcr, optimizing multiplex pcr, such species, mer counting this protocol, tetranychidae, mer counting, pcr

Abstract

This protocol uses a bioinformatics approach to extract species-specific primer candidate sequences (i.e., single primers unique to a specific species) from RNA-Seq datasets of spider mite species (Acari, Tetranychidae). Such species-specific primers may be used for practical species discrimination by optimizing multiplex PCR. Our primer design method is expected to be applicable to other taxa. The primer design methods are illustrated in Fig. 1 and 2.

Attachments



[Fig1.PDF](#)

155KB



[Fig2.PDF](#)

197KB

Image Attribution

Fig. 1 Primer design methods for the detection of spider mite species using *k*-mer counting.

Fig. 2 Workflow diagram for extracting species-specific primers from RNA-Seq data using *k*-mer counting.



Identification of species-specific primers

1 RNA-Sequencing

Total RNA was extracted from whole bodies of 100–200 adult females of the same population using an RNeasy Micro Kit (Qiagen, Valencia, CA, USA). Live female individuals for RNA samples and female individuals for voucher specimens were obtained from the same leaf discs and plants. The quantity and quality of the total RNA were evaluated using Agilent RNA ScreenTape System (Agilent Technologies, Santa Clara, CA, USA). The cDNA libraries were prepared from the total RNA using the TruSeq RNA sample prep kit (Illumina, San Diego, CA, USA), and the single ends were sequenced for 75 cycles on the NextSeq 500 sequencing platform (Illumina). All the reads were deposited in the DDBJ Sequence Read Archive.

2 *de novo* assembly

The workflow of the subsequent bioinformatics analysis is shown in Fig. 2. The sequence reads were trimmed by fastx_trimmer of the FASTX-Toolkit (parameter: -f 15) and by fastq_quality_trimmer (parameters: -t 28 and -l 40), and then filtered by fastq_quality_filter (parameters -q 28 and -p 80). The processed sequence reads were assembled into contigs per species by Bridger with the following command: Bridger --seqType fq --output [output_dir] --single [reads.fq] --CPU 16. Contigs with 95% or more similarity were judged as redundant and removed by CD-HIT. The open reading frames (ORFs) were identified by TransDecoder.

3 Identification of putative orthologous genes

A problem with using RNA-Seq data for primer design is that they don't include introns, which are present in genomic DNA. Introns in the primer sequence itself or between the primers could prevent amplification. Therefore, we extracted species-specific primers from intronless genes, which were identified from the exon annotations of the *T. urticae* genome (Grbić et al., 2011). The ORFs were annotated by FATE with the TBLASTN engine and other default parameters against the intronless genes of the *T. urticae* genome. An intronless genes were annotated, and orthologous ORFs were aligned by DIALIGN-TX with the -L option. Despite recent advances, existing ortholog detection methods still suffer from false-positive and false-negative rates. We therefore constructed a phylogenetic tree using RAXML to differentiate between orthologous and paralogous genes. A gene was classified as a paralog if its phylogenetic relationship distinctly deviated from the established relationships among known spider mite species in the family Tetranychidae. After removing paralogous genes, putative orthologous intronless genes were used for *k*-mer analysis.

Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, et al. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature*. 2011;479: 487–492. doi: 10.1038/nature10640.

4 ***k*-mer analysis**

Contigs clustered with CD-HIT, which were assigned to orthologous intronless genes, were used for the *k*-mer analysis. *K*-mer analysis with *k* set to 30 was performed by Jellyfish for each species. The command in Jellyfish is: `jellyfish count -m 30 -s 10000 -t 4 -o [output_file] [input_fasta]`.

5 **Extracting unique *k*-mers**

K-mers obtained from Jellyfish were filtered using the following requirements. Primer specificity for species-specific detection:

1. If a *k*-mer matched another *k*-mer within the same species or a *k*-mer of another species, the *k*-mer was removed.

T_m (primer melting temperature):

2. If the *T_m* value of a *k*-mer was less than 67 or greater than 82, the *k*-mer was removed.

GC Content: Preferably in the range of 40%–60%. Include 1–2 G or C bases at the 3' end to improve binding stability; however, avoid excessive GC content at the 3' terminus to prevent non-specific binding.

3. If the GC content of a full-length *k*-mer (30-mers) was less than 11 or greater than 17, the *k*-mer was removed.

4. If the GC content of the former or latter half of a *k*-mer (15-mers) was less than four or greater than 10, the *k*-mer was removed.

5. If both the 5' and 3' ends of a *k*-mer were A or T, the *k*-mer was removed.

6. If both the 5' and 3' ends of a *k*-mer contained three or more of G or C, the *k*-mer was removed.

7. If both the 5' and 3' ends of a *k*-mer contained three or more of A or T, the *k*-mer was removed.

Avoid Repeats and Runs: Prevent intra-primer and inter-primer complementarity to reduce primer-dimer formation.

8. If a *k*-mer contained four or more runs of a single base (e.g., AAAAA or GGGG), the *k*-mer was removed.

9. If a *k*-mer contained more than three times of dinucleotide repeats (e.g., ATATAT or CGCGCG), the *k*-mer was removed.

6 **Filtering *k*-mers**

The filtering of unique *k*-mers, which serve as candidates for species-specific primers, was performed using both Bowtie2 and BLASTN. Bowtie2 offers a significant advantage in computational speed, allowing for rapid initial filtering. In contrast, BLASTN permits more mismatches and supports finer matching criteria, providing greater flexibility and accuracy in sequence comparison. By first applying Bowtie2 for preliminary filtering and then refining the reduced set of *k*-mers with BLASTN, we optimize both computational performance and precision in the filtering process. Extracted unique *k*-mer sets were mapped to contigs of assemblies using Bowtie2 software. When a *k*-mer was mapped multiple times to contigs of the same species or

matched 26-mers or more to other species, the *k*-mer was removed. Then, BLASTN (blastn-short) searches were performed between the *k*-mers filtered by Bowtie2 (query) and contigs of assemblies after clustering with CD-HIT (database). When a *k*-mer aligned multiple times or 24–29 bp to a contig of the same species, the *k*-mer was removed. Furthermore, *k*-mers that matched contigs of other species by more than 25 bp were removed. The *k*-mers remaining after filtering by Bowtie2 and BLASTN were considered species-specific primer candidates.

PCR validation

7 Finding potential primer pairs from alignment

Computationally identified species-specific *k*-mers occasionally exhibited unintended sequence matches to other species during sequence alignment checks, reducing their specificity. For these reasons, PCR primers to be paired with species-specific primer candidates were designed from the alignment of the coding sequences (CDSs) and the 5' and 3' UTR sequences. CDSs were aligned using DIALIGN-TX, which considers amino acid sequences to improve alignment accuracy, as described in the section 'Identification of putative orthologous genes.' In contrast, 3' and 5' UTR sequences were aligned using MAFFT, a widely used alignment tool, as considering amino acid sequences is not necessary for these non-coding regions. Selection criteria included maintaining an appropriate distance for amplification and ensuring primer specificity within the transcriptome to minimize cross-amplification. This approach enabled the practical refinement of primer design beyond automated *k*-mer pairing. In some cases, the position and length of the species-specific primers were modified according to the alignment. To confirm the uniqueness of the species-specific primers, we conducted additional BLASTN searches to compare the species-specific primers with the contigs from assemblies after clustering with CD-HIT.

8 DNA extraction and PCR amplification

Total DNA was extracted from the whole bodies of individual female spider mites using a QIAamp DNA Micro Kit (Qiagen, Valencia, CA, USA). To ensure the quality of the extracted DNA and to provide a positive control for the PCR reactions, we tested the DNA using primer sets known to amplify spider mite DNA. Specifically, we targeted the cytochrome c oxidase subunit I (COI) gene of mitochondrial DNA and the 28S nuclear ribosomal RNA (rRNA) gene. The successful amplification of these markers confirms that the extracted DNA is of sufficient quality for reliable PCR analysis. A pair of primers, designed to amplify a single species, was tested against all species used in this study. When a gel band was observed for only a single species, the primer set was considered to be effective in detecting the species. PCR reactions were performed in a 20 μ L mixture containing 1 μ L of DNA template, 0.4 μ L of KOD FX Neo (1 U/ μ L, Toyobo, Osaka, Japan), 0.6 μ L of each primer (10 pmol/ μ L each), 4 μ L of 2 mM dNTPs (Toyobo), 10 μ L of

2× PCR Buffer for KOD FX Neo (Toyobo), and 3.4 µL of distilled water. The PCR cycling parameters were as follows: 2 min of denaturation at 94°C, 35 cycles of 10 sec at 98°C, and 1 min at 68°C. An additional 1 min at 68°C was allowed for final strand elongation. PCR products were visualized by electrophoresis on an agarose gel. A 100-bp DNA ladder (Takara Bio, Shiga, Japan) was used as a molecular size marker.

9 Sequencing

The PCR product was sequenced only when the target species was successfully amplified. PCR products were purified using Sephadryl S-300 High Resolution (GE Healthcare, Chicago, IL, USA) and directly sequenced. Sequencing was carried out in both directions using the amplifying primers with the BigDye Terminator Cycle Sequencing Kit v.3.1 (Applied Biosystems, Foster City, CA) and on an ABI 3130xl Genetic Analyzer (Applied Biosystems). Then, BLASTN searches were performed between the sequences of the PCR products (query) and the contigs that contain species-specific primers (database) to confirm sequence similarity. The sequences of the PCR products have been deposited in the DDBJ/EMBL/GenBank International Nucleotide Sequence Databases.